



Usando Data Science para Transformar  
Informação em Insight

John W. Foreman

# DATA SMART

“Data Smart mostra como é fácil implementar e entender métodos estatísticos modernos e algoritmos. Não é mais necessário passar por livros didáticos e artigos acadêmicos!”

—**Patrick Crosby**, Fundador do StatHat e o primeiro Diretor de Tecnologia na OkCupid



ALTA BOOKS  
EDITOR A



Usando Data Science para Transformar  
Informação em Insight

John W. Foreman

# DATA SMART

"Data Smart mostra como é fácil implementar e entender métodos estatísticos modernos e algoritmos. Não é mais necessário passar por livros didáticos e artigos acadêmicos!"

—**Patrick Crosby**, Fundador do StatHat e o primeiro Diretor de Tecnologia na OkCupid



# **A MAIORIA DAS PESSOAS ESTÁ TRATANDO DATA SCIENCE DE FORMA ERRADA.**

## **VEJA COMO FAZER CORRETAMENTE.**

Não quero te desiludir, mas os cientistas de dados não são feiticeiros místicos usuários da arte da magia. Data science é algo que você pode fazer. De verdade. Este livro mostra técnicas significativas, como elas funcionam, como usá-las, e como elas trazem benefícios para o seu negócio, seja ele pequeno ou grande. Não é apenas trabalhar com códigos ou tecnologias de base de dados. É transformar seus dados brutos em insight para que você possa operá-los da forma mais rápida e prática possível.

Erga as mangas e vamos lá.

Visite o site da editora em [www.altabooks.com.br](http://www.altabooks.com.br) e procure pelo título do livro para baixar as planilhas de cada capítulo, e acompanhe:

- O uso da inteligência artificial com o modelo linear geral, métodos ensemble e Naive Bayes
- Agrupamentos com k-means, modularidade de gráfico e k-means esférico
- Otimização matemática incluindo programação não-linear e algoritmos genéricos
- O trabalho com dados de séries temporais e previsões com suavização exponencial
- O uso da simulação Monte Carlo para quantificar ou endereçar o risco
- A detecção de valores atípicos em únicas e múltiplas dimensões
- A exploração da linguagem R voltada para data science

A compra deste conteúdo não prevê o atendimento e fornecimento de suporte técnico operacional, instalação ou configuração do sistema de leitor de ebooks. Em alguns casos, e dependendo da plataforma, o suporte poderá ser obtido com o fabricante do equipamento e/ou loja de comércio de ebooks.

# Data Smart

## Usando Data Science para Transformar Informação em Insight

John W. Foreman



# Data Smart — Usando Data Science para Transformar Informação em Insight

Copyright © 2018 da Starlin Alta Editora e Consultoria Eireli. ISBN: 978-85-508-433-0

*Translated from original Data Smart: Using Data Science to Transform Information into Insight. Copyright © 2014 by John Wiley & Sons, Inc. ISBN 978-1-118-66146-8. This translation is published and sold by permission of John Wiley & Sons, Inc., the owner of all rights to publish and sell the same. PORTUGUESE language edition published by Starlin Alta Editora e Consultoria Eireli, Copyright © 2016 by Starlin Alta Editora e Consultoria Eireli.*

Todos os direitos estão reservados e protegidos por Lei. Nenhuma parte deste livro, sem autorização prévia por escrito da editora, poderá ser reproduzida ou transmitida. A violação dos Direitos Autorais é crime estabelecido na Lei nº 9.610/98 e com punição de acordo com o artigo 184 do Código Penal.

A editora não se responsabiliza pelo conteúdo da obra, formulada exclusivamente pelo(s) autor(es).

**Marcas Registradas:** Todos os termos mencionados e reconhecidos como Marca Registrada e/ou Comercial são de responsabilidade de seus proprietários. A editora informa não estar associada a nenhum produto e/ou fornecedor apresentado no livro.

Edição revisada conforme o Acordo Ortográfico da Língua Portuguesa de 2009.

Publique seu livro com a Alta Books. Para mais informações envie um e-mail para [autoria@altabooks.com.br](mailto:autoria@altabooks.com.br)

Obra disponível para venda corporativa e/ou personalizada. Para mais informações, fale com [projetos@altabooks.com.br](mailto:projetos@altabooks.com.br)

**Erratas e arquivos de apoio:** No site da editora relatamos, com a devida correção, qualquer erro encontrado em nossos livros, bem como disponibilizamos arquivos de apoio se aplicáveis à obra em questão.

Acesse o site [www.altabooks.com.br](http://www.altabooks.com.br) e procure pelo título do livro desejado para ter acesso às erratas, aos arquivos de apoio e/ou a outros conteúdos aplicáveis à obra.

**Suporte Técnico:** A obra é comercializada na forma em que está, sem direito a suporte técnico ou orientação pessoal/exclusiva ao leitor.

A editora não se responsabiliza pela manutenção, atualização e idioma dos sites referidos pelos autores nesta obra.

## Produção Editorial

Editora Alta Books

## Gerência Editorial

Anderson Vieira

## Produtor Editorial

Claudia Braga

Thiê Alves

## Produtor Editorial (Design)

Aurélio Corrêa

## Marketing Editorial

Silas Amaro

[marketing@altabooks.com.br](mailto:marketing@altabooks.com.br)

## Vendas Atacado e Varejo

Daniele Fonseca  
Viviane Paiva  
comercial@altabooks.com.br

**Tradução**  
Welington Nascimento

**Copidesque**  
Vivian Sbravatti

**Revisão Gramatical**  
Samantha Batista

**Revisão Técnica**  
Ronaldo d'Avila Roenick

**Adaptação para formato e-Book**  
Daniel Vargas

**Dados Internacionais de Catalogação na Publicação (CIP) de acordo com ISBD**

F715d	Foreman, John W.
Data smart [recurso eletrônico] : usando data science para transformar informação em insight / John W. Foreman ; traduzido por Welington Carlos Santos do Nascimento. - Rio de Janeiro : Alta Books, 2018.	
448 p. ; il. ; 26.933 KB.	
Tradução de: Data Smart: Using Data Science to Transform Information into Insight Inclui índice. ISBN: 978-85-508-0433-0 (Ebook)	
1. Data smart. 2. Dados. I. Nascimento, Welington Carlos Santos do. III. Título.	
2018-1231	CDD 005.13 CDU 004.62

Elaborado por Wagner Rodolfo da Silva - CRB-8/9410

Rua Viúva Cláudio, 291 — Bairro Industrial do Jacaré  
CEP: 20970-031 — Rio de Janeiro - RJ  
Tels.: (21) 3278-8069 / 3278-8419  
[www.altabooks.com.br](http://www.altabooks.com.br) — [altabooks@altabooks.com.br](mailto:altabooks@altabooks.com.br)  
[www.facebook.com/altabooks](http://www.facebook.com/altabooks)

*Para minha esposa Lydia. O que você faz todos os dias é sensacional. Se não fosse por você, já teria perdido meus cabelos (e minha cabeça) anos atrás.*

# Sobre o Autor

**John W. Foreman** é Cientista de Dados Chefe no MailChimp.com. Também é um consultor de gestão que fez muitas análises para grandes empresas (Coca Cola, Royal Caribbean, Hotéis Grupo InterContinental) e para o governo americano (DoD, IRS, DHS, FBI).

Quando ele não está brincando com dados, John passa seu tempo caminhando, assistindo bastante televisão, comendo todo tipo de comida ruim e criando três meninos fedorentos.

# Sobre os Editores Técnicos

**Greg Jennings** é cientista de dados, engenheiro de software e cofundador da

ApexVis. Após completar o mestrado em Ciências de Materiais na Universidade da Virgínia, começou sua carreira com a equipe de análise Booz Allen Hamilton, onde iniciou uma equipe de implantação de análise preditiva e soluções de visualização de dados para planejar e programar problemas.

Após deixar o Booz Allen Hamilton, Greg cofundou sua primeira startup, Decision Forge, onde atuou como diretor de tecnologia e ajudou a desenvolver uma plataforma de mineração de dados baseada na web para um cliente do governo. Ele também trabalhou com uma grande organização de mídia para desenvolver um produto educacional que auxilia professoras no acesso a conteúdo direcionado para seus alunos e com uma startup baseada em McLean para ajudar a desenvolver aplicações de modelagem de audiência para otimizar campanhas de publicidade na web.

Após sair da Decision Forge, ele cofundou seu atual negócio ApexVis, especializado em ajudar empresas a conseguirem o valor máximo dos seus dados por meio da visualização padrão de dados e soluções analíticas de software. Ele vive em Alexandria, Virgínia, com sua esposa e duas filhas.

**Evan Miller** recebeu seu bacharelado em Física na Williams College em 2006 e atualmente é estudante de doutorado da Universidade de Chicago. Sua pesquisa inclui testes de especificação e métodos computacionais em econometria. Evan também é o autor do Wizard, um popular programa de Mac que realiza análises estatísticas.

# Agradecimentos

Este livro começou após uma quantidade improvável de pessoas verificarem meu blog analítico, Analytics Made Skeezy. Portanto, gostaria de agradecer esses leitores, bem como meus colegas de data science do Twitter por todo o apoio. E agradeço a Aarron Walter, Chris Mills e Jon Duckett por passarem a ideia deste livro para a Wiley baseados na premissa simples do meu blog.

Eu também gostaria de agradecer à equipe do MailChimp por tornar isso possível. Sem o apoio e a cultura aventureira no MailChimp, eu não teria confiança o bastante para fazer algo tão complicado quanto escrever um livro técnico enquanto trabalho e crio três filhos. Especificamente, eu não poderia ter feito isso sem a assistência diária de Neil Bainton e Michelle Riggin-Ransom. Além disso, eu estou em dívida com Ron Lewis, Josh Rosenbaum e Jason Travis por seus trabalhos na capa e no vídeo de marketing para o livro.

Agradeço a Carol Long da Wiley por acreditar em mim e a todos os editores por sua competência e trabalho árduo. Um enorme muito obrigado a Greg Jennings por fazer todas as planilhas!

Muito obrigado aos meus pais por lerem meu romance de ficção científica e não me mandarem parar de escrever.

# Sumário

## Introdução

O que Eu Estou Fazendo Aqui?

Uma Definição Prática de Data Science

Mas Espere, é Big Data?

Quem Sou Eu?

Quem É Você?

Sem Arrependimentos. Planilhas para Sempre.

Convenções

Vamos Seguir em Frente

## Tudo o que Você Sempre Quis Saber sobre Planilhas mas Tinha Medo de Perguntar

Alguns Exemplos de Dados

Movendo-se Rapidamente com o Botão de Controle

Copiando Fórmulas e Dados Rapidamente

Formatando Células

Colando Valores Especiais

Inserindo Gráficos

Encontrando os Menus Find e Replace

Fórmulas para Localizar e Retirar Valores

Usando VLOOKUP para Juntar Dados

Filtrando e Ordenando

Usando PivotTables (Tabelas Dinâmicas)

Usando Fórmulas Array

Resolvendo Coisas com o Solver

OpenSolver: Eu Gostaria que Não Precisássemos Fazer Isso, mas Precisamos

Resumindo

## Análise de Agrupamento Parte 1: Usando K-Means para Segmentar a Sua Clientela

Meninas dançam com Meninas, Meninos Coçam Seus Cotovelos

Caia na Real: Agrupamento K-Means de Assinantes em E-mail Marketing

Agrupamento K-Medians e Medidas de Distância Assimétricas  
Resumindo

## Naïve Bayes é a Incrível Leveza de Ser um Idiota

Quando Você Nomeia um Produto de Mandrill, Receberá Alguns Sinais ou Alguns Ruídos

A Introdução da Teoria da Probabilidade Mais Rápida do Mundo  
Usando a Regra de Bayes para Criar um Modelo de IA  
Vamos Começar a Festa do Excel

Resumindo

## Modelo de Otimização: porque aquele Suco de Laranja “Recém Espremido” Não irá se Misturar Sozinho

Por que Cientistas de Dados Deveriam Saber Otimização?

Comece com um Compromisso Simples

Direto da Plantação para o Seu Copo... com uma Parada em um Modelo de Mistura

Modelagem de Risco

Resumindo

## Análise de Grupo Parte II: Gráficos de Rede e Detecção de Comunidade

[O que é um Gráfico de Rede?](#)

[Visualizando um Gráfico Simples](#)

[Breve Introdução à Gephi](#)

[Construindo um Gráfico do Dado de Venda Atacadista de Vinho](#)

[Quanto Vale um Vértice? Pontuações e Penalidades em Modularidade de Gráfico](#)

[Vamos Agrupar!](#)

[Lá e de Volta Outra Vez: um Conto Gephi](#)

[Resumindo](#)

## [O Avô da Inteligência Artificial Supervisionada — Regressão](#)

[Espere, to Quê? Você Está Grávida?](#)

[Não se Engane](#)

[Prevendo Clientes Grávidas na RetailMart Usando Regressão Linear](#)

[Prevendo Clientes Grávidas na RetailMart Usando Regressão Logística](#)

[Para Mais Informações](#)

[Resumindo](#)

## [Modelos Ensemble: É Muita Pizza Ruim Junto](#)

[Usando os Dados do Capítulo 6](#)

[Bagging: Aleatorize, Treine e Repita](#)

[Boosting: Se Fizer Errado, Reinicie e Tente Novamente](#)

[Resumindo](#)

## [Forecasting: Respire Devagar; Você Não Pode Ganhar](#)

[O Mercado de Espadas Está a Mil](#)

[Conhecendo os Dados de Séries Temporais](#)

[Começando com Coluna com Suavização Exponencial Simples](#)

Talvez Você Tenha uma Tendência  
A Suavização Exponencial com Tendência Corrigida de Holt  
Suavização Exponencial Multiplicativa Holt-Winters  
Resumindo

Detecção de Valor Atípico: Só Porque Eles São Estranhos Não Significa que Não São Importantes

Os Valores Atípicos São (Más) Pessoas Também  
O Caso Fascinante de Hadlum vs Hadlum  
Terrível em Nada, Ruim em Tudo  
Resumindo

Trocando das Planilhas para R

Botando para Funcionar com R  
Praticando um Pouco de Data Science Real  
Resumindo

Conclusão

O que Aconteceu? Onde Estou?  
Antes que Você se Vá  
Seja Criativo e Mantenha Contato!

# Introdução

## O↑que↑Eu↑Estou↑Fazendo↑Aqui?

Você provavelmente ouviu falar do termo data science circulando pela mídia recentemente, em livros e jornais de negócios, e em conferências. Data science consegue convocar eleições para presidência, revelar mais sobre hábitos de compras do que você ousa contar para sua mãe, e prever quantos anos aquele burrito de chili com queijo terá tirado da sua vida.

Cientistas de dados, os praticantes de elite dessa arte, foram rotulados como “sexy” em um artigo recente da Harvard Business Review, embora aparentemente tal apelido seja como chamar unicórnio de sexy. Não há como verificar a declaração, mas se você pudesse me ver enquanto eu digito este livro com minha barba no pescoço e os olhos cansados de pai de três filhos, você saberia que sexy é um exagero.

Eu divago. A questão é que há um alvoroço sobre data science atualmente, e esse alvoroço está criando pressão em muitos negócios. Se você não está fazendo data science, você perderá a competição. Alguém surgirá com algum produto novo chamado “BláBláBláGráficoDataScience” e destruirá seu negócio.

Respire profundamente.

A verdade é que a maioria das pessoas está bastante errada sobre data science. Eles estão começando a comprar ferramentas e contratar consultores. Eles estão gastando todo seu dinheiro antes de saberem o que querem, porque uma ordem de compras parece passar por um progresso real em muitas empresas hoje em dia.

Ao ler este livro, você terá uma vantagem sobre esse palhaços, porque você aprenderá exatamente o que essas técnicas em data science são e como elas são usadas. Quando chegar no momento do planejamento, contratação e compra, você já saberá como identificar as oportunidades de data science dentro da sua própria organização.

O propósito deste livro é apresentar a você as práticas de data science de uma maneira confortável e coloquial. Quando terminar, eu espero que muita daquela ansiedade data science que você está sentindo seja substituída por excitação e ideias sobre como usar dados para levar seu negócio para o próximo nível.

## Uma↑Definição↑Prática↑de↑Data↑Science

Em parte, *data science* é sinônimo ou relacionado a termos como *business analytics*, *pesquisa operacional*, *business intelligence*, *competitive intelligence*, *análise de dados e modelagem* e *extração de conhecimento* (também chamado de *descobrimento em bases de dados/knowledge discovery in databases* ou KDD). É apenas um novo giro em algo que as pessoas vêm fazendo há muito tempo.

Houve uma mudança em tecnologia desde o auge desse termos. Avanços em hardware e software tornaram fácil e barato coletar, armazenar e analisar grandes quantidades de dados seja em vendas e dados de marketing, solicitações HTTP do seu web site, dados de suporte ao cliente, e por aí vai. Pequenos negócios e instituições não lucrativas agora podem participar do tipo de análise que anteriormente eram do alcance de grandes empresas.

Claro, enquanto data science é usada como um jargão para tudo hoje, data science é geralmente mais associada com técnicas de mineração de dados como inteligência artificial, agrupamento e detecção do valor atípico. Graças à proliferação de tecnologias baratas de dados de negócios transacionais, essas técnicas computacionais ganharam uma âncora em negócios nos últimos anos onde anteriormente era complicado usar em configurações de produção.

Neste livro, teremos uma visão ampla de data science. Esta é a definição a partir da qual trabalharei:

*Data science é a transformação de dados por meio da matemática e estatística em insights, decisões e produtos valiosos.*

Essa é uma definição *centrada em negócios*. É sobre um produto final útil e valioso derivado de dados. Por quê? Porque eu não estou nisso por motivos de pesquisa ou porque eu acho que dados têm um mérito estético. Eu faço data science para ajudar a minha organização a funcionar melhor e criar valor; se você está lendo isso, eu suspeito que esteja procurando algo parecido.

Com essa definição em mente, este livro abordará o alicerce de técnicas analíticas como otimização, previsão, e simulação, bem como tópicos “quentes” como inteligência artificial, gráficos de redes, agrupamentos e detecção do valor atípico.

Algumas dessas técnicas são tão antigas quanto a Segunda Guerra Mundial. Outras surgiram nos últimos 5 anos. E você verá que tempo não tem relação com dificuldade ou utilidade. Todas essas técnicas — sejam elas ou não a moda atual — são igualmente úteis no contexto de negócios certo.

E é por isso que você precisa entender como elas funcionam, como escolher a técnica correta para o problema certo, e como prototipar com elas. Existem muitas pessoas por aí que entendem uma ou duas dessas técnicas, mas o resto não está sob seus radares. Se tudo que eu tivesse na minha caixa de ferramentas fosse um martelo, eu provavelmente tentaria resolver quaisquer problemas batendo neles com muita força, não muito diferente do meu filho de dois anos.

É melhor ter algumas outras ferramentas à disposição.

## Mas↑Espere,↑e↑Big↑Data?

Você ouviu o termo *big data* até mais do que *data science*, provavelmente. Esse livro é sobre big data?

Isso depende de como você define big data. Se você define big data como computar simples estatísticas sumárias em lixo desestruturado armazenado em bancos de dados enormes, escaláveis horizontalmente, sem bancos de dados SQL, então não, este livro não é sobre big data.

Se você define big data como transformar dados de negócios transacionais em decisões e percepções usando análises de ponta (independente de onde aquele dado está armazenado), então sim, este é um livro sobre big data.

Este não é um livro que abordará tecnologias de bancos de dados, como MongoDB e HBase. Este não é um livro que abordará pacotes de codificação data science como Mahout, NumPy, numerosas bibliotecas R, e assim por diante. Existem outros livros sobre essas coisas.

Mas isso é bom. Esse livro ignora as ferramentas, o armazenamento e o código. Em vez disso, ele foca o máximo possível nas técnicas. Existem muitas pessoas por aí que pensam que armazenamento de dados e recuperação, com um pouco de limpeza e agregação misturados, constituem tudo que existe sobre big data.

Eles estão errados. Este livro o levará além dessa conversa fiada que você escuta sobre big data de representantes de vendas de softwares e blogueiros, para lhe mostrar o que realmente é possível com seus dados. E a parte legal é que para muitas dessas técnicas, seu conjunto de dados pode ser de qualquer tamanho, pequeno ou grande. Você não precisa ter um petabyte de dados e os custos que vêm com eles para prever os interesses da sua base de clientes. Se você tem um conjunto de dados grande, isso é ótimo, mas há alguns negócios que não o têm, não precisam dele e, provavelmente, nunca o criaram. Como meu açougue local. Mas isso não significa que seu e-mail marketing não poderia se beneficiar um pouco da detecção de grupo entre bacon e linguiça.

Se os livros de data science fossem treinos, este livro seria calistenia — sem pesos de equipamentos, sem circuitos. Uma vez que você entenda como implementar essas técnicas até com as ferramentas mais básicas, você se encontrará implementando-as em uma variedade de tecnologias, criando protótipos com elas facilmente, comprando de consultores os produtos de data science corretos, delegando a abordagem correta para seus desenvolvedores, e assim por diante.

# Quem↑Sou↑Eu?

Deixe-me parar um instante para lhe contar minha história. Ela percorrerá um longo caminho para a explicação do porquê de eu ensinar data science da maneira que faço. Muitas luas atrás, eu era um consultor de gestão. Eu trabalhava em problemas analíticos para organizações como o FBI, DoD, Coca-Cola Company, Grupo de Hotéis Intercontinental e Royal Caribbean International. E por todas essas experiências eu aprendi uma coisa — pessoas além de cientistas de dados precisam entender data science.

Trabalhei com gerentes que compraram simulações quando precisavam de um modelo de otimização. Eu trabalhei com analistas que apenas entendiam gráficos Gantt, então tudo precisava ser resolvido com gráficos Gantt. Como um consultor, não era difícil ganhar um cliente com qualquer papel branco velho e uma esperta plataforma PowerPoint, porque eles não podiam diferenciar IA de BI ou BI de BS.

O foco deste livro é aumentar o público daqueles que entendem e conseguem implementar técnicas de data science. Eu não estou tentando transformar você em um cientista de dados contra a sua vontade. Eu só quero que você esteja apto a integrar data science da melhor forma possível dentro da função na qual já é bom.

E isso me leva a quem você é.

# Quem↑É↑Você?

Não, eu não estive usando data science para espionar você. Eu não faço ideia de quem você é, mas obrigado por gastar algum dinheiro neste livro. Ou por apoiar a sua biblioteca local. Você pode fazer isso também.

Estes são alguns arquétipos (ou *personas* se você for do marketing) que eu tive em mente quando escrevi este livro. Talvez você seja:

- A vice-presidente de marketing que quer usar seus dados de negócios transacionais mais estrategicamente para fixar o preço

de produtos e segmentar clientes. Mas não entende as abordagens que seus desenvolvedores de software e consultores exagerados estão recomendando que teste.

- O analista de demanda de previsão que sabe que os dados históricos de compras de sua organização possuem mais conhecimento sobre seus clientes do que as projeções do próximo trimestre. Mas não sabe como extrair essa projeção.
- O CEO de uma start-up online que quer prever quando um cliente está mais propenso a ter interesse em comprar um item com base em suas compras passadas.
- A analista de business intelligence que vê dinheiro descendo ralo abaixo por causa dos custos com infraestrutura e cadeia de suprimentos que sua empresa acumula, mas não sabe como tomar decisões de economia de custos sistematicamente.
- O comerciante online que quer fazer mais com as interações gratuitas com clientes de sua empresa em e-mail, Facebook e Twitter, mas nesse momento elas apenas estão sendo lidas e salvas.

Tenho em mente que você é um leitor que seria diretamente beneficiado sabendo mais sobre data science mas não conseguiu ter um embasamento de todas as técnicas. O propósito deste livro é tirar todas as distrações ao redor de data science (o código, as ferramentas e as mentiras) e ensinar as técnicas usando casos de uso práticos que alguém com um semestre de álgebra linear ou cálculo na faculdade consiga entender. Presumindo que você não reprovou o semestre. Se reprovou, apenas leia mais lentamente e use a Wikipédia generosamente.

**Sem↑Arrependimentos.↑Planilhas↑para  
Sempre.**

Este não é um livro sobre codificação. Na verdade, estou dando a minha garantia “sem código” para você (pelo menos até o Capítulo 10). Por quê?

Porque eu não quero gastar centenas de páginas no início deste livro mexendo com Git, configurando variáveis de ambiente e fazendo a dança Emacs versus Vi.

Se você executa Windows e Microsoft Office quase exclusivamente. Se você trabalha para o governo e eles não permitem que você faça download e instale coisas de código aberto na sua caixa. Mesmo que MATLAB ou seu TI-83 tenham lhe assustado muito na faculdade, não precisa ter medo.

Você precisa saber escrever código para colocar a maioria dessas técnicas em configuração de produção automatizadas? Com certeza! Ou, pelo menos, alguém com quem você trabalhe precisa estar apto a lidar com códigos e tecnologias de armazenamento.

Você precisa saber escrever código para entender, distinguir e prototificar com essas técnicas? Absolutamente não!

É por isso que eu abordo cada técnica em software de planilha.

Agora, isso é tudo um pouco de mentira. O capítulo final deste livro na realidade se desloca para a linguagem de programação focada em data science, R. É para pessoas como você, que quer usar este livro como um ponto de partida para coisas mais profundas.

## Mas ↑ Planilhas ↑ São ↑ Tão ↑ Démodé!

Planilhas não são as ferramentas mais sexy existentes. Na verdade, elas são o Wilford Brimley vendendo Colonial-Penn do mundo de ferramenta analítica. Completamente não sexy. Desculpe, Wilford.

Mas esse é o motivo. As planilhas ficam fora do caminho. Elas permitem que você veja os dados e toque (ou, pelo menos, clique) os dados. Há uma liberdade aqui. Para aprender essas técnicas, você precisa de algo comum, algo que todos entendam, entretanto algo que permitirá que você se move rápido e leve enquanto aprende. Isso é uma planilha.

Repita comigo: “Eu sou humano. Eu tenho dignidade. Eu não deveria ter que escrever um trabalho map-reduce para aprender data science.”

E planilhas são ótimas para prototipagem! Você não está executando um modelo IA para o seu negócio varejista online a partir do Excel, mas isso não significa que não possa olhar para seus dados de compras, experimentar com recursos que preveem interesse do produto e prototificar um modelo de direcionamento. Na realidade, é o lugar perfeito para fazer isso.

## Use ↑Excel↑ou ↑LibreOffice

Todos os exemplos nos quais você trabalhá serão visualizados no Excel.

No website da editora ([www.altabooks.com.br](http://www.altabooks.com.br)) estão publicadas planilhas complementares para cada capítulo para que você possa acompanhar. Se você é realmente aventureiro, pode remover tudo menos os dados iniciais na planilha e replicar todos os trabalhos por si próprio.

Este livro é compatível com as versões do Excel 2007, 2010, 2011 para Mac e 2013. O Capítulo 1 abordará as diferenças de versões mais profundamente.

A maioria tem acesso ao Excel, e provavelmente já o usou para relatórios ou manutenção de dados do trabalho. Mas, se por alguma razão, você não possui uma cópia do Excel, pode comprá-la ou usar o LibreOffice ([www.libreoffice.org](http://www.libreoffice.org)).

O LibreOffice é código aberto, gratuito, e tem praticamente todas as mesmas funcionalidades que o Excel. Eu acho que seu solver nativo é preferível ao do Excel. Então, se você quiser seguir esse caminho neste livro, sinta-se à vontade.

### E O GOOGLE DRIVE?

Agora, alguns de vocês podem estar pensando se podem usar o Google Drive. É uma opção atraente já que está na nuvem e pode ser executado em seus dispositivos móveis bem como em seu computador. Mas isso não funcionará.

O Google Drive é ótimo para planilhas simples, mas para onde você está indo, o Google simplesmente não aguenta. Adicionar linhas e colunas no Drive é um incômodo constante, a implementação do Solver é horrível, e os gráficos nem mesmo têm linhas de tendência. Eu gostaria que fosse diferente.

## Convenções

Para ajudá-lo a obter o máximo do texto e a acompanhar o que está acontecendo, eu usei várias convenções ao longo do livro.

### CUIDADO

As caixas CUIDADO possuem informações importantes e não devem ser esquecidas pois são diretamente relevantes ao texto ao seu redor.

### NOTA

Notas abordam dicas, sugestões, truques e apartes sobre a discussão atual.

Frequentemente, neste texto eu referenciarei pequenos fragmentos de código do Excel, como este:

```
=CONCATENATE("THIS IS A FORMULA", " IN EXCEL!")
```

Nos *destacamos* termos novos e palavras importantes quando as apresentamos. Mostramos nomes de arquivos, URLs e fórmulas, assim:

[www.altabooks.com.br](http://www.altabooks.com.br)

## Vamos↑Seguir↑em↑Frente

No primeiro capítulo, eu preencherrei algumas lacunas no seu conhecimento de Excel. Após isso, entrarei em casos de uso. Até o final deste livro, você não somente saberá sobre mas também realmente terá experiência em implementar do zero as seguintes técnicas:

- Otimização usando programação linear e de inteiros
- Trabalhar com dados de séries temporais, detecção de tendências e padrões sazonais, e previsão com suavização exponencial
- Usar simulação Monte Carlo em otimização e cenários de previsão para quantificar e direcionar riscos
- Inteligência artificial usando modelo linear geral, funções de ligação logísticas, métodos de agrupamento, e naïve Bayes
- Medir distâncias entre clientes usando similaridade cosseno, criar gráficos kNN, calcular modularidade, e agrupar clientes
- Detectar valores atípicos em uma única dimensão com Teste de Tukey ou em múltiplas dimensões usando fatores atípicos
- Usar pacotes R para “subir nas costas” de outros analistas na condução dessas tarefas

Se qualquer uma dessas coisas parece excitante, leia!! Se alguma parece assustadora, eu prometo manter as coisas o mais claras e prazerosas possível.

De fato, eu prefiro clareza acima de exatidão matemática, então se é um acadêmico lendo isso, podem haver momentos nos quais você deveria fechar seus olhos e pensar na Inglaterra.

Sem mais delongas, então, vamos resolver esses números.

# 1

# Tudo↑o↑que↑Você↑Sempre↑Quis Saber↑sobre↑Planilhas↑mas↑Tinha Medo↑de↑Perguntar

Este livro acredita que você já possua um conhecimento em trabalhos com planilhas e eu vou presumir que você já saiba o básico. Se você nunca usou uma fórmula na vida, terá uma pequena dificuldade ao ler este livro. Eu recomendaria que você lesse algum livro *Para Leigos* ou algum outro tutorial de introdução ao Excel antes de mergulhar neste aqui.

Dito isso, mesmo que você seja um veterano em Excel, há algumas funcionalidades que aparecerão neste texto que você talvez não tenha usado antes. Não são difíceis; apenas coisas que notei que ninguém usa no Excel. Você percorrerá uma grande variedade de funcionalidades neste capítulo e o exemplo a esta altura pode parecer um pouco desconexo, mas você aprenderá o que puder aqui e, então, quando encontrá-lo mais adiante no livro, poderá retornar a este capítulo como referência.

Como Samuel L. Jackson disse em *Jurassic Park*, “Segurem-se bem!”

## DIFERENÇAS DE VERSÕES DO EXCEL

Como mencionado na introdução do livro, estes capítulos funcionam com Excel 2007, 2010, 2013, 2011 para Mac e LibreOffice. Infelizmente, em cada versão do Excel, a Microsoft mexe nas coisas sem motivo.

Por exemplo, as coisas na aba Layout em 2011 estão na aba View em outras versões. O Solver é o mesmo em 2010 e 2013, mas o desempenho é até melhor em 2007 e 2011 mesmo que a interface do Solver de 2007 seja grotesca.

As capturas de tela neste texto serão do Excel 2011. Se você possui uma versão mais recente ou mais antiga, suas interações poderão, por vezes, parecer diferentes — principalmente se tratando de coisas na barra de menu. Eu farei meu melhor para citar essas diferenças. Se não conseguir encontrar alguma coisa, o recurso de ajuda do Excel e o Google podem lhe ajudar.

A boa notícia é que quando estivermos na “parte planilha da planilha” tudo funciona da mesma forma.

## Alguns Exemplos de Dados

### NOTA

A pasta de trabalho usada neste capítulo, “Concessions.xlsx”, está disponível para download no site da editora, em [www.altabooks.com.br](http://www.altabooks.com.br), procurando pelo título do livro.

Imagine que você é muito malsucedido na vida, que mora com seus pais mesmo depois de adulto e trabalha na cantina do colégio em que estudava no ensino médio durante os jogos de basquete. (Eu juro que isso é apenas meio autobiográfico.)

Você tem uma planilha cheia com as vendas de ontem à noite e ela se parece com a Figura 1-1.

	A	B	C	D
1	Item	Category	Price	Profit
2	Beer	Beverages	\$ 4.00	50%
3	Hamburger	Hot Food	\$ 3.00	67%
4	Popcorn	Hot Food	\$ 5.00	80%
5	Pizza	Hot Food	\$ 2.00	25%
6	Bottled Water	Beverages	\$ 3.00	83%
7	Hot Dog	Hot Food	\$ 1.50	67%
8	Chocolate Dipped Cone	Frozen Treat	\$ 3.00	50%
9	Soda	Beverages	\$ 2.50	80%
10	Chocolate Bar	Candy	\$ 2.00	75%
11	Hamburger	Hot Food	\$ 3.00	67%
12	Beer	Beverages	\$ 4.00	50%
13	Hot Dog	Hot Food	\$ 1.50	67%
14	Licorice Rope	Candy	\$ 2.00	50%
15	Chocolate Dipped Cone	Frozen Treat	\$ 3.00	50%

**Figura 1-1:** Vendas da Cantina do Colégio

A Figura 1-1 mostra cada venda, qual item foi vendido, o tipo de comida ou bebida, o preço e a porcentagem da venda que será lucro.

## Movendo-se Rapidamente com o Botão de Controle

Se você quiser examinar os registros, é possível descer a planilha na tela com o botão de rolar do mouse, com um trackpad ou com a seta para baixo. Conforme você rola, ajuda manter a linha do cabeçalho travada no topo da tabela, para que você lembre o que cada coluna significa. Para fazer isso, escolha **Freeze Panes** ou **Freeze Top Row** na aba **"View"** no Windows (aba **"Layout"** no Mac 2001) conforme exibido na Figura 1-2).

Item	Category	Price	Profit
Beer	Beverages	\$ 4.00	50%
Hamburger	Hot Food	\$ 3.00	67%
Popcorn	Hot Food	\$ 5.00	80%
Pizza	Hot Food	\$ 2.00	25%
Bottled Water	Beverages	\$ 3.00	83%
Hot Dog	Hot Food	\$ 1.50	67%
Chocolate Dipped Cone	Frozen Treat	\$ 3.00	50%
Soda	Beverages	\$ 2.50	80%
Chocolate Bar	Candy	\$ 2.00	75%
Hamburger	Hot Food	\$ 3.00	67%

**Figura 1-2:** Congelando a linha superior

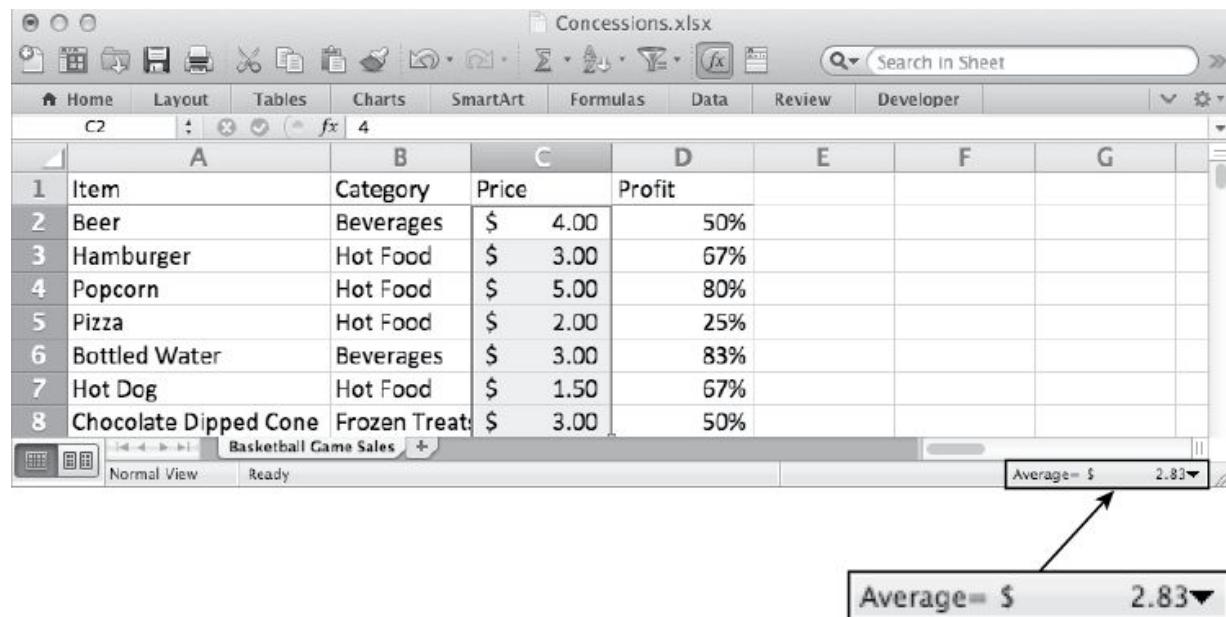
Para mover-se rapidamente para a parte inferior da planilha e ver quantas transações existem, você pode selecionar um valor em uma das

colunas preenchidas e pressionar  $\text{Ctrl} + \downarrow$  ( $\text{Command} + \downarrow$  no Mac). Você irá diretamente para a última célula preenchida naquela coluna. Nesta planilha, a linha final é 200. Observe também que você pode usar  $\text{Ctrl}/\text{Command}$  para saltar da esquerda para a direita na planilha da mesma forma.

Se você quiser tirar uma média de preços das vendas da noite, abaixo da coluna Price (preço), coluna C, você pode anotar a seguinte fórmula:

=AVERAGE (C2 : C200)

A média é \$2,83, então você não se aposentará com riqueza tão cedo. Alternativamente, você pode selecionar a última célula na coluna, C200, segurar Shift + Ctrl + ↑ para realçar toda a coluna e então selecionar o cálculo da Average na barra de status no canto inferior direito da planilha para ver um simples resumo das estatísticas (veja Figura 1-3). No Windows, você precisará dar um clique duplo na barra de status para selecionar a média se ela não estiver lá. No Mac, se sua barra de status estiver indisponível, clique no menu View e selecione “Status Bar” para ativá-la.



**Figura 1-3:** A média da coluna preço na barra de status

# Copiando Fórmulas e Dados Rapidamente

Talvez você queira ver seu lucro em dólares em vez de porcentagens. É possível adicionar um cabeçalho na coluna E chamado “Actual Profit”. Na E2, você precisa apenas multiplicar o preço e as colunas de lucro para obter isto:

=C2\*D2

Para beer é \$2,00. Não é preciso reescrever essa fórmula em todas as células na coluna. Em vez disso, o Excel permite que você selecione o canto inferior direito da célula e arraste a fórmula para onde desejar. As células referenciadas nas colunas C e D serão atualizadas com relação ao lugar no qual a fórmula foi copiada. Como no caso dos dados da cantina do colégio, se a coluna à esquerda for totalmente preenchida, pode-se dar um clique duplo no canto inferior direito da fórmula para que o Excel preencha a coluna inteira (veja a Figura 1-4). Tente fazê-lo, pois usarei muito isso no decorrer do livro e se pegar o jeito agora evitará muito estresse.

Agora, e se você não quiser que as células na fórmula sejam modificadas em relação ao alvo quando forem arrastadas e copiadas? Acrescente apenas um \$ à frente do que não quiser que mude.

Por exemplo, se modificou a fórmula em E2 para:

=C\$2\*D\$2

The screenshot shows a Microsoft Excel spreadsheet titled "Concessions.xlsx". The table has columns labeled A through E. Column A is "Item", B is "Category", C is "Price", D is "Profit", and E is "Actual Profit". Row 2 contains the formula =C2\*D2 in cell E2. The formula is being copied down to row 8, where it is shown as =C8\*D8. The status bar at the bottom indicates "Drag outside selection to extend series or fill; drag inside to clear".

	A	B	C	D	E
1	Item	Category	Price	Profit	Actual Profit
2	Beer	Beverages	\$ 4.00	50%	\$ 2.00
3	Hamburger	Hot Food	\$ 3.00	67%	
4	Popcorn	Hot Food	\$ 5.00	80%	
5	Pizza	Hot Food	\$ 2.00	25%	
6	Bottled Water	Beverages	\$ 3.00	83%	
7	Hot Dog	Hot Food	\$ 1.50	67%	
8	Chocolate Dipped Cone	Frozen Treat	\$ 3.00	50%	

**Figura 1-4:** Arrastando o canto para preencher uma fórmula

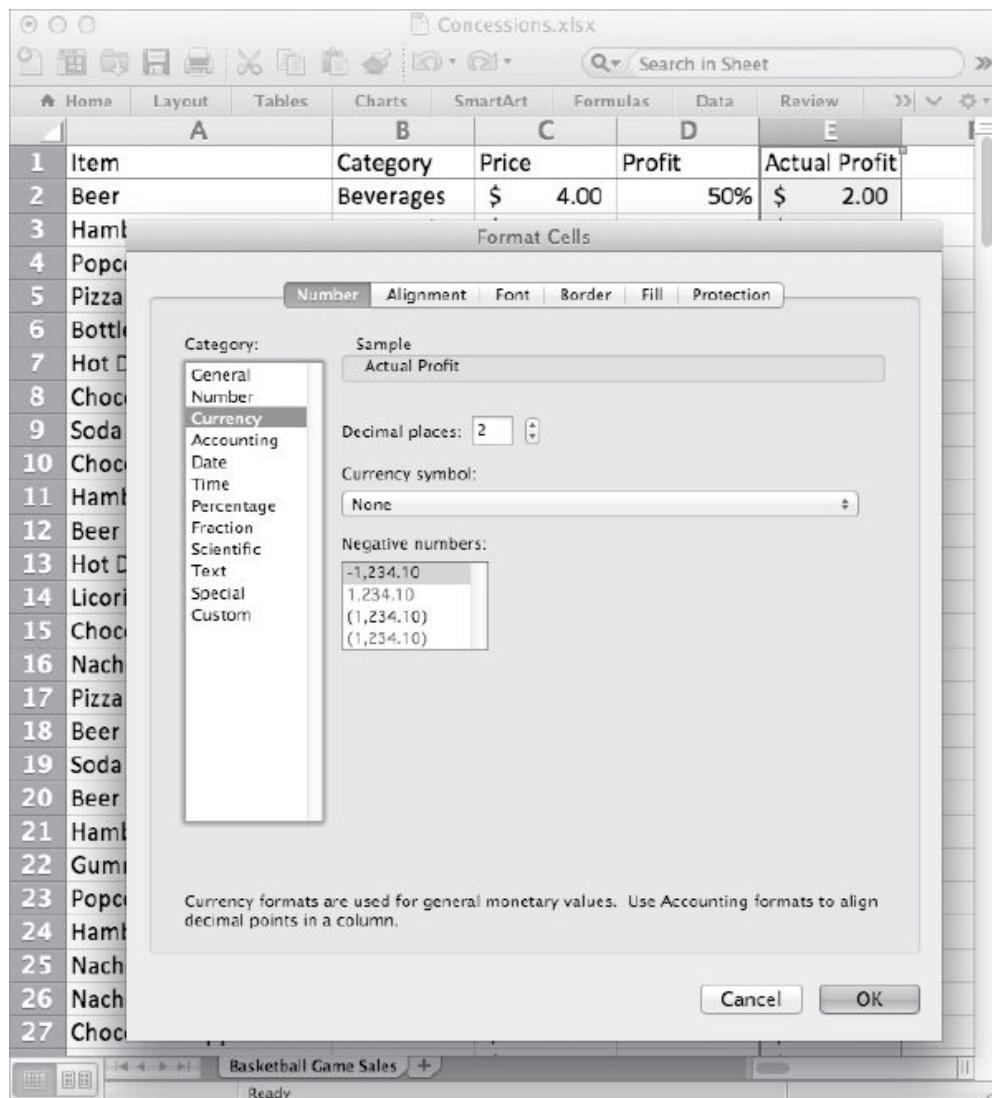
Então, quando copiar a fórmula, nada mudará. Ela continuará a referenciar a linha 2.

No entanto, se copiar a fórmula para a direita, C torna-se D, D torna-se E e assim por diante. Se não quiser tal comportamento, também será necessário adicionar um \$ à frente das referências da coluna. Isso é chamado de **referência absoluta**, o oposto de **referência relativa**.

## Formatando Células

O Excel oferece opções estáticas e dinâmicas para formatar valores. Olhe a coluna E, a coluna Actual Profit que você criou. Selecione-a clicando em sua barra cinza. Então clique com o botão direito na seleção e escolha Format Cells.

A partir do menu Format Cells, pode-se dizer ao Excel o tipo de número a ser encontrado na coluna E. Neste caso, você quer que seja Currency. E pode definir a quantidade de casas decimais. Deixe com duas casas decimais, como mostra a Figura 1-5. Em Format Cells também estão disponíveis opções para mudar as cores da fonte, o alinhamento de texto, as bordas, o preenchimento e assim por diante.



**Figura 1-5:** O menu Format Cells

Chegamos em um enigma. Mas se você quiser formatar apenas as células que possuem um certo valor ou uma variação de valores nelas? E se você quiser que aquela formatação mude com os valores?

Isso é chamado *formatação condicional* e este livro faz uso liberal dela.

Cancelar o menu Format Cells e navegue até a aba Home. Na seção Styles (Mac chama de Format), encontrará o botão Conditional Formatting (veja a Figura 1-6). Clique no botão para ver um menu suspenso de opções. A formatação condicional mais usada nesse texto é a Color Scales. Selecione uma escala para a coluna E e repare como cada célula na coluna é colorida de acordo com seu valor alto ou baixo.

	Item	Category	Price	Profit	Actual P
1	Beer	Beverages	\$ 4.00	50%	\$ 2.00
2	Hamburger	Hot Food	\$ 3.00	67%	\$ 2.01
3	Popcorn	Hot Food	\$ 5.00		
4	Pizza	Hot Food	\$ 2.00		
5	Bottled Water	Beverages	\$ 3.00		
6	Hot Dog	Hot Food	\$ 1.50		
7	Chocolate Dipped Cone	Frozen Treat	\$ 3.00		
8	Soda	Beverages	\$ 2.50		
9	Chocolate Bar	Candy	\$ 2.00		
10	Hamburger	Hot Food	\$ 3.00	67%	\$ 2.01
11	Beer	Beverages	\$ 4.00	50%	\$ 2.00
12	Hot Dog	Hot Food	\$ 1.50	67%	\$ 1.00
13					

**Figura 1-6:** Aplicando Conditional Formatting ao lucro

Para remover a formatação condicional, use a opções Clear Rules abaixo do menu Conditional Formatting.

## Colando Valores Especiais

É do seu próprio interesse não ter uma fórmula sem utilidade como você vê na Coluna E na Figura 1-4. Se estiver usando a fórmula `RAND()` para gerar um valor aleatório, por exemplo, ela muda toda vez que a planilha se autorrecalcula, o que é ótimo mas também pode ser extremamente irritante. A solução é copiar e colar novamente essas células na planilha como valores fixos.

Para converter fórmulas em valores, simplesmente copie a coluna preenchida com fórmulas (pegue a coluna E) e cole de volta usando a opção Paste Special (encontrada na aba Home abaixo da opção Paste no Windows e sob o menu Edit no Mac). Na janela Paste Special, escolha paste as values (veja Figura 1-7). Repare também que Paste Special lhe

permite *transpor* os dados de vertical para horizontal e vice-versa. Isso será bastante usado nos capítulos a seguir.

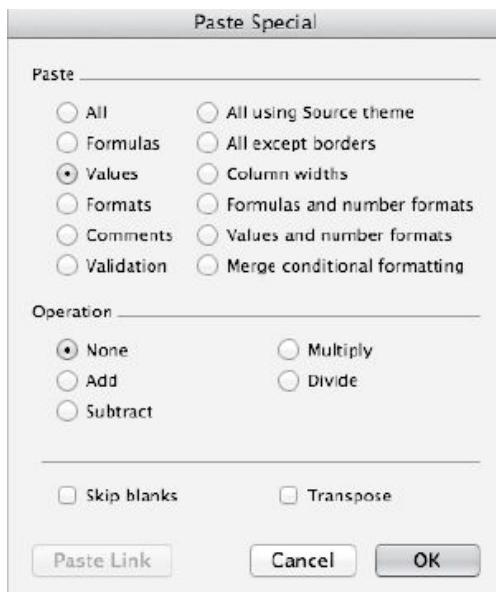


Figura 1-7: A janela Paste Special no Excel 2011

## Inserindo Gráficos

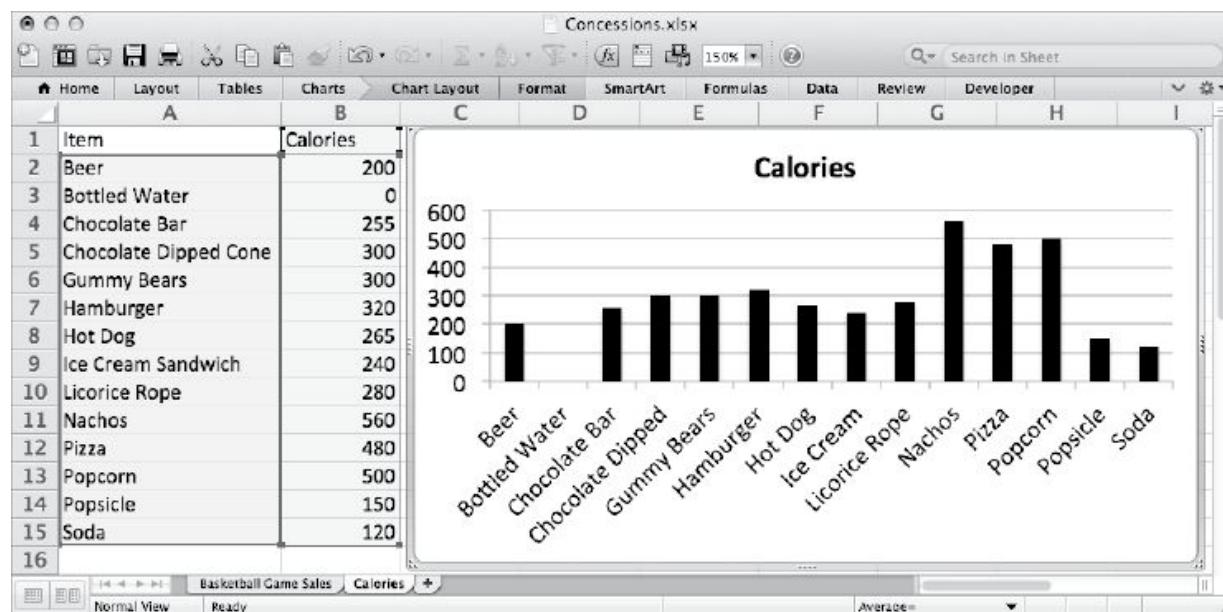
Na pasta de trabalho da cantina do colégio, também há uma aba chamada Calories com uma pequena tabela que mostra a contagem de calorias para cada item vendido. É possível modelar gráficos assim no Excel facilmente. Na aba Insert (Charts no Mac), há uma seção de gráficos que disponibiliza opções diferentes de visualização, como gráficos de barra, de linha e de pizza.

### NOTA

Neste livro, usaremos, em sua maioria, gráficos de coluna, de linha e de dispersão. Nunca seja pego usando um gráfico pizza. E, especialmente, nunca use os gráficos pizza 3D que o Excel oferece ou meu fantasma lhe assombrará pessoalmente quando eu morrer. Eles são feios, não comunicam bem os dados e os efeitos 3D possuem menos valores estéticos do que as pinturas de conchas na parede do meu dentista.

Realçando as colunas A:B na aba Calories, você pode selecionar um gráfico de Clustered Column para visualizar os dados. Brinque com o gráfico. Você pode clicar com o botão direito nas seções para apresentar menus de formatação. Por exemplo, ao clicar com o botão direito do mouse nas barras, é possível selecionar “Format Data Series...” no qual é possível modificar a cor de preenchimento das barras do padrão azul do Excel para qualquer tom que lhe agrade — preto, por exemplo.

Não há porque manter a legenda padrão, então você deveria selecioná-la e clicar em apagar para removê-la. Também é possível selecionar várias seções de texto no gráfico e aumentar o tamanho de suas fontes (o tamanho da fonte está sob a aba Home no Excel). O gráfico da Figura 1-8 é resultado disso.



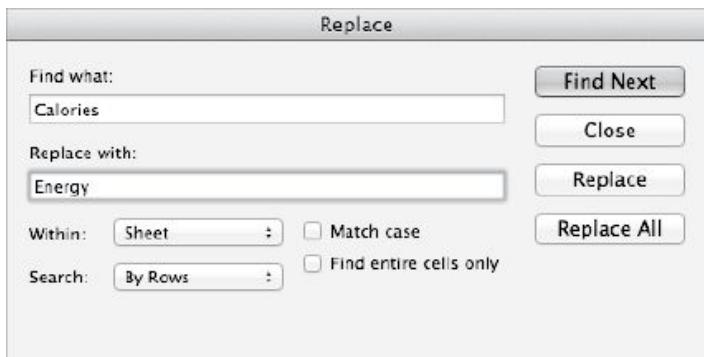
**Figura 1-8:** Inserindo um gráfico de colunas de calorias

## Encontrando os Menus Find e Replace

Você usará bastante localizar e substituir neste livro. No Windows, é possível pressionar Ctrl+F para abrir a janela Find (Ctrl+H para Replace) ou navegar até a aba Home e usar o botão Find na seção Edit. No Mac, há um campo de pesquisa na parte superior direita da planilha (pressione a

seta para baixo para o menu Replace), ou apenas pressione Cmd+F para ver o menu Find and Replace.

Apenas como teste, abra o menu substituir da planilha Calories. Pode-se substituir cada ocorrência da palavra “Calories” pela palavra “Energy” (veja a Figura 1-9) digitando as palavras na janela Find and Replace e selecionando Replace All.



**Figura 1-9:** Localizando e Substituindo

## Fórmulas para Localizar e Retirar Valores

Se eu não presumisse que você ao menos soubesse algumas fórmulas em Excel (`SUM`, `MAX`, `MIN`, `PERCENTILE` e assim por diante), nós ficaríamos aqui o dia todo. E eu quero começar logo. Mas existem algumas fórmulas muito utilizadas neste livro que você provavelmente ainda não usou, a não ser que tenha pesquisado muito o maravilhoso mundo das planilhas. Tais fórmulas lidam com *encontrar um valor em uma faixa e retornar sua localização* ou, em contrapartida, *encontrar uma localização em uma faixa e retornar seu valor*.

Quero falar um pouco sobre isso na aba Calories.

Às vezes você quer saber a localização de certo elemento na ordem de uma coluna ou linha. É o primeiro, segundo, terceiro? A fórmula `MATCH` resolve isso muito bem. Abaixo dos seus dados de calorias, nomeie A18 como **Match**. É possível implementar a fórmula em B18 para encontrar o item “Hamburger” na lista acima. Para usar a fórmula, forneça um valor

para pesquisar, uma faixa onde procurar e um 0 para forçá-la a lhe retornar à posição da palavra-chave:

```
=MATCH( "Hamburger" , A2 : A15 , 0 )
```

O resultado é 6, pois “Hamburger” é o sexto item na lista (veja a Figura 1-10).

Em seguida temos a fórmula `INDEX`. Nomeie A19 como **Index**.

Essa fórmula pega uma faixa de valores e um número de linha e de coluna e retorna o valor naquela localização. Por exemplo, você pode introduzir na fórmula `INDEX` nossa tabela de calorias A1:B15 e recuperar a contagem de calorias para bottled water, informe 3 linhas para baixo e 2 colunas a mais:

```
=INDEX( A1 : B15 , 3 , 2 )
```

Isso resulta em uma contagem de calorias de 0 como esperado (veja a Figura 1-10).

Outra fórmula que verá bastante nesse texto é a `OFFSET`. Prossiga e nomeie A20 como **Offset** e você pode brincar com a fórmula em B20.

Com essa fórmula, você fornece uma faixa que age como um cursor que é movido pelas linhas e colunas deslocadas (parecido com `INDEX` para o caso de único valor exceto o baseado em 0). Por exemplo, forneça à `OFFSET` uma referência ao topo esquerdo da planilha, A1, e então retire o valor 3 células abaixo produzindo uma linha deslocada de 3 e uma coluna deslocada de 0:

```
=OFFSET( A1 , 3 , 0 )
```

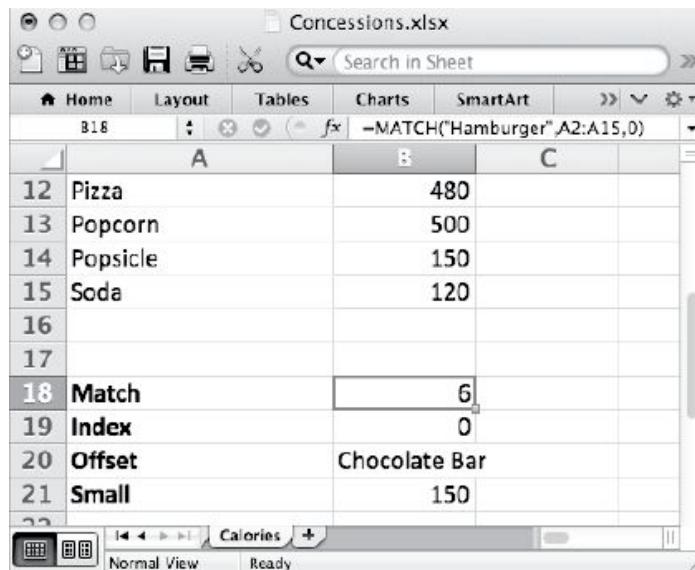
Isso retorna o nome do terceiro item na lista, “Chocolate Bar” Veja a Figura 1-10.

A última fórmula que quero ver nessa seção é a `SMALL` (sua correspondente se chama `LARGE` e funciona da mesma forma). Se possuir uma lista de valores e quiser retornar o terceiro menor, `SMALL` faz isso para você. Para ver como, nomeie A21 como **Small** e forneça a B21 uma lista de contagem de calorias e um índice de 3.

```
=SMALL( B2 : B15 , 3 )
```

Isso retorna um valor de 150 que é o terceiro menor depois de 0 (bottled water) e 120 (soda). Veja a Figura 1-10.

Bem, há mais uma fórmula usada para procurar valores que é como uma `MATCH` com esteroides, ela é a `VLOOKUP` (e sua correspondente horizontal é `HLOOKUP`). Essa possui sua própria seção em seguida pois ela é animal.



The screenshot shows a Microsoft Excel spreadsheet titled "Concessions.xlsx". The table has columns A, B, and C. Row 12 contains "Pizza" and "480". Row 13 contains "Popcorn" and "500". Row 14 contains "Popsicle" and "150". Row 15 contains "Soda" and "120". Row 16 is blank. Row 17 is blank. Row 18 contains "Match" and "6". Row 19 contains "Index" and "0". Row 20 contains "Offset" and "Chocolate Bar". Row 21 contains "Small" and "150". Row 22 is blank. The formula in cell B18 is `=MATCH("Hamburger",A2:A15,0)`. The formula in cell B19 is `=VLOOKUP("Hamburger",A2:C15,2,0)`. The formula in cell B20 is `=VLOOKUP("Hamburger",A2:C15,3,0)`. The formula in cell B21 is `=VLOOKUP("Hamburger",A2:C15,1,0)`.

	A	B	C
12	Pizza	480	
13	Popcorn	500	
14	Popsicle	150	
15	Soda	120	
16			
17			
18	Match	6	
19	Index	0	
20	Offset	Chocolate Bar	
21	Small	150	
22			

Figura 1-10: Fórmulas que você deve aprender

## Usando VLOOKUP para Juntar Dados

Retorne à sua aba Basketball Game Sales. Você ainda pode fazer referência a uma célula da aba anterior, Calories, simplesmente colocando o nome da aba e “!” à frente da célula referenciada. Por exemplo, `Calories!B2` é uma referência a calorias em beer, independente da planilha na qual esteja trabalhando.

Agora, e se você quisesse jogar os dados de calorias em uma coluna na planilha de vendas para que a cada item vendido a contagem de calorias estivesse listada ao lado? De alguma forma você teria que procurar a contagem de calorias para cada item vendido e colocá-la em uma coluna próxima da transação. Bom, acontece que há uma fórmula para isso chamada `VLOOKUP`.

Vá em frente e nomeie a Coluna F na planilha **Calories** com esse propósito. A célula F2 incluirá a contagem de calorias da primeira transação de beer na tabela Calories. Usando a fórmula VLOOKUP forneça o nome do item da célula A2, uma referência à tabela `Calories!$A$1:$B$15`, e à coluna deslocada de onde quer que seu valor de retorno seja lido, que é a segunda coluna:

```
=VLOOKUP(A2, Calories!$A$1:$B$15, 2, FALSE)
```

O FALSE ao final da fórmula VLOOKUP significa que você não aceitará resultados próximos a “Beer”. Se a fórmula não consegue encontrar “Beer” na tabela de calorias, ela retorna um erro.

Ao entrar com a fórmula, você pode ver que o dado 200 calorias é lido da tabela na aba Calories. Como você colocou o \$ na frente das referências da tabela na fórmula, é possível copiar essa fórmula na coluna abaixo dando um clique duplo no canto inferior direito da célula. *Voilà!* Como mostra a Figura 1-11, você possui contagens de calorias para cada transação.

	A	B	C	D	E	F
1	Item	Category	Price	Profit	Actual Profit	Calories
2	Beer	Beverages	\$ 4.00	50%	\$ 2.00	200
3	Hamburger	Hot Food	\$ 3.00	67%	\$ 2.00	320
4	Popcorn	Hot Food	\$ 5.00	80%	\$ 4.00	500
5	Pizza	Hot Food	\$ 2.00	25%	\$ 0.50	480
6	Bottled Water	Beverages	\$ 3.00	83%	\$ 2.50	0

Figura 1-11: Usando VLOOKUP para capturar as contagens de calorias.

## Filtrando e Ordenando

Agora que você tem as calorias, digamos que queira ver apenas aquelas transações da categoria Frozen Treats. O que deseja fazer é filtrar a planilha. Para fazer isso, primeiro selecione os dados no intervalo A1:F200. Você pode colocar o cursor em A1 e pressionar Shift+Ctrl+↓ e

então! . Um método ainda mais fácil é clicar no topo da coluna A e manter o botão pressionado conforme move o mouse pela coluna F para realçar todas as seis colunas.

Então, para colocar uma filtragem nessas seis colunas, pressione o botão Filter na seção Data na faixa de opções. Ele parece um funil cinza, conforme mostra a Figura 1-12.

1	Item	Category	Price	Profit	Actual Profit	Calories
196	Pizza	Hot Food	\$ 2.00	25%	\$ 0.50	480
197	Popsicle	Frozen Treat	\$ 3.00	83%	\$ 2.50	150
198	Chocolate Bar	Candy	\$ 2.00	75%	\$ 1.50	255
199	Bottled Water	Beverages	\$ 3.00	83%	\$ 2.50	0
200	Popsicle	Frozen Treat	\$ 3.00	83%	\$ 2.50	150
201						

Figura 1-12: Coloque a filtragem automática no intervalo selecionado.

Quando a filtragem automática é ativada, você pode clicar no menu suspenso que aparece na célula B1 e escolher para exibir somente certas categorias (neste caso, apenas transações Frozen Treats serão exibidas). Veja a Figura 1-13.

Depois de filtrar, realçar as colunas de dados permite que a barra de status no Excel lhe dê informações apenas sobre as células que permaneceram. Por exemplo, ao filtrar Frozen Treats, podemos realçar os valores na Coluna E e usar a barra de status para obter um total de lucro rápido daquela categoria. Veja a Figura 1-14.

Concessions.xlsx

The screenshot shows a Microsoft Excel spreadsheet titled "Concessions.xlsx". A filter dialog box is open over a table of data. The table has columns: Item, Category, Price, Profit, Actual Profit, and Calories. The filter dialog shows "Sort" options (A↓ Ascending, Z↓ Descending), a "By color:" dropdown set to "None", and a "Filter" section. In the "Filter" section, there is a dropdown set to "Equals 3.00", a dropdown set to "Frozen Treats", and radio buttons for "And" and "Or". Below these are checkboxes for "(Select All)", "Beverages", "Candy", and "Frozen Treats", with "Frozen Treats" checked. At the bottom right of the dialog is a "Clear Filter" button.

	A	B	C	D	E	F
1	Item	Category	Price	Profit	Actual Prof	Calories
127	Ice Cream Sandwich	Frozen Treat	\$ 3.00	67%	\$ 2.00	240
132	Ice Cream Sandwich	Frozen Treat	\$ 3.00	67%	\$ 2.00	240
150	Popsicle	Frozen Treat	\$ 2.50	150%	\$ 2.50	150
156	Popsicle	Frozen Treat	\$ 2.50	150%	\$ 2.50	150
164	Popsicle	Frozen Treat	\$ 2.50	150%	\$ 2.50	150
166	Popsicle	Frozen Treat	\$ 2.50	150%	\$ 2.50	150
169	Popsicle	Frozen Treat	\$ 2.50	150%	\$ 2.50	150
175	Popsicle	Frozen Treat	\$ 2.50	150%	\$ 2.50	150
176	Popsicle	Frozen Treat	\$ 2.50	150%	\$ 2.50	150
177	Popsicle	Frozen Treat	\$ 2.50	150%	\$ 2.50	150
179	Popsicle	Frozen Treat	\$ 2.50	150%	\$ 2.50	150
184	Popsicle	Frozen Treat	\$ 2.50	150%	\$ 2.50	150
190	Popsicle	Frozen Treat	\$ 2.50	150%	\$ 2.50	150
197	Popsicle	Frozen Treat	\$ 2.50	150%	\$ 2.50	150
200	Popsicle	Frozen Treat	\$ 2.50	150%	\$ 2.50	150

Figura 1-13: Filtrando item por categoria

Concessions.xlsx

The screenshot shows a Microsoft Excel spreadsheet titled "Concessions.xlsx". A summary dialog box is open at the bottom right, showing "Sum= \$ 69.00". The main table below it has rows for Popsicles, all categorized as "Frozen Treat". The "Calories" column shows values of 150 for each row. The "Actual Profit" column shows values of \$ 2.50 for each row. The "Sum" dialog box has an arrow pointing to the "Sum= \$ 69.00" text.

	A	B	C	D	E	F	G
1	Item	Category	Price	Profit	Actual Profit	Calories	
179	Popsicle	Frozen Treat	\$ 3.00	83%	\$ 2.50	150	
184	Popsicle	Frozen Treat	\$ 3.00	83%	\$ 2.50	150	
190	Popsicle	Frozen Treat	\$ 3.00	83%	\$ 2.50	150	
197	Popsicle	Frozen Treat	\$ 3.00	83%	\$ 2.50	150	
200	Popsicle	Frozen Treat	\$ 3.00	83%	\$ 2.50	150	

Figura 1-14: Resumindo uma categoria filtrada

A filtragem automática também permite que você ordene. Por exemplo, se você quer ordenar por lucro, simplesmente clique no menu filtragem automática na célula Profit (D1) e selecione Sort Ascending (ou “Smallest to Large” em algumas versões). Veja a Figura 1-15.

	A	B	C	D	E	F
1	Item	Category	Price	Profit	Actual Profit	Calories
8	Chocolate Dipped Cone	Frozen Treat:	\$ 3.00	50%	\$ 1.50	300
15	Chocolate Dipped Cone	Frozen Treat:	\$ 3.00	50%	\$ 1.50	300
27	Chocolate Dipped Cone	Frozen Treat:	\$ 3.00	50%	\$ 1.50	300
31	Chocolate Dipped Cone	Frozen Treat:	\$ 3.00	50%	\$ 1.50	300
34	Chocolate Dipped Cone	Frozen Treat:	\$ 3.00	50%	\$ 1.50	300
41	Chocolate Dipped Cone	Frozen Treat:	\$ 3.00	50%	\$ 1.50	300
48	Chocolate Dipped Cone	Frozen Treat:	\$ 3.00	50%	\$ 1.50	300
56	Chocolate Dipped Cone	Frozen Treat:	\$ 3.00	50%	\$ 1.50	300
67	Chocolate Dipped Cone	Frozen Treat:	\$ 3.00	50%	\$ 1.50	300
71	Chocolate Dipped Cone	Frozen Treat:	\$ 3.00	50%	\$ 1.50	300
81	Chocolate Dipped Cone	Frozen Treat:	\$ 3.00	50%	\$ 1.50	300
84	Ice Cream Sandwich	Frozen Treat:	\$ 3.00	67%	\$ 2.00	240
89	Ice Cream Sandwich	Frozen Treat:	\$ 3.00	67%	\$ 2.00	240
91	Ice Cream Sandwich	Frozen Treat:	\$ 3.00	67%	\$ 2.00	240
98	Ice Cream Sandwich	Frozen Treat:	\$ 3.00	67%	\$ 2.00	240
103	Ice Cream Sandwich	Frozen Treat:	\$ 3.00	67%	\$ 2.00	240
111	Ice Cream Sandwich	Frozen Treat:	\$ 3.00	67%	\$ 2.00	240

Figura 1-15: Organizando por Profit em ordem crescente

Para remover todos os filtros aplicados, volte à Category no menu de filtro e marque outras caixas, ou desative o botão de filtro na faixa de opções que pressionou primeiramente. Você verá que mesmo tendo todos os seus dados de volta, Frozen Treats ainda estará na ordem que dispôs.

O Excel também oferece a interface Sort para fazer ordenações mais complexas do que a filtragem automática. Para usar tal ferramenta, realce os dados a serem ordenados (segure A:F novamente) e selecione Sort na seção Sort & Filter da aba Data do Excel. Isso trará o menu de ordenação. No Mac, para chegar nessa janela, você deve pressionar a seta para baixo no botão de ordenação e selecionar Custom Sort...

No menu de ordenação, mostrado na Figura 1-16, é possível notar se seus dados possuem coluna de cabeçalho ou não, e, se possuem cabeçalhos como este exemplo, então você pode selecionar, por nome, as colunas a serem ordenadas.

Agora, a melhor parte dessa interface de ordenação é que abaixo do botão “Options...” é possível selecionar ordenação da esquerda para a direita em vez de coluna. Isso é algo que você não consegue fazer com filtragem automática. Do início ao fim deste livro, você precisará ordenar os dados aleatoriamente por colunas e linhas em dois passos rápidos, e essa interface será sua amiga. Por enquanto, apenas cancele já que os dados estão ordenados como você queria.

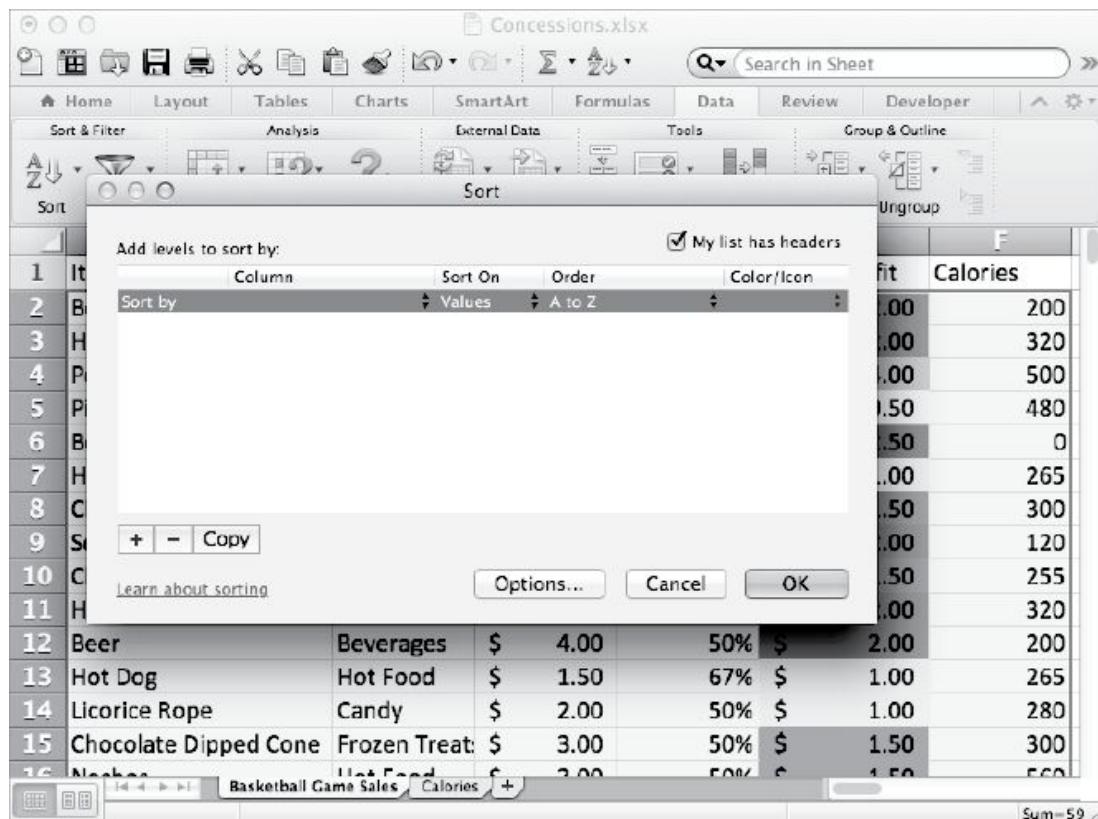


Figura 1-16: Utilizando o menu Sort

## Usando PivotTables (Tabelas Dinâmicas)

E se você quisesse saber a contagem total de cada tipo de item vendido? Ou o rendimento total por item?

Essas questões são semelhantes a consultas “agregar” ou “agrupar por” que você executaria em tradicionais bancos de dados SQL. Mas esses dados não estão em um banco, e sim em uma planilha. É aí que PivotTables vem ao resgate.

Assim como quando filtrou seus dados, comece selecionando os dados que quer manipular — neste caso, os dados estão na seleção A1:F:200. A partir da aba Insert (Data no Mac), é possível pressionar o botão PivotTable e selecionar para que o Excel crie uma nova planilha com PivotTable. Apesar de algumas versões do Excel permitirem que você insira uma PivotTable em uma planilha existente, o padrão é selecionar a opção nova planilha, a não ser que você tenha um excelente motivo para não fazer isso.

Nessa nova planilha, o PivotTable Builder será alinhado à direta da tabela (ele flutua no Mac). O builder (construtor) permite que você pegue as colunas dos dados originais selecionados e as use como filtro de relatório, nomes de colunas e linhas para agrupamentos, ou valores. Um filtro de relatório é similar ao filtro da seção anterior — ele permite que você selecione apenas um subconjunto dos dados, como Frozen Treats. Column Labels e Row Labels preenchem o relatório da PivotTable com valores distintos dos valores das colunas selecionadas.

No Windows, a PivotTable inicial estará completamente vazia, enquanto no Mac geralmente estará pré-preenchida com valores diferentes daqueles da primeira coluna selecionada descendo pelas linhas da tabela e valores diferentes dos da segunda coluna através das colunas. Se você estiver em um Mac, vá em frente e desmarque todas as caixas no Builder, para que possa trabalhar a partir de uma tabela vazia.

Agora, digamos que queira saber o rendimento total por item. Para chegar lá, arraste o bloco Item no Builder para a seção Rows e o bloco Price para a seção Values. Isso significa que trabalhará nos rendimentos pelo nome do item.

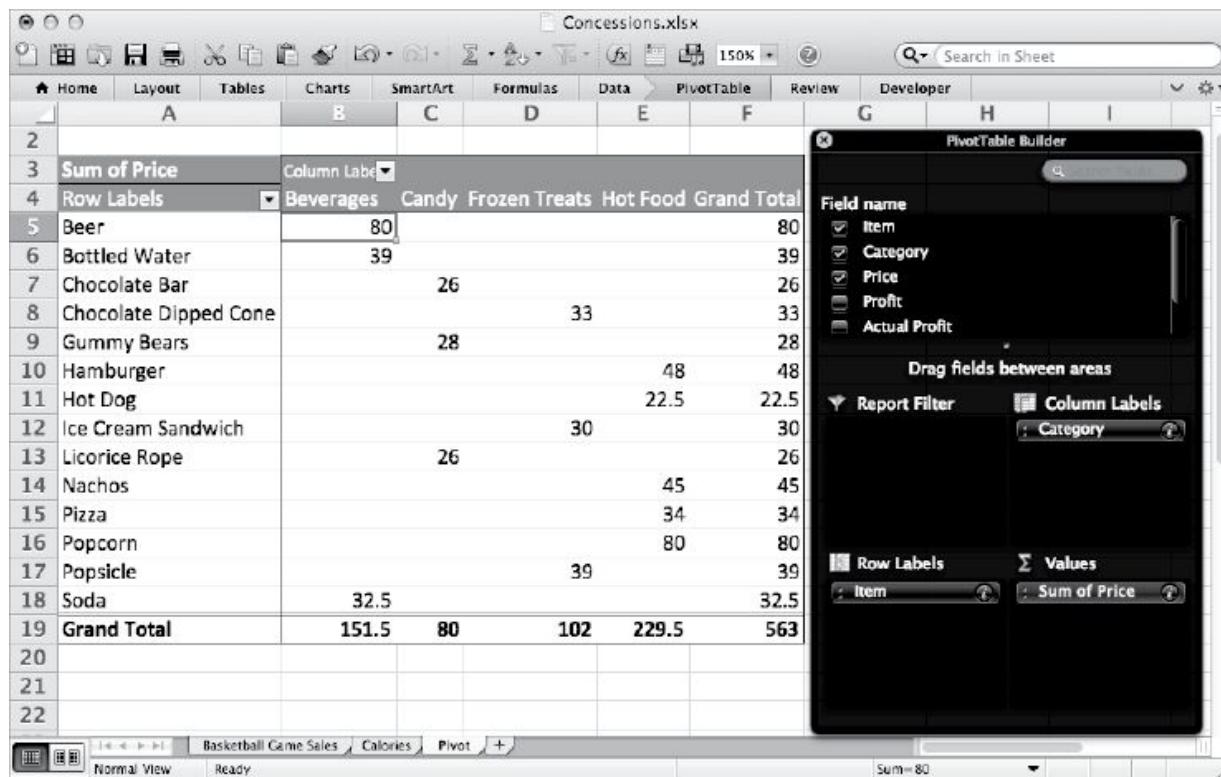
Inicialmente, todavia, a PivotTable é configurada para meramente contar o número de registro de preços que estão em um grupo. Por exemplo, existem 20 linhas de Beer. Veja a Figura 1-17.



Figura 1-17: O PivotTable Builder e uma contagem de vendas por item

Para examinar o rendimento será preciso mudar a conta para uma soma. Para fazer isso, no Windows, abra o menu suspenso no bloco Price na seção Values do Builder e selecione “Value Field Settings...”. No Mac, pressione o pequeno botão “i”. A partir de então, “sum” pode ser selecionado a partir de várias opções.

E se você quiser dividir essas somas em categorias? Para isso, arraste o bloco Category na seção Columns do builder. Isso gera a tabela exibida na Figura 1-18. Note que a PivotTable na figura totaliza automaticamente colunas e linhas para você.



**Figura 1-18:** Rendimento por item e categoria

E se quiser remover algo da tabela, apenas desmarque-o ou pegue o bloco da seção onde está e arraste-o para fora da planilha, como se estivesse jogando fora. Vá em frente e remova o bloco Category.

Uma vez que consiga o relatório desejado em uma PivotTable, é sempre possível selecionar os valores e colá-los em outra planilha para trabalhar posteriormente. Neste exemplo, você pode copiar a tabela (A5:B18 no Mac) e Paste Special (Colar Especial) seus valores em uma nova aba chamada Revenue By Item (veja a Figura 1-19).

Sinta-se à vontade para trocar vários nomes de colunas e linhas até que entenda o que está acontecendo. Por exemplo, tente chegar na contagem total de calorias por categoria usando uma PivotTable.

	Row Labels	Sum of Price
2	Beer	80
3	Bottled Water	39
4	Chocolate Bar	26
5	Chocolate Dipped Cone	33
6	Gummy Bears	28
7	Hamburger	48
8	Hot Dog	22.5
9	Ice Cream Sandwich	30
10	Licorice Rope	26
11	Nachos	45
12	Pizza	34
13	Popcorn	80
14	Popsicle	39
15	Soda	32.5
16		

Figura 1-19: A aba Revenue by Item criada ao colar valores de uma PivotTable

## Usando Fórmulas Array

Na planilha em questão, há uma aba chamada Fee Schedule. Acontece que o técnico O'Shaughnessy o deixaria operar a cantina do colégio se você conseguisse algum lucro para ele (talvez para custear seu hábito de compras de meias). A aba Fee Schedule mostra o percentual que ele recebe em cada item vendido.

Então quanto você deve a ele pelo jogo de ontem à noite? Para responder essa pergunta, é preciso multiplicar o rendimento total de cada item da PivotTable pela porcentagem do técnico e somar tudo.

Existe uma excelente fórmula para essa operação que fará toda a multiplicação e soma de uma vez. Nomeada sem criatividade nenhuma, é chamada **SUMPRODUCT**. Na célula E1 na planilha Revenue By Item, adicione o nome **Total Cut for Coach**. Na célula C2, determine a **SUMPRODUCT** do rendimento e as taxas adicionando essa fórmula:

```
=SUMPRODUCT(B2:B15, 'Fee Schedule'!B2:O2)
```

Ah, não. Há um erro; a célula apenas lê #value. O que está dando errado?

Mesmo que você tenha selecionado os dois intervalos de tamanhos iguais e colocado em SUMPRODUCT, a fórmula não consegue ver que os intervalos são iguais porque uma está na vertical e a outra na horizontal.

Felizmente, o Excel tem uma função para virar os arrays na direção correta. É chamada TRANSPOSE. A fórmula deve ser escrita da seguinte maneira:

```
=SUMPRODUCT(B2:B15, TRANSPOSE('Fee Schedule'!B2:O2))
```

O motivo de ainda causar erro é que toda fórmula em Excel, por padrão, retorna um valor único. Até TRANSPOSE retorna o primeiro valor no array transposto. Se você quer que *todo o array* seja retornado, precisa transformar TRANSPOSE em uma “fórmula array”, que significa exatamente o que você deve ter pensado. Fórmulas array retornam arrays, e não valores unitários.

Não é necessário mudar a forma como se escreve SUMPRODUCT para que isso aconteça. Tudo o que precisa fazer é pressionar Ctrl+Shift+Enter ao terminar de digitar a fórmula, em vez de pressionar Enter. No Mac, você usa Command+Return.

Sucesso! Como exibido na Figura 1-20, o novo cálculo lê \$57,60. Mas eu sugiro arredondar para \$50,00, porque de quantas meias o técnico realmente precisa?

	A	B	C	D	E
1	Row Labels	Sum of Price		Total Cut for Coach:	
2	Beer	80		57.6	
3	Bottled Water	39			
4	Chocolate Bar	26			

Figura 1-20: SUMPRODUCT com uma fórmula array

# Resolvendo Coisas com o Solver

Muitas das técnicas que você estudará neste livro podem ser resumidas em *modelos de otimização*. Um problema de otimização é um no qual você precisa tomar a melhor decisão (escolher os melhores investimentos, minimizar os custos da sua empresa, encontrar o cronograma com a menor quantidade de aulas pela manhã, e por aí vai). Portanto, em modelos de otimização, as palavras “minimize” e “maximize” aparecem muito quando articulamos um objetivo.

Em data science, muitas das práticas, seja em inteligência artificial, mineração de dados, ou previsão, são apenas alguns dados preparados mais um modelo adequado que na realidade é um modelo de otimização. Então faria sentido ensinar otimização primeiro. Mas aprender tudo o que existe sobre otimização é difícil para ser feito subitamente. Logo, você fará um estudo profundo sobre otimização no Capítulo 4 **após** resolver mais alguns problemas divertidos de aprendizado de máquina nos Capítulo 2 e 3. Para preencher as lacunas enquanto isso, será melhor se você praticar um pouco com otimização agora. Só uma amostra.

No Excel, problemas de otimização são resolvidos usando um Add-In que está disponível com o Excel, chamado Solver.

- No Windows, o Solver pode ser adicionado indo em File (no Excel 2007 é o botão à esquerda do botão do Windows no topo) Options Add-ins, e abaixo do menu suspenso Manage escolha Excel Add-ins e pressione Go. Marque a caixa do Solver Add-in e pressione em OK.
- No Mac, o Solver é adicionado indo em Tools Add-ins e selecionando Solver.xlam no menu.

Um botão Solver aparecerá na seção Analysis da aba Data em todas as versões.

Tudo certo! Agora que o Solver está instalado, eis um problema de otimização: Falaram que você precisa de 2.400 calorias por dia. Qual é o

menor número de itens que você pode comprar na cantina do colégio para chegar nessa quantidade? Obviamente, você poderia comprar 10 ice cream sandwiches com 240 calorias cada, mas existe uma forma de fazer isso com menos itens?

O Solver pode lhe dizer isso!

Para começar, faça uma cópia da planilha Calories, nomeie a planilha como **Calories-Solver**, e limpe tudo menos a tabela de calorias na cópia. Se você não sabe fazer uma cópia de uma planilha no Excel, simplesmente dê um clique duplo na aba que gostaria de copiar e selecione o menu Move or Copy. Isso gera a nova planilha exibida na Figura 1-21.

Para fazer o Solver funcionar, você precisa fornecer uma seleção de células que ele pode configurar com decisões. Neste caso, o Solver precisa decidir quanto comprar de cada item. Então, na Coluna C próxima à contagem de calorias, nomeie a coluna como **How many?** (ou o que desejar), e poderá permitir que o Solver armazene suas decisões nessa coluna.

O Excel considera células em branco como 0, então não é preciso preencher essas células para começar. O Solver fará isso para você.

	A	B
1	Item	Calories
2	Beer	200
3	Bottled Water	0
4	Chocolate Bar	255
5	Chocolate Dipped Cone	300
6	Gummy Bears	300
7	Hamburger	320
8	Hot Dog	265
9	Ice Cream Sandwich	240
10	Licorice Rope	280
11	Nachos	560
12	Pizza	480
13	Popcorn	500
14	Popsicle	150
15	Soda	120
		=SUM(C2:C15)

**Figura 1-21:** A planilha copiada para Calories-Solver

Na célula C16, some o número de itens acima a serem comprados como:

```
=SUM(C2:C15)
```

E abaixo disso você pode somar a contagem total de calorias desses itens (que você quer que eventualmente totalize 2.400) usando a fórmula SUMPRODUCT:

```
=SUMPRODUCT(B2:B15, C2:C15)
```

Isso gera a planilha inicial exibida na Figura 1-22.

Agora você está pronto para construir o modelo, então chame a janela do Solver pressionando o botão Solver na aba Data.

The screenshot shows a Microsoft Excel spreadsheet titled "Concessions.xlsx". The table has columns for "Item", "Calories", and "How many?". Row 16 contains the formula =SUMPRODUCT(B2:B15,C2:C15) in cell C17. Row 17 contains the results: Total Items: 0 and Total Calories: 0.

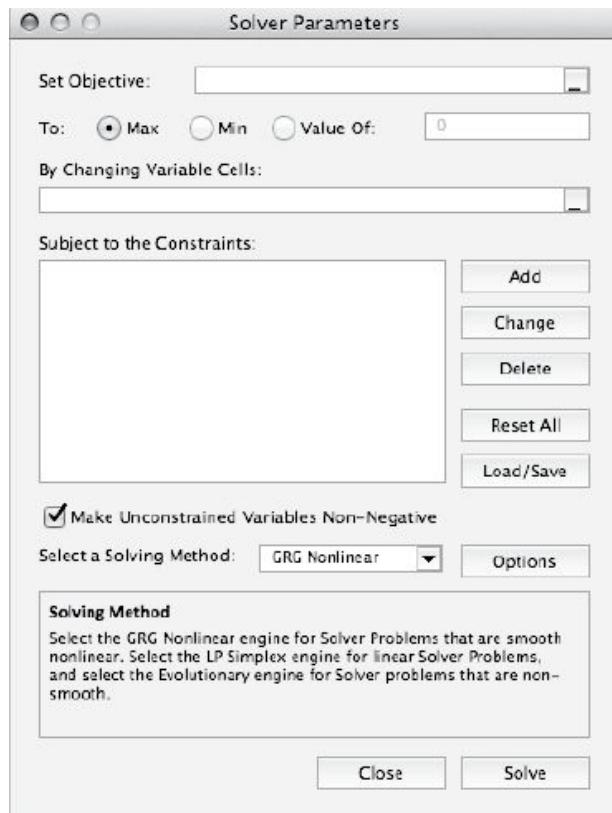
	A	B	C
1	Item	Calories	How many?
2	Beer	200	
3	Bottled Water	0	
4	Chocolate Bar	255	
5	Chocolate Dipped Cone	300	
6	Gummy Bears	300	
7	Hamburger	320	
8	Hot Dog	265	
9	Ice Cream Sandwich	240	
10	Licorice Rope	280	
11	Nachos	560	
12	Pizza	480	
13	Popcorn	500	
14	Popsicle	150	
15	Soda	120	
16		Total Items:	0
17		Total Calories:	0

**Figura 1-22:** Configurando as contagens de calorias e itens

#### NOTA

A janela do Solver, exibida na Figura 1-23 no Excel 2011, é bem similar no Excel 2010, 2011 e 2013. Na versão de 2007, a aparência é um pouco diferente, mas a única diferença considerável é que não existe caixa de seleção de algoritmo. No lugar disso, há uma caixa de marcação “Assume Linear Model” abaixo do menu Options. Aprenderemos sobre esses elementos mais tarde.

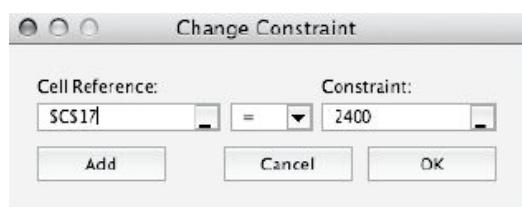
Os principais elementos ligados ao Solver para resolver um problema, conforme mostra a Figura 1-23, são uma célula objetiva, uma direção de otimização (maximização e minimização), algumas variáveis de decisão que podem ser modificadas pelo Solver, e algumas restrições.



**Figura 1-23:** A janela do Solver não inicializada

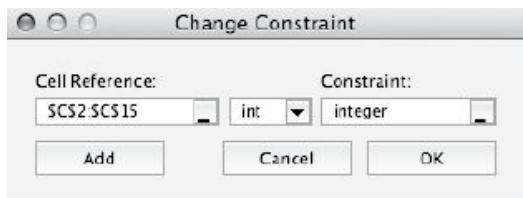
No seu caso, o objetivo é minimizar o total de itens na célula C16. As células que podem ser alteradas são as seleções de item em C2:C15. E as restrições são que C17, o total de calorias, precisa ser igual a 2.400. Além disso, precisaremos adicionar uma restrição para que nossas decisões contem números, então precisaremos marcar a caixa não-negativa (abaixo do menu de opções no Excel 2007) e acrescentar uma restrição de inteiro às decisões. Afinal, você não pode comprar 1,7 refrigerantes. Essas restrições de inteiros serão abordadas mais profundamente no Capítulo 4.

Para adicionar na restrição total de calorias, pressione o botão Add e coloque C17 igual a 2.400 conforme mostra a Figura 1-24.



**Figura 1-24:** Adicionando a restrição de calorias

Similarmente, adicione uma restrição configurando C2:C15 como inteiros conforme a Figura 1-25.



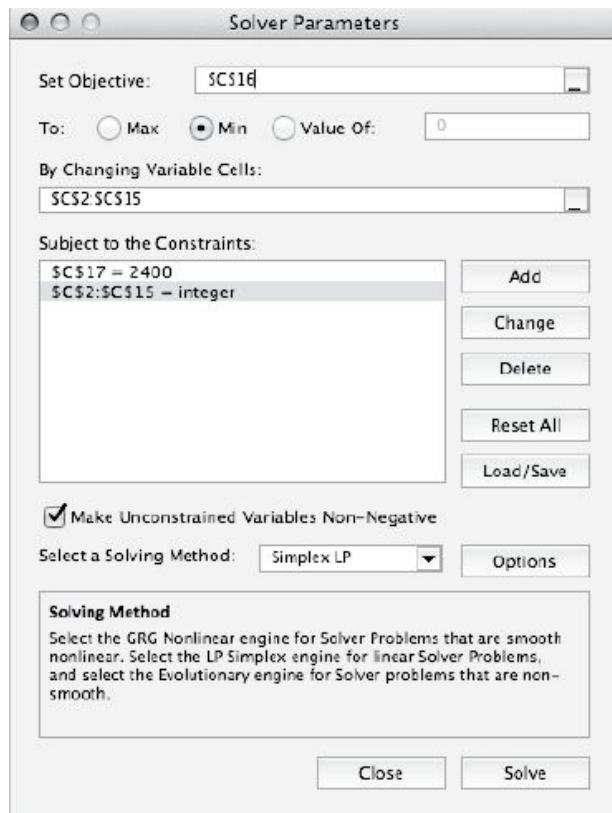
**Figura 1-25:** Adicionando uma restrição de inteiro

Pressione OK.

No Excel 2010, 2011 e 2013, certifique-se de que o método de solução esteja configurado como Simplex LP. Simplex LP é apropriado para este problema porque é *linear* (o “L” em LP significa linear, conforme verá no Capítulo 4). Com linear, quero dizer que o problema envolve somente combinações lineares de decisões de C2 à C15 (somas, produtos com constantes como contagens de calorias, etc.).

Se tivéssemos cálculos não-lineares no modelo (talvez a raiz quadrada de uma decisão, um logaritmo, ou uma função exponencial), então poderíamos usar um dos algoritmos que o Excel disponibiliza no Solver. O Capítulo 4 aborda isso em mais detalhes.

No Excel 2007, é possível denotar o problema como linear clicando em Assume Linear Model abaixo da tela Options. Sua configuração final deve aparecer como na Figura 1-26.



**Figura 1-26:** A configuração final do Solver para minimizar itens necessários para 2.400 calorias

Tudo certo! Vá em frente e pressione o botão Solve. O Excel deve encontrar uma solução quase de imediato. E a solução, como mostra a Figura 1-27, é 5. Agora, seu Excel deve selecionar 5 itens diferentes dos meus na tela, porém o mínimo é 5.

	A	B	C
1	Item	Calories	How many?
2	Beer	200	0
3	Bottled Water	0	0
4	Chocolate Bar	255	0
5	Chocolate Dipped Cone	300	0
6	Gummy Bears	300	0
7	Hamburger	320	0
8	Hot Dog	265	0
9	Ice Cream Sandwich	240	0
10	Licorice Rope	280	1
11	Nachos	560	2
12	Pizza	480	0
13	Popcorn	500	2
14	Popsicle	150	0
15	Soda	120	0
16	Total Items:		5
17	Total Calories:		2400
18			

The screenshot shows a Microsoft Excel spreadsheet titled "Concessions.xlsx". The data is organized in columns A, B, and C. Column A lists 15 food items from row 2 to row 16. Column B contains the calorie count for each item. Column C contains the number of items selected, which is 0 for most items except for Licorice Rope (1), Nachos (2), Popcorn (2), and Soda (0). Row 17 contains two formulas: "SUMPRODUCT(B2:B15,C2:C15)" in cell C17 and "=C17" in cell B17. The formula in C17 calculates the total calories (2400) by multiplying the calorie count by the number of items selected for each item. The formula in B17 copies the value from C17 into B17. The status bar at the bottom indicates "Calories-Solver".

Figura 1-27: A seleção de item otimizada

## OpenSolver: Eu Gostaria que Não Precisássemos Fazer Isso, mas Precisamos

Este livro foi projetado originalmente para trabalhar completamente com o Solver embutido do Excel. Entretanto, as funcionalidades foram *removidas* do Solver em versões mais recentes por razões misteriosas e inesperadas.

Isso significa que enquanto este livro inteiro trabalha usando o Solver no Excel 2007 e Excel 2011 para Mac, no Excel 2010 e 2013 o Solver embutido ocasionalmente reclamará que o modelo de otimização linear é muito grande (avisarei antecipadamente neste livro quando um modelo ficar complexo assim).

Por sorte, existe uma excelente ferramenta chamada OpenSolver disponível para as versões do Excel para Windows que é dedicada a essa deficiência. Com o OpenSolver, pode-se ainda construir o modelo na interface regular do Solver, mas o OpenSolver disponibiliza um botão que você pressiona para usar seu algoritmo de implementação Simplex LP, que é incrivelmente rápido.

Para configurar o OpenSolver, navegue até <http://OpenSolver.org> – em inglês, e faça o download do arquivo zip. Descompacte esse arquivo em uma pasta e, sempre que quiser resolver um modelo complexo, apenas coloque-o em uma planilha como de costume e dê um clique duplo no arquivo OpenSolver.xlam, que lhe dará uma seção OpenSolver na aba Data do Excel. Pressione o botão Solve para resolver o modelo existente. Como exibido na Figura 1-28, eu apliquei o OpenSolver ao modelo da seção anterior no Excel 2013 e ele compra cinco fatias de pizza.

The screenshot shows a Microsoft Excel 2013 window with the following details:

- Excel Title Bar:** Book1 - Excel
- Ribbon:** The **DATA** tab is selected.
- OpenSolver Extension:** An "OpenSolver" ribbon extension is present on the right side of the ribbon, with the "OpenSolver" button highlighted.
- Tooltip for OpenSolver:**
  - Solve optimization model:** Solve an existing Solver model on the active worksheet by constructing the model's equations and then calling the Coin-DR CBC optimization engine. OpenSolver can only solve linear models.
  - OpenSolver**
  - Tell me more**
- Worksheet Content:**

A	B	C	D	E	F	G	H
1 Item		Calories	How many?				
2 Beer		200	0				
3 Bottled Water		0	0				
4 Chocolate Bar		255	0				
5 Chocolate Dipped Cone		300	0				
6 Gummy Bears		300	0				
7 Hamburger		320	0				
8 Hot Dog		265	0				
9 Ice Cream Sandwich		240	0				
10 Licorice Rope		280	0				
11 Nachos		560	0				
12 Pizza		480	5				
13 Popcorn		500	0				
14 Popsicle		150	0				
15 Soda		120	0				
16	Total Items:	5					
17	Total Calories:	2400					
18							
19							
- Bottom Status Bar:** READY, zoom controls, and 100%.

**Figura 1-28:** OpenSolver\compra\pizza\como\um\louco

## Resumindo

Bem, você aprendeu a navegar e selecionar intervalos rapidamente, utilizar referências absolutas, colar valores especiais, usar VLOOKUP e outras fórmulas, aprendeu a ordenar e filtrar dados, criar PivotTables e tabelas, executar fórmulas array e aprendeu como e quando utilizar o Solver.

Eis um fato que pode ser deprimente ou engraçado, dependendo da sua perspectiva. Eu conheci consultores de gestão em empresas proeminentes que recebiam salários excelentes fazendo o que eu chamo de “consultoria em dois passos”:

- 1.** Fale sobre coisas sem sentido com os clientes (esportes, férias, churrasco ... não que exista algo sem sentido quando o assunto é carne defumada).
- 2.** Resuma dados em Excel.

É possível que você não saiba tudo sobre futebol universitário (eu certamente não sei), mas se você internalizar esse capítulo, você tirará o passo 2 de letra.

Mas você não está aqui para se tornar um consultor de gestão. Você está aqui para se aprofundar em data science, e isso começa no próximo capítulo no qual começaremos com um pouco de aprendizado de máquina não supervisionado.

## 2

# Análise↑de↑Agrupamento↑Parte↑1: Usando↑K-Means↑para↑Segmentar a↑Sua↑Clientela

**E**u trabalho na indústria de e-mail marketing de um web site chamado MailChimp.com. Nós ajudamos os clientes a enviar e-mails informativos ao seu público, e toda vez que alguém usa o termo “e-mail blast” (enviar um único e-mail para muitas pessoas ao mesmo tempo), uma parte de mim morre.

Por quê? Porque endereços de e-mail não são mais caixas-pretas que você arremessa para explodirem como granadas. Não, em e-mail marketing (como em muitas outras formas de abordagem on-line, incluindo tweets, posts do Facebook e campanhas do Pinterest), um negócio recebe feedback em como seu público está se envolvendo *em nível individual* por meio de rastreamento de cliques, compras online, compartilhamento social e aí por diante. Esses dados não são apenas barulho. Eles caracterizam o seu público. Mas para os não iniciados, eles podem parecer grego. Ou esperanto.

Como você pega um conjunto de dados transacionais dos seus clientes (ou público, usuários, assinantes, cidadãos e assim por diante) e usa isso para entendê-los? Quando você lida com muitas pessoas, fica difícil entender cada cliente pessoalmente, principalmente se todos eles possuem maneiras diferentes de se relacionar com você. Mesmo que fosse possível entender todos em um nível pessoal, seria difícil trabalhar com isso.

Você precisa pegar essa base de clientes e encontrar um meio-termo entre “explodir” todo mundo como se eles fossem entidades anônimas e entender tudo sobre todos para criar um marketing personalizado para

cada receptor. Uma forma de acertar esse equilíbrio é usar *agrupamento* para criar *segmentação de mercado* dos seus clientes para que você possa anunciar para segmentos da sua base com conteúdo objetivo, promoções, etc.

*Análise de grupo* é a prática de agrupar vários objetos e separá-los em grupos de objetos similares. Explorando esses diferentes grupos — determinando como eles são parecidos e diferentes — você pode aprender muito sobre a pilha amorfa anterior que seus dados eram. E tal compreensão pode ajudá-lo a tomar melhores decisões em um nível que será mais detalhado do que antes.

Deste modo, o agrupamento é chamado de *mineração exploratória de dados*, pois essas técnicas de agrupamento ajudam a separar as relações em grandes conjuntos de dados que são muito difíceis de identificar a olho. E revelar os relacionamentos em sua população é útil para as indústrias, seja para recomendar filmes baseados nos hábitos do povo em um grupo de preferência, identificar locais com maior índice de crimes em áreas urbanas ou agrupar retornos relacionados a investimentos financeiros para assegurar um portfólio diversificado que cubra os grupos.

Um dos meus usos favoritos de agrupamento é o de imagem — juntar arquivos de imagens que “são parecidos” no computador. Por exemplo, em serviços de compartilhamento de fotos como o Flickr, um usuário gera muito conteúdo e pode acabar tendo fotos demais para navegar de maneira simples. Mas usando técnicas de agrupamento, é possível agrupar imagens similares e permitir que os usuários naveguem entre esses grupos antes de se aprofundarem.

---

## APRENDIZADO DE MÁQUINA SUPERVISIONADO VERSUS NÃO SUPERVISIONADO

Por definição, na mineração exploratória de dados, você não sabe antecipadamente o que está procurando. Você é um explorador. Como a Dora. Você pode até conseguir articular quando dois clientes são idênticos e quando eles são diferentes, mas você não sabe o melhor caminho para segmentar sua

base de clientes. Então quando você pede ao computador para segmentar seus clientes para você, isso é chamado de *aprendizado de máquina não supervisionado*, porque você não está “supervisionando” — dizendo ao computador como fazer seu trabalho.

Isso é um contraste com *aprendizado de máquina supervisionado*, que geralmente aparece quando a inteligência artificial tem o papel principal. Se eu sei que quero dividir os clientes em dois grupos — digamos “possíveis compradores” e “não possíveis compradores” — e eu forneço exemplos de históricos de tais clientes ao computador e digo a ele para atribuir todas as novas direções a um desses dois grupos, isso é supervisionado.

Se em vez disso eu falar, “isto é o que eu sei sobre meus clientes e é assim que deve-se medir se eles são similares ou diferentes. Diga-me o que é interessante”, isso é não supervisionado.

---

Este capítulo considera o tipo mais comum de agrupamento, chamado *agrupamento k-means*, que surgiu na década de 1950 e desde então tornou-se uma técnica de agrupamento confiável para exploração de conhecimento em bases de dados (KDD, Knowledge Discovery in Databases) em indústrias e no governo.

K-means não é a técnica mais rigorosa matematicamente falando. Ele nasce a partir do tipo de praticidade e senso comum que você pode ver na comida caseira. A comida caseira não tem o pedigree esnobe da cozinha francesa, mas ela promove total satisfação às vezes. A análise de grupo com k-means, como você verá em breve, é parte matemática, parte conto de fadas. Mas sua simplicidade intuitiva é parte da atração.

Para ver como funciona, você começará com um simples exemplo.

## Meninas↑dançam↑com↑Meninas,↑Meninos Coçam↑Seus↑Cotovelos

O objetivo em agrupamento k-means é pegar alguns pontos no espaço e colocá-los em grupos k (em que k é qualquer número que você queira escolher). Esses grupos k são definidos por um ponto no centro, como

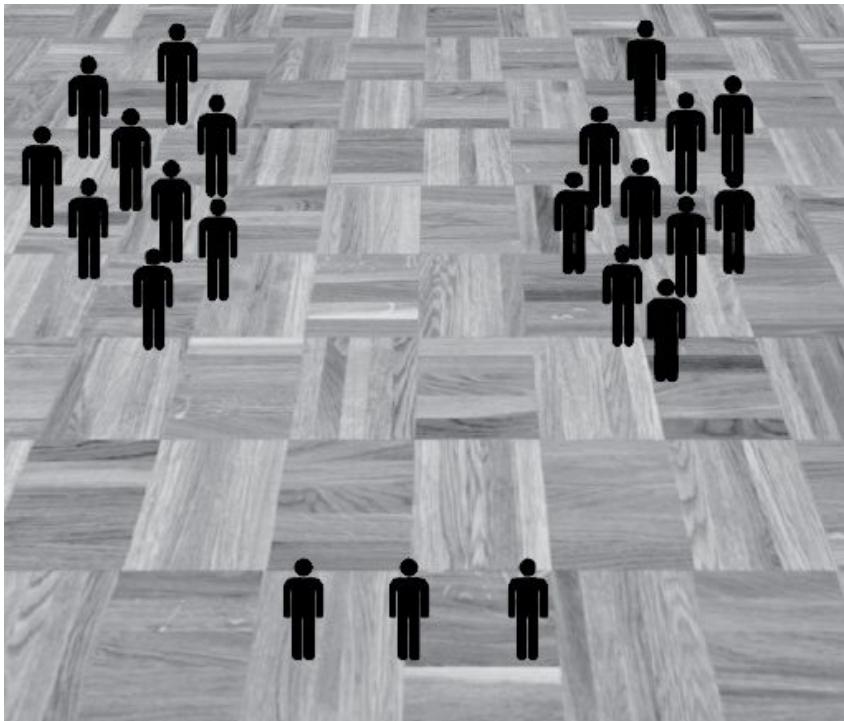
uma bandeira na lua que diz, “ei, este é o centro do meu grupo. Junte-se a mim se você estiver mais próximo a essa bandeira do que a alguma outra”. Esse centro de grupo (formalmente chamado de *grupo do centroide*) é a *média (mean)* de onde k-means recebe seu nome.

Use como exemplo um baile de ensino médio. Se você bloqueou o terror dos bailes do ensino médio da sua mente, eu peço desculpas por trazer à tona tais memórias dolorosas.

Aqueles presentes no Baile da Escola de Ensino Médio McAcne, romanticamente chamado de “Gala Sob o Mar”, estão espalhados pela pista conforme mostra a Figura 2-1. Eu até adicionei um piso parquet na imagem para ajudar na ilusão.

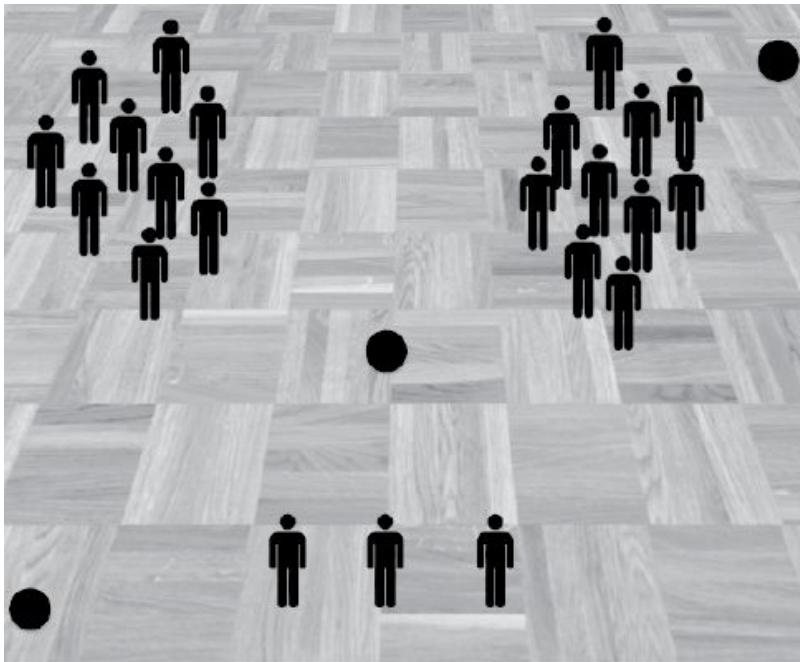
E esta é uma amostra das músicas que esses jovens líderes do mundo livre dançarão desajeitadamente para que você possa ouvir no Spotify:

- Styx: Come Sail Away
- Everything But the Girl: Missing
- Ace of Bass: All that She Wants
- Soft Cell: Tainted Love
- Montell Jordan: This is How We Do It
- Eiffel 65: Blue



**Figura 2-1:** Alunos da Escola McAcne arrasando na pista de dança

Agora, o agrupamento k-means exige que você especifique em quantos grupos você quer colocar os alunos presentes. Vamos escolher três grupos para começar (posteriormente neste capítulo observaremos como escolher  $k$ ). O algoritmo plantará três bandeiras na pista de dança, começando com alguma possível solução inicial, tal como a retratada na Figura 2-2, onde você tem três médias iniciais espalhadas pela pista, denotadas por círculos pretos.



**Figura 2-2:** Centros iniciais do grupo posicionados

Em agrupamento k-means, dançarinos são designados ao grupo mais próximo a eles, então entre dois centros de grupos na pista, você pode desenhar uma linha de demarcação, pela qual se um dançarino estiver em um lado da linha ele está em um grupo, mas se estiver do outro lado, o grupo muda (veja a Figura 2-3).

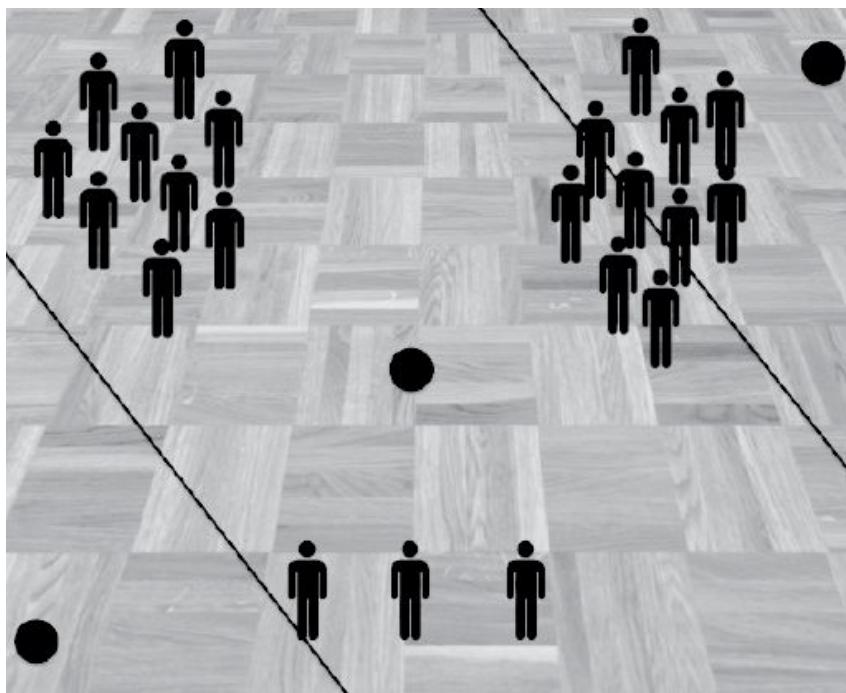
Usando essas linhas de demarcação, é possível atribuir dançarinos a seus grupos e sombreá-las adequadamente, como na Figura 2-4. Esse diagrama, que divide o espaço em polítopos baseados em quais regiões são atribuídas a quais centros de grupos por distância, é chamado de *diagrama de Voronoi*.

Agora, a tarefa inicial não parece correta, não é? Você separou o espaço de uma forma estranha, deixando o grupo da parte esquerda inferior vazio e muitas pessoas na extremidade do grupo da parte superior direita.

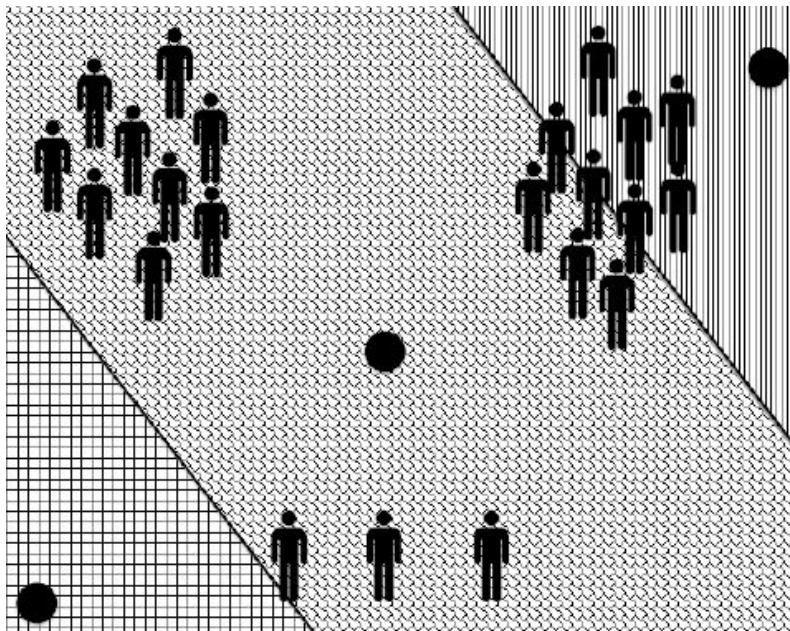
O algoritmo de agrupamento k-means desliza esses três centros de grupos pela pista de dança até conseguir o melhor ajuste.

Como é medido o “melhor ajuste”? Bom, cada pessoa presente está a uma certa distância do seu grupo. Qualquer organização de centros de

grupo que minimize a distância entre os presentes de seus centros é a melhor.



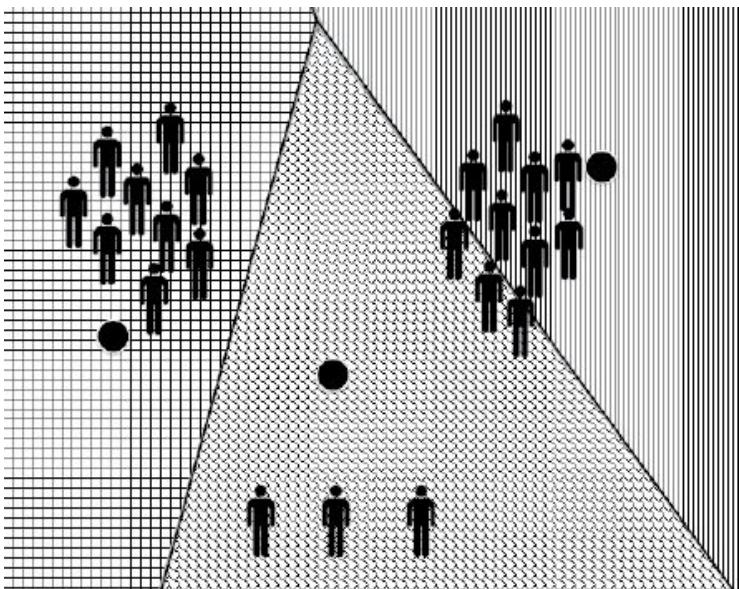
**Figura 2-3:** Linhas †denotam †as extremidades †dos grupos



**Figura 2-4:** Atribuições †de grupos †feitas †por regiões †sombreadas †no †diagrama †de †Voronoi

Agora, como mencionei no Capítulo 1, a palavra “minimizar” é uma dica que você precisará de um modelo de otimização para posicionar melhor seus centros de grupo. Então, você usará o Solver para movimentar os centros de grupo neste capítulo. A forma como o Solver posicionará os centros corretamente é movimentando-os de forma inteligente e iterativa, monitorando as boas disposições encontradas e combinando-as (literalmente acasalando-as como cavalos de corrida) para conseguir a melhor disposição.

Então, enquanto o diagrama na Figura 2-4 parece muito ruim, o Solver pode eventualmente concluir que os centros devem ser algo como a Figura 2-5. Isso faz a distância média entre cada dançarino e seu centro diminuir um pouco.



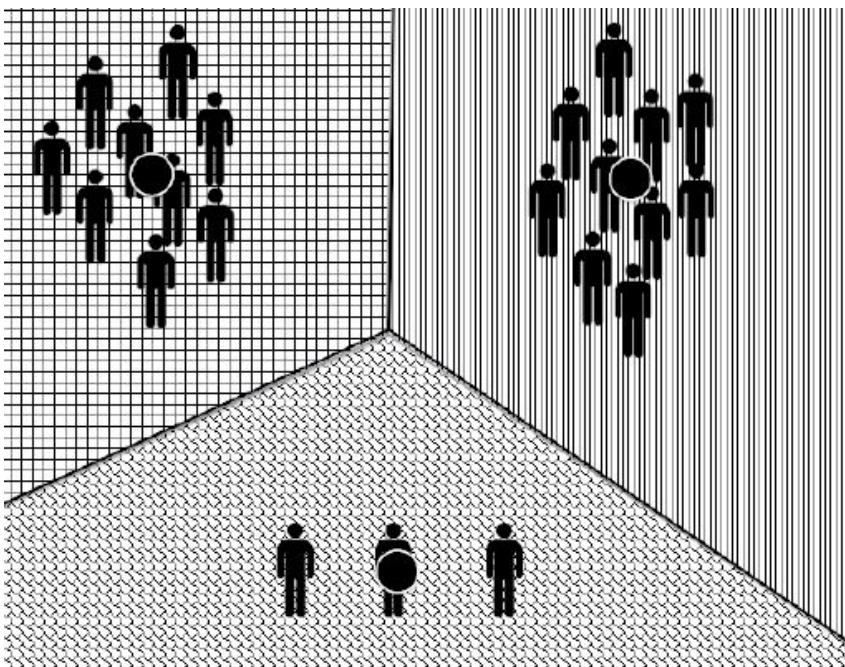
**Figura 2-5:** Movendo os centros um pouco

No entanto, eventualmente, o Solver descobre que os centros deveriam ser posicionados no meio dos três grupos de dançarinos conforme exibe a Figura 2-6.

Legal! É assim que um grupo ideal se parece. Os centroides do grupo estão nos centros de cada grupo de dançarinos, minimizando a distância média entre dançarino e o centro mais próximo. E agora que você possui

um agrupamento, é possível ir para a parte divertida: tentar entender o que os grupos *significam*.

Se você investigou as cores dos cabelos dos dançarinos, tendências políticas, velocidade de corridas, os grupos podem não fazer muito sentido. Mas no momento que você avalia os gêneros e as idades dos alunos presentes em cada grupo, você começa a ver alguns temas em comum. O grupo pequeno na parte inferior é todo de pessoas idosas — eles devem ser os acompanhantes de dança. O grupo da esquerda é todo de homens jovens, e o grupo da direita é de mulheres jovens. Todos estão com muito receio de dançar uns com os outros.



**Figura 2-6:** O excelente agrupamento de três médias do baile McAcne

Tudo bem! Então k-means permitiu que você segmentasse essa população de alunos presentes no baile e correlacionasse descritores dos alunos presentes com associação de grupo para entender o *porquê* por trás das atribuições.

Agora, você provavelmente está dizendo a si mesmo, “Sim, mas isso é estúpido. Eu já sabia a resposta antes de começar.” Você está certo. Nesse exemplo, você sabia. O motivo de não ser um problema, é que você já

consegue resolver isso apenas *olhando* para os pontos. Tudo está em um espaço bidimensional, por isso é muito fácil de agrupar com os seus olhos.

Mas e se você fosse a uma loja que vendesse milhares de produtos? Alguns clientes compraram um ou dois no ano passado. Outros compraram dezenas. E os itens comprados variam de cliente para cliente.

Como você os agrupa em suas “pistas de dança”? Bom, sua pista de dança não está em um espaço bidimensional ou tridimensional. Ela está em um espaço de compra de produtos de mil dimensões no qual um cliente ou comprou ou não comprou o produto em cada dimensão. Rapidamente, entenda, um problema de agrupamento pode exceder os limites da “Mark I Eyeball” (termo militar que refere-se ao reconhecimento visual sem aparelhos como binóculo ou radar), como meus amigos militares gostam de dizer.

## Caia na Real: Agrupamento K-Means de Assinantes em E-mail Marketing

Vamos continuar com um caso mais substancial. Eu sou um cara de e-mail marketing, então usarei um exemplo de [MailChimp.com](#), onde eu trabalho. Mas esse mesmo exemplo funcionaria com dados de compras a varejo, dados de conversão de anúncios, dados de mídia social e assim por diante. Ele funciona basicamente com qualquer tipo de dado onde você se aproxima dos clientes com material de marketing e eles decidem se escolhem você.

## O Empório Atacadista de Vinho Joey Bag O'Donuts

Vamos imaginar que você mora em Nova Jersey onde você gerencia o Empório Atacadista de Vinho Joey Bag O'Donuts. É um negócio de importação e exportação focado em levar vinho a granel para os Estados Unidos e vender para lojas de vinho e bebidas selecionadas pelo país. O

negócio funciona com Joey Bags viajando pelo mundo a procura de incríveis ofertas para grandes quantidades de vinho. Joey envia tudo para Jersey e é seu trabalho vender tudo isso para lojas com lucro.

Você alcança clientes de inúmeras maneiras — uma página no Facebook, Twitter, até a ocasional mala direta — mas o e-mail informativo desperta a maioria dos negócios. No ano passado, você enviou um informativo por mês. Geralmente, há duas ou três ofertas de vinho em cada e-mail, talvez uma seria Champanhe, e outra Malbec. Algumas ofertas são surpreendentes, com 80% ou mais de desconto. No total, você ofereceu 32 ofertas esse ano, das quais todas ocorreram muito bem.

Mas só porque as coisas estão indo bem, não significa que você não possa fazer melhor. Seria legal se você entendesse um pouco mais os clientes. Claro, você pode pesquisar uma compra em particular — como uma pessoa com o sobrenome Adams comprou algum Espumante em julho com 50% de desconto — mas você não consegue informar se isso é porque ele gostou de o requisito mínimo para compra ser uma caixa com seis garrafas, do preço ou que ainda não tinha passado do pico de amadurecimento.

Seria melhor se você pudesse segmentar a lista em grupos baseados em interesses. Então, seria possível customizar o informativo para cada segmento e talvez despertar mais negócios. Qualquer oferta que você achar que combina melhor com o segmento pode ir no assunto ou vir primeiro no informativo. Esse tipo de direcionamento pode resultar em impacto nas vendas.

Mas como você segmenta a lista? Onde você começa?

Essa é uma oportunidade para deixar o computador segmentar a lista para você. Usando o agrupamento k-means, é possível encontrar os melhores segmentos e tentar entender *porque* eles são os melhores.

## O↑Conjunto↑de↑Dados↑Inicial

## NOTA

A pasta de trabalho do Excel usada nesse capítulo, “WineKMC.xlsx”, está disponível para download no web site da editora, [www.altabooks.com.br](http://www.altabooks.com.br). Essa pasta de trabalho inclui todos os dados iniciais se quiser trabalhar a partir deles. Ou você pode apenas ler usando as planilhas que eu coloquei na pasta de trabalho.

Para começar, você tem duas interessantes fontes de dados:

- O metadado onde cada oferta é salva em uma planilha, incluindo varietal, quantidade mínima de garrafas por compra, desconto de varejo, se o vinho passou do seu ponto, e país e estado de origem. Esse dado fica armazenado em uma aba chamada OfferInformation, como consta da Figura 2-7.
- Você também sabe quais clientes compraram quais ofertas, então você pode transferir essa informação de MailChimp para a planilha com o metadado da oferta em uma aba chamada Transactions. Esse dado transacional, como mostra a Figura 2-8, é simplesmente representado como o cliente que fez a compra e qual oferta ele comprou.

WineKMC.xlsx

The screenshot shows a Microsoft Excel spreadsheet titled "WineKMC.xlsx". The active sheet is named "OfferInformation". The table has columns labeled A through G. Column A contains row numbers from 1 to 33. Column B contains "Offer #", Column C contains "Campaign", Column D contains "Varietal", Column E contains "Minimum Qty (kg)", Column F contains "Discount (%)", and Column G contains "Origin". Column H contains "Past Peak" which is set to FALSE for all rows. The data includes various wine types like Malbec, Pinot Noir, Espumante, Champagne, Cabernet Sauvignon, Prosecco, Merlot, Chardonnay, and Pinot Grigio, from different origins such as France, Oregon, New Zealand, Chile, Australia, South Africa, Italy, California, Germany, and others.

	A	B	C	D	E	F	G
1	Offer #	Campaign	Varietal	Minimum Qty (kg)	Discount (%)	Origin	Past Peak
2	1	January	Malbec	72	56	France	FALSE
3	2	January	Pinot Noir	72	17	France	FALSE
4	3	February	Espumante	144	32	Oregon	TRUE
5	4	February	Champagne	72	48	France	TRUE
6	5	February	Cabernet Sauvignon	144	44	New Zealand	TRUE
7	6	March	Prosecco	144	86	Chile	FALSE
8	7	March	Prosecco	6	40	Australia	TRUE
9	8	March	Espumante	6	45	South Africa	FALSE
10	9	April	Chardonnay	144	57	Chile	FALSE
11	10	April	Prosecco	72	52	California	FALSE
12	11	May	Champagne	72	85	France	FALSE
13	12	May	Prosecco	72	83	Australia	FALSE
14	13	May	Merlot	6	43	Chile	FALSE
15	14	June	Merlot	72	64	Chile	FALSE
16	15	June	Cabernet Sauvignon	144	19	Italy	FALSE
17	16	June	Merlot	72	88	California	FALSE
18	17	July	Pinot Noir	12	47	Germany	FALSE
19	18	July	Espumante	6	50	Oregon	FALSE
20	19	July	Champagne	12	66	Germany	FALSE
21	20	August	Cabernet Sauvignon	72	82	Italy	FALSE
22	21	August	Champagne	12	50	California	FALSE
23	22	August	Champagne	72	63	France	FALSE
24	23	September	Chardonnay	144	39	South Africa	FALSE
25	24	September	Pinot Noir	6	34	Italy	FALSE
26	25	October	Cabernet Sauvignon	72	59	Oregon	TRUE
27	26	October	Pinot Noir	144	83	Australia	FALSE
28	27	October	Champagne	72	88	New Zealand	FALSE
29	28	November	Cabernet Sauvignon	12	56	France	TRUE
30	29	November	Pinot Grigio	6	87	France	FALSE
31	30	December	Malbec	6	54	France	FALSE
32	31	December	Champagne	72	89	France	FALSE
33	32	December	Cabernet Sauvignon	72	45	Germany	TRUE

Figura 2-7: Os detalhes das últimas 32 ofertas

	A	B
1	Customer Last Name	Offer #
2	Smith	2
3	Smith	24
4	Johnson	17
5	Johnson	24
6	Johnson	26
7	Williams	18
8	Williams	22
9	Williams	31
10	Brown	7
11	Brown	29
12	Brown	30

Figura 2-8: Uma lista de ofertas compradas por cliente

## Determinando o que Medir

Aqui temos um dilema. No problema do baile da escola do ensino médio, medir distâncias entre dançarinos e grupos foi fácil, certo? Apenas arrebente na fita métrica!

Mas o que você faz neste problema?

Você sabe que houve 32 ofertas promovidas no ano passado, e você tem uma lista das 324 compras na aba Transactions, separada por cliente. Mas para medir a distância entre cada cliente e um centro de grupo, você precisa posicioná-los nesse espaço de 32 ofertas. Em outras palavras, é preciso entender as ofertas *que eles não aceitaram*, e criar uma matriz de ofertas por clientes, na qual cada cliente tem sua coluna 32 ofertas repleta de uns para as ofertas que eles aceitaram e de zeros para as que não.

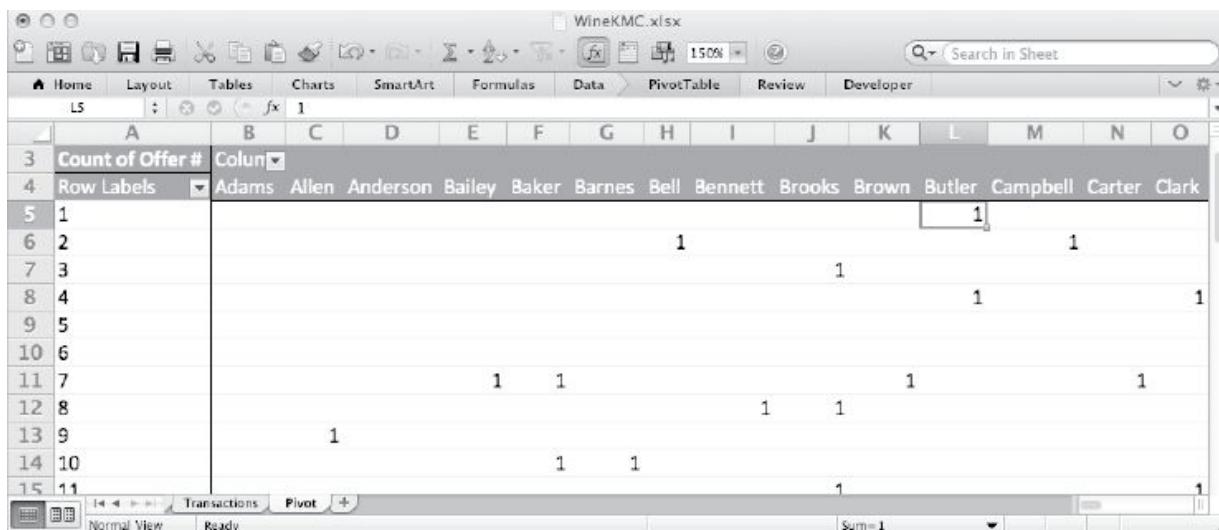
Em outras palavras, você precisa pegar essa aba Transactions orientada por fileiras e transformá-la em uma matriz com clientes em colunas e

ofertas em fileiras. E a melhor forma para criar tal matriz é usando PivotTable.

#### NOTA

Para uma introdução sobre PivotTables, veja o Capítulo 1.

Então faça o seguinte. Na aba Transactions, realce as colunas A e B e insira uma PivotTable. Usando o PivotTable Builder, simplesmente selecione ofertas como linhas, clientes como colunas e faça uma contagem das ofertas para os valores. Essa conta será 1 se um par cliente/oferta estava presente nos dados originais e 0 caso contrário (o 0 acaba sendo uma célula em branco nesse caso). A PivotTable resultante está retratada na Figura 2-9.



**Figura 2-9:** PivotTable de ofertas versus clientes

Agora que você tem as compras em forma de matriz, copie a aba OfferInformation e a nomeie como **Matrix** (matriz). Nessa nova planilha, cole os valores do PivotTable (não é necessário copiar e colar o número de ofertas, porque já está nas informações de oferta) na nova aba começando na coluna H. Você ficará com uma versão melhorada da

matriz que consolidou as descrições de ofertas com os dados de compras, como mostra a Figura 2-10.

	A	B	C	D	E	F	G	H	I	J	K
1	Offer #	Campaign	Varietal	Minimum Qty (kg)	Discount (%)	Origin	Past Peak	Adams	Allen	Anderson	Baile
4	3	February	Espumante	144	32	Oregon	TRUE				
5	4	February	Champagne	72	48	France	TRUE				
6	5	February	Cabernet Sauvignon	144	44	New Zealand	TRUE				
7	6	March	Prosecco	144	86	Chile	FALSE				
8	7	March	Prosecco	6	40	Australia	TRUE				1
9	8	March	Espumante	6	45	South Africa	FALSE				
10	9	April	Chardonnay	144	57	Chile	FALSE		1		
11	10	April	Prosecco	72	52	California	FALSE				
12	11	May	Champagne	72	85	France	FALSE				
13	12	May	Prosecco	72	83	Australia	FALSE				
14	13	May	Merlot	6	43	Chile	FALSE				
15	14	June	Merlot	72	64	Chile	FALSE				
16	15	June	Cabernet Sauvignon	144	19	Italy	FALSE				
17	16	June	Merlot	72	88	California	FALSE				
18	17	July	Pinot Noir	12	47	Germany	FALSE				
19	18	July	Espumante	6	50	Oregon	FALSE	1			
20	19	July	Champagne	12	66	Germany	FALSE				
21	20	August	Cabernet Sauvignon	72	82	Italy	FALSE				
22	21	August	Champagne	12	50	California	FALSE				
23	22	August	Champagne	72	63	France	FALSE				
24	23	September	Chardonnay	144	39	South Africa	FALSE				
25	24	September	Pinot Noir	6	34	Italy	FALSE		1		
26	25	October	Cabernet Sauvignon	72	59	Oregon	TRUE				
27	26	October	Pinot Noir	144	83	Australia	FALSE			1	
28	27	October	Champagne	72	88	New Zealand	FALSE	1			
29	28	November	Cabernet Sauvignon	12	56	France	TRUE				
30	29	November	Pinot Grigio	6	87	France	FALSE	1			1
31	30	December	Malbec	6	54	France	FALSE	1			
32	31	December	Champagne	72	89	France	FALSE				
33	32	December	Cabernet Sauvignon	72	45	Germany	TRUE				

**Figura 2-10:** Descrição de oferta e dado de compra combinados em uma única matriz

## PADRONIZANDO SEUS DADOS

Neste capítulo, cada dimensão dos seus dados é o mesmo tipo de dado de compra binária. Mas em muitos problemas de agrupamento, esse não é o caso. Visualize um cenário no qual as pessoas são agrupadas baseadas em altura, peso e salário. Esses três tipos de dados estão todos em escalas diferentes. Altura pode variar de 1,52 m a 2,03 m enquanto peso pode variar de 45 kg a 136 kg.

Nesse contexto, medir a distância entre clientes (como dançarinos na pista de dança) torna-se complicado. Então é comum **padronizar** cada coluna de dados subtraindo a média e dividindo pela medida da dispersão que encontraremos no Capítulo 4 conhecida como desvio padrão. Isso coloca cada coluna na mesma escala, centralizada ao redor de 0.

Enquanto nossos dados no Capítulo 2 não exigem uma padronização, é possível vê-la em ação no capítulo de identificação do valor atípico, Capítulo 9.

## Comece com Quatro Grupos

Tudo certo, agora você tem todos os seus dados consolidados em um formato único e aproveitável. Para começar a agrupar, você precisa escolher  $k$ , que é o número de grupos no algoritmo de agrupamento  $k$ -means. Geralmente, a abordagem em  $k$ -means é tentar vários valores diferentes para  $k$  (chegarei em como escolher entre eles posteriormente), mas, para começar, você precisa escolher apenas um.

Para começar, deve-se escolher um número de grupos que esteja dentro da estimativa com a qual está disposto a trabalhar. Você não criará 50 grupos e enviará 50 campanhas de anúncios direcionadas para algumas pessoas em cada grupo. Isso anula o propósito do exercício em primeiro lugar. Você quer algo pequeno nesse caso. Para esse exemplo, então, comece com quatro — em um mundo ideal, talvez sua lista seria dividida em quatro grupos perfeitamente compreensíveis de 25 clientes cada (isso não é provável).

Tudo bem então, se você fosse dividir os clientes em quatro grupos, quais seriam os melhores quatro grupos para isso?

Em vez de bagunçar a bonita aba Matrix, copie os dados para uma nova aba e chame-a de 4MC. É possível inserir quatro colunas após Past Peak nas colunas H até K que serão os centros de grupos. Para inserir uma coluna, dê um clique com o botão direito na Coluna H e selecione Insert. Uma coluna será adicionada à esquerda. Nomeie esses grupos como Cluster 1 até Cluster 4. Você também pode colocar alguma formatação condicional neles para que quando cada centro de grupo for indicado seja possível ver como eles diferem.

A aba 4MC aparecerá como na Figura 2-11.

Esses centros de grupos são todos 0 nesse momento. Mas, tecnicamente, eles podem ser o que você quiser, e o que você gostaria de ver é que eles, como no caso do baile da escola, se distribuíssem para minimizar as distâncias entre cada cliente e seu centro de grupo mais próximo.

Então, obviamente, esses centros de grupos terão valores entre 0 e 1 para cada oferta já que todos os vetores clientes são binários.

Mas o que significa medir a distância entre um centro de grupo e um cliente?

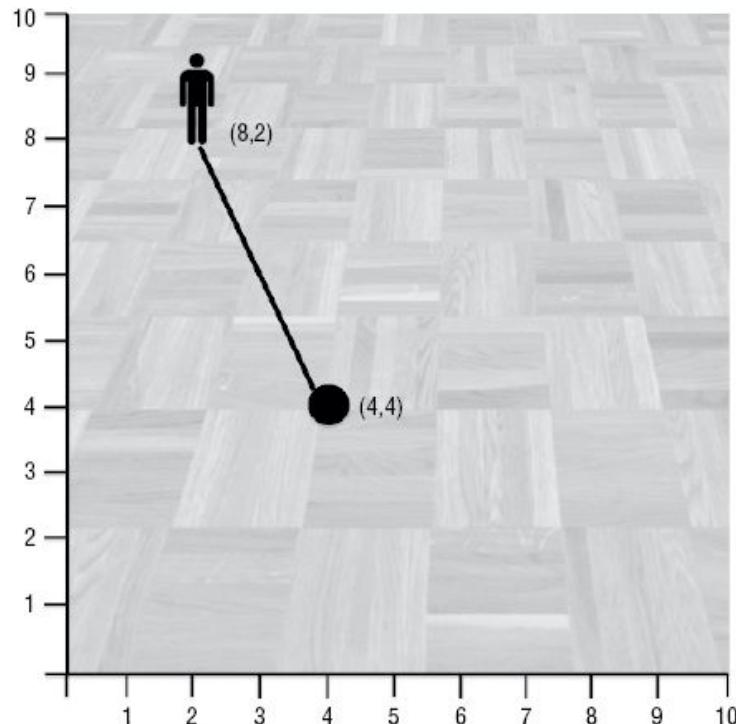
Distância Euclidiana: Medindo Distâncias em Linha Reta

Agora você tem uma única coluna por cliente, então como você mede a distância da pista de dança entre eles? Bom, o termo oficial para isso é “em linha reta”, medir distância de fita métrica é a *Distância Euclidiana*.

Vamos voltar para o problema da pista de dança para entender como computar isso.

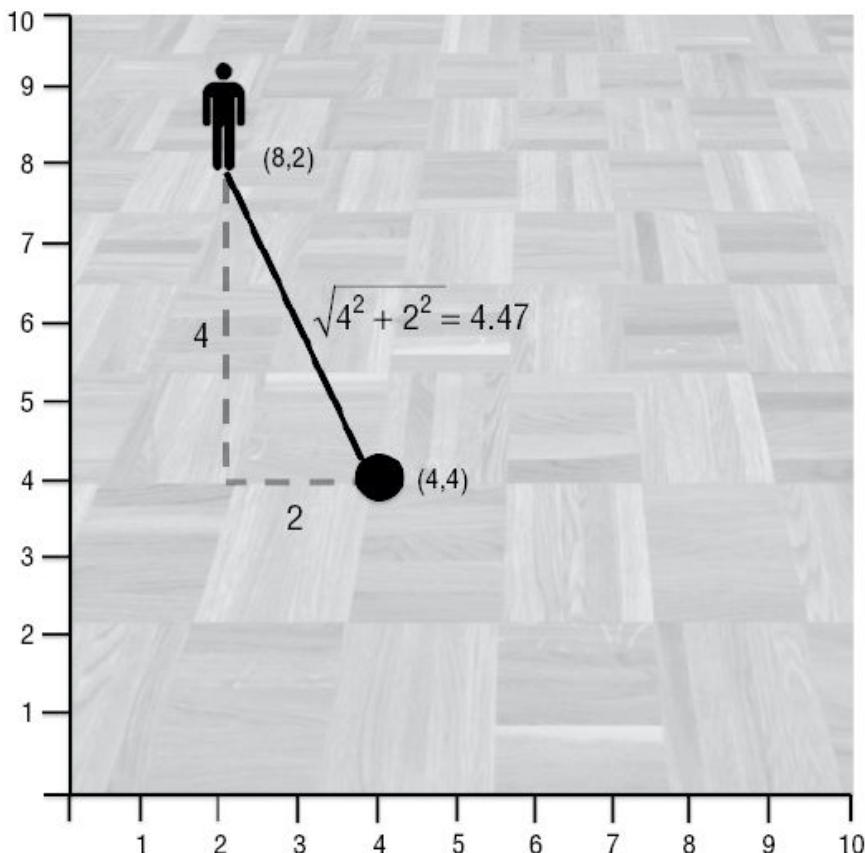
Estabelecerei um eixo horizontal e um vertical na pista de dança, e na Figura 2-12, é possível ver que você tem um dançarino em  $(8, 2)$  e um centro de grupo em  $(4, 4)$ . Para computar a distância Euclidiana entre eles, você deve lembrar do teorema de Pitágoras que aprendeu na escola.

**Figura 2-11:** Centros de grupos em branco localizados na aba 4MC



**Figura 2-12:** Um dançarino em (8, 2) e um centro de grupo em (4, 4)

Esses dois pontos estão a  $8 - 4 = 4$  unidades de distância em direção vertical. Eles estão a  $4 - 2 = 2$  unidades de distância em direção horizontal. Então, pelo teorema de Pitágoras, a distância quadrática entre esse dois pontos é  $4^2 + 2^2 = 16 + 4 = 20$  unidades. Logo, a distância entre eles é a raiz quadrada de 20, que é aproximadamente 4,47 pés (veja a Figura 2-13).



**Figura 2-13:** A distância Euclidiana é a raiz quadrada da soma das distâncias quadráticas em cada sentido único

No contexto dos assinantes de informativos, você tem mais do que duas dimensões, mas o mesmo conceito se aplica. A distância entre um cliente e um centro de grupo é calculada tirando a diferença entre os dois pontos de cada oferta, elevando-os ao quadrado, somando-os, e tirando a raiz quadrada.

Por exemplo, digamos que na aba 4MC você queria tirar a distância Euclidiana entre o centro de Cluster 1 na coluna H e as compras do cliente Adams na coluna L.

Na célula L34, abaixo das compras de Adams, pode-se tirar a diferença do vetor de Adams e o centro do grupo, elevá-la o quadrado, somá-la e tirar a raiz quadrada da soma, usando a seguinte fórmula array (note as referências absolutas que permitem que você arraste essa fórmula para a direita ou para baixo sem que a referência do centro do grupo mude):

```
{=SQRT(SUM((L$2:L$33-$H$2:$H$33)^2))}
```

Você deve usar uma fórmula array (entre com a fórmula e pressione Ctrl + Shift + Enter ou Cmd + Return no Mac como disposto no Capítulo 1) porque a porção da fórmula ( $L2:L33 - H2:H33$ )<sup>2</sup> precisa saber passar item a item tirando as diferenças e elevando-as ao quadrado. O resultado final, no entanto, é um número individual: 1,732 nesse caso (veja a Figura 2-14). Isso faz sentido pois Adams aceitou três ofertas, mas o centro de grupo inicial é todo 0, e a raiz quadrada de 3 é 1,732.

The screenshot shows a Microsoft Excel spreadsheet titled "WineKMC.xlsx". The formula bar at the top contains the formula: `=SQRT(SUM((L$2:L$33-$H$2:$H$33)^2))`. The main table has columns labeled E through M. Column E is "Discount (%)", column F is "Origin", column G is "Past Peak", and columns H through L are "Cluster 1", "Cluster 2", "Cluster 3", and "Cluster 4" respectively. Row 1 contains the headers: "Discount (%)", "Origin", "Past Peak", "Cluster 1", "Cluster 2", "Cluster 3", "Cluster 4", "Adams", and "Allen". Rows 20 through 33 list various wine entries with their discount percentages, origin countries, and past peak values. Row 34 is highlighted and contains the text "Distance to Cluster 1". The cell L34 contains the value "1.732". The status bar at the bottom right shows "Sum=1.732".

	E	F	G	H	I	J	K	L	M
1	Discount (%)	Origin	Past Peak	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Adams	Allen
20	66	Germany	FALSE						
21	82	Italy	FALSE						
22	50	California	FALSE						
23	63	France	FALSE						
24	39	South Africa	FALSE						
25	34	Italy	FALSE						
26	59	Oregon	TRUE						
27	83	Australia	FALSE						
28	88	New Zealand	FALSE						
29	56	France	TRUE						
30	87	France	FALSE					1	
31	54	France	FALSE					1	
32	89	France	FALSE						
33	45	Germany	TRUE						
34			Distance to Cluster 1					1.732	

Figura 2-14: A distância entre Adams e Cluster 1.

Na planilha exibida na Figura 2-14, eu congelei os painéis (veja o Capítulo 1) entre as colunas G e H e nomeei a fileira 34 em G34 como Distance to Cluster 1 apenas para acompanhar as coisas quando você rola para a direita.

## Distâncias e Atribuição de Grupos para Todos!

Então agora você sabe como calcular a distância entre um vetor de compra e um centro de grupo.

É hora de adicionar os cálculos de distância para Adams a outros centros arrastando a célula L34 para baixo até L37 e modificando a referência do centro de grupo ***manualmente*** da coluna H à I, J e K nas células ascendentes. Você fica com as seguintes 4 fórmulas em L34:L37:

```
{=SQRT(SUM((L$2:L$33-$H$2:$H$33)^2))}  
{=SQRT(SUM((L$2:L$33-$I$2:$I$33)^2))}  
{=SQRT(SUM((L$2:L$33-$J$2:$J$33)^2))}  
{=SQRT(SUM((L$2:L$33-$K$2:$K$33)^2))}
```

Como você usou referências absolutas (o sinal \$ nas fórmulas; veja o Capítulo 1 para mais detalhes) para os centros de grupos, você pode arrastar L34:L37 até DG34:DG37 para calcular distâncias entre cada cliente e todos os quatro centros de grupos. Também, na coluna G, nomeie as fileiras 35 até 37 como Distance to Cluster 2, e assim por diante. Essas novas distâncias estão representadas na Figura 2-15.

WineKMC.xlsx								
<a href="#">Home</a> <a href="#">Layout</a> <a href="#">Tables</a> <a href="#">Charts</a> <a href="#">SmartArt</a> <a href="#">Formulas</a> <a href="#">Data</a> <a href="#">Review</a> <a href="#">Developer</a>								
DG37								
E	F	G	DC	DD	DE	DF	DG	
1	Discount (%)	Origin	Past Peak	Williams	Wilson	Wood	Wright	Young
26	59	Oregon	TRUE					
27	83	Australia	FALSE					
28	88	New Zealand	FALSE					1
29	56	France	TRUE					
30	87	France	FALSE					
31	54	France	FALSE			1		
32	89	France	FALSE		1			1
33	45	Germany	TRUE					1
34			Distance to Cluster 1	1.732	1.414	2.000	2.000	2.449
35			Distance to Cluster 2	1.732	1.414	2.000	2.000	2.449
36			Distance to Cluster 3	1.732	1.414	2.000	2.000	2.449
37			Distance to Cluster 4	1.732	1.414	2.000	2.000	2.449
40								

**Figura 2-15:** Cálculos de distância de cada cliente para cada grupo

Então, você sabe a distância de cada cliente para todos os quatro grupos. Sua atribuição de grupo é para o mais próximo, que pode ser calculado em dois passos.

Primeiro, voltando ao cliente Adams na coluna L, calcularemos a distância mínima para o centro de grupo na célula L38. Que é apenas:

```
=MIN(L34:L37)
```

E então, para determinar qual grupo corresponde à distância mínima, você pode usar a fórmula `MATCH` (veja o Capítulo 1 para mais detalhes). Colocando a seguinte fórmula `MATCH` em L39, você pode determinar qual índice de célula na sequência L34 a L37 contando a partir de 1 corresponde com a distância mínima:

```
=MATCH(L38,L34:L37,0)
```

Nesse caso, a distância mínima é um empate entre todos os quatro grupos, então `MATCH` escolhe o primeiro (L34) retornando índice 1 (veja a Figura 2-16).

É possível arrastar essas duas fórmulas pela planilha até DG38:DG39 também. Adicione Minimum Cluster Distance e Assigned Cluster na Coluna G como nomes para as fileiras 38 e 39 apenas para manter as coisas organizadas.

	G	H	I	J	K	L	M	N	O	P	
1	Past Peak	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Adams	Allen	Anderson	Bailey	Baker	Ba
28	FALSE							1			
29	TRUE										
30	FALSE						1				
31	FALSE						1				1
32	FALSE										1
33	TRUE										
34	Distance to Cluster 1					1.732	1.414	1.414	1.414	2.000	
35	Distance to Cluster 2					1.732	1.414	1.414	1.414	2.000	
36	Distance to Cluster 3					1.732	1.414	1.414	1.414	2.000	
37	Distance to Cluster 4					1.732	1.414	1.414	1.414	2.000	
38	Minimum Cluster Distance					1.732	1.414	1.414	1.414	2.000	
39	Assigned Cluster					1	1	1	1	1	

Figura 2-16: Combinações de grupo adicionadas na planilha

## Resolvendo Centros de Grupo

Agora você tem cálculos de distâncias e atribuições de grupos na planilha. Para fixar os centros de grupo em suas melhores localizações, é preciso encontrar os valores nas colunas H até K que minimizem a distância total entre os clientes e os grupos atribuídos a eles denotados na fileira 39 abaixo de cada cliente.

E se você leu o Capítulo 1, você sabe exatamente o que pensar quando escuta a palavra *minimizar*: Esse é um passo de otimização, e um passo de otimização significa usar o Solver.

Para usar o Solver, você precisa de uma célula objetiva, então, na célula A36, somaremos todas as distâncias entre clientes e suas atribuições de grupos:

=SUM(L38 : DG38)

Essa soma das distâncias dos clientes de seus centros de grupos mais próximos é exatamente a função objetiva encontrada anteriormente quando agrupamos na pista de dança da Escola de Ensino Médio McAcne. Mas a distância Euclidiana com suas potências e raízes quadradas é altamente não linear, então você precisa usar o evolucionário método de solução em vez do método simplex para estabelecer os centros de grupo.

No Capítulo 1, você usou o algoritmo simplex. Simplex é mais rápido quando é permitido, mas não é possível quando você está elevando ao quadrado, calculando a raiz quadrada, ou senão, tirando funções não lineares das suas decisões. Da mesma forma, OpenSolver (apresentado no Capítulo 1), que usa uma implementação de simplex em esteroides, não é útil aqui.

Nesse caso, o algoritmo evolucionário construído no Solver usa uma combinação de pesquisa aleatória e um bom recurso “reprodutor” para encontrar boas soluções similares a como a evolução funciona em contextos biológicos.

## NOTA

Para um tratamento completo de otimização, veja o Capítulo 4.

Note que você tem tudo o que precisa para montar um problema no Solver:

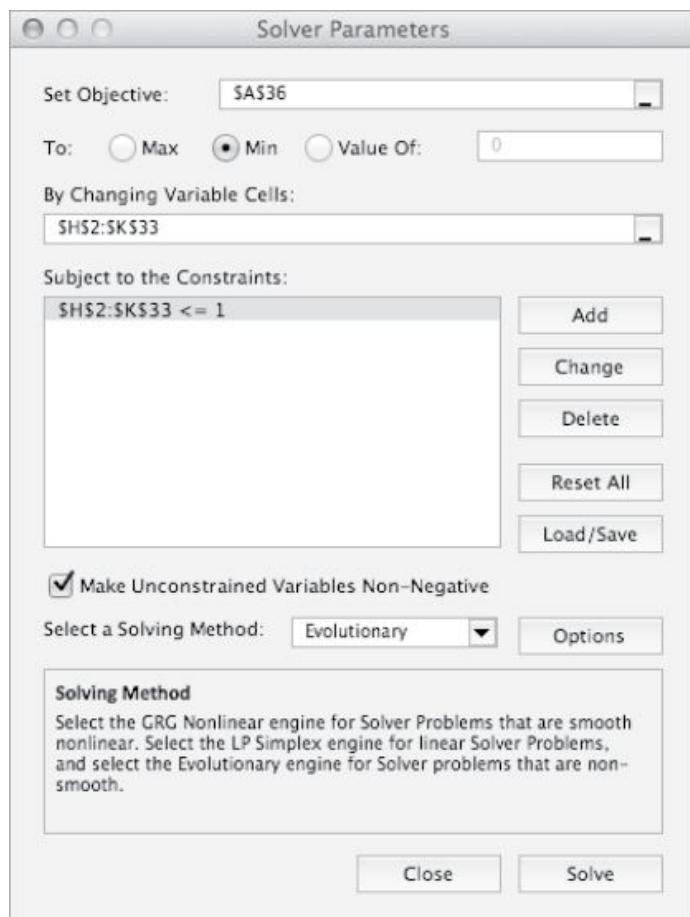
- **Objetivo:** Minimizar as distâncias totais entre clientes e seus centros de grupos (A36).
- **Variáveis de decisão:** Os valores de oferta de cada fileira dentro do centro do grupo (H2:K33).
- **Restrições:** Centros de grupos deveriam ter valores entre 0 e 1.

Abra o Solver e elabore os requerimentos. Você configurará o Solver para minimizar A36 mudando H2:K33 com a restrição de que H2:K33 seja  $\leq 1$  assim como todos os vetores de ofertas. Certifique-se que as variáveis estejam selecionadas como não negativas e que o evolutionary solver esteja escolhido. Veja a Figura 2-17.

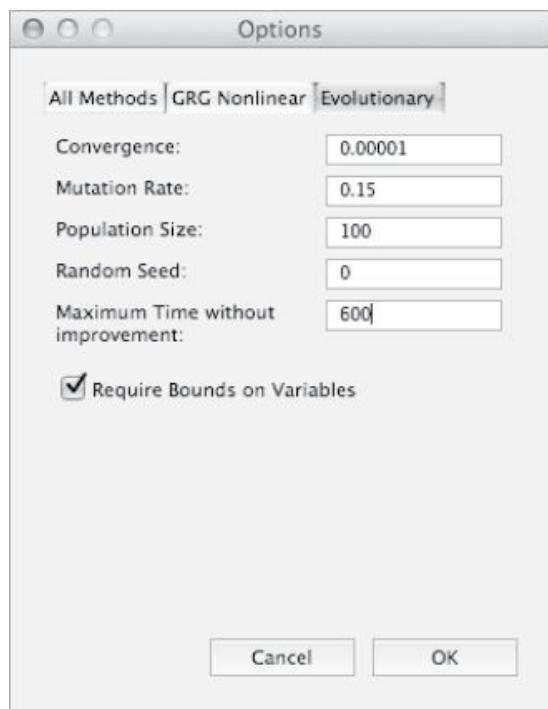
Além disso, configurar esses grupos não é simples para o Solver, então você deveria reforçar algumas das opções do evolutionary solver pressionando o botão options na janela do Solver e marcando a aba evolutionary. É útil aumentar o parâmetro Maximum Time Without Improvement em algum lugar ao norte de 30 segundos, dependendo de quanto tempo você quer esperar para o Solver terminar. Na Figura 2-18, eu defini o meu em 600 segundos (10 minutos). Dessa forma, eu posso configurar o Solver para rodar e sair para jantar. E se alguma vez você quiser finalizar o Solver antecipadamente, apenas pressione Escape e então saia com a melhor solução que ele encontrou até o momento.

Se você está curioso, os mecanismos interiores do evolutionary solver são abordados em maiores detalhes no Capítulo 4 e em

<http://www.solver.com> — conteúdo em inglês.



**Figura 2-17:** A configuração do Solver para agrupamentos de 4 centros



**Figura 2-18:** A aba de configurao do evolutionary solver

Pressione Solve e veja o Excel fazer o resto at que o algoritmo evolucionrio converja.

## Atribuindo Sentido aos Resultados

Uma vez que o Solver lhe d os melhores centros de grupos, a diverso comea. Voc pode explorar os grupos para entendimento! Ento, na Figura 2-19, s possvel ver que o Solver calculou uma distncia total perfeita de 140,7, e os quatro centros de grupo, graas a formatao condicional, parecem diferentes.

Note que seus centros de grupo podem parecer diferentes da planilha disponibilizada com o livro porque o algoritmo evolucionrio utiliza nmeros aleatrios e no fornece a mesma resposta toda vez. Os grupos podem ser fundamentalmente diferentes ou, mais provvel, eles podem estar em uma ordem diferente (por exemplo, o meu Cluster 1 est muito perto do seu Cluster 4, e assim por diante).

Como voc colou as descries das ofertas nas colunas B at G quando iniciou a aba, pode ler os detalhes das ofertas na Figura 2-19 que parecem importantes para os centros de grupos.

The screenshot shows a Microsoft Excel spreadsheet titled "WineKMC.xlsx". The data is organized into several sections:

- Offer Data:** Rows 29 to 33 show wine offers with columns for Offer #, Campaign, Varietal, Min. Qty, Discount (%), Origin, Past Peak, and four columns for Cluster 1 through Cluster 4, followed by columns for Adams, Allen, Anderson, and Bailey.
- Distance Calculations:** Rows 34 to 38 calculate distances from the offers to the four clusters. Row 34 contains formulas for "Distance to Cluster 1" through "Distance to Cluster 4". Row 35 is labeled "Total Distance" with the value "140.7" in row 36.
- Summary Statistics:** Row 37 calculates the "Minimum Cluster Distance". Row 38 shows the "Assigned Cluster" for each offer, with values 2, 3, 1, and 2 respectively. Row 39 is blank.

The "Developer" tab is selected in the ribbon. The formula bar shows the formula for cell A36: =SUM(L38:O38). The status bar at the bottom right indicates "Sum=140.7".

Offer #	Campaign	Varietal	Min. Qty	Discount (%)	Origin	Past Peak	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Adams	Allen	Anderson	Bailey
29	28	November Cabernet	12	56	France	TRUE	0.026	0.017	0.090	0.030				
30	29	November Pinot Grig	6	87	France	FALSE	0.012	0.619	0.043	0.038	1			
31	30	December Malbec	6	54	France	FALSE	0.020	0.729	0.079	0.136	1			1
32	31	December Champag	72	89	France	FALSE	0.023	0.027	0.211	0.259				
33	32	December Cabernet	72	45	Germany	TRUE	0.093	0.013	0.053	0.125				
34							Distance to Cluster 1				2.166	1.939	0.740	1.924
35	Total Distance						Distance to Cluster 2				1.044	1.886	1.865	1.042
36	140.7						Distance to Cluster 3				1.691	1.339	1.428	1.359
37							Distance to Cluster 4				2.012	1.731	1.781	1.749
38							Minimum Cluster Distance				1.044	1.339	0.740	1.042
39							Assigned Cluster				2	3	1	2

**Figura 2-19:** Os quatro centros de grupos ideais

Para o Cluster 1 na coluna H, a formatação condicional chama as ofertas 24, 26, 17, e para um grau menor, 2. Lendo os detalhes daquelas ofertas, a principal coisa que elas têm em comum é: **Todas Elas São Pinot Noir.**

Se você olhar para a coluna I, todas as células verdes têm uma quantidade mínima baixa em comum. Esses são os compradores que não querem ter que comprar em grande quantidade para conseguir uma oferta.

Mas serei honesto; os dois últimos centros de grupos são um tanto difíceis de interpretar. Bom, e se, em vez de interpretar o centro de grupo, você investigar os membros do grupo e determinar quais ofertas eles gostam? Isso pode ser mais esclarecedor.

## Conseguindo as Melhores Ofertas por Grupo

Então, em vez de considerar quais dimensões estão mais próximas de 1 para um centro de grupo, vamos ver quem está atribuído a cada grupo e quais ofertas eles preferem.

Para fazer isso, começaremos fazendo uma cópia da aba OfferInformation e chamando-a de 4MC — TopDealsByCluster. Nessa nova aba, nomeie as colunas H até K como 1, 2, 3 e 4 (veja a Figura 2-20).

A	B	C	D	E	F	G	H	I	J	K
Offer #	Offer date	Product	Minimum Qt	Discount	Origin	Past Peak	1	2	3	4
2	1 January	Malbec	72	56 France		FALSE				
3	2 January	Pinot Noir	72	17 France		FALSE				
4	3 February	Espumante	144	32 Oregon		TRUE				
5	4 February	Champagne	72	48 France		TRUE				
6	5 February	Cabernet Sauvignon	144	44 New Zealand		TRUE				

**Figura 2-20:** Criando uma aba para contar ofertas populares por grupo

De volta à aba 4MC, você tem atribuições de grupo listadas (1-4) na linha 39. Tudo o que precisa fazer para conseguir as contagens de ofertas por grupo é verificar o título da coluna na aba 4MC —

TopDealsByCluster nas colunas H até K, veja quem de 4MC foi atribuído àquele grupo usando a linha 39, e então some seus valores para cada linha de oferta. Isso lhe dará o total de clientes de um grupo determinado que aceitaram a oferta.

Comece com a célula H2, isto é, a conta de clientes em Cluster 1 que aceitaram a oferta #1, a oferta do Malbec de janeiro. Você quer somar L2:DG2 na aba 4MC mas somente para aqueles clientes que estão no Cluster 1, e esse é um clássico uso para a fórmula SUMIF. A fórmula é assim:

```
=SUMIF('4MC'!$L$39:$DG$39, '4MC -  
TopDealsByCluster'!H$1, '4MC'!$L2:$DG2)
```

A declaração de SUMIF funciona com o fornecimento de alguns valores para verificar na primeira seção '4MC'!\$L\$39:\$DG\$39, que são verificados contra o 1 no cabeçalho da coluna ('4MC - TopDealsByCluster'!H\$1), e então, para qualquer combinação, você soma a linha 2 especificando '4MC'!\$L2:\$DG2 na terceira seção da fórmula.

Note que você usou referências absolutas (o \$ na fórmula) na frente de tudo na linha da atribuição de grupo, na frente do número da linha para os nossos cabeçalhos de coluna, e na frente da letra da coluna para nossas ofertas aceitas. Fazendo essas referências absolutas, é possível arrastar essa fórmula até o intervalo H2:K33 para conseguir as contagens de ofertas para todos os centros de grupos e combinações de ofertas, conforme exibido na Figura 2-21. Você pode colocar alguma formatação condicional nessas colunas para torná-las mais legíveis.

Selecionando as colunas de A à K e usando auto-filter (veja o Capítulo 1), você pode tornar esses dados ordenáveis. Ordenando de alto para baixo na coluna H, é possível ver quais ofertas são mais populares no Cluster 1 (veja a Figura 2-22).

Como salientado anteriormente, as quatro melhores ofertas para esse grupo são todas Pinot. Essas pessoas assistiram a *Sideways – Entre Umas e Outras* muitas vezes. Quando você ordena no Cluster 2, torna-

se perfeitamente claro que esses são compradores de baixa quantidade (veja a Figura 2-23).

Mas quando você ordena no Cluster 3, as coisas não são tão claras. Há mais do que um punhado de ofertas excelentes. A redução entre transações dentro e transações fora não é tão extrema. Mas as mais populares parecem ter menos coisas em comum — os descontos são muito bons, cinco das seis melhores ofertas são espumantes in natura, e a França está em três das quatro melhores ofertas. Mas nada é conclusivo (veja a Figura 2-24).

Quanto ao Cluster 4, esses caras realmente gostaram da oferta de Champanhe de agosto por algum motivo. Além disso, cinco das seis melhores ofertas são da França, e nove das dez melhores ofertas são de alta quantidade (veja a Figura 2-25). Talvez esse seja o Cluster de inclinação por francês de alta quantidade? A superposição entre os grupos 3 e 4 é um pouco perturbadora.

Isso leva a uma questão: 4 é o número correto para k em agrupamento k-means? Talvez não. Mas como você saberia?

The screenshot shows a Microsoft Excel spreadsheet titled "WineKMC.xlsx". The formula bar displays the formula: =SUMIF(\$4:\$11\$39:\$D\$39,\$D\$39,\$4:\$11\$1:\$4:\$11\$2:\$D\$2). The table has columns F, G, H, I, J, K, L and rows 1 through 7. Column F is labeled "Origin" and column G is labeled "Past Peak". Columns H, I, J, K, and L represent different clusters. The data shows that France is the most frequent origin across all clusters, particularly in cluster 4 where it has the highest count of 6. Other origins like Oregon, New Zealand, and Chile also appear in the table.

F	G	H	I	J	K	L
1	Origin	Past Peak	1	2	3	4
2	France	FALSE	0	0	4	6
3	France	FALSE	4	0	4	2
4	Oregon	TRUE	0	0	2	4
5	France	TRUE	0	0	7	5
6	New Zealand	TRUE	0	0	2	2
7	Chile	FALSE	0	0	5	7

Figura 2-21: Total de cada oferta quebrado por grupo

WineKMC.xlsx

	A	B	C	D	E	F	G	H
1	Offer #	Offer date	Product	Minimum	Discount	Origin	Past Peak	1
2		24 September	Pinot Noir	6	34 Italy	FALSE	12	
3		26 October	Pinot Noir	144	83 Australia	FALSE	8	
4		17 July	Pinot Noir	12	47 Germany	FALSE	7	
5		2 January	Pinot Noir	72	17 France	FALSE	4	
6		1 January	Malbec	72	56 France	FALSE	0	
7		3 February	Espumante	144	32 Oregon	TRUE	0	
8		4 February	Champagne	72	48 France	TRUE	0	

Figura 2-22:↑Ordenar↑no↑Cluster↑1↑—↑Pinot,↑Pinot,↑Pinot!

WineKMC.xlsx

	A	B	C	D	E	F	G	I
1	Offer #	Offer date	Product	Minimum	Discount	Origin	Past Peak	2
2		30 December	Malbec	6	54 France	FALSE	16	
3		29 November	Pinot Grigio	6	87 France	FALSE	15	
4		7 March	Prosecco	6	40 Australia	TRUE	12	
5		8 March	Espumante	6	45 South Africa	FALSE	11	
6		18 July	Espumante	6	50 Oregon	FALSE	11	
7		13 May	Merlot	6	43 Chile	FALSE	6	
8		24 September	Pinot Noir	6	34 Italy	FALSE	0	
9		26 October	Pinot Noir	144	83 Australia	FALSE	0	
10		17 July	Pinot Noir	12	47 Germany	FALSE	0	

Figura 2-23:↑Ordenar↑no↑Cluster↑2↑—↑insignificante

WineKMC.xlsx

	A	B	C	D	E	F	G	J
1	Offer #	Offer date	Product	Minimum	Discount	Origin	Past Peak	3
2	31	December	Champagne	72	89	France	FALSE	10
3	4	February	Champagne	72	48	France	TRUE	7
4	9	April	Chardonnay	144	57	Chile	FALSE	7
5	11	May	Champagne	72	85	France	FALSE	7
6	8	March	Espumante	6	45	South Africa	FALSE	6
7	27	October	Champagne	72	88	New Zealand	FALSE	6
8	26	October	Pinot Noir	144	83	Australia	FALSE	5
9	6	March	Prosecco	144	86	Chile	FALSE	5
10	10	April	Prosecco	72	52	California	FALSE	5
11	14	June	Merlot	72	64	Chile	FALSE	5
12	16	June	Merlot	72	88	California	FALSE	5
13	7	March	Prosecco	6	40	Australia	TRUE	4
14	2	January	Pinot Noir	72	17	France	FALSE	4
15	1	January	Malbec	72	56	France	FALSE	4
16	20	August	Cabernet Sauvignon	72	82	Italy	FALSE	4
17	28	November	Cabernet Sauvignon	12	56	France	TRUE	4
18	12	May	Prosecco	72	83	Australia	FALSE	3
19	23	September	Chardonnay	144	39	South Africa	FALSE	3
20	25	October	Cabernet Sauvignon	72	59	Oregon	TRUE	3
21	32	December	Cabernet Sauvignon	72	45	Germany	TRUE	3
22	30	December	Malbec	6	54	France	FALSE	2
23	29	November	Pinot Grigio	6	87	France	FALSE	2

Figura 2-24: Ordenar no Cluster 3 é um pouco bagunçado

	A	B	C	D	E	F	G	K
1	Offer #	Offer date	Product	Minimum	Discount	Origin	Past Pea	4
2		22 August	Champagne	72	63	France	FALSE	21
3		31 December	Champagne	72	89	France	FALSE	7
4		6 March	Prosecco	144	86	Chile	FALSE	7
5		11 May	Champagne	72	85	France	FALSE	6
6		1 January	Malbec	72	56	France	FALSE	6
7		4 February	Champagne	72	48	France	TRUE	5
8		14 June	Merlot	72	64	Chile	FALSE	4
9		30 December	Malbec	6	54	France	FALSE	4
10		3 February	Espumante	144	32	Oregon	TRUE	4
11		15 June	Cabernet Sau	144	19	Italy	FALSE	4
12		9 April	Chardonnay	144	57	Chile	FALSE	3
13		8 March	Espumante	6	45	South Africa	FALSE	3
14		27 October	Champagne	72	88	New Zealand	FALSE	3
15		7 March	Prosecco	6	40	Australia	TRUE	3
16		25 October	Cabernet Sau	72	59	Oregon	TRUE	3
17		19 July	Champagne	12	66	Germany	FALSE	3
18		26 October	Pinot Noir	144	83	Australia	FALSE	2
19		10 April		72	52	California	FALSE	2

Figura 2-25: Ordenar no Cluster 4 — esses caras só gostam de chamarhe em agosto?

## A Silhueta: Uma Boa Maneira de Deixar Diferentes Valores de K Competirem

Não há nada de errado com apenas fazer atribuição k-means para alguns valores de k até você encontrar algo que faça lhe algum sentido intuitivo. Claro, talvez a razão de um dado k não “ler bem” não seja porque k está errado mas porque a informação oferecida está omitindo algo que ajudaria a descrever melhor esses grupos.

Então existe outra maneira (que não seja olhar para os grupos) de aprovar ou não um valor de k em particular?

Existe — computando uma pontuação para seus grupos chamada *silhueta*. O legal sobre silhueta é que ela é relativamente agnóstica ao valor de k, então você consegue comparar diferentes valores de k usando uma única pontuação.

## *A Silhueta em Alto Nível: Quão Longe Estão Seus Vizinhos de Você?*

É possível comparar a distância média entre cada cliente e seus amigos no grupo que foram atribuídos com a distância média para os clientes no grupo com o próximo centro mais próximo.

Se eu estou muito mais próximo a pessoas no meu grupo do que pessoas no grupo vizinho, essas pessoas são um bom grupo para mim, certo? Mas e se as pessoas do próximo grupo mais próximo forem quase tão boas para mim quanto meus irmãos de grupo? Bom, então minha atribuição de grupo é um tanto duvidosa, não é?

Uma maneira formal de escrever esse valor é:

*(Distância média para quem está no grupo vizinho mais próximo — Distância média para quem está no meu grupo)/O máximo daquelas duas médias*

O denominador no cálculo mantém o valor entre -1 e 1.

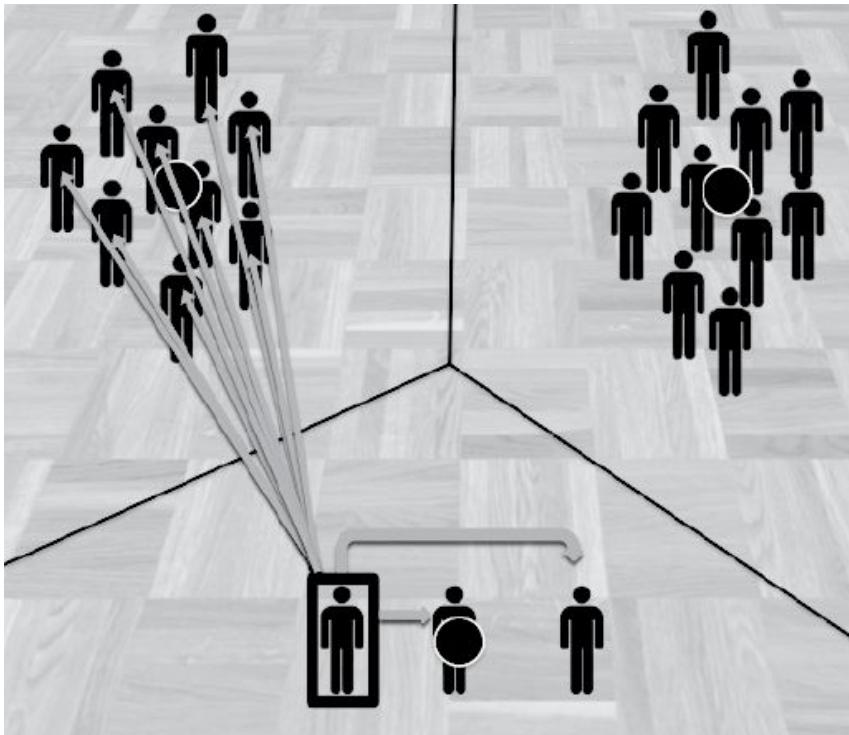
Pense sobre essa fórmula. Conforme os residentes do próximo grupo mais próximo se distanciam mais e mais (mais inadequados para mim), o valor se aproxima de 1. E se as duas distâncias médias são quase a mesma? Então o valor se aproxima de 0.

Pegar a média desse cálculo para cada cliente lhe dá a silhueta. Se a silhueta for 1, está perfeita. Se for 0, os grupos estão inadequados. Se for menos que 0, muitos clientes estão melhores passando um tempo em outro grupo, o que é uma bomba.

E para valores diferentes de k, é possível comparar siluetas para ver se você está melhorando.

Para ver esse conceito mais claramente, volte para o exemplo do baile da escola. A Figura 2-26 exibe uma ilustração dos cálculos de distância usados na formação da silhueta. Note que uma das distâncias do acompanhante de outros dois acompanhantes está sendo comparada às distâncias do próximo grupo mais próximo, que é a turma de meninos da escola.

Agora, os outros dois acompanhantes estão bem mais próximos do que a horda de adolescentes esquisitos, então isso tornaria o cálculo da proporção da distância bem maior do que 0 para esse acompanhante.



**Figura 2-26:** As distâncias consideradas para a contribuição do acompanhante para o cálculo da silhueta

### *Criando uma Matriz de Distância*

Para implementar a silhueta, há uma importante parte de dados que você precisa: a distância entre os clientes. E enquanto os centros de grupos podem se movimentar, a distância entre dois clientes nunca muda. Então você pode apenas criar uma única aba **Distances** e usá-la em todos os seus cálculos de silhueta independente do valor de k ou de onde aqueles centros terminam.

Vamos começar criando uma planilha nova chamada **Distances** e colar nela os clientes nas linhas e nas colunas. Uma célula na matriz armazenará a distância entre o cliente na linha e o cliente na coluna. Para colar os clientes pelas linhas, copie H1:DC1 da aba Matrix e use Paste

Special para colar os valores, certificando-se de escolher a opção Transpose na janela Paste Special.

É preciso observar onde os clientes estão na aba Matrix, então numere os clientes de 0 a 99 em ambas direções. Vamos colocar esses números na coluna A e na linha 1, então insira linhas e colunas vazias para a esquerda e acima dos nomes que você já colou clicando com o botão direito na coluna A e na linha 1 e inserindo uma nova linha 1 e uma nova coluna A.

#### NOTA

Para sua informação, existem muitas formas de colocar aquela enumeração 0–99 no Excel. Por exemplo, você pode digitar os primeiros 0, 1, 2, 3, e então realçá-los e arrastar o canto inferior da seleção pelo resto dos clientes. O Excel entenderá e estenderá a contagem. A matriz vazia resultante está exibida na Figura 2-27.

Considere a célula C3, que é a distância entre Adams e Adams, em outras palavras, entre Adams e ele mesmo. Ela deveria ser 0, correto? Você não pode ficar mais perto de você do que você mesmo.

Então como você calcula isso? Bom, a coluna H na aba Matrix mostra o vetor de oferta de Adams. Para calcular a distância Euclidiana entre Adams e ele mesmo, apenas subtraia a coluna H da coluna H, eleve as diferenças ao quadrado, some-as, e calcule a raiz quadrada.

Mas como você arrasta o cálculo por todas as células na matriz? Eu detestaria digitá-las manualmente. Isso levaria uma eternidade. O que você precisa usar é a fórmula `OFFSET` na célula C3 (veja o Capítulo 1 para uma explicação de `OFFSET`).

A fórmula `OFFSET` assimila uma série de células; neste caso, faz do vetor de oferta de Adams `Matrix!$H$2:$H$33`, e move toda a série um determinado número de linhas e colunas na direção que você especificar.

Então, por exemplo, `OFFSET(Matrix!$H$2:$H$33, 0, 0)` é apenas o vetor de oferta de Adams porque você está movendo a série original 0 linhas

abaixo e 0 colunas à direita.

A	B	C	D	E	F	G	H
1		0	1	2	3	4	
2		Adams	Allen	Anderson	Bailey	Baker	Barnes
3	0	Adams					
4	1	Allen					
5	2	Anderson					
6	3	Bailey					
7	4	Baker					
8	5	Barnes					
9	6	Bell					
10	7	Bennett					
11	8	Brooks					
12	9	Brown					

**Figura 2-27:** A tabela Distance sem preenchimento

Mas `OFFSET(Matrix!$H$2:$H$33, 0, 1)` é a coluna de oferta de Allen.

`OFFSET(Matrix!$H$2:$H$33, 0, 2)` é de Anderson, e assim por diante.

E é aqui que aqueles índices 0–99 na linha 1 e coluna A serão úteis. Por exemplo:

```
{=SQRT(SUM((OFFSET(Matrix!$H$2:$H$33, 0, Distances!C$1) -  
OFFSET(Matrix!$H$2:$H$33, 0, Distances!$A3))^2))}
```

Essa é a distância entre Adams e ele mesmo. Note que você está arrastando `Distances!C$1` para a coluna deslocada no primeiro vetor de oferta e `Distances!$A3` para a coluna deslocada no segundo vetor de oferta.

Desta forma, quando você arrasta esse cálculo para o outro lado e para baixo da planilha, tudo fica ancorado ao vetor de oferta de Adams, mas a fórmula `OFFSET` altera os vetores pela quantidade apropriada usando os índices na coluna A e linha 1. Assim, pegará os dois vetores de ofertas

apropriados para os clientes que você gosta. A Figura 2-28 mostra a matriz de distância preenchida.

Além disso, lembre-se que assim como na aba 4MC, essas distâncias são fórmulas array.

		A	B	C	D	E	F	G	H
1			0	1	2	3	4		
2			Adams	Allen	Anderson	Bailey	Baker	Barnes	
3	0	Adams	0.000	2.236	2.236	1.732	2.646	2.646	
4	1	Allen	2.236	0.000	2.000	2.000	2.449	2.449	
5	2	Anderson	2.236	2.000	0.000	2.000	2.449	2.449	
6	3	Bailey	1.732	2.000	2.000	0.000	2.000	2.449	
7	4	Baker	2.646	2.449	2.449	2.000	0.000	2.000	
8	5	Barnes	2.646	2.449	2.449	2.449	2.000	0.000	
9	6	Bell	2.646	2.449	1.414	2.449	2.828	2.828	
10	7	Bennett	1.732	2.000	2.000	2.000	2.449	2.449	
11	8	Brooks	2.646	2.449	2.449	2.449	2.828	2.449	
12	9	Brown	1.414	2.236	2.236	1.000	2.236	2.646	

Figura 2-28:A matriz de distância completa

### Implementando a Silhueta no Excel

Tudo bem, agora que você tem a aba Distances, pode criar outra aba chamada 4MC Silhouette para o cálculo final da silhueta.

Para começar, vamos copiar os clientes e suas atribuições comunitárias da aba 4MC e Paste Especial os nomes dos clientes na coluna A e as atribuições na B (não esqueça de marcar a caixa Transpose na janela Paste Especial).

Em seguida, pode usar a aba Distances para calcular a distância média entre cada cliente e aqueles em um grupo em particular. Então nomeie as Colunas de C a F como Distance from People **in 1** até Distance from People in 4.

Na minha pasta de trabalho, Adams foi alocado no Grupo 2, então calcule na célula C2 a distância entre ele e todos os clientes no Grupo 1. É preciso procurar os clientes e verificar quais estão no Grupo 1 e então medir suas distâncias de Adams na linha 3 da aba Distances.

Parece um caso para a fórmula AVERAGEIF:

```
=AVERAGEIF('4MC'!$L$39:$DG$39,1,Distances!$C3:$CX3)
```

AVERAGEIF verifica as atribuições de grupos e as corresponde com o Grupo 1 antes de tirar a média das distâncias apropriadas de C3:CX3.

Para as colunas de D a F, as fórmulas são as mesmas menos o Grupo 1, que é substituído por 2, 3, e 4 na fórmula. Você pode dar um clique duplo nessas fórmulas e copiá-las para todos os clientes, rendendo a tabela exibida na Figura 2-29.

		Distance from people in 1	Distance from people in 2	Distance from people in 3	Distance from people in 4
1	Name	Community			
2	Adams	2	2.358	1.495	2.318
3	Allen	3	2.134	2.215	1.980
4	Anderson	1	0.957	2.215	2.097
5	Bailey	2	2.134	1.554	2.080
6	Baker	3	2.562	2.429	2.346

Figura 2-29: Distância média entre cada cliente e todos os clientes em todos os grupos

Na coluna G, pode-se calcular o grupo de clientes mais próximo usando a fórmula MIN. Por exemplo, para Adams, é simplesmente:

```
=MIN(C2:F2)
```

E, na coluna H, pode-se calcular o segundo grupo de clientes mais próximo usando a fórmula SMALL (o 2 na fórmula é para segundo lugar):

```
=SMALL(C2:F2, 2)
```

Da mesma maneira, pode-se calcular a distância para sua própria comunidade de membros (que provavelmente é a mesma da coluna G, mas nem sempre) na coluna I dessa maneira:

=INDEX(C2:F2, B2)

A fórmula INDEX é usada para calcular a distância apropriada da coluna em C até F usando o valor de atribuição em B como índice.

E para o cálculo da silhueta, também é preciso a distância para o grupo de clientes mais próximo que *não* estão no seu grupo, que provavelmente é a coluna H, mas nem sempre. Para chegar nisso na coluna J, verifica-se a distância do seu próprio grupo em I com o grupo mais próximo em G, e, se eles combinarem, o valor é H. Senão, é G.

=IF(I2=G2, H2, G2)

Copiando todos esses valores, você obterá a planilha exibida na Figura 2-30.

Name	Community	Distance from people in 1		Distance from people in 2		Distance from people in 3		Distance from people in 4		Second Closest	My Cluster	Neighboring Cluster
		in 1	in 2	in 3	in 4	Closest	Closest					
Adams	2	2.358	1.495	2.318	2.688	1.495	2.318	1.495	2.318	2.318	2.318	
Allen	3	2.134	2.215	1.980	2.476	1.980	2.134	1.980	2.134	2.134	2.134	
Anderson	1	0.957	2.215	2.097	2.558	0.957	2.097	0.957	2.097	0.957	2.097	
Bailey	2	2.134	1.554	2.080	2.462	1.554	2.080	1.554	2.080	1.554	2.080	
Baker	3	2.562	2.429	2.346	2.703	2.346	2.429	2.346	2.429	2.346	2.429	
Barnes	4	2.562	2.631	2.423	2.345	2.345	2.423	2.345	2.423	2.345	2.423	
Bell	1	1.075	2.621	2.405	2.907	1.075	2.405	1.075	2.405	1.075	2.405	

**Figura 2-30:** As distâncias médias para as pessoas no meu próprio grupo e para o grupo mais próximo no qual ele não esteja

Uma vez que colocou os valores juntos, acrescentar a silhueta para um cliente específico na coluna K é simples:

= (J2 - I2) / MAX(J2, I2)

Você pode simplesmente copiar a fórmula planilha abaixo para conseguir essas proporções para cada cliente.

Você notará que, para alguns clientes, esses valores estão mais próximos de 1. Por exemplo, o valor da silhueta para Anderson na minha solução de agrupamento é 0,544 (veja a Figura 2-31). Nada mal! Mas para outros clientes, como Collins, o valor é, na realidade, menor que 0,

insinuando que, em condições normais, Collins estaria melhor no grupo vizinho do que em seu atual. Coitado.

Agora, é possível tirar a média desses valores para conseguir a quantia da silhueta final. No meu caso, como exibido na Figura 2-31, é 0,1492, o que parece muito mais próximo de 0 do que de 1. Isso é desanimador, mas não totalmente surpreendente. Afinal, dois dos quatro grupos estavam muito instáveis quando você tentou interpretá-los com as descrições de ofertas.

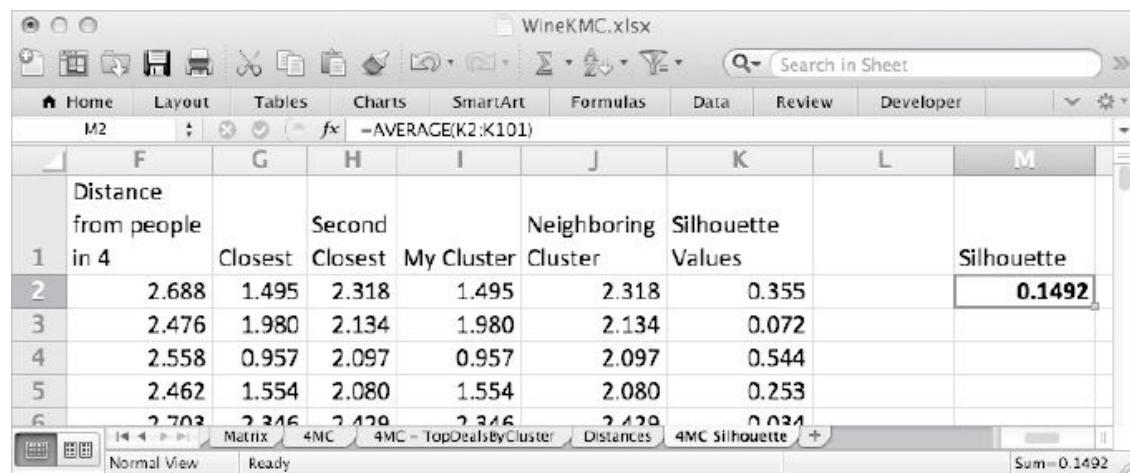


Figura 2-31: A silhueta final para o agrupamento de 4 médias

Tudo bem. E agora?

Claro, a silhueta é 0,1492. Mas o que isso significa? Como você usa isso? Você testa outros valores de k! Então, pode usar a silhueta para ver se está melhorando.

## Que Tal Cinco Grupos?

Tente aumentar para 5 e veja o que acontece.

Estas são as boas notícias: como você já fez quatro grupos, não precisa começar as planilhas do zero. Não é necessário fazer nada com a planilha Distances. Ela está pronta.

Comece criando uma cópia da aba 4MC e chamando-a de 5MC. Tudo o que precisa fazer é adicionar um quinto grupo na planilha e trabalhá-lo nos seus cálculos.

Primeiro, vamos clicar com o botão direito na coluna L e inserir uma nova coluna chamada Cluster 5. É também preciso inserir uma linha Distance to Cluster 5 clicando com o botão direito na linha 38 e selecionando Insert. É possível copiar a linha Distance to Cluster 4 para a linha 38 e mudar a coluna de K para L, para criar a linha Distance to Cluster 5. Quanto às linhas Minimum Cluster Distance e Assigned Cluster, referências à linha 37 devem ser revisadas para 38 para incluir a nova distância de grupo.

Você terá, ao final, a planilha exibida na Figura 2-32.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Offer #	Campaign	Varietal	Minimum	Discount (%)	Origin	Past Peak	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Adams	Allen	Anderson
25	24	September	Pinot Noir	6	34	Italy	FALSE	0.941	0.043	0.017	0.035				1
26	25	October	Cabernet Sa	72	59	Oregon	TRUE	0.025	0.034	0.083	0.114				
27	26	October	Pinot Noir	144	83	Australia	FALSE	0.690	0.030	0.090	0.130				1
28	27	October	Champagne	72	88	New Zealand	FALSE	0.010	0.021	0.087	0.141				1
29	28	November	Cabernet Sa	12	56	France	TRUE	0.026	0.017	0.090	0.030				
30	29	November	Pinot Grigio	6	87	France	FALSE	0.012	0.619	0.043	0.038				1
31	30	December	Malbec	6	54	France	FALSE	0.020	0.729	0.079	0.136				1
32	31	December	Champagne	72	89	France	FALSE	0.023	0.027	0.211	0.259				
33	32	December	Cabernet Sa	72	45	Germany	TRUE	0.093	0.013	0.053	0.125				
34							Distance to	Cluster 1				2.166	1.939	0.740	
35		Total Distance					Distance to	Cluster 2				1.044	1.886	1.865	
36		140.6					Distance to	Cluster 3				1.691	1.339	1.428	
37							Distance to	Cluster 4				2.012	1.731	1.781	
38							Distance to	Cluster 5				1.732	1.414	1.414	
39							Minimum Cluster Distance					1.044	1.339	0.740	
40							Assigned Cluster					2	3	1	

Figura 2-32: A tabela de agrupamento de 5 médias

## Resolvendo para Cinco Grupos

Abrindo o Solver, você só precisa modificar \$H\$2:\$K\$33 para \$H\$2:\$L\$33 tanto nas variáveis de decisão quanto nas seções de restrição para incluir o novo quinto grupo. Todo o restante permanece igual.

Pressione Solve e deixe esse novo problema executar.

Na minha execução, o Solver terminou com uma distância total de 135,1, conforme mostra a Figura 2-33.

A	B	C	D	E	F	G	H	I	J	K	L
Offer #	Campaign	Varietal	Minimum Qty	Discount (%)	Origin	Past Peak	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
1	January	Malbec	72	56	France	FALSE	0.006	0.007	0.001	0.216	0.102
2	January	Pinot Noir	72	17	France	FALSE	0.266	0.003	0.000	0.080	0.101
3	February	Espumante	144	32	Oregon	TRUE	0.007	0.010	0.013	0.191	0.024
4	February	Champagne	72	48	France	TRUE	0.011	0.004	0.012	0.159	0.159
5	February	Cabernet Sauvignon	144	44	New Zealand	TRUE	0.011	0.010	0.000	0.093	0.081
6	March	Prosecco	144	86	Chile	FALSE	0.008	0.010	0.013	0.255	0.102
7	March	Prosecco	6	40	Australia	TRUE	0.011	0.607	0.000	0.113	0.113
8	March	Espumante	6	45	South Africa	FALSE	0.011	0.320	1.000	0.127	0.001
9	April	Chardonnay	144	57	Chile	FALSE	0.009	0.010	0.000	0.080	0.209
10	April	Prosecco	72	52	California	FALSE	0.011	0.005	0.012	0.084	0.085
11	May	Champagne	72	85	France	FALSE	0.006	0.009	0.000	0.263	0.166
12	May	Prosecco	72	83	Australia	FALSE	0.011	0.003	0.002	0.110	0.071
13	May	Merlot	6	43	Chile	FALSE	0.011	0.205	0.016	0.010	0.005
14	June	Merlot	72	64	Chile	FALSE	0.007	0.011	0.000	0.156	0.123
15	June	Cabernet Sauvignon	144	19	Italy	FALSE	0.011	0.002	0.005	0.149	0.048
16	June	Merlot	72	88	California	FALSE	0.011	0.008	0.001	0.012	0.105
17	July	Pinot Noir	12	47	Germany	FALSE	0.611	0.001	0.008	0.004	0.009
18	July	Espumante	6	50	Oregon	FALSE	0.010	0.475	0.028	0.057	0.027
19	July	Champagne	12	66	Germany	FALSE	0.008	0.008	0.000	0.116	0.052
20	August	Cabernet Sauvignon	72	82	Italy	FALSE	0.011	0.008	0.003	0.033	0.100
21	August	Champagne	12	50	California	FALSE	0.011	0.010	0.005	0.085	0.048
22	August	Champagne	72	63	France	FALSE	0.007	0.009	0.004	1.000	0.004
23	September	Chardonnay	144	39	South Africa	FALSE	0.011	0.007	0.008	0.077	0.072
24	September	Pinot Noir	6	34	Italy	FALSE	1.000	0.011	0.004	0.005	0.009
25	October	Cabernet Sauvignon	72	59	Oregon	TRUE	0.011	0.010	0.008	0.099	0.082
26	October	Pinot Noir	144	83	Australia	FALSE	0.719	0.008	0.000	0.033	0.147
27	October	Champagne	72	88	New Zealand	FALSE	0.010	0.011	0.021	0.152	0.112
28	November	Cabernet Sauvignon	12	56	France	TRUE	0.010	0.011	0.000	0.068	0.100
29	November	Pinot Grigio	6	87	France	FALSE	0.005	0.679	0.044	0.008	0.048
30	December	Malbec	6	54	France	FALSE	0.006	0.760	0.021	0.182	0.051
31	December	Champagne	72	89	France	FALSE	0.008	0.006	0.013	0.310	0.239
32	December	Cabernet Sauvignon	72	45	Germany	TRUE	0.000	0.003	0.004	0.039	0.065
33	32										
34							Distance to Cluster 1				
35	Total Distance						Distance to Cluster 2				
36	135.1						Distance to Cluster 3				

Figura 2-33: Os otimizados grupos de 5 médias

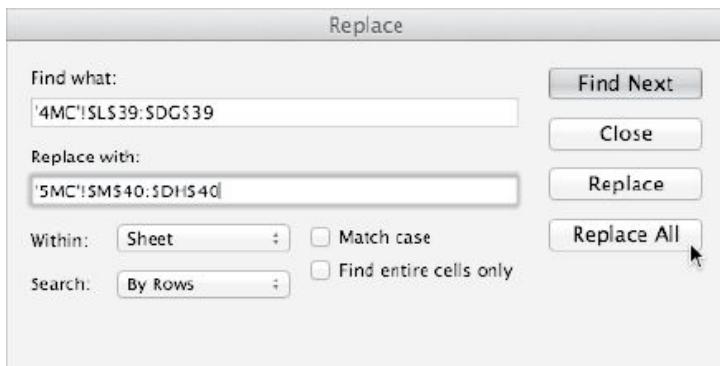
## Conseguindo as Melhores Ofertas para Todos os Cinco Grupos

Tudo bem. Vamos ver como você se saiu.

Você pode criar uma cópia da aba 4MC — TopDealsByCluster e renomeá-la para **5MC — TopDealsByCluster**, mas precisará revisar algumas das fórmulas para fazê-la funcionar.

Primeiramente, é preciso assegurar que essa planilha esteja ordenada por Offer# na coluna A. Então nomeie a coluna L com um 5 e arraste as fórmulas de K até L. Você também deveria selecionar as colunas de A a L e reaplicar o filtro automático para tornar as aquisições de ofertas do Cluster 5 ordenáveis.

Tudo nessa planilha está atualmente apontando para a aba 4MC, então está na hora de usar o velho Find and Replace. As atribuições de grupos na aba 5MC estão deslocadas uma linha para baixo e uma coluna para a direita, então a referência para '4MC'!\$L\$39:\$D\$39 nas fórmulas SUMIF deveriam se tornar '5MC'!\$M\$40:\$DH\$40. Conforme ilustra a Figura 2-34, pode-se usar Find and Replace para mudar isso.



**Figura 2-34** Modificando as atribuições de grupo de 4 médias para 5 médias

#### NOTA

Lembre-se de que seus resultados serão diferentes dos meus devido ao solver evolucionário.

Ordenando no Cluster 1, você claramente tem seu grupo Pinot Noir novamente (veja a Figura 2-35).

WineKMC.xlsx

	A	B	C	D	E	F	G	H
1	Offer #	Offer date	Product	Minimum	Discount	Origin	Past Peak	1
2	24	September	Pinot Noir	6	34	Italy	FALSE	12
3	26	October	Pinot Noir	144	83	Australia	FALSE	8
4	17	July	Pinot Noir	12	47	Germany	FALSE	7
5	2	January	Pinot Noir	72	17	France	FALSE	4
6	1	January	Malbec	72	56	France	FALSE	0
7	3	February	Espumante	144	32	Oregon	TRUE	0
8	4	February	Champagne	72	48	France	TRUE	0
9	5	February	Cabernet Sauvignon	144	44	New Zealand	TRUE	0
10	6	March	Prosecco	144	86	Chile	FALSE	0
11	7	March	Prosecco	6	40	Australia	TRUE	0

Figura 2-35: Ordenar no Cluster 1 — Pinot Noir saindo pelas orelhas

O Cluster 2 é o grupo de compradores de pouca quantidade (veja a Figura 2-36).

WineKMC.xlsx

	A	B	C	D	E	F	G	I
1	Offer #	Offer date	Product	Minimum	Discount	Origin	Past Peak	2
2	30	December	Malbec	6	54	France	FALSE	15
3	29	November	Pinot Grigio	6	87	France	FALSE	13
4	7	March	Prosecco	6	40	Australia	TRUE	12
5	18	July	Espumante	6	50	Oregon	FALSE	10
6	8	March	Espumante	6	45	South Africa	FALSE	7
7	13	May	Merlot	6	43	Chile	FALSE	5
8	24	September	Pinot Noir	6	34	Italy	FALSE	0
9	26	October	Pinot Noir	144	83	Australia	FALSE	0
10	17	July	Pinot Noir	12	47	Germany	FALSE	0
11	2	January	Pinot Noir	72	17	France	FALSE	0

Figura 2-36: Ordenar no Cluster 2 — apenas pequenas quantidades, por favor

Quanto ao Cluster 3, esse faz minha cabeça doer. Parece existir apenas um Espumante Sul-africano, que é importante por algum motivo (veja a Figura 2-37).

WineKMC.xlsx

	A	B	C	D	E	F	G	J
1	Offer #	Offer date	Product	Minimum	Discount	Origin	Past Peak	3
2	8	March	Espumante	6	45	South Africa	FALSE	10
3	29	November	Pinot Grigio	6	87	France	FALSE	2
4	18	July	Espumante	6	50	Oregon	FALSE	2
5	30	December	Malbec	6	54	France	FALSE	1
6	13	May	Merlot	6	43	Chile	FALSE	1
7	3	February	Espumante	144	32	Oregon	TRUE	1
8	4	February	Champagne	72	48	France	TRUE	1
9	6	March	Prosecco	144	86	Chile	FALSE	1
10	10	April	Prosecco	72	52	California	FALSE	1
11	27	October	Champagne	72	88	New Zealand	FALSE	1

Figura 2-37: Ordenar no Cluster 3 → Espumante é tão importante assim?

Os clientes do Cluster 4 estão interessados em grandes quantidades, essencialmente ofertas francesas com bons descontos. Pode até mesmo existir uma propensão a vinhos espumantes. Esse grupo é difícil de ler; tem muita coisa acontecendo (veja a Figura 2-38).

	A	B	C	D	E	F	G	K
1	Offer #	Offer date	Product	Minimum	Discount	Origin	Past Peak	4
2	22	August	Champagne	72	63	France	FALSE	21
3	31	December	Champagne	72	89	France	FALSE	7
4	6	March	Prosecco	144	86	Chile	FALSE	6
5	1	January	Malbec	72	56	France	FALSE	5
6	11	May	Champagne	72	85	France	FALSE	5
7	30	December	Malbec	6	54	France	FALSE	4
8	3	February	Espumante	144	32	Oregon	TRUE	4
9	4	February	Champagne	72	48	France	TRUE	4
10	14	June	Merlot	72	64	Chile	FALSE	4
11	15	June	Cabernet Sau	144	19	Italy	FALSE	4
12	8	March	Espumante	6	45	South Africa	FALSE	3
13	27	October	Champagne	72	88	New Zealand	FALSE	3
14	7	March	Prosecco	6	40	Australia	TRUE	3
15	19	July	Champagne	12	66	Germany	FALSE	3
16	10	April	Prosecco	72	52	California	FALSE	2
17	2	January	Pinot Noir	72	17	France	FALSE	2
18	9	April	Chardonnay	144	57	Chile	FALSE	2
19	12	May	Prosecco	72	83	Australia	FALSE	2
20	21	August	Champagne	12	50	California	FALSE	2
21	23	September	Chardonnay	144	39	South Africa	FALSE	2
22	25	October	Cabernet Sau	72	59	Oregon	TRUE	2
23	28	November	Cabernet Sau	12	56	France	TRUE	2
24	18	July	Espumante	6	50	Oregon	FALSE	1
25	26	October						1

Figura 2-38: Ordenar no Cluster 4 — todo tipo de interesses

Ordenar no Cluster 5 lhe fornece resultados parecidos aos do Cluster 4, embora grandes quantidade e descontos altos pareçam ser os fatores primordiais (veja a Figura 2-39).

## Computando a Silhueta para Agrupamento de 5 Médias

Você pode estar imaginando se cinco grupos foram melhor do que quatro. De uma perspectiva inicial, não parece ter muita diferença. Vamos computar a silhueta para cinco grupos e ver o que o computador pensa.

Comece fazendo uma cópia de 4MC Silhouette e renomeando-a para 5MC Silhouette. Em seguida, clique com o botão direito na coluna G,

insira uma nova coluna, e nomeie-a Distance From People in 5. Arraste a fórmula de F2 para G2, mude a verificação de grupo de 4 para 5, e então dê um clique duplo no canto inferior direito da célula para dispará-la planilha abaixo.

	A	B	C	D	E	F	G	L
1	Offer #	Offer date	Product	Minimum	Discount	Origin	Past Peak	5
2	31 December	Champagne	72	89	France	FALSE	9	
3	11 May	Champagne	72	85	France	FALSE	8	
4	9 April	Chardonnay	144	57	Chile	FALSE	8	
5	4 February	Champagne	72	48	France	TRUE	7	
6	26 October	Pinot Noir	144	83	Australia	FALSE	6	
7	6 March	Prosecco	144	86	Chile	FALSE	5	
8	1 January	Malbec	72	56	France	FALSE	5	
9	14 June	Merlot	72	64	Chile	FALSE	5	
10	27 October	Champagne	72	88	New Zealand	FALSE	5	
11	20 August	Cabernet Sau	72	82	Italy	FALSE	5	
12	16 June	Merlot	72	88	California	FALSE	5	
13	7 March	Prosecco	6	40	Australia	TRUE	4	
14	10 April	Prosecco	72	52	California	FALSE	4	
15	2 January	Pinot Noir	72	17	France	FALSE	4	
16	25 October	Cabernet Sau	72	59	Oregon	TRUE	4	
17	28 November	Cabernet Sau	12	56	France	TRUE	4	
18	12 May	Prosecco	72	83	Australia	FALSE	3	
19	23 September	Chardonnay	144	39	South Africa	FALSE	3	
20	5 February	Cabernet Sau	144	44	New Zealand	TRUE	3	
21	32 December	Cabernet Sau	72	45	Germany	TRUE	3	
22	30 December	Malbec	6	54	France	FALSE	2	
23	15 June	Cabernet Sau	144	19	Italy	FALSE	2	
24	19 July	Champagne	12	66	Germany	FALSE	2	
25	21 August	Champagne	12	10	California	FALSE	2	

Figura 2-39: Ordenar no Cluster 5 — grande quantidade

Como na seção anterior, você precisará de Find and Replace de

'4MC' !\$L\$39:\$DG\$39 para '5MC' !\$M\$40:\$DH\$40.

Nas células H2, I2, e J2, deve-se incluir as distâncias para as pessoas no Cluster 5 em seus cálculos, então quaisquer séries que parem em F2 devem ser expandidas para incluir G2. Você pode ressaltar H2:J2 e dar um duplo clique no canto inferior direito para enviar esses cálculos atualizados planilha abaixo.

Finalmente, você precisa copiar e Paste Special valores das atribuições de grupo na linha 40 da aba 5MC para a coluna B na aba 5MC Silhouette. Isso significa que deve marcar o botão Transpose ao usar Paste Special.

Uma vez que tenha revisado a planilha, deverá obter algo parecido com o que está na Figura 2-40.

The screenshot shows a Microsoft Excel spreadsheet titled "WineKMC.xlsx". The active sheet is "5MC Silhouette". The table has columns for "Distance from people in 5", "Second Closest", "My Community", "Neighboring Community", "Silhouette Values", and "Silhouette". Row 1 contains the column headers. Rows 2 through 8 contain data points. Row 8 is highlighted in yellow. The formula bar shows "=AVERAGE(L2:L101)". The status bar at the bottom right says "Sum= 0.134".

	G	H	I	J	K	L	M	N
1	Distance from people in 5	Second Closest	My Community	Neighboring Community	Silhouette Values			Silhouette
2	2.371	1.434	2.031	1.434	2.031	0.294		0.134
3	2.017	1.975	2.017	2.017	1.975	-0.021		
4	2.135	0.957	2.033	0.957	2.033	0.529		
5	2.124	1.483	1.975	1.483	1.975	0.249		
6	2.381	2.381	2.405	2.381	2.405	0.010		
7	2.468	2.285	2.405	2.285	2.405	0.050		
8	2.521	1.075	2.481	1.075	2.481	0.567		

**Figura 2-40:** A silhueta para agrupamento de 5 médias

Bom, isso é deprimente, não é? A silhueta não é tão diferente. Em 0,134, é na verdade um pouco pior! Mas não é tanto uma surpresa após explorar os grupos. Em ambos os casos, você tinha três grupos que realmente faziam sentido. Os outros eram barulhentos. Talvez você devesse ir em outra direção e tentar  $k=3$ ? Se quiser experimentar isso, eu deixo como exercício para que tente sozinho.

Em vez disso, vamos refletir um pouco sobre o que pode estar dando errado aqui para gerar esses grupos ruidosos e desconcertantes.

## Agrupamento K-Medians e Medidas de Distância Assimétricas

Geralmente, fazer agrupamento k-means com distância Euclidiana está bom, mas você encontrou alguns problemas que muitos que fazem

agrupamento de dados esparsos (seja em vendas no varejo ou classificação de texto ou bioinformática) frequentemente encontram.

## Usando↑Agrupamento↑K-Medians

O primeiro problema óbvio é que seus centros de grupos são decimais ainda que o vetor de oferta do cliente seja de zeros e uns. O que 0,113 realmente significa? Eu quero centros de grupos que comprometam-se com uma oferta ou não!

Se você modificar o algoritmo de agrupamento para usar apenas valores apresentados nos vetores de oferta dos clientes, isso é chamado de agrupamento *K-medians*, em vez de agrupamento k-means.

Se você quisesse ficar com distância Euclidiana, tudo o que precisaria fazer seria adicionar uma restrição binária, (`bin`) no Solver para todos os seus centros de grupos.

Mas se tornar seus centros de grupos binários, a distância Euclidiana é o que você quer?

## Obtendo↑uma↑Distância↑Métrica↑Mais Apropriada

Quando as pessoas mudam de k-means para k-medians, elas normalmente param de usar a distância Euclidiana e começam a usar algo chamado *distância Manhattan*.

Embora um corvo possa voar do ponto A ao B em uma linha reta, um táxi em Manhattan tem que ficar no espaço de ruas retas; ele só pode ir para o norte, sul, leste e oeste. Então, enquanto na Figura 2-13 você viu que a distância entre um dançarino da escola de ensino médio e seu centro de grupo era aproximadamente 4,47, sua distância Manhattan era de 6 unidades (isso são 4 unidades abaixo mais 2 unidades à frente).

Em termos de dados binários, como os dados de compras, a distância Manhattan entre um centro de grupo e um vetor de compra do cliente é apenas a conta das incompatibilidades. Se o centro de grupo tem um 0 e

eu tenho um 0, naquela direção há uma distância de 0, e se você não igualou 0 e 1, tem uma distância de 1 naquela direção. Somando-as, você obtém a distância total, que é apenas o número de incompatibilidades. Ao trabalhar com dados binários como esses, a distância Manhattan é também comumente chamada de *distância de Hamming*.

### **A Distância Manhattan Resolve os Problemas?**

Antes de você mergulhar de cabeça em fazer agrupamento k-medians usando a distância Manhattan, pare e pense sobre os dados de compra.

O que significa quando os clientes aceitam uma oferta? Isso significa que eles realmente queriam aquele produto!

O que significa quando os clientes não aceitam uma oferta? Significa que eles não queriam o produto tanto quanto eles queriam aquele que compraram? Um sinal negativo é tão forte quanto o positivo? Talvez eles tenham gostado da Champanhe mas já tinham muitas em estoque. Talvez apenas não tenham visto seu e-mail informativo naquele mês. Existem inúmeras razões para alguém não fazer alguma coisa, mas existem poucas razões para o fazerem.

Em outras palavras, você deveria se preocupar com as compras, não com as não-compras.

Uma forma elegante de falar isso é dizer que existe uma “assimetria” nos dados. Os 1s valem mais do que os 0s. Se um cliente corresponde a outro em três 1s, isso é mais importante do que corresponder a outro cliente em três 0s. O que cheira mal, no entanto, é que enquanto 1s são tão importantes, existem pouquíssimos deles nos dados — consequentemente, o termo “esparsos”.

Mas pense sobre o que significa para um cliente estar próximo de um centro de grupo em uma perspectiva Euclidiana. Se eu tenho um cliente com um 1 para uma oferta e um 0 para outra, ambas são importantes para calcular se um cliente está perto de um centro de grupo.

O que você precisa fazer é um *cálculo de distância assimétrica*. E para dados transacionais codificados binários, como essas compras de vinhos,

existem muitos bons.

Talvez o *cálculo de distância assimétrica* para dados 0-1 mais utilizado seja algo chamado *distância cosseno*.

### ***A Distância Cosseno Não É Assustadora, Apesar da Trigonometria***

A forma mais fácil de explicar a distância cosseno é explicando seu oposto: *similaridade cosseno*.

Digamos que você tenha dois vetores de compras binários bidimensionais  $(1, 1)$  e  $(1, 0)$ . No primeiro vetor, ambos os produtos foram comprados, enquanto, no segundo, apenas o primeiro produto foi adquirido. É possível visualizar esses dois vetores de compra no espaço e ver que eles têm um ângulo de 45 graus entre si (veja a Figura 2-41). Vá em frente, use o transferidor e verifique.

Você pode dizer que eles possuem uma similaridade cosseno de  $\cos(45 \text{ graus})=0,707$ . Mas por quê?

Acontece que o cosseno de um ângulo entre dois vetores de compra binários é igual a:

*A contagem de compras correspondentes nos dois vetores dividida pelo produto da raiz quadrada do número de compras no primeiro vetor vezes a raiz quadrada do número de compras no segundo vetor.*

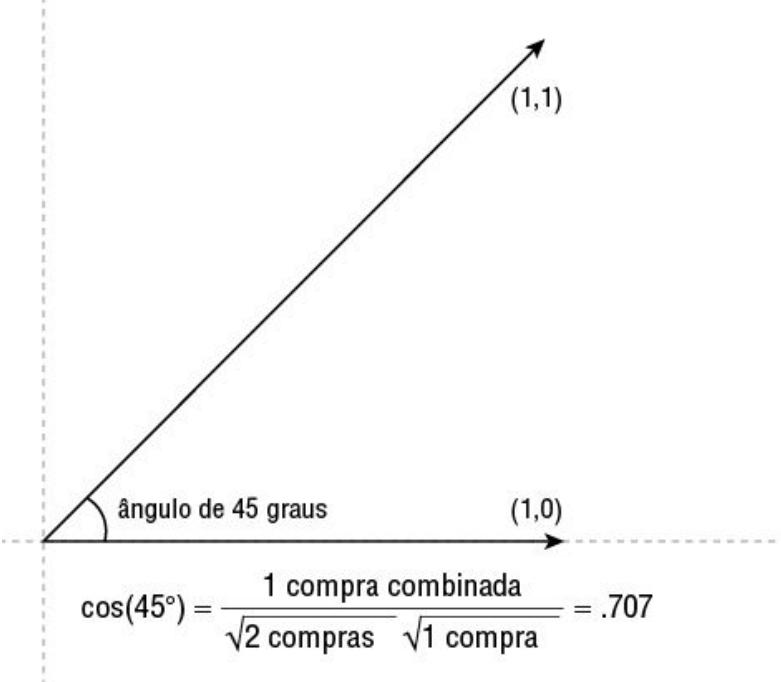
No caso dos dois vetores  $(1, 1)$  e  $(1, 0)$ , eles têm uma compra correspondente, então o cálculo é 1 dividido pela raiz quadrada de 2 (2 ofertas aceitas), vezes a raiz quadrada de uma oferta aceita. E o resultado é 0,707 (veja a Figura 2-41).

Por que esse cálculo é tão legal?

Três razões:

- O numerador no cálculo conta números de compras correspondentes apenas, então essa é uma medida assimétrica, que é o que você está procurando.

- Ao dividir pela raiz quadrada do número de compras em cada vetor, você está justificando o fato de que um vetor **onde tudo é comprado**, chame-o de vetor de compras promíscuo, está ainda mais distante de outro vetor do que o que corresponde nas mesmas ofertas e não aceitou muitas outras. Você quer combinar vetores cujos gostos correspondam, não um vetor que inclui o gosto do outro.
- Para dados binários, esse valor de similaridade varia entre 0 e 1, onde dois vetores não chegam a 1 a não ser que suas compras sejam idênticas. Isso significa que  $1 - \text{similaridade}$  cosseno pode ser usada como a distância métrica chamada distância cosseno, que também varia entre 0 e 1.



**Figura 2-41:** Uma ilustração de similaridade cosseno em dois vetores de compras binários

## Colocando Tudo no Excel

Está na hora de experimentar agrupamento k-medians com distância cosseno no Excel.

## NOTA

O agrupamento com distância cosseno também é algumas vezes conhecido como k-means esférico. No Capítulo 10, você verá k-means esférico em R.

Por razões de consistência, continue usando k=5.

Comece fazendo uma cópia da aba 5MC e nomeando-a 5MedC. Como os centros de grupo precisam ser binários, você pode apagar o que o Solver deixou lá.

Os únicos itens que precisa modificar (além de adicionar a restrição binária no Solver para k-medians) são os cálculos de distância nas linhas 34 a 38. Comece na célula M34, que é a distância entre Adams e o centro de Cluster 1.

Para contar as correspondências de ofertas entre Adams e Cluster 1, você precisa usar SUMPRODUCT nas duas colunas. Se nenhum ou ambos tiverem 0s, eles não obtêm nada por aquela linha, mas se ambos têm um 1, essa correspondência será totalizada por SUMPRODUCT (já que 1 vezes 1 é 1 no final das contas).

Quanto a resolver a raiz quadrada do número de ofertas aceitas em um vetor, isso é apenas um SQRT aplicado a SUM do vetor. Assim, a equação da distância total pode ser escrita como:

```
=1-SUMPRODUCT(M$2:M$33,$H$2:$H$33)/  
(SQRT(SUM(M$2:M$33))*SQRT(SUM($H$2:$H$33)))
```

Note o 1- no início da fórmula, que muda de similaridade cosseno para distância. Além disso, diferente da distância Euclidiana, o cálculo da distância cosseno não exige o uso de fórmulas array.

No entanto, quando coloca isso na célula M34, deve adicionar uma verificação de erro no caso do centro de grupo ser todo de 0s:

```
=IFERROR(1-SUMPRODUCT(M$2:M$33,$H$2:$H$33)/  
(SQRT(SUM(M$2:M$33))*SQRT(SUM($H$2:$H$33))),1)
```

Adicionar a fórmula `IFERROR` evita que você tenha uma situação de divisão por 0. Se, por algum motivo, o Solver selecionar um centro de grupo todo de 0s, então você pode considerar que aquele centro tem uma distância de 1 para tudo (1 sendo a maior distância possível nessa configuração binária).

Você pode então copiar M34 até M38 e mudar as referências da coluna H para I, J, K, e L respectivamente. Assim como no caso da distância Euclidiana, usa-se referências absolutas (\$) na fórmula para que possa arrastá-la pela planilha sem que as colunas do centro de grupo mudem.

Isso gera a planilha 5MedC (veja a Figura 2-42) que é notavelmente similar à aba anterior 5MC.

Agora, para encontrar os grupos, você precisar abrir o Solver e mudar a restrição `<=1` de H2:L33 para ler como binário ou restrição `bin.`

Pressione Solve. Você pode descansar por meia hora enquanto o computador encontra os melhores grupos. Agora, notará claramente que os centros de grupos são todos binários, então, igualmente, a formatação condicional passa para duas tonalidades, o que é muito mais extremo.

## As 5 Melhores Ofertas para os 5 Grupos de Médias

Quando o Solver finaliza, você fica com cinco centros de grupo, cada qual com alguns 1s, indicando quais ofertas são preferidas por aquele grupo. Na minha execução do Solver, eu terminei com um valor objetivo de 42,8, embora o seu possa certamente variar (veja a Figura 2-43).

WineKMC.xlsx

**Figura 2-42:** A tabela 5MedC ainda não otimizada

WineKMC.xlsx

**Figura 2-43:** O grupo de cinco médias

Vamos entender esses grupos usando as mesmas técnicas de contagem de ofertas que você usou em k-means. Para fazer isso, a primeira coisa que precisa fazer é criar uma cópia de 5MC — aba TopDealsByCluster e renomeá-la para 5MedC — TopDealsByCluster.

Nessa aba, tudo o que precisa para fazê-la funcionar é encontrar e substituir 5MC por 5MedC. Como o layout das linhas e colunas entre essas duas planilhas é idêntico, todos os cálculos são transferidos uma vez que a referência da planilha é modificada.

Agora, seus grupos podem estar um pouco diferentes do que os meus tanto na ordem quanto na composição por causa do algoritmo evolucionário, mas espero que não tão substancialmente. Vamos passear pelos meus grupos um por vez para ver como o algoritmo particionou os clientes.

Ordenando no Cluster 1, é evidente que esse é o grupo de pouca quantidade (veja a Figura 2-44).

	A	B	C	D	E	F	G	H
1	Offer #	Offer date	Product	Minimum	Discount	Origin	Past Peak	1
2		29 November	Pinot Grigio	6	87 France	FALSE		16
3		30 December	Malbec	6	54 France	FALSE		16
4		7 March	Prosecco	6	40 Australia	TRUE		15
5		8 March	Espumante	6	45 South Africa	FALSE		15
6		18 July	Espumante	6	50 Oregon	FALSE		13
7		13 May	Merlot	6	43 Chile	FALSE		6
8		10 April	Prosecco	72	52 California	FALSE		2
9		3 February	Espumante	144	32 Oregon	TRUE		1
10		6 March	Prosecco	144	86 Chile	FALSE		1
11		12 May	Prosecco	72	83 Australia	FALSE		1
12		21 August	Champagne	12	50 California	FALSE		1
13		28 November	Cabernet Sauvignon	12	56 France	TRUE		1
14		1 January	Malbec	72	56 France	FALSE		0
15		2 January	Pinot Noir	72	17 France	FALSE		0
16								0

Figura 2-44: Ordenar no Cluster 1 — clientes de baixa quantidade

O Cluster 2 gravou os clientes que compram apenas vinho espumante. Champanhe, Prosecco e Espumante dominam os 11 melhores lugares no

grupo (veja a Figura 2-45). É interessante notar que a abordagem k-means não demonstrou tão claramente o grupo espumante com K igual a 4 ou 5.

O Cluster 3 é o nosso grupo Francófilo. As cinco melhores ofertas são todas francesas (veja a Figura 2-46). Eles não sabem que os vinhos californianos são melhores?

	A	B	C	D	E	F	G	I
1	Offer #	Offer date	Product	Minimum	Discount	Origin	Past Peak	2
2		6 March	Prosecco	144	86	Chile	FALSE	6
3		4 February	Champagne	72	48	France	TRUE	6
4		22 August	Champagne	72	63	France	FALSE	6
5		27 October	Champagne	72	88	New Zealand	FALSE	6
6		19 July	Champagne	12	66	Germany	FALSE	5
7		31 December	Champagne	72	89	France	FALSE	5
8		7 March	Prosecco	6	40	Australia	TRUE	4
9		8 March	Espumante	6	45	South Africa	FALSE	4
10		3 February	Espumante	144	32	Oregon	TRUE	4
11		21 August	Champagne	12	50	California	FALSE	2
12		10 April	Prosecco	72	52	California	FALSE	1
13		29 November	Pinot Grigio	6	87	France	FALSE	0
14		30 December	Malbec	6	54	France	FALSE	0
15		18 July	Espumante	6	50	Oregon	FALSE	0
16		12 May						0

**Figura 2-45:** Ordenar no Cluster 2 — nem todos que brilham são vampiros

	A	B	C	D	E	F	G	J
1	Offer #	Offer date	Product	Minimum	Discount	Origin	Past Peak	3
2		22 August	Champagne	72	63	France	FALSE	10
3		31 December	Champagne	72	89	France	FALSE	7
4		1 January	Malbec	72	56	France	FALSE	7
5		11 May	Champagne	72	85	France	FALSE	6
6		30 December	Malbec	6	54	France	FALSE	5
7		9 April	Chardonnay	144	57	Chile	FALSE	5
8		14 June	Merlot	72	64	Chile	FALSE	4
9		4 February	Champagne	72	48	France	TRUE	2
10		10 April	Prosecco	72	52	California	FALSE	2
11		28 November	Cabernet Sauvignon	12	56	France	TRUE	2
12		2 January	Pinot Noir	72	17	France	FALSE	2
13		23 September	Chardonnay	144	39	South Africa	FALSE	2
14		8 March	Espumante	6	45	South Africa	FALSE	1
15		3 February	Espumante	144	32	Oregon	TRUE	1
16		21 August	Champagne	17	50	California	FALSE	1

**Figura 2-46:** Ordenar no Cluster 3 — Francófilos

No Cluster 4, todas as ofertas são de alto volume. E as melhores ofertas foram as que tiveram os maiores descontos e não passaram do pico (Figura 2-47).

	A	B	C	D	E	F	G	K
1	Offer #	Offer date	Product	Minimum	Discount	Origin	Past Peak	4
2		11 May	Champagne	72	85	France	FALSE	6
3		20 August	Cabernet Sauvignon	72	82	Italy	FALSE	6
4		22 August	Champagne	72	63	France	FALSE	5
5		31 December	Champagne	72	89	France	FALSE	5
6		9 April	Chardonnay	144	57	Chile	FALSE	5
7		14 June	Merlot	72	64	Chile	FALSE	5
8		15 June	Cabernet Sauvignon	144	19	Italy	FALSE	5
9		25 October	Cabernet Sauvignon	72	59	Oregon	TRUE	5
10		6 March	Prosecco	144	86	Chile	FALSE	5
11		16 June	Merlot	72	88	California	FALSE	5
12		4 February	Champagne	72	48	France	TRUE	4
13		12 May	Prosecco	72	83	Australia	FALSE	4
14		5 February	Cabernet Sauvignon	144	44	New Zealand	TRUE	4
15		32 December	Cabernet Sauvignon	72	45	Germany	TRUE	4
16		26 October	Pinot Noir	144	83	Australia	FALSE	3
17		28 November	Cabernet Sauvignon	12	56	France	TRUE	2
18		23 September	Chardonnay	144	39	South Africa	FALSE	2
19		27 October	Champagne	72	88	New Zealand	FALSE	2
20		1 January	Malbec	72	56	France	FALSE	1
21		30 December	Malbec	6	54	France	FALSE	1
22		10 April	Prosecco	72	52	California	FALSE	1
23		29 November	Pinot Grigio	6	87	France	FALSE	1
24		2 January	Pinot Noir	72	17	France	FALSE	0

Figura 2-47: Ordenar no Cluster 4 grande quantia para 19 ofertas em sequência

O Cluster 5 é o grupo Pinot Noir mais uma vez (veja a Figura 2-48).

Isso parece bem mais claro, não parece? Isso é porque no caso k-medians, usando a medida de distância assimétrica como distância cosseno, você pode agrupar clientes baseando-se em seus interesses mais do que nos desinteresses. E isso é o que realmente importa.

Que diferença uma medida de distância faz!

Então, agora você pode pegar essas cinco atribuições de grupos, importá-las de volta a MailChimp.com como um campo mesclado na lista de e-mails, e usar os valores para customizar seu e-mail marketing por grupo. Isso deve ajudá-lo a direcionar melhor os clientes e impulsionar as vendas.

	A	B	C	D	E	F	G	L
1	Offer #	Offer date	Product	Minimum	Discount	Origin	Past Peak	5
2		24 September	Pinot Noir	6	34	Italy	FALSE	12
3		26 October	Pinot Noir	144	83	Australia	FALSE	11
4		2 January	Pinot Noir	72	17	France	FALSE	8
5		17 July	Pinot Noir	12	47	Germany	FALSE	7
6		1 January	Malbec	72	56	France	FALSE	2
7		11 May	Champagne	72	85	France	FALSE	1
8		28 November	Cabernet Sauvignon	12	56	France	TRUE	1
9		23 September	Chardonnay	144	39	South Africa	FALSE	1
10		27 October	Champagne	72	88	New Zealand	FALSE	1
11		10 April	Prosecco	72	52	California	FALSE	1
12		20 August	Cabernet Sauvignon	72	82	Italy	FALSE	0
13		22 August	Champagne	72	63	France	FALSE	0
14		31 December	Champagne	72	89	France	FALSE	0

Figura 2-48: Ordenar pelo Cluster 5 — linha principal Pinot Noir

## Resumindo

Este capítulo abordou várias coisas boas. Para resumir, você viu:

- Distância Euclidiana
- Agrupamento k-means usando Solver para otimizar os centros
- Como entender os agrupamentos uma vez que os têm
- Como calcular a silhueta de uma determinada execução k-means
- Agrupamento k-medians
- Distância Manhattan/Hamming
- Similaridade e Distância cosseno

Se você conseguiu passar por esse capítulo, deveria sentir-se confiante não somente em como agrupar dados, mas também em quais questões podem ser respondidas em negócios por meio de agrupamento, e como preparar seus dados para agrupamento.

O agrupamento k-means já existe há décadas e definitivamente é o lugar para começar para alguém que pretende segmentar e buscar conhecimentos dos dados de seus clientes. Mas não é a técnica de

agrupamento mais “atual”. No Capítulo 5, você explorará usando gráficos de redes para encontrar comunidades de clientes dentro desse mesmo conjunto de dados. Você irá até mesmo em uma viagem fora do Excel, bem rapidamente, para visualizar os dados.

Se deseja se aprofundar em agrupamento k-means, lembre-se que o Excel sustenta até 200 variáveis de decisão no Solver, então você precisa atualizar para um Solver *não linear* (por exemplo, Premium Solver disponível em Solver.com — conteúdo em inglês, ou apenas migrar para um Solver não linear no LibreOffice) para agrupar em dados com muitas dimensões e um alto valor de k.

A maioria dos softwares estatísticos oferecem capacidades de agrupamento. Por exemplo, R vem com a função `kmeans()`; entretanto, capacidades do pacote `fastcluster`, que inclui k-medians e uma variedade de funções de distância, é preferível. No Capítulo 10, você verá o pacote `skmeans` para realizar k-means esférico.

# 3

## Naïve Bayes e a incrível leveza de ser um idiota

No capítulo anterior, você começou trabalhando um pouco com aprendizado não supervisionado. Viu agrupamento k-means, que é como o nugget de frango do mundo de mineração de dados: simples, intuitivo e útil. Delicioso também.

Neste capítulo, você passará de não supervisionado para modelos de inteligência artificial supervisionados, treinando um *modelo naïve Bayes*, que é, por falta de uma metáfora melhor, também um nugget de frango, embora supervisionado.

Como mencionado no Capítulo 2, em inteligência artificial supervisionada, você “treina” um modelo para fazer previsões usando dados que já foram classificados. O uso mais comum de naïve Bayes é para *classificação de documento*. Esse e-mail é spam ou um item desejado? Esse é um tweet feliz ou irritado? Essa chamada telefônica por satélite interceptada pode ser catalogada para investigações adicionais pelos espiões? Você fornece “dado para treinamento”, ou seja, exemplos classificados, desses documentos para o algoritmo de treinamento, e então, mais adiante, o modelo pode classificar novos documentos nessas categorias usando seu próprio conhecimento.

O exemplo que verá neste capítulo é um muito querido para mim. Deixe-me explicar.

Quando você nomeia um Produto de Mandrill, receberá algum sinal ou algum ruído

Recentemente, a empresa para a qual eu trabalho, MailChimp, lançou um produto chamado [Mandrill.com](http://Mandrill.com) — conteúdo em inglês. Ele tem o logotipo mais assustador que eu vi nos últimos tempos (veja a Figura 3-1).

Mandrill é um produto transacional de e-mail para desenvolvedores de software que querem que seus aplicativos enviem e-mails únicos, recibos, redefinições de senha, e qualquer outra coisa que seja individual. Por ele permitir que você rastreie aberturas e cliques de e-mails transacionais individuais, você pode até mesmo conectá-lo à sua conta de e-mail pessoal e rastrear se seus parentes estão realmente visualizando aquelas imagens do seu gato que você vive enviando para eles. (Acredite em um cientista de dados — eles não estão.)



**Figura 3-1:** O logotipo indutor de transe do Mandrill

Mas desde que Mandrill foi lançado, uma coisa tem me incomodado perpetuamente. Enquanto um “MailChimp” é algo que nós inventamos, um mandril, também um primata, está na terra faz um tempo. E eles são bem populares. Darwin chamou o traseiro colorido do mandril de “extraordinário”.

Isso significa que se você for no Twitter e quiser ver alguns tweets mencionando o produto Mandrill, recebe algo parecido com o que vê na Figura 3-2. O tweet inferior é sobre o novo módulo ligando a linguagem de programação Perl ao Mandrill. Esse é relevante. Mas os dois acima são sobre Spark Mandrill do jogo de Super Nintendo Megaman X e uma banda chamada Mandrill.

The screenshot shows three tweets from Twitter:

- Ryan Seguin** (@kerish42) posted 1h ago: "I liked a @YouTube video from @smoothmcgroove youtu.be/hyx9-kWYjDI?a Megaman X - Spark **Mandrill** Acapella" with a link to media.
- KZKO The Vibe** (@KzkoTheVibe) posted 1h ago: "Git It All - **Mandrill** rdo.to/KZKO #nowplaying #listenlive" with an "Expand" link.
- CPAN New Modules** (@cpan\_new) posted 2h ago: "WebService-Mandrill 0.3 by LEV - metacpan.org/release/LEV/We..." with a "View summary" link.

**Figura 3-2:** Três tweets, dos quais apenas um importa

Que nojo.

Mesmo que você gostasse de Megaman X quando era adolescente, muitos desses tweets não são relevantes para a sua pesquisa. De fato, existem mais tweets sobre a banda mais o jogo mais o animal mais outros usuários do Twitter com “mandrill” combinado em seus nomes do que existem sobre Mandrill.com. Isso é muito ruído.

Então é possível criar um modelo que possa distinguir o sinal do ruído? Um modelo de IA pode alertar você somente sobre o produto Mandrill?

Esse é um clássico problema de classificação de documento. Se um documento, tal como um tweet Mandrill, pode pertencer a múltiplas classes (sobre Mandrill.com, sobre outras coisas), para qual classe ele deveria ir?

E a forma mais habitual de atacar esse problema é usar um *modelo bag of words* combinado com o classificador naïve Bayes. Um modelo bag of words trata documentos como uma coleção de palavras desordenadas. “John comeu Little Debbie” é o mesmo que “Debbie comeu Little John”;

ambos são tratados como uma coleção de palavras {"comeu", "Debbie", "John", "Little"}.

Um classificador naïve Bayes recebe um conjunto em treinamento desses sacos de palavras que já estão classificados. Por exemplo, pode-se abastecê-lo com alguns sacos de palavras sobre-o-aplicativo-Mandrill e alguns de palavras sobre-outros-mandrills e treiná-los para distinguir entre os dois. Então, no futuro, poderá abastecê-lo com um saco de palavras desconhecido, e ele classificará para você.

E é isso o que construirá neste capítulo — um classificador de documento naïve Bayes que trata tweets mandrill como sacos de palavras e lhe retorna uma classificação. E isso será muito divertido. Por quê?

Porque naïve Bayes é geralmente chamado de “Bayes de idiota”. Como será possível ver, você fará muitas suposições negligentes e idiotas sobre seus dados, e ainda assim funciona! É como o respingo de tinta de modelos de IA, e por ser tão simples e fácil de implementar (pode ser feito em 50 linhas de código), empresas usam isso o tempo todo para serviços simples de classificação. Pode-se usá-lo para classificar e-mails da empresa, transcrições do serviço ao consumidor, artigos da AP, ocorrências policiais, documentos médicos, críticas de filmes, seja o que for!

Agora, antes de começar a implementar isso no Excel (que é bem simples), você terá que aprender um pouco de teoria da probabilidade. Desculpe. Se você se perde na matemática, passe para a implementação e veja como tudo simplesmente se resolve.

## A↑Introdução↑de↑Teoria↑da↑Probabilidade Mais↑Rápida↑do↑Mundo

Nas próximas duas seções, usarei a notação  $p()$  para falar sobre probabilidade.

Por exemplo:

$p(\text{o próximo filme do Michael Bay será terrível}) = 1$

$$p(\text{John Foreman um dia será vegano}) = 0,0000001$$

Desculpe, é extremamente improvável que eu desista da linguiça defumada Conecuh — a única coisa que eu gosto que vem do Alabama.

## Probabilidades↑Condicionais↑Totalizadoras

Agora, os dois exemplos anteriores são simples probabilidades, mas você trabalhará muito com *probabilidades condicionais* neste capítulo. Esta é uma probabilidade condicional:

$$p(\text{John Foreman será vegano} \mid \text{você pagará \$1B}) = 1$$

Embora as chances de eu virar vegano sejam extremamente baixas, a probabilidade de eu virar vegano caso você me pague um bilhão de dólares é 100 %. Aquela barra vertical | na declaração é usada para separar o evento ao qual ele está condicionado.

Como você concilia a probabilidade vegana total de 0.000001 com a probabilidade condicional virtualmente assegurada? Bom, pode-se usar a *lei da probabilidade total*. Essa lei é a probabilidade de eu me tornar vegano sendo igual à soma das probabilidades de eu virar vegano *condicionadas a* todos os possíveis casos vezes as probabilidades de eles acontecerem:

$$p(\text{vegano}) = p(\$1B) * p(\text{vegano} \mid \$1B) + p(\text{sem \$1B}) * p(\text{vegano} \mid \text{sem \$1B}) = 0,0000001$$

A probabilidade total é a soma ponderada de todas as probabilidades condicionais multiplicada pela probabilidade daquela condição. E a probabilidade da condição de você me pagar um bilhão de dólares é 0 (tenho certeza de que é uma suposição segura). O que significa que  $p(\text{sem \$1B})$  é 1, então você obtém:

$$p(\text{vegano}) = 0 * p(\text{vegano} \mid \$1B) + 1 * p(\text{vegano} \mid \text{sem \$1B}) = 0,0000001$$

$$p(\text{vegano}) = 0 * 1 + 1 * 0,0000001 = 0,0000001$$

# Probabilidade Conjunta, a Regra de Cadeia e Independência

Outro conceito em teoria de probabilidade é o de **probabilidade conjunta**, que é uma maneira elegante de dizer “e”. Lembre-se dos seus dias de vestibular.

Esta é a probabilidade de eu almoçar em um restaurante mexicano hoje:

$$p(\text{John come tacos}) = 0,2$$

Eu como uma vez por semana. E esta é a probabilidade de eu escutar alguma música eletrônica brega hoje:

$$p(\text{John escuta música brega}) = 0,8$$

É altamente provável.

Então quais são as chances de eu **fazer ambos** hoje? Isso é chamado de probabilidade conjunta, e é escrito como a seguir:

$$p(\text{John come tacos, John escuta música brega})$$

Apenas separa-se os dois eventos com uma vírgula.

Agora, nesse caso esses eventos são **independentes**. Isso significa que escutar não afeta comer e vice-versa. Dada essa independência, você pode então multiplicar essas duas probabilidades para obter a probabilidade conjunta:

$$p(\text{John come tacos, John escuta música brega}) = 0,2 * 0,8 = 0,16$$

Isso às vezes é chamado de **regra de multiplicação de probabilidade**. Note que a probabilidade conjunta é menor do que a probabilidade de um deles acontecer, o que faz muito sentido. Ganhar na loteria no dia que você é atingido por um raio é bem menos provável de acontecer do que um dos eventos isolados.

Uma forma de ver isso é utilizando-se da **regra de cadeia de probabilidade**, que é algo assim:

$$p(\text{John come tacos, John escuta música brega}) = p(\text{John come tacos}) * p(\text{John escuta música brega} | \text{John come tacos})$$

A probabilidade conjunta é a probabilidade de um evento acontecer vezes a probabilidade de outro evento acontecer dado o primeiro evento acontecer. Mas como esses dois eventos são independentes, a condição não importa. Escutarei música eletrônica brega a mesma quantidade de vezes independente do almoço, então:

$$p(\text{John escuta música brega} \mid \text{John come tacos}) = p(\text{John escuta música brega})$$

Isso reduz a configuração da regra de cadeia para simplesmente:

$$p(\text{John come tacos, John escuta música brega}) = p(\text{John come tacos}) * p(\text{John escuta música brega}) = 0,16$$

## O que Acontece em uma Situação Dependente?

Apresentarei outra probabilidade, a probabilidade de eu escutar Depeche Mode hoje:

$$p(\text{John escuta Depeche Mode}) = 0,3$$

Há uma chance de 30% de eu escutar um pouco de DM hoje. Não julgue. Agora eu tenho dois eventos que possuem dependências um do outro: escutar Depeche Mode e escutar música eletrônica brega. Por quê? Porque Depeche mode é eletrônico brega. Isso significa que:

$$p(\text{John escuta música brega} \mid \text{John escuta Depeche Mode}) = 1$$

Se eu escutar Depeche Mode hoje, há uma chance de 100 % de eu escutar música eletrônica brega. É uma tautologia. Como Depeche Mode é brega, a probabilidade de eu escutar música eletrônica brega por escutar Depeche Mode deve ser 1.

E isso significa que quando eu quero calcular a probabilidade conjunta deles, eu não apenas obterei o produto das duas probabilidades. Usando a regra de cadeia:

$$p(\text{John escuta música brega, John escuta DM}) = p(\text{John escuta Depeche Mode}) * p(\text{John escuta música brega} \mid \text{John escuta Depeche Mode})$$

$$Depeche\ Mode) = 0,3 * 1 = 0,3$$

## Regra de Bayes

Como eu defini Depeche Mode como eletrônico brega, a probabilidade de eu escutar eletrônico brega *dado* a eu escutar Depeche Mode é 1. Mas e o contrário? Você ainda não tem a probabilidade para esta declaração:

$$p(John\ escuta\ Depeche\ Mode \mid John\ escuta\ música\ brega)$$

Afinal, existem outros grupos de eletrônico por aí. Kraftwerk, alguém? O novo álbum do Daft Punk, talvez?

Bom, um gentil senhor chamado Bayes inventou esta regra:

$$p(música\ brega) * p(DM \mid música\ brega) = p(DM) * p(música\ brega \mid DM)$$

Essa regra permite que você relate a probabilidade de um evento condicional à probabilidade de quando o evento e a condição estão invertidos.

Então, reorganizando os termos, nós podemos isolar a probabilidade que não sabemos (a probabilidade de eu escutar Depeche Mode dado a eu escutar música brega):

$$p(DM \mid música\ brega) = p(DM) * p(música\ brega \mid DM) / p(música\ brega)$$

A fórmula precedente é a maneira como você encontra a **Regra de Bayes** na maioria das vezes. É meramente uma forma de rotacionar as probabilidades condicionais. Quando você sabe uma probabilidade condicional que vai em uma direção, e já sabe as probabilidades totais do evento e a condição, você pode inverter tudo:

Ao inserir valores, obterá:

$$p(DM \mid música\ brega) = 0,3 * 1 / 0,8 = 0,375$$

Eu normalmente tenho uma chance de 30% de escutar Depeche Mode em qualquer dia. No entanto, se eu escutarei algum tipo de música eletrônica brega hoje, as probabilidades de escutar Depeche Mode saltam para 37,5% dado àquele fato. Legal!

# Usando a Regra de Bayes para Criar um Modelo de IA

Tudo bem, está na hora de deixar meu gosto musical para trás e pensar nesse problema de tweet Mandrill. Você tratará cada tweet como um saco de palavras, o que significa que desmembrará cada tweet em palavras (geralmente chamadas de *tokens*) com base em espaços e pontuação. Existem duas classes de tweets — chamadas de **aplicativos** para tweets Madrill.com e **outros** para o restante:

Você se preocupa com essas duas probabilidades:

$$p(\text{aplicativo} | \text{palavra}_1, \text{palavra}_2, \text{palavra}_3, \dots)$$

$$p(\text{outros} | \text{palavra}_1, \text{palavra}_2, \text{palavra}_3, \dots)$$

Essas são as probabilidades de um tweet ser sobre o aplicativo ou sobre outra coisa dado a vermos as palavras “palavra1”, “palavra2”, “palavra3” etc.

A implementação padrão de um modelo naïve Bayes classifica um novo documento baseado em quais dessas duas classes é mais provável consideradas as palavras. Em outras palavras, se:

$$p(\text{aplicativo} | \text{palavra}_1, \text{palavra}_2, \text{palavra}_3, \dots) > p(\text{outros} | \text{palavra}_1, \text{palavra}_2, \text{palavra}_3, \dots)$$

então você terá um tweet sobre o aplicativo Mandrill.

Essa regra de decisão — que seleciona a classe que é mais provável dadas as palavras — é chamada de *regra máxima a posteriori (regra MAP)*.

Mas como você calcula essas duas probabilidades? O primeiro passo é usar a Regra de Bayes neles. Usando a Regra de Bayes, pode-se reescrever a probabilidade condicional do aplicativo da seguinte forma:

$$p(\text{aplicativo} | \text{palavra}_1, \text{palavra}_2, \dots) = (\text{aplicativo}) p(\text{palavra}_1, \text{palavra}_2, \dots | \text{aplicativo}) / p(\text{palavra}_1, \text{palavra}_2, \dots)$$

Da mesma forma, você obtém:

$$p(\text{outros} \mid \text{palavra1, palavra2, ...}) = (outros) p(\text{palavra1, palavra2, ...} \mid \text{outros}) / p(\text{palavra1, palavra2, ...})$$

Mas repare que ambos os cálculos têm o mesmo denominador:

$$p(\text{palavra1, palavra2, ...})$$

Essa é apenas a probabilidade de receber essas palavras em um documento em geral. Como essa quantidade não muda com base na classe, você pode sair da comparação MAP, o que significa que você apenas se importa com quais desses dois valores é maior:

$$p(\text{aplicativo}) p(\text{palavra1, palavra2, ...} \mid \text{aplicativo})$$

$$p(\text{outros}) p(\text{palavra1, palavra2, ...} \mid \text{outros})$$

Mas como você calcula a probabilidade de receber um saco de palavras devido a ser um tweet **aplicativo** ou um tweet outros?

É aqui que as coisas se tornam idiotas!

Suponha que as probabilidades dessas palavras estarem no documento são independentes uma da outra. Então você obtém:

$$p(\text{aplicativo}) p(\text{palavra1, palavra2, ...} \mid \text{aplicativo}) = p(\text{aplicativo})$$

$$p(\text{palavra1} \mid \text{aplicativo}) p(\text{palavra2} \mid \text{aplicativo}) p(\text{palavra3} \mid \text{aplicativo}) \dots$$

$$p(\text{outros}) p(\text{palavra1, palavra2, ...} \mid \text{outros}) = p(\text{outros}) p(\text{palavra1} \mid \text{outros}) p(\text{palavra2} \mid \text{outros}) p(\text{palavra3} \mid \text{outros}) \dots$$

Essa suposição de independência permite que você quebre aquela probabilidade condicional conjunta do saco de palavras dada a classe em probabilidades de palavras individuais dada a classe.

E por que isso é idiota? Porque palavras não são independentes umas das outras em um documento!

Se você estivesse classificando e-mails spam e tivesse duas palavras no documento — “erétil” e “disfunção” — isso presumiria:

$$p(\text{erétil, disfunção} \mid \text{spam}) = p(\text{erétil} \mid \text{spam}) p(\text{disfunção} \mid \text{spam})$$

Mas isso é idiota, não é? Isso é ingênuo, porque se eu lhe falasse que eu recebi um e-mail spam com a palavra “disfunção” nele e eu pedisse para

você adivinhar qual foi a palavra seguinte, você quase certamente diria “erétil”. Há uma dependência aí que está sendo ignorada descaradamente.

O engraçado é que, de alguma forma, essa idiotice não importa para muitas aplicações práticas. Isso é devido a regra MAP não se importar de fato se você calculou suas probabilidades corretamente; ela apenas se importa com qual probabilidade calculada erroneamente é maior. E, presumindo a independência das palavras, você está inserindo todo tipo de erro naquele cálculo, mas, pelo menos, esse desleixo é geral. As comparações usadas na regra MAP tendem a sair na mesma direção que iriam se você tivesse aplicado os mais elegantes conhecimentos linguísticos ao modelo.

## Probabilidades↑de↑Classes↑de↑Alto↑Nível↑São Frequentemente↑Presumidas↑como↑Iguais

Então, para recapitular, no caso do aplicativo Mandrill, você quer classificar tweets baseado em qual destes dois valores é maior:

$$p(\text{aplicativo}) \quad p(\text{palavra1} | \text{aplicativo}) \quad p(\text{palavra2} | \text{aplicativo}) \\ p(\text{palavra3} | \text{aplicativo})$$

$$p(\text{outros}) \quad p(\text{palavra1} | \text{outros}) \quad p(\text{palavra2} | \text{outros}) \quad p(\text{palavra3} | \text{outros})$$

Então o que são  $p(\text{aplicativo})$  e  $p(\text{outros})$ ? Você pode fazer logon no Twitter e ver que  $p(\text{aplicativo})$  realmente é aproximadamente 20%. 80% dos tweets usando a palavra mandrill são sobre outras coisas.

Embora isso seja verdade agora, pode mudar com o tempo, e eu preferiria obter muito mais tweets classificados como tweets aplicativo (falsos positivos) do que filtrar alguns relevantes (falsos negativos), então eu presumirei minhas probabilidades como 50/50. Você verá essa suposição constantemente em classificação naïve Bayes no mundo real, especialmente em filtro spam onde a porcentagem de e-mail que é spam muda com o tempo e pode ser difícil de medir globalmente.

Mas se você presumir que ambos  $p(\text{aplicativo})$  e  $p(\text{outros})$  são 50%, quando comparar os dois valores usando a regra de decisão MAP, você pode muito bem desistir deles. Assim, é possível classificar um tweet como relacionado ao aplicativo se:

$$p(\text{palavra1} \mid \text{aplicativo}) p(\text{palavra2} \mid \text{aplicativo}) \dots \geq p(\text{palavra1} \mid \text{outros}) p(\text{palavra2} \mid \text{outros}) \dots$$

Mas como você calcula a probabilidade de uma palavra dada a classe onde ela está? Por exemplo, contemple a seguinte probabilidade:

$$p(\text{"spark"} \mid \text{aplicativo})$$

Para compreender isso, pode-se puxar um conjunto de tweets em treinamento para o aplicativo, indicá-los como palavras, contar as palavras, e descobrir a porcentagem dessas palavras serem “spark”. Provavelmente será 0%, uma vez que a maioria dos tweets mandrill com “spark” são sobre videogames.

Pare por um momento e contemple esse ponto. Para construir um modelo de classificação naïve Bayes, é preciso apenas rastrear as frequências de palavras relacionadas ao aplicativo e não relacionadas ao aplicativo. Isso não é difícil!

## Mais<sup>↑</sup>Algumas<sup>↑</sup>Bugigangas

Agora, antes de você começar no Excel, deve resolver dois obstáculos reais de implementação naïve Bayes em Excel ou em qualquer linguagem de programação:

- Palavras Raras
- Underflow de ponto-decimal flutuante

### *Lidando com Palavras Raras*

O primeiro é o problema das **palavras raras**. E se você receber um tweet que deve classificar, mas há a palavra “Tubal-cain” nele? Baseado em dados passados no conjunto em treinamento, talvez uma ou ambas as classes nunca viram essa palavra antes. Onde isso acontece muito no

Twitter é com URLs reduzidas, já que cada tweet novo de uma URL deve ter uma codificação diferente, nunca vista antes.

Pode-se presumir que:

$$p(\text{"Tubal-cain"} | \text{aplicativo}) = 0$$

Mas então você receberia:

$$p(\text{"Tubal-cain"} | \text{aplicativo}) p(\text{palavra2} | \text{outros}) p(\text{palavra3} | \text{outros}) \\ \dots = 0$$

Tubal-cain efetivamente “zera” o cálculo inteiro da probabilidade.

Em vez disso, suponha que você já viu “Tubal-cain” uma vez. Você pode fazer isso para todas as palavras raras.

Mas espere — isso é injusto com as palavras que você *realmente já viu uma vez antes*. Tudo bem, então acrescente 1 a elas, também.

Mas isso é injusto com as palavras que você já viu duas vezes. Tudo bem, então acrescente 1 a elas, também.

Isso é chamado de *suavização aditiva (additive smoothing)*, e é frequentemente usado para acomodar palavras não vistas até esta data em modelos de sacos de palavras.

### *Lidando com Underflow de Ponto-decimal Flutuante*

Agora que você já viu as palavras raras, o segundo problema com o qual temos que lidar é o underflow de ponto-decimal flutuante.

Muitas dessas palavras são raras, então você acaba ficando com probabilidades bem pequenas. Nessas dados, a maioria das probabilidades de palavra será menos de 0,001. E por conta da suposição de independência, você multiplicará essas probabilidades de palavras particulares de forma conjunta.

E se você tiver um tweet de 15 palavras com todas as probabilidades abaixo de 0,001? Você ficaria com um valor pequeno na comparação MAP, tal como  $1 \times 10^{-45}$ . Agora, na verdade, o Excel pode tratar um número tão pequeno quanto  $1 \times 10^{-45}$ . Ele se perde em algum lugar nas centenas de 0s após o ponto decimal. Então, para classificar tweets, você

provavelmente estaria bem. Mas para documentos maiores (e.g. e-mails, artigos de notícias), pequenos números podem causar danos nos cálculos.

Apenas para ficar do lado seguro, é preciso encontrar uma forma de não fazer a avaliação MAP diretamente:

$$p(\text{palavra1} \mid \text{aplicativo}) p(\text{palavra2} \mid \text{aplicativo}) \dots \geq p(\text{palavra1} \mid \text{outros}) p(\text{palavra2} \mid \text{outros}) \dots$$

Pode-se resolver esse problema usando a função logarítmica (o logaritmo natural no Excel está disponível utilizando-se a fórmula `LN`).

Eis um fato matemático divertido para você. Digamos que tenha um produto:

$$0,2 * 0,8$$

Se tirar o valor logarítmico dele, o seguinte é verdadeiro:

$$\ln(0,2 * 0,8) = \ln(0,2) + \ln(0,8)$$

E quando você tira o logaritmo natural de qualquer valor entre 0 e 1, em vez de receber um pequeno decimal, você recebe um grande número negativo. Então, pode tirar o logaritmo natural de cada uma das probabilidades e somá-los para conduzir a uma comparação máxima a posteriori. Isso resulta em um valor que o computador não colocará para fora.

Se está um pouco confuso, não se preocupe. Isso se tornará muito claro no Excel.

# Vamos ↑ Começar ↑ a ↑ Festa ↑ do ↑ Excel

## NOTA

A pasta de trabalho do Excel neste capítulo, “Mandrill.xlsx”, está disponível para download no web site do livro em [www.altabooks.com.br](http://www.altabooks.com.br), procurando pelo título do livro. Essa pasta de trabalho inclui todos os dados iniciais se quiser trabalhar a partir daí. Ou você pode apenas ler usando as planilhas que eu já coloquei na pasta de trabalho.

Na pasta de trabalho deste capítulo, chamada de Mandril.xlsx, há duas abas de dados de entrada para começar. Uma aba, AboutMandrill, contém 150 tweets, um por linha, relativos a Mandrill.com. A outra aba, AboutOther, contém 150 tweets sobre outras coisas relacionadas a mandrill.

Eu só quero dizer antes de você começar — bem-vindo ao mundo de *processamento de linguagem natural* (PLN). O processamento de linguagem natural se preocupa com mastigar texto escrito por humanos e cuspir conhecimento. E isso quase sempre significa preparar aquele conteúdo escrito por humanos (como tweets) para consumo do computador. Está na hora de se preparar.

## Removendo↑Pontuação↑Estranha

O primeiro passo em criação de um saco de palavras a partir de um tweet é marcar as palavras sempre que houver um espaço entre elas. Mas antes de dividir as palavras onde existir um espaço vazio, deve-se deixar tudo em minúsculo e substituir a maioria da pontuação com espaço já que pontuação em tweets nem sempre é significante. O motivo de se deixar tudo em minúsculo é porque as palavras “e-mail” e “E-mail” não são significantemente diferentes.

Então, adicione esta fórmula nas duas abas de tweets na célula B2:

=LOWER(A2)

Isso deixará em letra minúscula o primeiro tweet. Em C2, remova quaisquer pontos. Você não quer mutilar as URLs, então remova

qualsquer pontos com um espaço depois deles usando o comando  
**SUBSTITUTE:**

=SUBSTITUTE(B2, " . ", " ")

Essa fórmula substitui a string “. ” por um espaço único “ ”.

Você também pode indicar a célula D2 na célula C2 e substituir quaisquer dois pontos com um espaço depois deles por um espaço único:

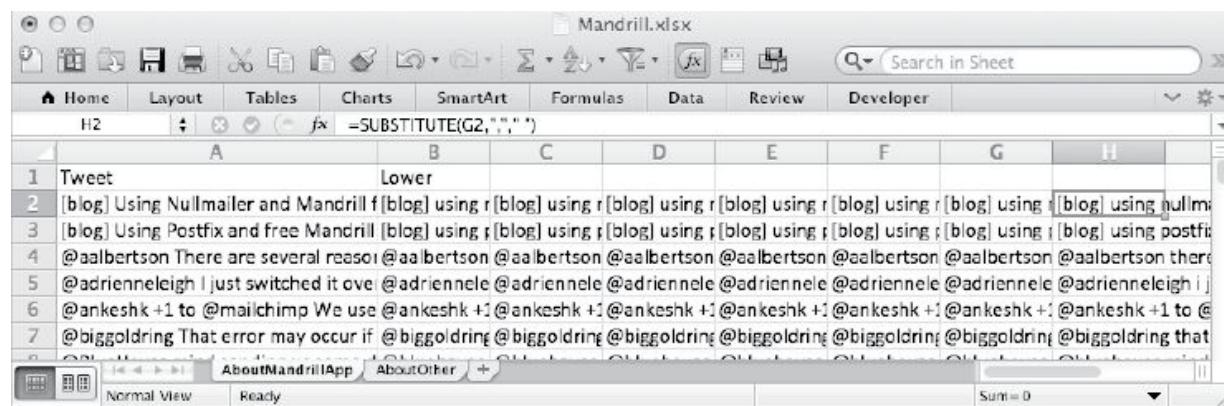
=SUBSTITUTE(C2, " : ", " ")

Nas células E2 até H2, você deve fazer substituições parecidas com as strings “?”, “!”, “;” e “,”:

```
=SUBSTITUTE(D2, "?", " ")  
=SUBSTITUTE(E2, "!", " ")  
=SUBSTITUTE(F2, ";", " ")  
=SUBSTITUTE(G2, ",", " ")
```

Não é preciso adicionar um espaço após a pontuação nas quatro fórmulas anteriores porque elas não aparecem em URLs (especialmente em links reduzidos) com tanta frequência.

Ressalte as células B2:H2 em ambas as abas e dê um clique duplo nas fórmulas para enviá-las para baixo até a linha 151. Isso lhe dá duas abas como as exibidas na Figura 3-3.



**Figura 3-3:** Dados preparados do tweet

# Separando↑Espaços

Em seguida, crie duas novas abas e chame-as de AppTokens e OtherTokens.

Você precisa contar quantas vezes cada palavra é usada em todos os tweets de uma categoria. Isso significa que é preciso ter todas as palavras dos tweets em uma única coluna. É seguro assumir que cada tweet não contém mais de 30 palavras (sinta-se livre para expandir isso para 40 ou 50 se quiser), então você extrairá um token (símbolo) de um tweet por linha, isso significa que é preciso ter  $150 \times 30 = 4.500$  linhas.

Para começar, nessas duas abas nomeie A1 como Tweet.

Realce A2:A4501 e use Paste Special para colar os valores de tweets da coluna H das duas abas iniciais. Isso lhe dará uma lista de tweets processados, como exibido na Figura 3-4. Note que, como você está colando 150 tweets em 4.500 linhas, o Excel repete tudo automaticamente para você.

Isso significa que, se você extrair a primeira palavra do primeiro tweet na linha 2, esse mesmo tweet é repetido para extrair a segunda palavra dele na linha 152, então o terceiro na linha 302, e assim por diante.

	Tweet
4498	we're unifying your mandrill and mailchimp data   mailchimp email n
4499	whaaat i didn't know @mailchimp had an email delivery api service t
4500	would like to send emails for welcome password resets payment no!
4501	zapier makes mandrill integration easy   mandrill email platform blog
4502	
4503	

**Figura 3-4:** A planilha inicial AppTokens

Na coluna B, é preciso indicar a posição de cada espaço sucessivo entre as palavras em um tweet. Pode-se nomear essa coluna algo como Space Position. Por não existir espaço no início de cada tweet, comece

colocando um 0 em A2:A151 para indicar que as palavras começam no primeiro caractere de cada tweet.

Começando em B152 quando os tweets repetem pela primeira vez, pode-se calcular o espaço seguinte como segue:

```
=FIND(" ",A152,B2+1)
```

A fórmula `FIND` pesquisará o tweet do próximo espaço vazio começando com o caractere após o espaço anterior referenciado na célula B2, que é 150 células acima. Veja a Figura 3-5.

	A	B
1	Tweet	Space Position
149	whaaat i didn't know @mailchimp had an email delivery api service t	0
150	would like to send emails for welcome password resets payment not	0
151	zapier makes mandrill integration easy   mandrill email platform blog	0
152	[blog] using nullmailer and mandrill for your ubuntu linux server outb	7
153	[blog] using postfix and free mandrill email service for smtp on ubunt	7
154	@aalbertson there are several reasons emails go to spam mind subm	12
155	@adrienneleigh i just switched it over to mandrill let's see if that imp	15
156	@ankeshk11 to @mailchimp who use mailchimp for marketing emails	0

**Figura 3-5:** A posição do espaço da segunda palavra no tweet na linha 152

No entanto, note que essa fórmula gerará um erro uma vez que acabarem os espaços se houver menos palavras do que as 30 que você planejou, então, para acomodar isso, é preciso envolver a fórmula em uma declaração `IFERROR` e apenas retornar um mais o tamanho do tweet para indicar a posição após a última palavra.

```
=IFERROR(FIND(" ",A152,B2+1),LEN(A152)+1)
```

Você pode então dar um clique duplo nessa fórmula para enviá-la planilha abaixo até A4501. Isso produzirá a planilha exibida na Figura 3-6.

	A	B	C	D	E
1	Tweet	Space Position			
149	whaaat i didn't know @mailchimp had an emai	0			
150	would like to send emails for welcome passwo	0			
151	zapier makes mandrill integration easy   mand	0			
152	[blog] using nullmailer and mandrill for your u	=IFERROR(FIND(" ",A152,B2+1),LEN(A152)+1)			
153	[blog] using postfix and free mandrill email ser	7			

**Figura 3-6:** Posições de cada espaço no tweet

Em seguida, na coluna C, comece a extrair tokens individuais dos tweets. Nomeie a coluna C como Token, e, começando na célula C2, puxe a palavra apropriada do tweet usando a função MID. MID recebe uma string de texto, uma posição inicial e o número de caracteres para tirar. Então, em C2, seu texto está em A2, a posição inicial é uma após o último espaço ( $B2 + 1$ ), e o tamanho é a diferença entre a posição do espaço subsequente na célula B152 e a posição do espaço atual em B2 menos 1 (lembrando que tweets idênticos são deslocados por 150 linhas).

Isso produz a seguinte fórmula:

MID(A2, B2+1, B152-B2-1)

Agora, mais uma vez, você pode entrar em algum aperto no final da string quando acabarem as palavras. Então, se houver um erro, transforme o token em “.”, assim ele será fácil de ser ignorado posteriormente:

=IFERROR(MID(A2, B2+1, B152-B2-1), ".")

Agora é possível dar um clique duplo na fórmula e enviá-la planilha abaixo para referenciar cada tweet, como indicado na Figura 3-7.

Acrescente uma coluna Length na coluna D, e, na célula D2, calcule o tamanho do token em C2, assim:

=LEN(C2)

Você pode clicar duas vezes na fórmula e enviá-la para o restante da planilha. Esse valor permite que você encontre e apague qualquer token com três caracteres ou menos, que tendem a ser insignificantes no geral.

	A	B	C	D	E	F
1	<b>Tweet</b>	<b>Space Position</b>	<b>Token</b>			
149	whaaat i didn't know @mailchimp had an em:	0	whaaat			
150	would like to send emails for welcome passwo	0	would			
151	zapier makes mandrill integration easy   manc	0	zapier			
152	[blog] using nullmailer and mandrill for your u	7	=IFERROR(MID(A152,B152+1,B302-B152-1),".")			
153	[blog] using postfix and free mandrill email ser	7	using			
154	@aalbertson there are several reasons emails	12	there			
155	for this reason you can use mandrill to send emails	12	for			

Figura 3-7: Tokens de todos os tweets

### NOTA

Normalmente, nesse tipo de tarefa de processamento de linguagem natural, em vez de abandonar todas as palavras pequenas, uma lista de palavras vazias (stop words) para o idioma específico (inglês neste caso) seriam removidas. Palavras vazias são palavras que possuem pouco conteúdo léxico, que é como um conteúdo nutricional, para modelos de sacos de palavras.

Por exemplo, “because” ou “instead” podem ser palavras vazias, pois elas são comuns e não fazem muito para distinguir um tipo de documento de outro. As palavras vazias mais comuns em inglês são curtas, como “a”, “and”, “the”, etc., que é o motivo de, neste capítulo, você pegar a rota mais fácil, contudo mais Draconiana, para remover palavras curtas apenas de tweets.

Se seguir esses passos, você terá a planilha AppTokens exibida na Figura 3-8 (a planilha OtherTokens é idêntica exceto pelos tweets copiados para a coluna A).

	A	B	C	D
1	Tweet	Space Position	Token	Length
2	[blog] using nullmailer and mandrill for your u	0	[blog]	6
3	[blog] using postfix and free mandrill email ser	0	[blog]	6
4	@aalbertson there are several reasons emails	0	@aalbertson	11
5	@adrienneleigh i just switched it over to mand	0	@adriennele	14

**Figura 3-8:** AppTokens com seus respectivos tamanhos

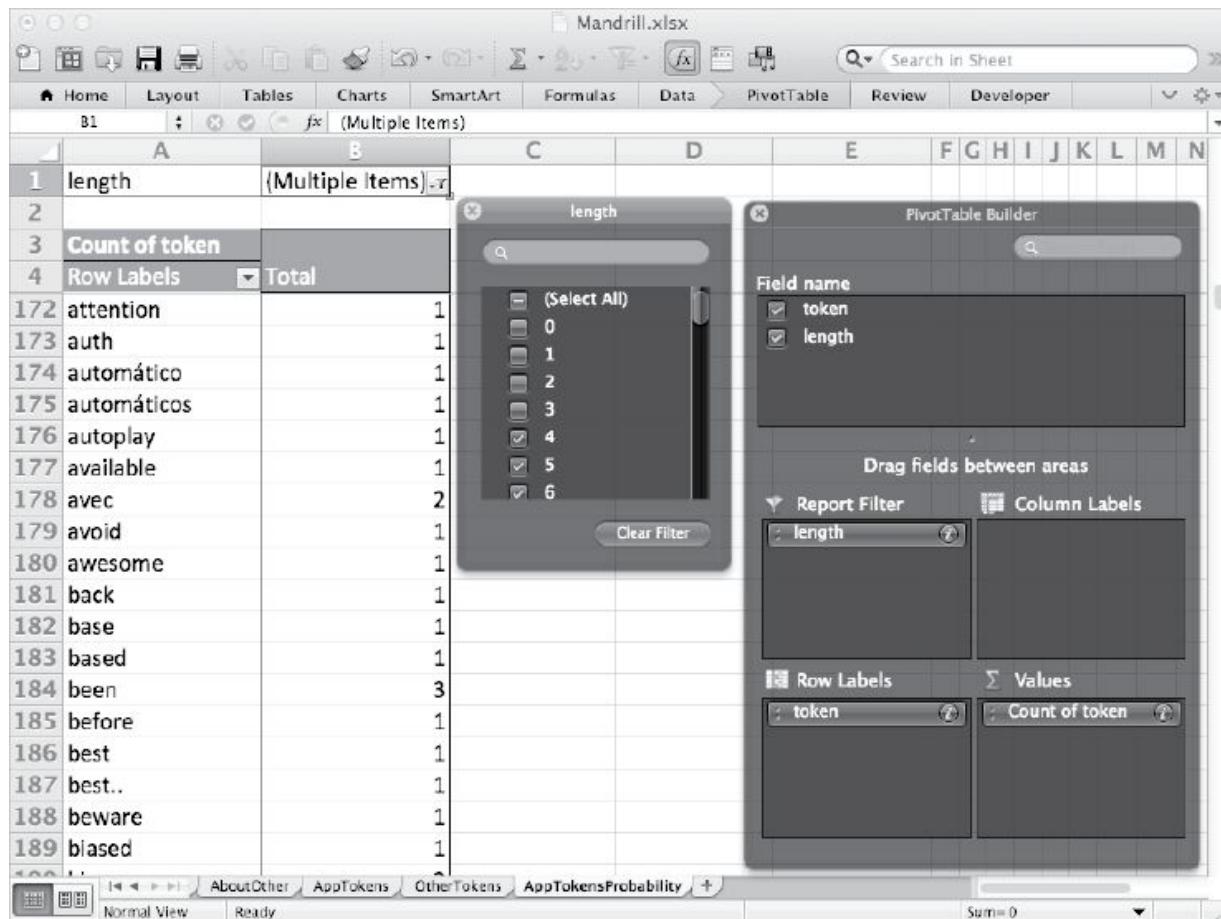
## Contando Tokens e Calculando Probabilidades

Agora que você referenciou seus tweets, está pronto para calcular a probabilidade condicional de um token,  $p(token | classe)$ .

Para fazer isso, é preciso determinar quantas vezes cada token é usado. Comece com a aba AppTokens selecionando o token e o tamanho de seleção C1:D4501 e inserindo os dados em uma PivotTable. Renomeie a aba pivot table criada para AppTokensProbability.

No PivotTable Builder, filtre por tamanho de token, faça com que os tokens nomeiem as linhas e defina o valor como uma contagem de cada token na caixa de valores. Isso fornece a configuração do Builder exibida na Figura 3-9.

No pivot atual, exiba o menu length filter e desmarque tokens de length 0, 1, 2 ou 3. (No Windows você deve instruir o Excel para Select Multiple Items [Selecionar Múltiplos Itens] no menu suspenso.) Isso também está exibido na Figura 3-9.



**Figura 3-9:** Configuração do PivotTable Builder para contagem de tokens

Agora você tem apenas os maiores tokens de cada tweet, todos contados.

Você pode fixar as probabilidades em cada token, mas antes de passar os números, aplique o conceito de suavização aditiva referidos mais cedo nesse capítulo, adicionando um para cada token.

Nomeie a coluna C como Add One To Everything, e defina  $C5 = B5+1$  ( $C4 = B4+1$  no Windows, onde Excel constrói pivot tables uma linha acima só para irritar este livro). Você pode dar um clique duplo na fórmula e desce-la pela página.

Como adicionou um em tudo, também precisará de uma nova contagem total de tokens. Então, na parte inferior da tabela (linha 828 na aba AppTokensProbability), configure a célula para somar as contagens

acima dela. Mais uma vez, note que se você está no Windows tudo é uma linha acima (C4:C825 para a seleção de somatório):

=SUM(C5 : C827)

Na coluna D, pode-se calcular a probabilidade de cada token como sua conta na coluna C dividida pela contagem total de tokens. Nomeie a coluna D como  $P(\text{Token}|\text{App})$ . A probabilidade para o primeiro token em D5 (D4 no Windows) é calculada assim:

=C5/C\$828

Note a referência absoluta à contagem total de tokens. Isso permite que você dê um clique duplo na fórmula e a envie até a coluna D. Então, na coluna E (chame-a de  $\ln(P)$ ), pode tirar o logaritmo natural de probabilidade em D5 como a seguir:

=LN(D5)

Envie isso para baixo da planilha, agora você tem os valores que precisa para a regra MAP. Veja a Figura 3-10.

	A	B	C	D	E
1	length	(Multi-Select Items)			
2					
3	Count of token		Add One To		
4	Row Labels	Total	Everything	P(Token App)	LN(P)
822	you'd	1	2	0.000829876	-7.094234846
823	you'll	1	2	0.000829876	-7.094234846
824	you're	6	7	0.002904564	-5.841471877
825	your	11	12	0.004979253	-5.302475377
826	zapier	1	2	0.000829876	-7.094234846
827	сервис	1	2	0.000829876	-7.094234846
828	Grand Total	1587	2410		

**Figura 3-10:** As probabilidades registradas para os tokens de aplicativos

Além disso, crie uma aba idêntica usando os tokens que não são de aplicativos chamada OtherTokensProbabilities.

## E↑Agora↑Nós↑Temos↑um↑Modelo!↑Vamos↑Usá-lo

Diferente do modelo de regressão (que você encontrá no Capítulo 6), não há passo de otimização aqui. Sem o Solver, sem ajuste do modelo. Um modelo naïve Bayes nada mais é que essas duas tabelas de probabilidades condicionais.

Essa é uma das razões para programadores amarem esse modelo. Não há passo complicado de ajuste de modelo — eles apenas pegam alguns tokens e os somam. E você pode salvar aquele dicionário de tokens no disco para usar mais tarde. É extremamente fácil.

Tudo certo, então agora que o modelo naïve Bayes está treinado, você pode usá-lo. Na aba TestTweets da pasta de trabalho, você encontrará 20 tweets, 10 sobre o aplicativo e 10 sobre outros mandrills. Você preparará esses tweets, referenciá-los (dessa vez, fará as referências de uma maneira um pouco diferente), calculará suas probabilidades logarítmicas para ambas as classes e determinará qual classe é mais provável.

Para começar, copie as células B2:H21 de AboutMandrillApp e cola-as em D2:J21 da aba TestTweets para preparar os tweets. Isso gera a planilha exibida da Figura 3-11.

**Figura 3-11:** Teste de tweets preparados

Em seguida, crie uma aba chamada TestPredictions. Na aba, cole as colunas Number e Class de TestTweets. Nomeie a coluna C como Prediction, que preencherá com os valores de classe previstos. Então nomeie a coluna D como Tokens, e, em D2:D21, cole os valores da coluna J da aba TestTweets. Isso lhe dá a planilha da Figura 3-12.

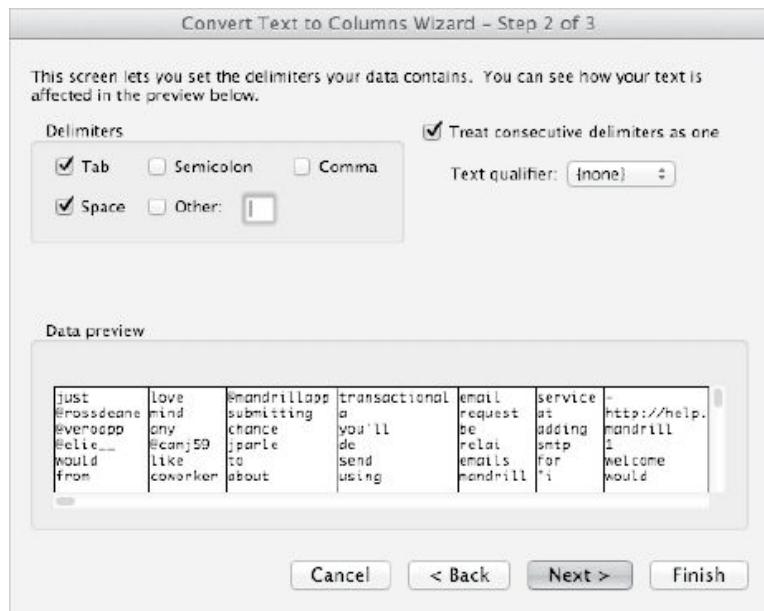
Number	Class	Prediction	Tokens
1	APP		just love @mandrillapp transactional email service - http://mandrill.com sorry @sendgrid
2	APP		@rossdeane mind submitting a request at http://help.man
3	APP		@veroapp any chance you'll be adding mandrill support to
4	APP		@elie__ @camj59 jparle de relai smtp 1 million de mail ch
5	APP		would like to send emails for welcome password resets p
6	APP		from coworker about using mandrill "i would entrust emai
7	APP		@mandrill realised i did that about 5 seconds after hitting
8	APP		holy shit it's here http://www.mandrill.com/
9	APP		our new subscriber profile page activity timeline aggregat
10	APP		@mandrillapp increases scalability ( http://bit.ly/14myvuh
11	APP		the beets rt @missmya #nameanamazingband mandrill
12	OTHER		rt @luiscondoval fernando vergas mandrill mexican pride n
13	OTHER		tt @luiscondoval fernando vergas mandrill mexican pride n

**Figura 3-12:** A aba TestPredictions

Diferente de quando construiu as tabelas de probabilidades, você não quer corresponder esses tokens com os tweets. Você quer avaliar cada tweet separadamente, e isso simplifica a referênciação.

Para começar, ressalte os tweets em D2:D21 e escolha Text to Columns na aba Data da faixa de opções do Excel. No assistente Convert Text to Columns que aparece, selecione Delimited e pressione Next.

Na segunda tela do assistente, especifique Tab e Space como delimitadores. Você também pode escolher Treat Consecutive Delimiters As One e certificar que Text Qualifier está fixado em {none}. Isso gera a configuração exibida na Figura 3-13.



**Figura 3-13: O assistente de configuração Text to Columns**

Pressione Finish. Isso ordena os tweets em colunas até a coluna AI (veja a Figura 3-14).

	A	B	C	D	E	F	G	H	I	J
1	Number	Class	Prediction	Tokens						
2	1	APP		just love @mandrillap transactional email service -						
3	2	APP		@rossdeane mind submitting a request at http://help.n						
4	3	APP		@veroapp any chance you'll be adding mandrill						
5	4	APP		@elie__ @camj59 jparle de relai smtp 1						
6	5	APP		would like to send emails for welcome						
7	6	APP		from coworker about using mandrill "i would						
8	7	APP		@mandrill realised i did that about 5						
9	8	APP		holy shit it's here http://www.mandrill.com/						
10	9	APP		our new subscriber profile page activity timeline						
11	10	APP		@mandrillap increases scalability ( http://bit.ly/ ) then						
12	11	OTHER		the beets rt @missmya #nameanami mandrill						
13	12	OTHER		rt @luissandovfernando vargas mandrill mexican pride						

**Figura 3-14: Os tokens de test tweets**

Abaixo dos tokens, começando na coluna D na linha 25, você deveria procurar pelas probabilidades para cada token. Para tal, pode usar a função VLOOKUP (veja o Capítulo 1 para saber mais sobre VLOOKUP), começando com a célula D25:

```
=VLOOKUP(D2 , AppTokensProbability!$A$5 : $E$827 , 5 , FALSE)
```

A função VLOOKUP pega o token correspondente de D2 e tenta encontrá-lo na coluna A na aba AppTokensProbability. Quando encontra o token, lookup pega o valor da coluna E.

Mas isso não é o bastante, porque é preciso lidar com as palavras raras que não estão na tabela lookup — esses tokens receberão um valor N/A de VLOOKUP. Conforme abordado anteriormente, essas palavras raras deveriam receber uma probabilidade de 1 dividida pela contagem total de token na célula B828 na aba AppTokensProbability.

Para lidar com essas palavras raras, apenas insira VLOOKUP em uma verificação ISNA e reduza a probabilidade logarítmica da palavra rara se necessário:

```
IF (ISNA (VLOOKUP (D2 , AppTokensProbability!$A$5:$E$827 , 5 , FALSE)) ,  
LN (1/AppTokensProbability!$C$828) , VLOOKUP (D2 , AppTokensProbability!  
$A$5:$E$827 , 5 , FALSE) )
```

A única coisa que essa solução ainda não resolveu são os pequenos tokens que você quer jogar fora. Como essas probabilidades logarítmicas serão somadas, você pode definir qualquer probabilidade logarítmica de token como zero (isso é semelhante a definir a probabilidade em 1 em ambos os lados, ou seja, jogar fora).

Para fazer isso, apenas envolva a fórmula toda em mais uma declaração IF que verifica o tamanho:

```
=IF (LEN (D2) <=3 , IF (ISNA (VLOOKUP (D2 , AppTokensProbability!  
$A$5:$E$827 , 5 , FALSE)) , LN (1/AppTokensProbability!$C$828) ,  
VLOOKUP (D2 , AppTokensProbability!$A$5:$E$827 , 5 , FALSE)))
```

Note que referências absolutas são usadas na aba AppTokensProbability para que você arraste a fórmula.

Como os tokens tweet alcançam até a coluna AI, você pode arrastar a fórmula de D25 até AI44 para marcar cada token. Isso lhe dá a planilha exibida na Figura 3-15.

**Mandrill.xlsx**

	A	B	C	D	E	F	G	H	I
25		1		-5.3024754	-6.6887697	-5.1483247	-5.3024754	-4.4915452	-5.3024754
26		2		-7.0942348	-5.3024754	-5.3894868	0	-4.9541687	0 -4.
27		3		-7.0942348	0	-7.0942348	-7.0942348	0	-7.0942348 -3.
28		4		-6.6887697	-5.9956226	-7.0942348	0	-7.0942348	-5.7079405
29		5		-5.9956226	-5.9956226	0	-5.1483247	-5.3894868	0 -7.
30		6		-5.5901574	-7.0942348	-5.4847969	-5.3024754	-3.2335051	0 -5.
Sum				-5.302475377					

**Figura 3-15:** As probabilidades logarítmicas de aplicativos atribuídas aos tokens

Começando na célula D48, você pode usar a mesma fórmula de D25, com a diferença de que ela deveria referenciar a aba OtherTokensProbability, e a variação na aba de probabilidades muda para \$A\$5:\$E\$810 no VLOOKUP com a contagem de tokens total ficando em \$C\$811.

Isso então fornece a planilha mostrada na Figura 3-16.

**Mandrill.xlsx**

	A	B	C	D	E	F	G	H	I
48		1		-5.668402	-6.5156999	-7.6143121	-7.6143121	-7.6143121	-6.921165
49		2		-7.6143121	-6.921165	-7.6143121	0	-7.6143121	0 -7.
50		3		-7.6143121	0	-7.6143121	-7.6143121	0	-7.6143121 -3.
51		4		-7.6143121	-7.6143121	-7.6143121	0	-7.6143121	-7.6143121
52		5		-6.2280178	-5.8225527	0	-6.921165	-7.6143121	0 -7.
53		6		-5.8225527	-7.6143121	-6.0048742	-7.6143121	-3.1034526	0 -6.
Sum				-5.668401997					

**Figura 3-16:** Os dois conjuntos de probabilidades logarítmicas atribuídas aos testes de tweets

Na coluna C, pode-se somar cada linha de probabilidade, rendendo a planilha exibida na Figura 3-17. Por exemplo, C25 é simplesmente:

=SUM (D25 : AI25)

	A	B	C	D	E	F	G	H	I	J
24			Sum of conditional probabilities							
25	1		-65.538	-5.3024754	-6.6887697	-5.1483247	-5.3024754	-4.4915452	-5.3024754	
26	2		-74.483	-7.0942348	-5.3024754	-5.3894868	0	-4.9541687	0	-4.651887
27	3		-44.883	-7.0942348	0	-7.0942348	-7.0942348	0	-7.0942348	-3.233505
28	4		-109.77	-6.6887697	-5.9956226	-7.0942348	0	-7.0942348	-5.7079405	
29	5		-82.763	-5.9956226	-5.9956226	0	-5.1483247	-5.3894868	0	-7.094234
30	6		-58.475	-5.5901574	-7.0942348	-5.4847969	-5.3024754	-3.2335051	0	-5.995622
31	7		-50.884	-6.6887697	-7.0942348	0	0	-5.5901574	-5.4847969	

**Figura 3-17:** Soma de probabilidades logarítmicas condicionais de tokens

Na célula C2, você pode classificar o primeiro tweet simplesmente comparando seus resultados abaixo nas células C25 e C48 usando a seguinte declaração IF:

```
=IF(C25>C48, "APP", "OTHER")
```

Copiando essa fórmula até C21, você recebe todas as classificações, conforme exibe a Figura 3-18.

Recebe 19 de 20! Nada mal. Se você olhar para o único tweet que não foi classificado, a linguagem é bem vaga — os resultados estão quase empatados.

E é isso aí. Modelo construído, previsões feitas.

## Resumindo

Esse capítulo é muito pequeno comparado aos outros neste livro. Por quê? Porque naïve Bayes é fácil! E é por isso que as pessoas o amam. Naïve Bayes parece fazer algum tipo de mágica complexa quando, na verdade, ele apenas depende do computador ter uma boa memória de com que frequência cada token nos dados em treinamento apareceu na classe.

Existe um provérbio que diz: “A experiência é o pai do conhecimento e a memória a mãe.” Em nenhum lugar isso é tão verdadeiro quanto com

naïve Bayes. Toda sua sabedoria deriva de uma combinação de dados passados e armazenamento com um pouco de fita adesiva matemática.

	A	B	C	D	E	
1	Number	Class	Prediction	Tokens		
2		1 APP	APP	just	love	@ma
3		2 APP	APP	@rossdeane	mind	subm
4		3 APP	APP	@veroapp	any	chanc
5		4 APP	APP	@elie_	@camj59	jparle
6		5 APP	APP	would	like	to
7		6 APP	APP	from	coworker	about
8		7 APP	APP	@mandrill	realised	i
9		8 APP	APP	holy	shit	it's
10		9 APP	APP	our	new	subsc
11		10 APP	APP	@mandrillap	increases	scalal
12		11 OTHER	OTHER	the	beets	rt
13		12 OTHER	OTHER	rt	@luissandOv	fern
14		13 OTHER	OTHER	photo	oculi-ds	mand
15		14 OTHER	OTHER	@mandrill	me	neith
16		15 OTHER	OTHER	@mandrill	n	/
17		16 OTHER	OTHER	megaman	x	-
18		17 OTHER	OTHER	@angeluserr	storm	eagle
19		18 OTHER	OTHER	gostei	de	um
20		19 OTHER	APP	what	is	2-yea
21		20 OTHER	OTHER	120 years		of

**Figura 3-18:** Tweets de teste classificados

Naïve Bayes empresta a si mesmo particularmente bem para implementações simples no código. Por exemplo, esta é uma implementação C#:

<http://msdn.microsoft.com/en-us/magazine/jj891056.aspx> — conteúdo em inglês.

Esta é uma pequena versão que alguém postou online em Python:

<http://www.mustapps.com/spamfilter.py> — conteúdo em inglês.

E uma em Ruby:

<http://blog.saush.com/2009/02/11/naive-bayesian-classifiers-and-ruby/> — conteúdo em inglês.

Umas das ótimas coisas sobre esse tipo de modelo é que ele funciona bem mesmo quando há muitos **atributos** (entradas de modelos de IA) com os quais você está fazendo previsões (no caso desses dados, cada palavra era um atributo). Mas dito isso, lembre-se que um simples modelo de saco de palavras tem algumas desvantagens. Principalmente, a parte naïve do modelo pode causar problemas. Eu darei um exemplo.

Suponha que eu construa um classificador naïve Bayes que tenta classificar tweets sobre filmes em “positivos” e “negativos”. Quando alguém diz algo como:

*O novo filme do Michael Bay é uma pilha de lixo misógino, cheio de explosões e péssimas atuações, que não quer dizer nada. E eu, individualmente, amei a viagem!*

O modelo entenderá isso corretamente? Você possui vários tokens negativos seguidos de um token positivo no final.

Como um modelo de saco de palavras joga fora a estrutura do texto e presume-se que os tokens estão desordenados, isso poderia ser um problema. Muitos modelos naïve Bayes realmente aceitam frases em vez de palavras individuais como tokens. Isso ajuda a contextualizar um pouco as palavras (e torna a suposição naïve ainda mais absurda... mas quem se importa!). Você precisa de mais dados em treinamento para fazer esse trabalho porque o espaço de possíveis frases é maior do que o espaço de possíveis palavras.

Para algo como essa crítica de filmes, você pode precisar de um modelo que realmente se importe com a posição de uma palavra na crítica. Qual frase “teve a última palavra”? Incorporar esse tipo de informação imediatamente retira o simples conceito de saco de palavras.

Mas isso é uma minuciosidade. Naïve Bayes é uma ferramenta de Inteligência Artificial versátil e direta. É fácil de criar protótipos e testar. Então você pode testar outra ideia de modelagem com naïve Bayes, e, se funcionar bem o bastante, excelente. Se parece promissora mas é pobre, você pode passar para algo mais robusto, como um modelo ensemble (abordado no Capítulo 7).

## 4

# Modelo↑de↑Otimização:↑porque↑aquele↑Suco↑de Laranja↑“Recém↑Espremido”↑Não↑Irá↑se Misturar↑Sozinho

**A** *Business Week* recentemente publicou um artigo sobre como a Coca-Cola Company utiliza um grande modelo analítico para determinar como misturar sucos de laranjas para criar o perfeito produto não derivado de concentrados.

Eu estava discutindo o artigo com alguns amigos e um deles falou algo como, “Mas você jamais poderia fazer isso com um modelo de *inteligência artificial*!”

Eles estavam certos. Você não pode. Porque a Coca-Cola não usa um modelo de inteligência artificial. Ela usa um modelo de otimização. Qual é a diferença?

Um modelo de inteligência artificial prevê o resultado de um processo analisando suas entradas. Não é isso que a Coca-Cola faz. A Coca-Cola não precisa prever o resultado de quando combinam o suco A com o suco B. Ela precisa decidir qual combinação de suco A, B, C, D e etc., comprar e misturar. A Coca-Cola pega alguns dados e algumas regras de negócios (seu inventário, sua demanda, suas especificações, e assim por diante) e **decide** como misturar um produto. Essas decisões permitem que a Coca-Cola misture sucos com forças e fraquezas complementares (talvez um seja doce demais e outro não seja doce o bastante) para conseguir exatamente o sabor certo pelo mínimo custo e o máximo lucro.

Não há um resultado que precise de predição. O modelo consegue mudar o futuro. O modelo de otimização é o Arminianismo de análise para o Calvinismo de Inteligência Artificial. Livre arbítrio, baby! (Desculpe, essa é a última piada teológico-histórica neste livro.)

Empresas de todas as indústrias usam modelos de otimização diariamente para responder questões como estas:

- Como eu programo meus funcionários de call center para acomodar seus pedidos de férias, saldo de hora extra e eliminar sucessivos turnos noturnos para qualquer um dos funcionários?
- Quais oportunidades de perfuração de petróleo eu exploro para maximizar o retorno ao mesmo tempo que mantendo o risco sob controle?
- Quando faço novos pedidos para a China e como faço com que eles sejam enviados com o mínimo de custo e a antecipação das demandas?

A otimização é a prática de formular matematicamente um problema de negócios e então encontrar a melhor solução para essa representação matemática. E, como visto no Capítulo 1, esse objetivo sempre é a minimização ou a maximização em que a “melhor solução” pode significar o que você quiser — menor custo, maior lucro ou menor probabilidade de o levar à prisão.

A forma mais utilizada e compreendida de otimização matemática, chamada de *programação linear*, foi desenvolvida secretamente pela União Soviética no final da década de 1930 e adquiriu força

com seu uso extensivo na Segunda Guerra Mundial para planejamento de transporte e alocação de recursos a fim de minimizar custos e riscos e maximizar os danos ao inimigo.

Neste capítulo, entrarei em detalhes na parte *linear* de programação linear. A parte de *programação* é remanescente de uma terminologia de guerra e não tem nada a ver com programação de computador. Apenas ignore.

Este capítulo aborda as otimizações linear e inteira, e um pouco de otimização não-linear. Ele foca em como formular problemas de negócios em uma linguagem com a qual o computador possa resolvê-los. O capítulo também aborda em alto nível como o método de otimização de padrão industrial construído na ferramenta Solver do Excel ataca esses problemas e chega nas melhores soluções.

## Porque Cientistas de Dados Deveriam Saber Otimização?

Se você assistir a vários filmes como James Bond ou *Missão Impossível*, notará que eles frequentemente têm uma grande sequência de ação antes dos créditos iniciais. Nada chama mais atenção do que uma explosão.

Os capítulos anteriores sobre mineração de dados e inteligência artificial eram apenas isso — nossas explosões. Mas agora, como em qualquer bom filme de ação, o enredo deve continuar. No Capítulo 2, você usou um pouco de modelagem de otimização para encontrar a melhor localização para grupos centroides. Neste capítulo, você mergulhará profundamente em otimização e ganhará muita experiência em como formular modelos a fim de resolver problemas de negócios.

A inteligência artificial está fazendo sucesso atualmente pelo seu uso em empresas de tecnologia e startups. A otimização, por outro lado, parece mais ser uma prática de negócios das empresas na lista Fortune 500 (Fortune Global 500 é a classificação das 500 maiores corporações em todo o mundo). Fazer a reengenharia da sua cadeia de suprimentos para reduzir custos com combustível da sua frota pode ser tudo, menos sexy. Mas a otimização, seja reduzindo a gordura ou aproveitando ao máximo economias de escala, é *fundamental* para gerir um negócio eficaz.

E quando nós falamos de data science, a verdade é que a otimização é fundamental aí também. Como verá neste livro, a otimização não é apenas uma prática analítica valiosa para entender por iniciativa própria, mas qualquer praticante de data science digno da profissão precisará utilizá-la para implementar outras técnicas de data science. Apenas neste livro, a otimização faz uma participação em outros quatro capítulos:

- Determinando os melhores centros de grupos em agrupamento k-means conforme visto no Capítulo 2
- Maximizando a modularidade para detecção comunitária (Capítulo 5)
- Treinando coeficientes para um modelo de IA (ajustando uma regressão no Capítulo 6)
- Configurando parâmetros de suavização em um modelo de previsão (Capítulo 8)

Problemas de otimização estão inseridos em toda parte de data science, então você precisa dominar a resolução deles antes de continuar.

## Comece com um Compromisso Simples

Esta seção começa discutindo os dois recursos favoritos dos economistas — armas e manteiga. O ano é 1941 e você foi lançado de paraquedas atrás das linhas inimigas onde assumiu a identidade de um Jérémie (ou Ameline) Galiendo, um produtor de leite francês.

Seu trabalho diurno: ordenhar vacas e vender manteiga cremosa para a população local.

Seu trabalho noturno: fabricar e vender metralhadoras para a resistência francesa.

Seu trabalho é complexo e cheio de riscos. Você foi retirado do QG e deixado sozinho para administrar a fazenda e não ser capturado pelos nazistas. Você tem somente o dinheiro suficiente para se sustentar enquanto produz armas e manteiga; você deve ser autossustentável durante a guerra. Não pode perder a fazenda e seu disfarce junto com ela.

Após sentar e pensar sobre sua condição, você encontrou uma maneira de caracterizar sua situação ao redor de três elementos:

- **O objetivo:** Você recebe 195 dólares (ou, reais, embora honestamente meu Excel esteja configurado em dólares, e não vou mudá-lo pelos números aqui) em renda de cada metralhadora que você vende para seu contato, Pierre. Você recebe \$150 de cada tonelada de manteiga que você vende no mercado. Você precisa trazer renda o suficiente para manter a fazenda.
- **As decisões:** Você precisa descobrir qual composição de armas e toneladas de manteiga produzir por mês a fim de maximizar o lucro total.
- **As restrições:** A produção de uma tonelada de manteiga custa \$100, e a de uma metralhadora, \$150. Seu orçamento de \$1.800 por mês é destinado à produção de novos produtos para venda. Você também precisa armazenar essas coisas em seu porão de 21 metros cúbicos. As armas ocupam 0,5 metros cúbicos quando embaladas, e uma tonelada de manteiga ocupa 1,5 metros cúbicos. Você não pode armazenar a manteiga em outro lugar ou ela estragará. Você não pode armazenar as armas em outro lugar ou será capturado pelos nazistas.

## Representando o problema como um polítopo

Esse problema da forma como foi definido é chamado de *programa linear*. Um programa linear é caracterizado como um conjunto de decisões que precisam ser tomadas para otimizar um objetivo levando em consideração algumas restrições, nas quais ambos, restrições e objetivos, são *lineares*. Nesse caso, linear significa que qualquer equação no problema pode apenas somar decisões, subtrair decisões, multiplicar decisões por constantes, ou alguma combinação disso tudo.

Em programação linear, você não pode empurrar suas decisões para as funções não-lineares, que pode incluir:

- Multiplicar decisões em conjunto (armas **vezes** manteiga não pode ser usado em nenhum lugar)
- Enviar uma variável de decisão por um tipo de lógica de verificação, como uma instrução `IF` (“Se você armazenar apenas manteiga no porão, então você consegue apertar um pouco e aumentar a capacidade para 22 metros cúbicos.”)

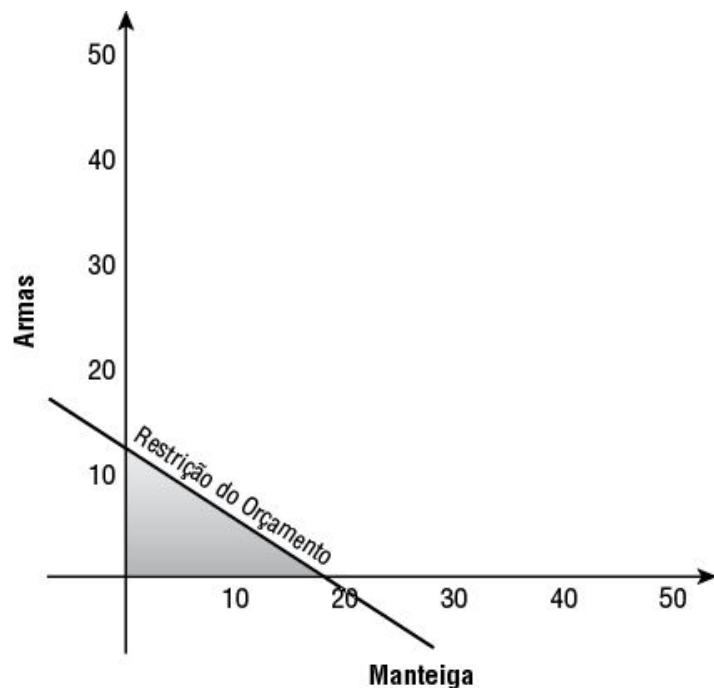
Como verá neste capítulo, as restrições produzem criatividade.

Agora, de volta ao problema. Comece fazendo um gráfico da “região viável” para esse problema. A região viável é um conjunto de possíveis soluções. Você pode não produzir nenhuma arma e nenhuma manteiga? Claro, é viável. Não maximizará a renda, mas é viável. Pode produzir 100 armas e 1.000 toneladas de manteiga? Não, não no orçamento, e não no porão. Não é viável.

Certo, então por onde você começa seu gráfico? Você não pode produzir quantidades negativas de armas e manteiga. Isso não é física teórica. Logo, você está lidando com o primeiro quadrante no plano x-y.

Em termos de orçamento, com \$150 a unidade você pode fazer 12 armas do orçamento de \$1.800. Com \$100 a tonelada, você pode fazer 18 toneladas de manteiga.

Então, se você fizesse um gráfico da restrição de orçamento como uma linha no plano x-y, ele passaria das 12 armas e 18 toneladas de manteiga. Como exibido na Figura 4-1, a região viável é um triângulo de valores positivos em que você pode produzir, no máximo, 12 armas e 18 toneladas de manteiga, ou alguma combinação linear dos dois extremos.

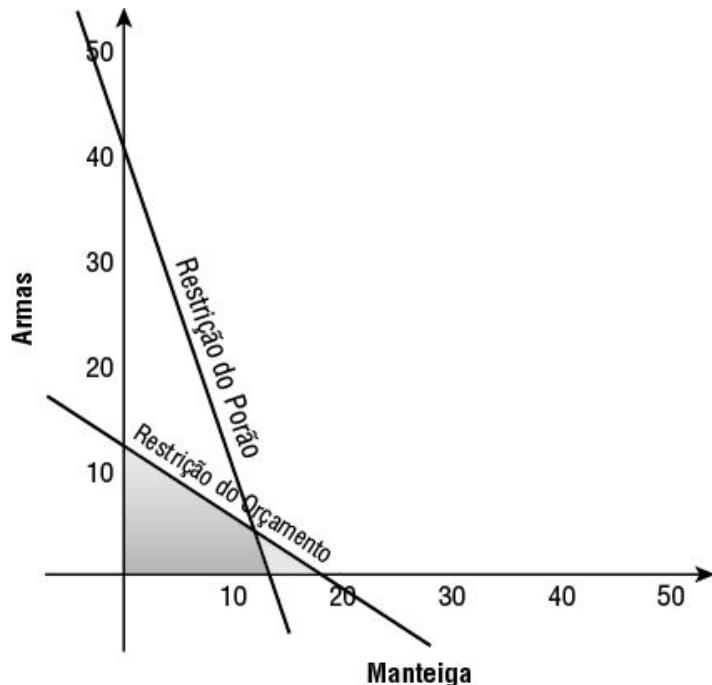


**Figura 4-1:** A restrição de orçamento torna a região viável um triângulo

Agora, esse triângulo é mais conhecido como um polítopo. Um polítopo nada mais é do que uma forma geométrica com lados achatados. Você provavelmente já ouviu o termo **polígono**. Bom, um polígono é apenas um polítopo em um espaço bidimensional. Se você possui uma grande pedra de um anel de noivado em sua mão....Bam! O diamante é um polítopo.

Todos os programas lineares podem expressar suas regiões como polítopos. Alguns algoritmos, como você verá momentaneamente, exploram esse fato para chegar rapidamente em soluções para problemas de programação linear.

Em relação ao problema em questão, está na hora de considerar a segunda restrição — o porão. Se você produzisse apenas armas, seria capaz de embalar 42 delas no porão. Por outro lado, poderia colocar no máximo 14 toneladas de manteiga no porão. Então, adicionando essa restrição ao polítopo, você remove parte da região viável, como mostra a Figura 4-2.



**Figura 4-2:** A restrição do porão corta um pedaço da região viável

## Resolvendo ao Deslizar a Superfície de Nível

Agora que você determinou a região viável, pode começar a fazer a pergunta, “Onde naquela região está a melhor combinação armas/manteiga?”

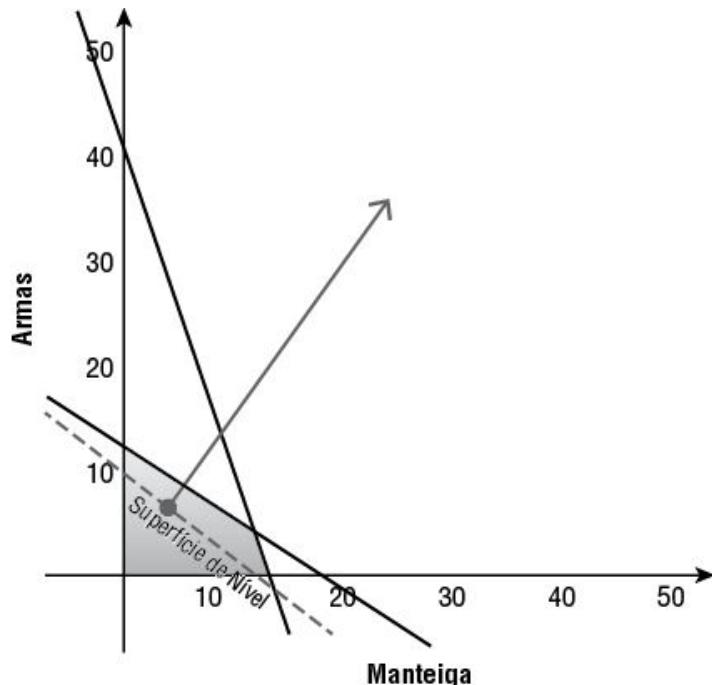
Para responder essa questão, comece definindo algo chamado *superfície de nível*. Uma superfície de nível para o seu modelo de otimização é uma região no polítopo onde todos os pontos produzem a mesma renda.

Como sua função de renda é  $\$150 \cdot \text{Manteiga} + \$195 \cdot \text{Armas}$ , cada superfície de nível pode ser definida pela linha  $\$150 \cdot \text{Manteiga} + \$195 \cdot \text{Armas} = C$ , em que  $C$  é uma quantidade fixa de renda.

Considere o caso em que  $C$  é \$1950. Para a superfície de nível  $\$150 \cdot \text{Manteiga} + \$195 \cdot \text{Armas} = \$1950$ , ambos os pontos  $(0, 10)$  e  $(13, 0)$  existem na superfície de nível como qualquer combinação de armas e manteiga em que  $\$150 \cdot \text{Manteiga} + \$195 \cdot \text{Armas}$  resultam em \$1950. Essa superfície de nível está representada na Figura 4-3.

Usando essa ideia de superfície de nível, você poderia pensar em resolver o problema de maximização de renda deslizando a superfície de nível na direção do aumento de renda (isso é perpendicular à própria superfície de nível) até o *último momento possível antes de você sair da região viável*.

Na Figura 4-3, uma superfície de nível está representada com uma linha tracejada, enquanto a seta e a linha tracejada juntas representam a sua função objetiva.



**Figura 4-3:** A superfície de nível é a função objetiva para a optimização de renda

## O Método Simplex: Procurando no Entorno dos Vértices

Para reiterar, se você quer saber quais pontos viáveis são os melhores, pode apenas deslizar aquela superfície de nível na direção do aumento de renda. Bem na margem antes da superfície de nível sair do polítopo, é onde os melhores pontos estariam. E esta é a parte legal sobre isso:

*Um desses melhores pontos otimizados na margem sempre será um vértice do polítopo.*

Prossiga e confirme isso na Figura 4-3. Coloque um lápis na superfície de nível e mova-a para cima e à direita na direção do aumento de renda. Percebe como ela deixa o polítopo em um vértice?

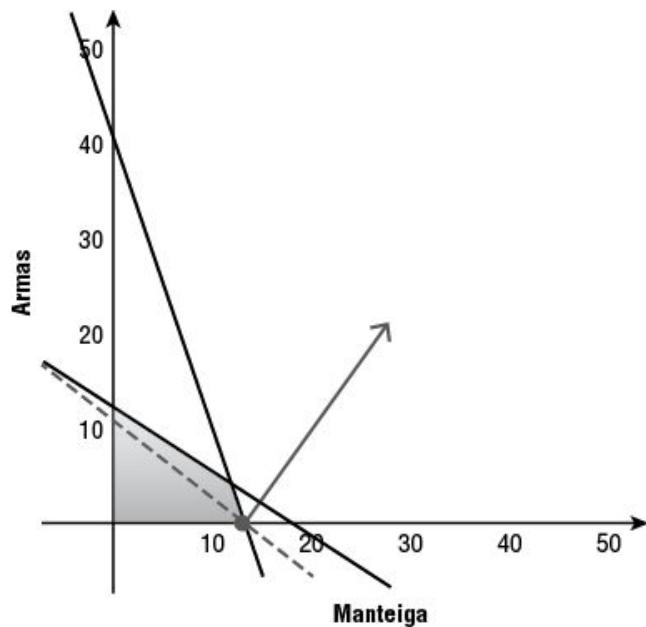
Por que isso é legal? Bem, o polítopo na Figura 4-3 possui um número infinito de soluções viáveis. Pesquisar o espaço inteiro seria horrível. Até as extremidades têm um número infinito de pontos! Mas existem apenas quatro vértices, e há uma excelente solução em uma delas. Chances bem melhores.

Acontece que há um algoritmo que foi criado para verificar vértices. E mesmo em problemas com centenas de milhares de decisões, é muito eficaz. O algoritmo é chamado de **método simplex**.

Basicamente, o método simplex começa em um vértice do polítopo e desliza pelas arestas do polítopo que beneficiam o objetivo. Quando ele chega em uma aresta cujos limites de saída são todos prejudiciais ao objetivo, bem, esse vértice é o melhor.

No caso da venda de armas e manteiga, suponha que você comece no ponto  $(0, 0)$ . É um vértice, mas possui 0 em renda. Certamente você pode melhorar.

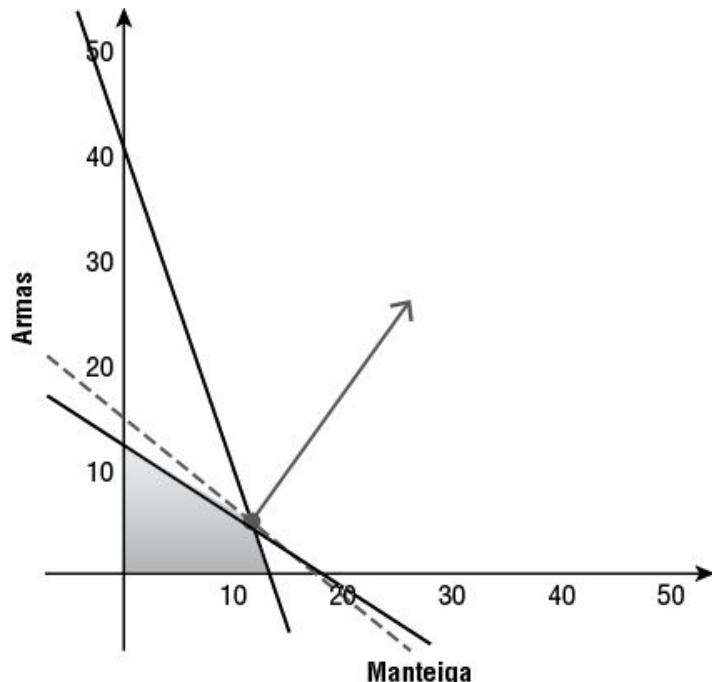
Como visto na Figura 4-3, o canto inferior do polítopo aumenta a renda conforme você vai para a direita. Então, deslizando pelo canto inferior do polítopo nessa direção, é possível chegar na aresta  $(14, 0)$  — 14 toneladas de manteiga e nenhuma arma produzirá 2.100 dólares (veja a Figura 4-4).



**Figura 4-4:** Testando o vértice tutto manteiga

A partir da aresta tudo manteiga, você pode deslizar pela aresta da restrição porão de armazenamento na direção do aumento de renda. O próximo vértice que você atinge será  $(12,9, 3,4)$ , que lhe dá uma tímida renda de 2.600 dólares. Todos os vértices saindo da aresta levam aos piores nós, portanto você terminou.

Como exibido na Figura 4-5, essa é a melhor!



**Figura 4-5:** Localizado o melhor vértice

Trabalhando no Excel

Antes de você deixar esse simples problema para trás por algo um pouco mais difícil, eu quero construí-lo e resolvê-lo no Excel. A primeira coisa que precisa fazer em uma pasta de trabalho nova no Excel é criar espaço para o objetivo e para as variáveis de decisão, então nomeie a célula B2 como o ponto onde a renda total irá e as células B4:C4 como a sequência onde as decisões de produção irão.

Abaixo das seções objetivo e decisão, adicione informações de tamanho e preço para armas e manteiga, os limites no espaço de armazenamento e orçamento, e a contribuição de cada item para a renda.

A planilha incompleta deve se parecer com a Figura 4-6.

The screenshot shows a Microsoft Excel spreadsheet titled "LPIntro.xlsx". The data is organized into several rows and columns:

	A	B	C	D
1	Revenue			
2				
3		Guns	Butter (tons)	
4	Purchase Amount			
5				
6		Guns	Butter (tons)	Limit
7	Storage	0.5	1.5	21
8	Price	\$ 150	\$ 100	\$ 1,800
9	Revenue	\$ 195	\$ 150	

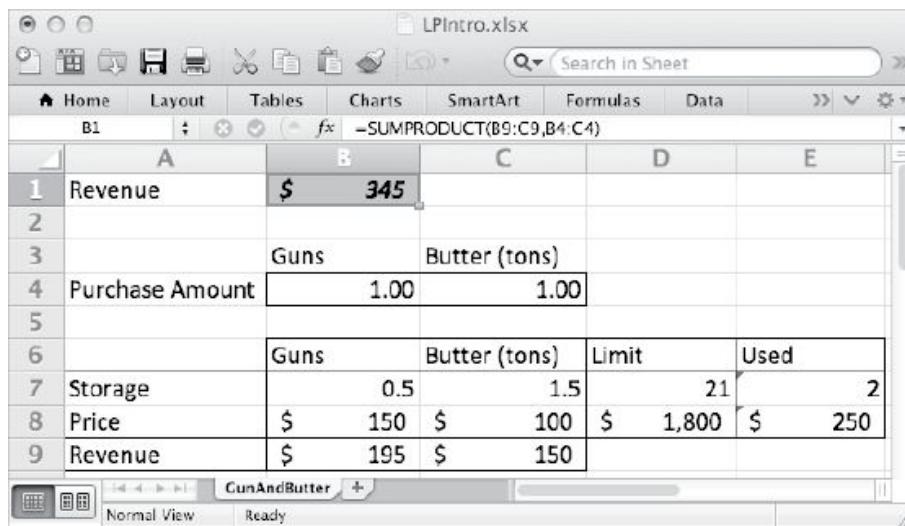
Figura 4-6: Dados de armas e manteiga inseridos, carinhosamente, no Excel

Para esses dados, você precisa adicionar vários cálculos, especificamente, os cálculos de restrições e de renda. Na Coluna E, próximo às células Limit, é possível multiplicar a quantidade de armas e manteiga produzidas vezes seus respectivos tamanhos e preços, e somá-las em uma coluna Used. Por exemplo, em E7, deve-se colocar quanto espaço é utilizado no porão usando a fórmula:

=SUMPRODUCT(B4:C4, B7:C7)

Repare que essa fórmula é linear porque somente o intervalo B4:C4 é um intervalo de decisão. O outro intervalo apenas aloja coeficientes de armazenamento. Você pode fazer o mesmo cálculo para reunir a quantidade total gasta em armas e manteiga.

Para a função objetiva, você precisa apenas de uma SUMPRODUCT das quantidades adquiridas na linha 4 com sua renda na linha 9. Colocar a solução viável, como 1 arma, 1 tonelada de manteiga, nas células de decisão, produz uma planilha como a representada na Figura 4-7.

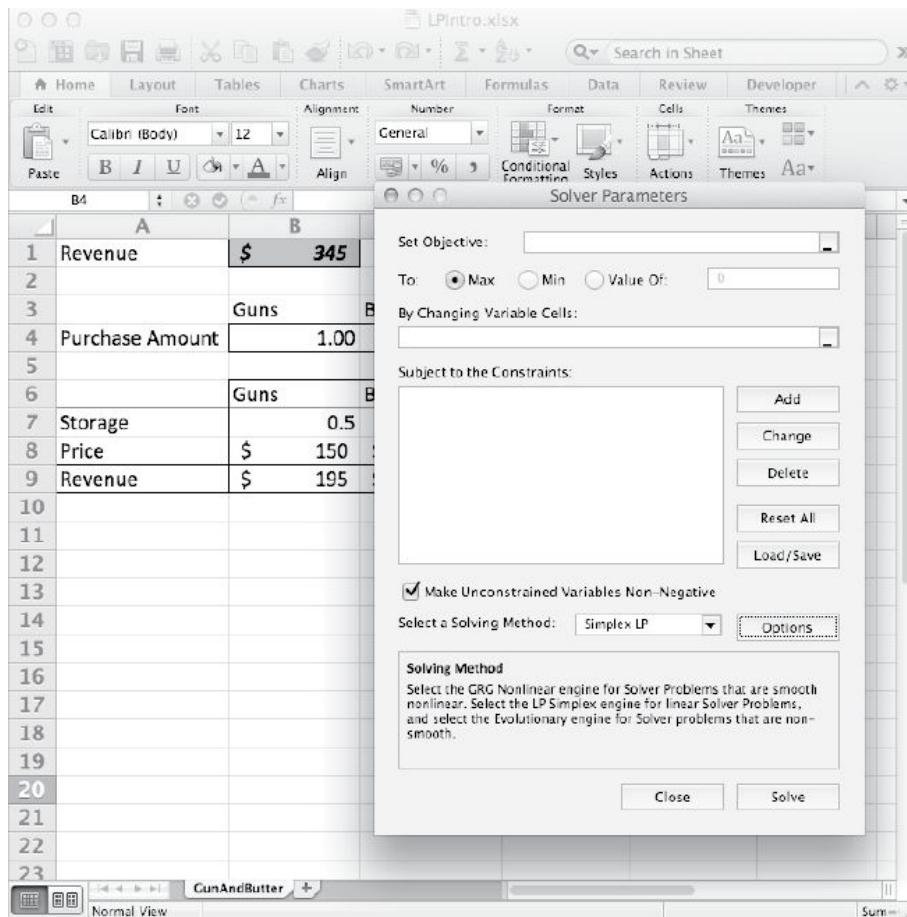


**Figura 4-7:** Cálculos de renda e restrições dentro do problema das armas e manteiga

Tudo bem, então como fazer com que o Excel defina as variáveis para seus valores ideais? Para fazer isso, use o Solver! Comece abrindo uma janela vazia do Solver (exibida na Figura 4-8). Para mais sobre como adicionar o Solver no Excel, veja o Capítulo 1.

Como simulamos anteriormente neste capítulo, você precisa dar um objetivo, decisões e restrições ao Solver. O objetivo é a célula de renda criada em B1. Além disso, certifique-se de ter escolhido o botão Max já que você está maximizando, não minimizando, a renda. Se estiver trabalhando em um problema com custo e risco na função objetiva, você usaria a opção Min.

As decisões estão em B4:C4. Após adicioná-las à seção “By Changing Variable Cells”, a janela do Solver estará parecida com a Figura 4-9.

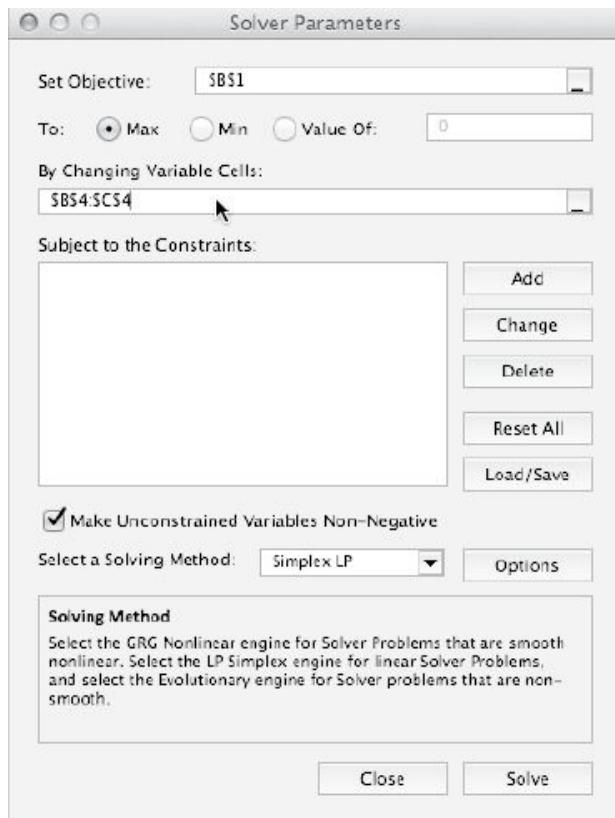


**Figura 4-8:**A janela do Solver

Em relação às restrições, existem duas que é necessário adicionar. Comece com a restrição do porão de armazenamento. Clique no botão Add próximo à seção constraints. Ao preencher a pequena caixa de diálogo, você precisa indicar que a célula E7 deve ser menor ou igual ( $\leq$ ) à célula D7 (veja a Figura 4-10). A quantidade de espaço que você está usando deve ser menor que o limite.

#### NOTA

Repare que o Solver acrescentará referências absolutas (\$) em tudo na sua formulação. Não importa que o Solver faça isso. Honestamente, eu não sei o motivo de ele fazer isso já que você não pode arrastar as fórmulas no contexto



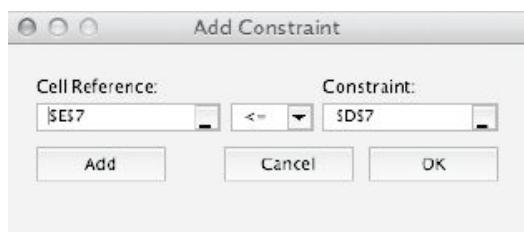
**Figura 4-9:** Objetivo e decisões no Solver

#### NOTA

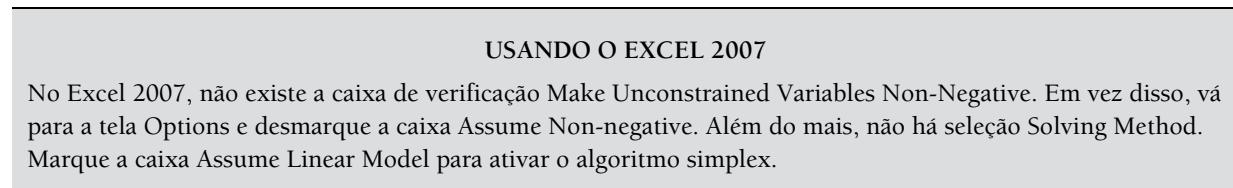
Antes de pressionar OK, observe os outros modelos de restrições que o Solver oferece. Além de  $\leq$ ,  $\geq$  e  $=$ , há algumas muito boas, especificamente, int, bin e dif. Essas restrições diferentes podem ser colocadas em células para torná-las inteiras, binárias (0 ou 1), ou “tudo diferente”. Lembre-se da restrição int. Voltaremos a ela em um segundo.

Pressione OK para adicionar a restrição e, então, adicione a restrição de orçamento da mesma forma ( $E8 \leq D8$ ). Confirme também se a caixa Make Unconstrained Variables Non-Negative está marcada para certificar-se de que a produção de armas e manteiga não se torne negativa por algum motivo. (Uma outra solução é adicionar apenas uma restrição  $B4:C4 \geq 0$ , mas a caixa de verificação facilita isso.)

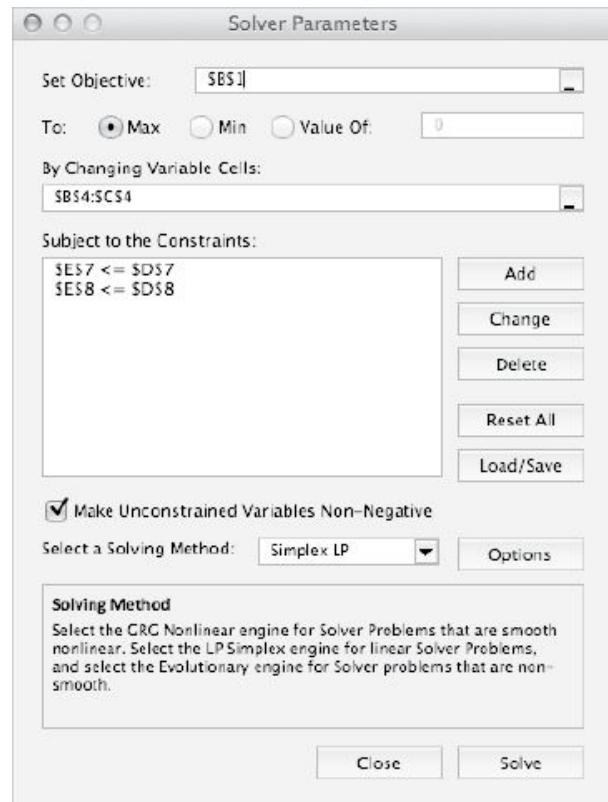
Agora, a partir de Select a Solving Method, assegure-se de que o algoritmo Simplex LP esteja selecionado. Você está pronto para continuar (veja a Figura 4-11).



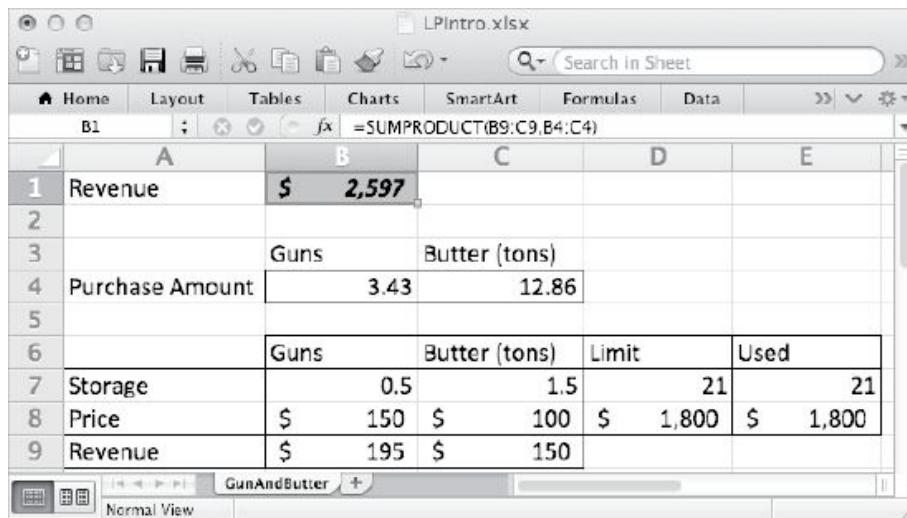
**Figura 4-10:** A caixa de diálogo Add Constraint



Quando pressionar Solve, o Excel rapidamente encontra a solução para o problema e exibe uma caixa de diálogo para que você a veja. Você pode aceitar a solução encontrada ou restaurar os valores nas células de decisão (veja a Figura 4-12). Ao pressionar OK para aceitar a solução, você veria que são 3,43 armas e 12,86 toneladas de manteiga como você fez no gráfico (veja a Figura 4-13).



**Figura 4-12:** O Solver avisa quando o problema foi resolvido



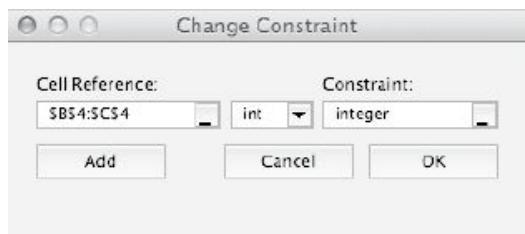
**Figura 4-13:** Pasta de trabalho Armas e Manteiga otimizada

### Mas Você Não Pode Fabricar 3,43 Armas

Agora, nosso alter ego está provavelmente gritando, “*Caramba!*” Por quê? Porque você não pode fabricar 43% de uma arma. E eu reconheço isso.

Ao trabalhar com programas lineares, as soluções fracionárias às vezes podem ser um aborrecimento. Se você estivesse produzindo armas e manteiga em milhões, o decimal poderia ser ignorado sem muito perigo de impraticabilidade e mudanças na renda. Mas, para esse problema, os números são pequenos o bastante para onde você realmente precisa que o Solver os transforme em inteiros.

Então, voltando para a janela do Solver, adicione uma restrição para forçar as células de decisão B4:C4 a serem inteiras (veja a Figura 4-14). Clique em OK para retornar à janela Solver Parameters.



**Figura 4-14:** Tornando as decisões de armas e manteiga inteiros

Abaixo da seção Options próximo a Simplex LP, certifique-se de que a caixa Ignore Integer Constraints não esteja marcada. Pressione OK.

Clique em Solve e uma nova solução aparece. Em 2.580 dólares, você só perdeu aproximadamente 17 dólares. Nada mal! Repare que, ao forçar as decisões a serem inteiras, você nunca pode melhorar, apenas piorar, devido ao fato de você estar estreitando as soluções possíveis.

As armas foram promovidas a 4 enquanto a manteiga caiu para 12. E, embora o orçamento esteja completamente utilizado, observe que você conseguiu 1 metro cúbico extra de armazenamento no porão.

Então por que não tornar suas decisões sempre inteiras? Bem, às vezes, você não precisa delas. Por exemplo, se você está misturando líquidos, as frações funcionam bem.

Além disso, por debaixo dos panos, o algoritmo que o Solver usa na verdade muda quando os inteiros são inseridos, fazendo com que a performance caia. O algoritmo que o Solver utiliza quando encontra a restrição inteira ou binária é chamado “Branch and Bound”, e, em alto padrão, ele deve executar o algoritmo simplex diversas vezes nos pedaços do seu problema original, procurando por soluções viáveis de inteiros a cada passo.

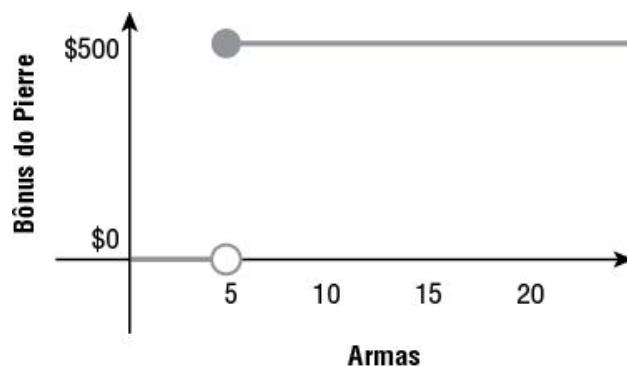
### Vamos Fazer o Problema Não-Linear Por Diversão

Apesar de você ter adicionado uma restrição inteira às decisões, o problema básico em questão ainda é linear.

E se você recebesse um bônus de \$500 do seu contato Pierre se você pudesse lhe trazer 5 ou mais armas por mês? Bom, você pode colocar uma instrução *IF* na função de renda para verificar a produção de arma na célula B4:

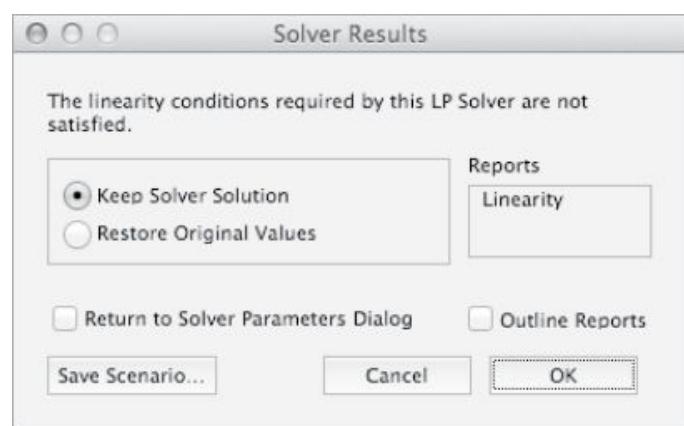
```
=SUMPRODUCT(B9:C9,B4:C4) + IF(B4>=5,500,0)
```

Uma vez que você defina a instrução *IF*, a função objetiva se torna não-linear. Fazendo o gráfico da instrução *IF* na Figura 4-15, você pode facilmente ver a grande descontinuidade não-linear em 5 armas.



**Figura 4-15:** Um gráfico do bônus de R\$500 de Pierre

Se você fosse abrir o Solver e usar Simplex LP novamente para resolver esse problema, o Excel educadamente reclamaria que “the linearity conditions required by this LP Solver are not satisfied” (veja a Figura 4-16).



**Figura 4-16:** O Excel não permite que você coloque variáveis de decisão por uma declaração IF ao usar o Simplex LP

Por sorte, o Solver proporciona outros dois algoritmos para resolver esse problema, chamados de algoritmos “Evolucionário” e “GRG Não Linear”. Você fará uma tentativa com a abordagem evolucionária aqui, com a qual está familiarizado por tê-la utilizado no Capítulo 2. (No Excel 2007, como não há caixa de seleção de algoritmo, deixar a caixa Assume Linear Model **desmarcada** ativará um algoritmo de otimização não-linear.)

A maneira como o algoritmo evolucionário funciona é adequadamente modelada em como a evolução funciona em biologia:

- Gera um conjunto de soluções iniciais (como um “patrimônio genético”), alguns viáveis e alguns inviáveis.
- Cada solução tem algum nível de aptidão para sobrevivência.
- As soluções produzidas por cruzamentos são componentes selecionados e combinados a partir de duas ou mais soluções.
- As soluções **sofrem mutações** para criar soluções novas.
- Novas soluções são geradas em algumas **pesquisas locais** procurando a melhor solução atual para a população.
- A **seleção** ocorre quando candidatas a soluções de baixo desempenho selecionadas são retiradas do patrimônio genético.

Repare que essa abordagem não exige que aquela estrutura de problema seja linear, quadrática, ou outra coisa, especificamente. De certo modo, o problema pode ser tratado sem uma questão definida.

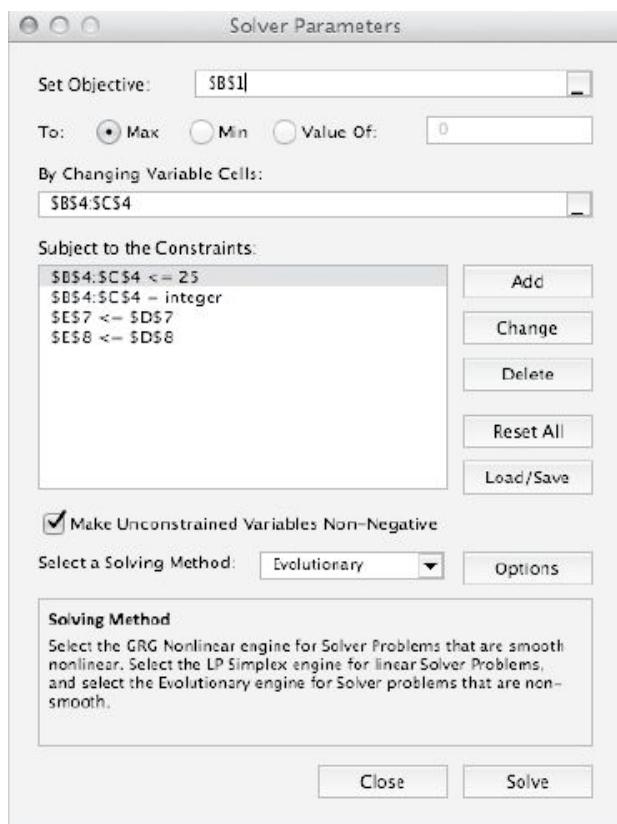
Isso significa que, ao modelar um programa linear no Excel, você está limitado a coisas como os sinais  $+$ / $-$ , as fórmulas SUM e AVERAGE, e a fórmula SUMPRODUCT, na qual apenas um intervalo contém decisões. Mas com o solver evolucionário, as escolhas da sua fórmula se moldam para o que seu coração quer, incluindo estas úteis funções não lineares:

- Verificações lógicas:
  - IF
  - COUNTIF
  - SUMIF
- Funções estatísticas
  - MIN
  - MAX
  - MEDIAN
  - LARGE
  - NORMDIST, BINOMDIST, e assim por diante
- Funções de consulta:
  - VLOOKUP
  - HLOOKUP
  - OFFSET
  - MATCH
  - INDEX

Agora eu sei que você está pegando o ritmo, então deixe-me reduzir essa animação só um pouco. Existem alguns problemas com o solver evolucionário:

- Ele não fornece garantias de que pode encontrar a solução ideal. Tudo o que ele faz é acompanhar a melhor solução em uma população até o tempo acabar, até a população não ter mudado o bastante por um tempo em virtude de continuidade, ou até que você finalize o Solver com a tecla Esc. Você pode modificar esses “critérios de interrupção” na seção evolutionary algorithm options do Solver do Excel.
- O solver evolucionário pode ser bem lento. Com regiões possíveis complexas, ele falha com frequência, incapaz de encontrar até mesmo um bom ponto de partida.
- A fim de fazer o algoritmo evolucionário funcionar bem no Excel, você deveria especificar os limites rígidos para cada variável de decisão. Se você tem uma decisão que é mais ou menos ilimitada, é necessário escolher um número alto e limitá-la.

No que diz respeito a esse último aspecto, para o problema das armas e da manteiga, você deveria adicionar uma restrição em que ambas as decisões devessem ficar abaixo de 25, proporcionando uma nova configuração representada na Figura 4-17.



**Figura 4-17:** Formulação para o solver evolucionário

Pressione OK e então Solve. O algoritmo dá o pontapé inicial e eventualmente deve encontrar uma solução de 6 armas e 9 toneladas de manteiga. Então o algoritmo evolucionário decidiu aceitar o bônus de 500 dólares do Pierre. Ótimo! Mas observe que até mesmo em um problema pequeno, isso

demorou. Aproximadamente 30 segundos no meu laptop. Pense no que isso pode significar para um modelo de produção.

## Há um Monstro no Final Deste Capítulo

Certo, então esse é um problema imaginário. Na próxima seção, demonstrarei os poderes do Solver em algo mais robusto. Você também passará um tempo aprendendo a modelar funções não-lineares (como os 500 dólares de bônus de armas do Pierre) em formas lineares, para que ainda possa usar o rápido algoritmo Simplex LP.

Se você está ansioso pelo próximo tópico, agora você sabe a maioria das coisas necessárias para obter sucesso nos capítulos seguintes. Fique pelo menos até a seção Se-Então e a Restrição “Big M” deste capítulo para aprender o que precisa para o Capítulo 5 sobre agrupamento em gráficos. Ou, melhor ainda, espere e trabalhe em todos os problemas restantes aqui! Mas esteja avisado, as duas últimas regras de negócios modeladas neste capítulo são monstruosas.

### OUTRAS FERRAMENTAS

Grandes modelos não se ajustam bem no Excel. A versão do Solver que vem junto com o Excel permite apenas 100 – 200 variáveis de decisão e restrições, dependendo da versão em que esteja funcionando. Isso limitará o tamanho dos problemas atacáveis neste livro.

Se você quer evoluir no Excel, pode comprar uma versão maior do Solver da Frontline Systems. Melhor ainda, se você estiver em uma caixa de diálogo no Windows, use o OpenSolver como nas últimas seções deste capítulo. O OpenSolver, apresentado no Capítulo 1, chama um solver de código livre chamado COIN Branch and Cut (<http://www.coin-or.org> — em inglês) que é excelente para problemas de otimização médios. Eu usei o OpenSolver em centenas de milhares de variáveis de maneira eficaz.

Outros mecanismos de programação linear pesados incluem Gurobi e CPLEX. Eu geralmente recomendo aos desenvolvedores e pessoas que gostam dos seus softwares “na nuvem” que confiram Gurobi, ou CPLEX, propriedade da IBM, uma solução corporativa confiável.

A interação com essas ferramentas de força industrial ocorre de todas as formas. Por exemplo, o CPLEX vem embutido em um ambiente chamado OPL no qual você pode escrever modelos em uma linguagem especializada com excelentes ligações em planilhas. Há vários ganchos em linguagens de programação para incluir esses algoritmos e modelos dentro dos sistemas de produção.

Minha ferramenta favorita para conectar em solvers pesados como CPLEX e Gurobi é chamada AIMMS ([www.AIMMS.com](http://www.AIMMS.com) em inglês). O software permite que você construa modelos de otimização e junte uma interface de usuário neles sem precisar escrever código. Além disso, o software consegue se comunicar com planilhas e bancos de dados.

Para o restante deste livro, você ficará colado com Excel e Solver, mas apenas saiba que há muitos ambientes de modelagem de alta tecnologia por aí para resolver problemas maiores, caso suas necessidades cresçam além do que o Excel suporta.

Direto da Plantação para o Seu Copo... com uma Parada em um Modelo de Mistura

## NOTA

A pasta de trabalho utilizada neste capítulo, “OrangeJuiceBlending.xlsx”, está disponível para download na página da editora, em [www.altabooks.com.br](http://www.altabooks.com.br), procurando pelo nome do livro. Essa pasta de trabalho inclui todos os dados iniciais se quiser trabalhar a partir deles, ou você pode apenas ler usando as planilhas.

Quando você era uma criança, houve um dia em que alguém lhe explicou que o Papai Noel não existia, ele não passava de um homem com pele rosada fantasiado no shopping.

Bom, hoje eu destruirei outra crença: seu suco de laranja especial não concentrado não foi espremido à mão. Na verdade, a polpa provavelmente é de diferentes laranjas, e o suco foi tirado de diferentes tinas e misturado de acordo com os modelos matemáticos para assegurar que cada jarra possua o mesmo sabor da anterior.

O sabor consistente no suco de laranja não é algo que qualquer um pode produzir durante todo o ano. Na Flórida, as laranjas não estão em época o ano inteiro. Elas estão maduras em épocas e variedades diferentes. Tire a fruta cedo demais e ela terá um sabor “verde”. Compre frutas de outro país em que esteja na época, e o suco pode ter um outro sabor ou ser mais doce. Os consumidores demandam consistência. Isso pode ser fácil para a Del Valle, mas como conseguir isso de um monte de jarras de suco de laranjas recém-espremidas e muito frescas?

## Você Usa um Modelo de Mistura

No famoso programa de TV *Downton Abbey*, o milionário Lord Grantham investe todo o dinheiro da família em um investimento de ferrovia. É arriscado. E ele perde bastante. Aparentemente, no início de 1900, diversificação não era um conceito popular.

Ao fazer a média de risco e retorno de um portfólio de investimento por meio de investimentos múltiplos, as probabilidades de lucrar aumentam, mas as probabilidades de você falir também. Essa mesma abordagem se aplica à produção de suco de laranja hoje.

O suco pode ser consumido em todo o mundo, de diferentes laranjas em diferentes temporadas. Cada produto tem especificações diferentes — alguns podem ser um pouco mais ácidos, outros um pouco mais adstringentes, e outros doces demais. Misturando esse “portfólio” de sucos, um único sabor consistente pode ser mantido.

Esse é o problema no qual você trabalhará nesta seção. Como construir um modelo de mistura que reduz o custo enquanto mantém a qualidade, e qual tipo de chave pode ser lançada nos trabalhos que precisariam ser matematicamente formulados pelo caminho?

## Vamos Começar com Algumas Especificações

Digamos que você seja um analista trabalhando na JuiceLand e seu chefe, o Sr. Juice R. Landingsly III (sua empresa é cheia de nepotismo), pediu para você planejar a compra de suco dos seus fornecedores para janeiro, fevereiro e março do próximo ano. Junto com essa tarefa, o Sr. Landingsly lhe entrega uma planilha de especificações dos seus fornecedores contendo o país de origem e varietal, a quantidade disponível para comprar nos próximos três meses e o preço e custo de entrega por 1.000 galões.

A planilha de especificações classifica a cor do suco em uma escala de um a dez e três componentes de sabor:

- **Índice Brix/Acid:** *Brix* é a medida de doçura no suco, então o índice Brix/Acid é uma medida de doçura a amargura, que, na verdade, é do que realmente se trata o suco de laranja.
- **Acidez (%):** *Acid* é como a porcentagem de suco é quebrada individualmente, porque em um certo ponto, não importa o quanto doce o suco é, ainda é muito ácido.
- **Adstringência (escala 1-10):** *Adstringency* é uma medida da qualidade “verde” do suco. É aquele sabor amargo, não-maduro e de planta que você pode colocar. Essa escala é acessada por um painel de sabores no qual há uma escala de 1-10.

Todas essas especificações são representadas em uma planilha de especificação como mostra a Figura 4-18.

		Qty Available (1,000 Gallons)	Brix / Acid Ratio	Acid (%)	Astringency (1-10 Scale)	Color   1-10 Scale)	Price (per 1K Gallons)	Shipping
1	Varietal	Region						
2	Hamlin	Brazil	672	10.5	0.60%	3	\$ 500.00	\$ 100.00
3	Mosambi	India	400	6.5	1.40%	7	\$ 310.00	\$ 150.00
4	Valencia	Florida	1200	12	0.95%	3	\$ 750.00	
5	Hamlin	California	168	11	1.00%	3	\$ 600.00	\$ 60.00
6	Gardner	Arizona	84	12	0.70%	1	\$ 600.00	\$ 75.00
7	Sunstar	Texas	210	10	0.70%	1	\$ 625.00	\$ 50.00
8	Jincheng	China	588	9	1.35%	7	\$ 440.00	\$ 120.00
9	Berna	Spain	168	15	1.10%	4	\$ 600.00	\$ 110.00
10	Verna	Mexico	300	8	1.30%	8	\$ 300.00	\$ 90.00
11	Biondo Commune	Egypt	210	13	1.30%	3	\$ 460.00	\$ 130.00
12	Belladonna	Italy	180	14	0.50%	3	\$ 505.00	\$ 115.00

Figura 4-18: A planilha de especificações da compra de suco de laranja bruto

Qualquer suco que você escolher comprar será entregue na sua unidade de mistura em tanques grandes, assépticos e refrigerados, seja por navio cargueiro ou ferrovia. É por isso que não há custo de envio para as laranjas Flórida Valéncia — a unidade de mistura é localizada no bosque da Flórida (onde, antigamente, plantavam-se todas as laranjas necessárias).

Dê uma olhada nas especificações exibidas na Figura 4-18. O que se pode dizer sobre elas? O suco está vindo de uma seleção internacional de variedades e localizações.

Alguns sucos, tais como os do México, são baratos mas um pouco sem gosto. No caso do México, a adstringência é muito alta. Em outros casos, como das laranjas Sunstar do Texas, o suco é mais doce e menos adstringente, porém, o custo é maior.

Qual suco comprar para os próximos meses depende de algumas considerações:

- Se você está reduzindo custos, você pode comprar o que quiser?
- Quanto suco você precisa?
- Quais são os sabores e colorações para cada lote?

## Voltando para a Consistência

Por meio dos testes de sabor e várias entrevistas com clientes, JuiceLand determinou qual sabor seu suco de laranja deveria ter e como deveria parecer. Qualquer desvio fora da variação permissível dessas especificações são mais prováveis de rotular o suco como genérico, barato, ou, ainda, **de concentrado**. Eca.

O Sr. Landingsly III expõe as exigências para você:

- Ele quer o menor plano de custo para a compra em janeiro, fevereiro e março a fim de atender uma demanda projetada de 600.000 galões de suco em janeiro e fevereiro e 700.000 galões em março.
- JuiceLand entrou em acordo com o estado da Flórida que fornece à empresa os incentivos de taxas assim que a empresa comprar pelo menos 40% do seu suco dos produtores de Flórida Valência. Esse acordo não pode ser violado sob nenhuma circunstância.
- O índice Brix/Land (BAR) deve ficar entre 11,5 e 12,5 na mistura de cada mês.
- O nível de acidez deve permanecer entre 0,75 e 1%.
- O nível de adstringência deve ficar em 4 ou menos.
- A cor deve permanecer entre 4,5 e 5,5. Não muito aguado, não muito espesso.

Rapidamente, coloque estes itens em um rascunho de uma fórmula LP:

- **Objetivo:** Minimizar custos de compra
- **Decisões:** Reunir a quantidade de suco para comprar por mês
- **Restrições:**
  - Demanda
  - Fornecimento
  - Exigências Flórida Valência
  - Sabor
  - Cor

## Inserindo os Dados no Excel

Para modelar o problema no Excel, a primeira coisa que você precisa é criar uma nova aba para colocar a fórmula. Chame-a de **Optimization Model**.

Na célula A2, abaixo de **Total Cost**, coloque um espaço reservado para o objetivo.

Abaixo disso, na célula A5, cole tudo da aba Specs, mas insira quatro colunas entre as colunas Region e Qty Available para abrir caminho para as variáveis de decisão bem como seus totais por linha.

As primeiras três colunas serão nomeadas como January, February e March, e a quarta será nomeada com a soma delas, Total Ordered. Na coluna Total Ordered, você precisa somar as três células à esquerda, então, por exemplo, no caso das laranjas brasileiras Hamlin, a célula F6 contém:

=SUM(C6 : E6)

Você pode arrastar a célula F6 para baixo até a célula F16. Ao colocar alguma formatação condicional no intervalo C6:E16, a planilha resultante se parecerá com a que está na Figura 4-19.

Varietal	Region	January	February	March	Total Ordered	SPECS			Astringency	Color (1-10 Scale)	Price (per 1K Gallons)	Shipping
						Qty Available	Brix / Acid Ratio	Acid (%)				
Hamlin	Brazil	0.0	0.0	0.0	0.0	672	10.5	0.60%	3	3	\$ 500	\$ 100
Mosambi	India	0.0	0.0	0.0	0.0	400	6.5	1.40%	7	1	\$ 310	\$ 150
Valencia	Florida	0.0	0.0	0.0	0.0	1200	12	0.95%	3	3	\$ 750	\$ -
Hamlin	California	0.0	0.0	0.0	0.0	168	11	1.00%	3	5	\$ 600	\$ 60
Gardner	Arizona	0.0	0.0	0.0	0.0	84	12	0.70%	1	5	\$ 600	\$ 75
Sunstar	Texas	0.0	0.0	0.0	0.0	210	10	0.70%	1	5	\$ 625	\$ 50
Jincheng	China	0.0	0.0	0.0	0.0	588	9	1.35%	7	3	\$ 440	\$ 120
Berna	Spain	0.0	0.0	0.0	0.0	168	15	1.10%	4	8	\$ 600	\$ 110
Verna	Mexico	0.0	0.0	0.0	0.0	300	8	1.30%	8	3	\$ 300	\$ 90
Biondo Comir	Egypt	0.0	0.0	0.0	0.0	210	13	1.30%	3	5	\$ 460	\$ 130
Belladonna	Italy	0.0	0.0	0.0	0.0	180	14	0.50%	3	9	\$ 505	\$ 115

Figura 4-19: Configurando a planilha de mistura

Abaixo dos campos de compra mensal, adicione alguns campos para compra mensal e custos de envio. Para janeiro, coloque o custo de compra mensal na célula C17 como segue:

=SUMPRODUCT (C6 : C16 , \$L6 : \$L16)

Mais uma vez, como apenas a coluna C é uma variável de decisão, esse cálculo é linear. Da mesma forma, é preciso adicionar o seguinte cálculo em C18 para calcular os custos de envio para o mês:

=SUMPRODUCT (C6 : C16 , \$M6 : \$M16)

Arrastando essas fórmulas pelas colunas D e E, toda a sua compra e custos de envio serão calculados. É possível configurar a função objetiva na célula A2 como a soma de C17:E18. A planilha resultante está representada na Figura 4-20.

Varietal	Region	January	February	March	Total Ordered	SPECS			Astringency	Color (1-10 Scale)	Price (per 1K Gallons)	Shipping
						Qty Available	Brix / Acid Ratio	Acid (%)				
Hamlin	Brazil	0.0	0.0	0.0	0.0	672	10.5	0.60%	3	3	\$ 500	\$ 100
Mosambi	India	0.0	0.0	0.0	0.0	400	6.5	1.40%	7	1	\$ 310	\$ 150
Valencia	Florida	0.0	0.0	0.0	0.0	1200	12	0.95%	3	3	\$ 750	\$ -
Hamlin	California	0.0	0.0	0.0	0.0	168	11	1.00%	3	5	\$ 600	\$ 60
Gardner	Arizona	0.0	0.0	0.0	0.0	84	12	0.70%	1	5	\$ 600	\$ 75
Sunstar	Texas	0.0	0.0	0.0	0.0	210	10	0.70%	1	5	\$ 625	\$ 50
Jincheng	China	0.0	0.0	0.0	0.0	588	9	1.35%	7	3	\$ 440	\$ 120
Berna	Spain	0.0	0.0	0.0	0.0	168	15	1.10%	4	8	\$ 600	\$ 110
Verna	Mexico	0.0	0.0	0.0	0.0	300	8	1.30%	8	3	\$ 300	\$ 90
Biondo Comir	Egypt	0.0	0.0	0.0	0.0	210	13	1.30%	3	5	\$ 460	\$ 130
Belladonna	Italy	0.0	0.0	0.0	0.0	180	14	0.50%	3	9	\$ 505	\$ 115
Monthly Cost		\$ -	\$ -	\$ -	\$ -							
Shipping		\$ -	\$ -	\$ -	\$ -							

Figura 4-20: Cálculos de custo adicionados à planilha de mistura

Agora adicione os cálculos que você precisa para satisfazer a demanda e as restrições Flórida Valência. Na linha 20, some a quantidade total de suco adquirida para aquele mês e, na linha 21, coloque os níveis requisitados de 600, 600 e 700, respectivamente, nas colunas C até E.

Para o total que Valência pediu da Flórida, mapeie C8:E8 para as células C23:E23 e coloque os 40% da demanda total (240, 240, 280) abaixo dos valores.

Isso tem como resultado a planilha exibida na Figura 4-21.

Agora que você terminou a função objetiva, as variáveis de decisão e o fornecimento, a demanda e os cálculos Valência, tudo o que resta são os cálculos de sabor e cor baseados no que você encomendou.

Vamos lidar com o índice Brix/Acid primeiro. Na célula B27, coloque o mínimo BAR da mistura, que é 11,5. Então, na célula C27, você pode usar `SUMPRODUCT` das encomendas de Janeiro (coluna C) com suas especificações Brix/Acid na coluna H, divididas pela **demande total**, para obter o índice Brix/Acid.

#### CUIDADO

*Não divida pelo total encomendado, pois essa é uma função das suas variáveis de decisão! Decisões divididas por decisões são altamente não-lineares.*

PURCHASE DECISIONS					
Varietal	Region	January	February	March	Total Ordered
Hamlin	Brazil	0.0	0.0	0.0	0
Mosambi	India	0.0	0.0	0.0	0
Valencia	Florida	0.0	0.0	0.0	0
Hamlin	California	0.0	0.0	0.0	0
Gardner	Arizona	0.0	0.0	0.0	0
Sunstar	Texas	0.0	0.0	0.0	0
Jincheng	China	0.0	0.0	0.0	0
Berna	Spain	0.0	0.0	0.0	0
Verna	Mexico	0.0	0.0	0.0	0
Blondo Comm	Egypt	0.0	0.0	0.0	0
Belladonna	Italy	0.0	0.0	0.0	0
Monthly Cost Price	\$ -	\$ -	\$ -	\$ -	
Shipping	\$ -	\$ -	\$ -	\$ -	
Total Ordered		0.0	0.0	=SUM(E6:E16)	
Total Required		600	600	700	
Valencia Ordered		0.0	0.0	0.0	
Valencia Required		240	240	280	

**Figura 4-21:** Demanda e cálculos Valência adicionados

Apenas lembre-se, você definirá a quantidade total encomendada igual à demanda projetada como uma restrição, então não há motivo para não apenas dividir pela demanda ao receber a média BAR da mistura. Logo, a célula C27 fica desta forma:

```
=SUMPRODUCT(C$6:C$16,$H$6:$H$16)/C$21
```

Você pode arrastar tal fórmula para a direita até a coluna E. Na coluna F, terminará a linha digitando o BAR máximo de 12,5. Pode-se repetir esses passos para definir os cálculos para acidez,

adstringência e cor nas linhas de 28 até 30. A planilha resultante está representada na Figura 4-22.

## Configurando o Problema no Solver

Tudo bem, então você tem todos os dados e cálculos necessários para configurar o problema de mistura no Solver. A primeira coisa a se fazer no Solver é a função de custo total em A2 que você está minimizando.

As variáveis de decisão são as quantidades de compras mensais de cada varietal localizada no intervalo de células C6:E16. Novamente, essas decisões não podem ser negativas, então certifique-se de que a caixa Make Unconstrained Variables Non-Negative esteja marcada (e que Assume Linear Model esteja marcada no Excel 2007).

OrangeJuiceBlending.xlsx					
IFERROR					
13	Berna	Spain	0.0	0.0	0.0
14	Verna	Mexico	0.0	0.0	0.0
15	Biondo	Comrr Egypt	0.0	0.0	0.0
16	Belladonna	Italy	0.0	0.0	0.0
17	Monthly Cost	Price	\$ -	\$ -	\$ -
18		Shipping	\$ -	\$ -	\$ -
19					
20	Total Ordered		0.0	0.0	0.0
21	Total Required		600	600	700
22					
23	Valencia Ordered		0.0	0.0	0.0
24	Valencia Required		240	240	280
25					
26	Quality Const	Minimum			Maximum
27	BAR	11.5	=SUMPRODUCT(C\$6:C\$16,\$H\$6:\$H\$16)/C\$21		
28	ACID	0.0075	0	0	0.01
29	ASTRINGENCY	0	0	0	4
30	COLOR	4.5	0	0	5.5

Figura 4-22: Adicionando restrições de sabor e cor à planilha

Quando se trata de adicionar restrições, esse problema realmente desvia-se do exemplo de armas e manteiga. Há muitas delas.

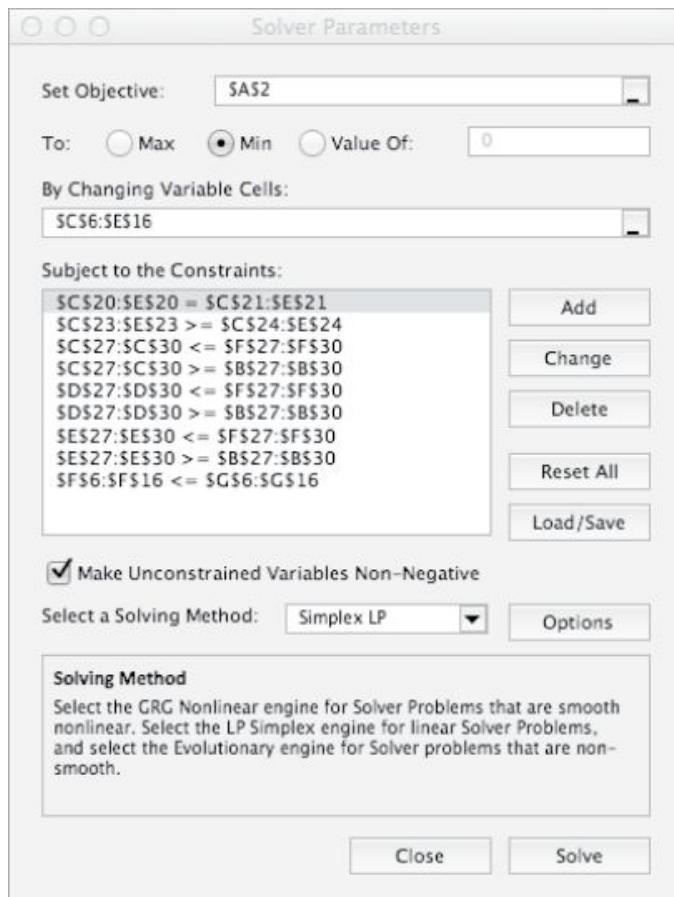
A primeira restrição é que os pedidos na linha 20 devem se igualar à demanda na linha 21 para cada mês. Da mesma forma, os pedidos Flórida Valência na linha 23 devem ser maiores ou iguais à quantidade requisitada na linha 24. Além disso, a quantidade total encomendada de cada local, calculada em F6:F16, deveria ser menor ou igual ao que está disponível em G6:G16.

Com restrições de fornecimento e demanda adicionadas, você precisa acrescentar as restrições de sabor e cor.

Agora, o Excel não permite que você coloque uma restrição em duas séries de tamanhos diferentes, então se você inserir C27:E30 ≥ B27:B30, ele não entenderá como deverá manipular isso. (Eu acho

isso extremamente irritante.) Em vez disso, deve-se adicionar restrições para as colunas C, D e E individualmente. Por exemplo, para pedidos de janeiro você tem  $C27:C30 \geq B27:B30$  e  $C27:C30 \leq F27:F30$ . E o mesmo se aplica para fevereiro e março.

Após acrescentar todas essas restrições, certifique-se de que Simplex LP seja o método de resolução escolhido. A fórmula final deve se parecer com a Figura 4-23.



**Figura 4-23:** O diálogo Solver preenchido para o problema de mistura

Ao resolver, obtém-se um custo ideal de 1,23 milhões de dólares nos custos de compras (veja a Figura 4-24). Repare como as compras Flórida Valência encontram seus limites inferiores. Obviamente, essas laranjas não são o melhor negócio, mas o modelo está sendo forçado a contentar-se com as taxas propostas. A segunda laranja mais popular é a Verna do México, que é barata e de má qualidade, se não péssima. O modelo equilibra esse suco amargo e ácido com misturas de Belladonna, Biondo Commune e Gardner, que são mais amenas, mais doces e superiores na cor. Bem genial!

## Reduzindo Seus Padrões

Animado, você traz o plano de mistura ideal ao seu gerente, o Sr. Landingsly III. Você explica como chegou na sua resposta, e ele olha com suspeita. Mesmo falando que é excelente, ele quer que você corte mais 5% do custo. Ele dá sua opinião supostamente sem sentido usando analogias esportivas sobre “jogar todos os sets” e “dar 110%”.

De nada adianta argumentar contra analogias esportivas. Se 1.170.000 dólares é o ponto ideal, então que seja. Você explica que jamais chegará nisso dentro dos limites de qualidade atuais e ele dá um

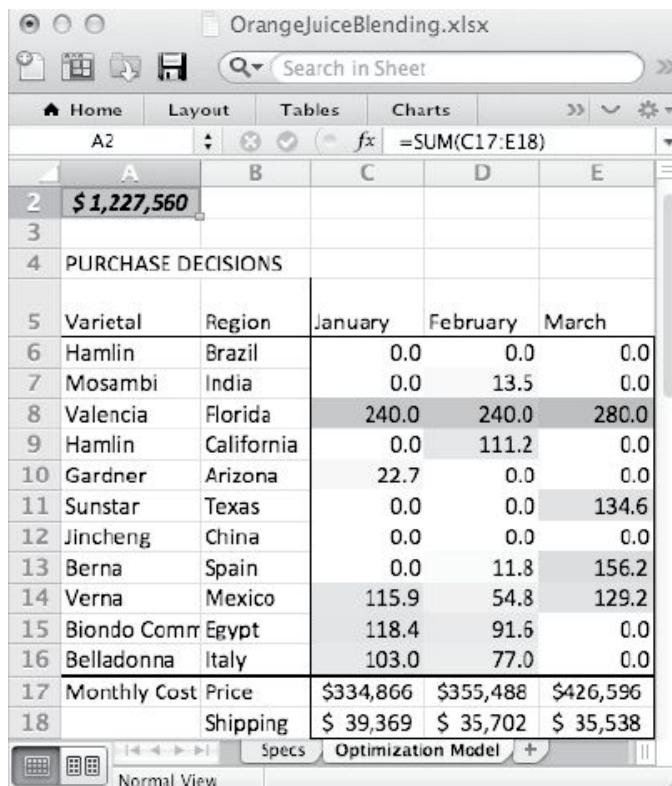
leve grunhido e diz para “quebrar as regras um pouco”.

Hmmm...

Você retorna à sua planilha perturbado.

Como você consegue a melhor mistura para um custo de 1.170.000?

Após a conversa franca com o Sr. Landingsly, o custo não é mais um objetivo. É uma restrição! Então qual é o objetivo?



PURCHASE DECISIONS				
Varietal	Region	January	February	March
Hamlin	Brazil	0.0	0.0	0.0
Mosambi	India	0.0	13.5	0.0
Valencia	Florida	240.0	240.0	280.0
Hamlin	California	0.0	111.2	0.0
Gardner	Arizona	22.7	0.0	0.0
Sunstar	Texas	0.0	0.0	134.6
Jincheng	China	0.0	0.0	0.0
Berna	Spain	0.0	11.8	156.2
Verna	Mexico	115.9	54.8	129.2
Biondo Comr	Egypt	118.4	91.6	0.0
Belladonna	Italy	103.0	77.0	0.0
Monthly Cost	Price	\$334,866	\$355,488	\$426,596
	Shipping	\$ 39,369	\$ 35,702	\$ 35,538

Figura 4-24: A solução para o problema de mistura de suco de laranja

Seu novo objetivo baseado nos grunhidos do seu chefe parece ser encontrar a solução que **menos degrada a qualidade** para 1,17 milhões de dólares. E a forma de implementar isso é jogar uma variável de decisão no modelo a fim de liberar um pouco as restrições de qualidade.

Vá em frente e copie a aba Optimization Model para uma nova planilha, chamada Relaxed Quality. Você não precisa mudar muito para fazer isso.

Tire um momento e pense em como você pode mudar as coisas para acomodar o novo objetivo de qualidade simplificada e restrição de custo. Não prossiga até sua cabeça doer!

Tudo bem.

A primeira coisa a fazer é usar 1.170.000 como o limite de custo na célula B2 perto do antigo objetivo. Além disso, copie e cole **valores** do antigo mínimo e máximo para sabor e cor nas colunas H e I, respectivamente. E na coluna G nas linhas 27 até 30, adicione uma nova variável de decisão chamada % **Relaxed**.

Agora considere como poderia usar a decisão de relaxamento Brix/Acid na célula G27 para relaxar o limite inferior de 11,5. Atualmente, a faixa admissível de Brix/Acid é de 11,5 a 12,5, que é uma largura de 1. Então uma aplicação de 10% na parte inferior da restrição daria o mínimo de 11,4.

Seguindo essa abordagem, substitua o mínimo em B27 com esta fórmula:

=H27-G27\*(I27-H27)

Isso pega o antigo mínimo, agora em H27, e subtrai dele a porcentagem de relaxamento vezes a distância do antigo máximo do antigo mínimo (I27 menos H27). Você pode copiar essa fórmula até a linha 30. Da mesma forma, implemente o máximo de relaxamento na coluna F.

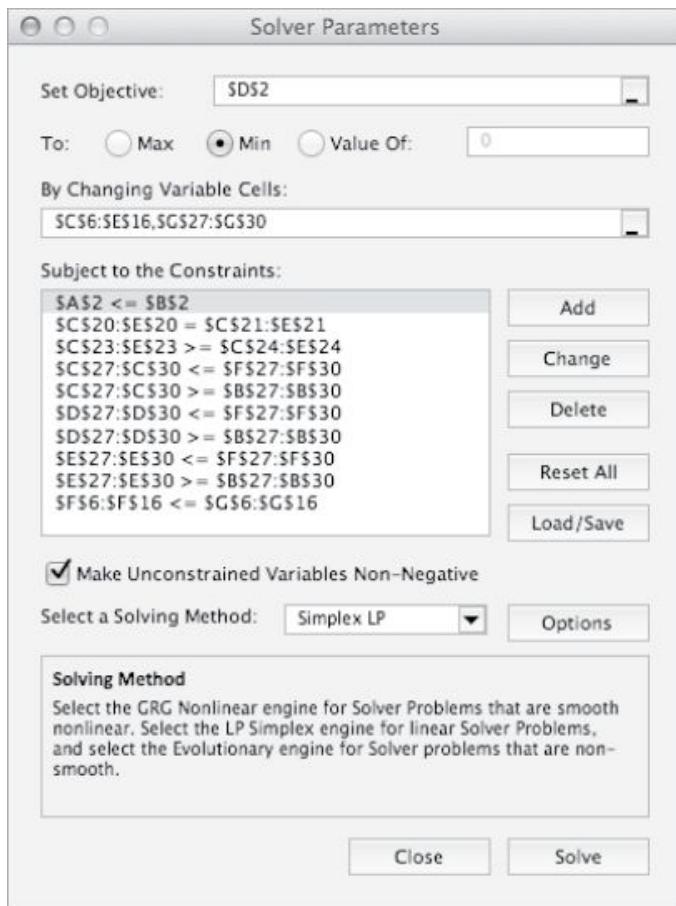
Para o objetivo, use a média das decisões simplificadas em G27:G30. Ao colocar esse cálculo na célula D2, a nova planilha se parecerá com a Figura 4-25.

PURCHASE DECISIONS		SPECS						
Varietal	Region	January	February	March	Total Ordered	Available	Brix / Acid Ratio	Acid (%)
Hemlin	Brazil	0.0	0.0	0.0	0.0	672	10.5	0.50%
Mosambi	India	0.0	0.0	0.0	0.0	400	6.5	1.40%
Valencia	Florida	0.0	0.0	0.0	0.0	1200	12	0.95%
Hemlin	California	0.0	0.0	0.0	0.0	168	11	1.00%
Gardner	Arizona	0.0	0.0	0.0	0.0	84	12	0.70%
Sunstar	Texas	0.0	0.0	0.0	0.0	210	10	0.70%
Jincheng	China	0.0	0.0	0.0	0.0	588	9	1.35%
Berna	Spain	0.0	0.0	0.0	0.0	168	15	1.10%
Verna	Mexico	0.0	0.0	0.0	0.0	300	8	1.30%
Biondo Comune	Egypt	0.0	0.0	0.0	0.0	210	13	1.30%
Belladonna	Italy	0.0	0.0	0.0	0.0	180	14	0.50%
Monthly Cost Totals:		\$ -	\$ -	\$ -				
Shipping		\$ -	\$ -	\$ -				
Total Ordered		0.0	0.0	0.0				
Total Required		600	500	700				
Valencia Ordered		0.0	0.0	0.0				
Valencia Required		240	240	280				
Quality Constraints		Minimum		Maximum	% Relaxed	Minimum	Maximum	
BAR		11.5	0	0	12.5	0	11.5	12.5
ACID		0.0075	0	0	0.01	0	0.0075	0.01
ASTRINGENCY		0	0	0	4	0	0	4
COLOR		4.5	0	0	5.5	0	4.5	5.5

Figura 4-25: O modelo de qualidade simplificada

Abra o Solver e modifique o objetivo para minimizar o relaxamento da média dos limites de qualidade calculados na célula D2. Você também precisa adicionar G27:G30 à lista de variáveis de decisão e fixar o custo em A2 como menor ou igual ao limite em B2. Essa nova fórmula está exibida na Figura 4-26.

Então, recapitulando, você transformou seu objetivo de custo anterior em uma restrição com um limite superior. Você também transformou suas restrições de qualidade rígidas em leves que podem ser simplificadas alterando G27:G30. Seu objetivo em D2 é minimizar a quantidade média que deve degradar a qualidade pelas suas especificações. Pressione Solve.



**Figura 4-26:** A implementação do Solver do modelo de qualidade simplificada

O Excel descobre que com uma média de relaxamento de 35% em cada extremidade dos limites, uma solução pode ser alcançada a fim de atender a restrição de custo, como mostra a Figura 4-27.

Agora que você tem o modelo configurado, algo que pode ser feito é fornecer mais informações do que foi solicitado pelo Sr. Landingsly. Você sabe que para 1,23 milhões é possível obter uma degradação de qualidade de 0%, então por que não reduzir o custo em incrementos de mais ou menos 20 mil e ver no que resulta a degradação de qualidade? Para 1,21 milhões é 5%, para 1,19 milhões é 17%, e assim por diante, incluindo 35%, 54%, 84% e 170%. Se tentar reduzir abaixo de 1,1 milhão o modelo se torna inviável.

Ao criar uma nova aba chamada Frontier, é possível colocar todas essas soluções e fazer um gráfico para ilustrar as compensações entre custo e qualidade (veja a Figura 4-28). Para inserir um gráfico como o da Figura 4-28, simplesmente realce as duas colunas de dados na planilha Frontier e insira um gráfico Smoothed Line Scatter na seleção Scatter do Excel (veja o Capítulo 1 para mais informações sobre inserção de gráficos).

OrangeJuiceBlending.xlsx

Purchase Decisions				
Varietal	Region	January	February	March
Hamlin	Brazil	79.3	77.6	90.6
Mosambi	India	73.5	48.8	56.9
Valencia	Florida	240.0	240.0	280.0
Hamlin	California	0.0	0.0	0.0
Gardner	Arizona	0.0	0.0	0.0
Sunstar	Texas	0.0	0.0	0.0
Jincheng	China	0.0	0.0	0.0
Berna	Spain	23.4	0.0	0.0
Verna	Mexico	78.0	102.5	119.5
Biondo Commune	Egypt	51.9	73.0	85.2
Belladonna	Italy	54.0	58.2	67.9
Monthly Cost Totals:		\$ 330,976	\$ 327,616	\$ 382,218
		Price	\$ 41,494	\$ 40,475
		Shipping	\$ 47,221	

Figura 4-27: A solução para o modelo de qualidade simplificada



Figura 4-28: Criando um gráfico do compromisso entre custo e qualidade

## Remoção do Esquilo Morto: A Fórmula Minimax

Se olharmos para a solução de qualidade simplificada de um limite de custo de 1,17 milhões, há um problema em potencial. Claro, a média de relaxamento pelos limites de sabor e gosto é de 35%, mas para cor é 80% e para o índice Brix/Acid é 51%. A média esconde essa variabilidade.

O que você deveria fazer nessa situação é *minimizar o relaxamento ao máximo* pelos quatro limites de qualidade. Esse problema é comumente chamado de problema “minimax” porque é a minimização de um máximo, e é engraçado falar isso rápido. Minimax, minimax, minimax.

Mas como fazer isso? Se fizer a função objetiva  $\text{MAX}(G27:G30)$ , será não-linear. Você poderia tentar com o solver evolucionário, mas levaria uma eternidade. Acontece que existe uma forma de modelar esse problema não-linear de maneira linear.

Primeiro, copie o modelo simplificado em uma nova aba chamada **Minimax Relaxed Quality**.

Agora, quantos de vocês tiveram que pegar e se livrar de um animal morto? No verão passado, encontrei um esquilo morto no meu sótão extremamente quente em Atlanta, e o cheiro deixou muita gente corajosa de pernas bambas.

Como eu me livrei daquele esquilo?

Eu me recusei a tocá-lo ou lidar com ele diretamente.

Em vez disso, eu o levantei com uma pá e o pressionei com um cabo de vassoura. Foi como pegá-lo com aquelas pinças gigantes de salada ou pauzinhos para comida japonesa. Finalmente, esse movimento de pinça teve o mesmo efeito de pegar o esquilo com minhas próprias mãos, mas menos nojento.

Você pode lidar com o cálculo  $\text{MAX}(G27:G30)$  da mesma forma que eu lidei com o esquilo morto. Como você não está mais computando a média de G27:G30, pode limpar o objetivo em D2. É aí que você computaria a função  $\text{MAX}()$ , mas pode deixar a célula em branco. De alguma forma, ela precisa ser elevada ao máximo sem precisar de contato direto.

É assim que pode ser feito:

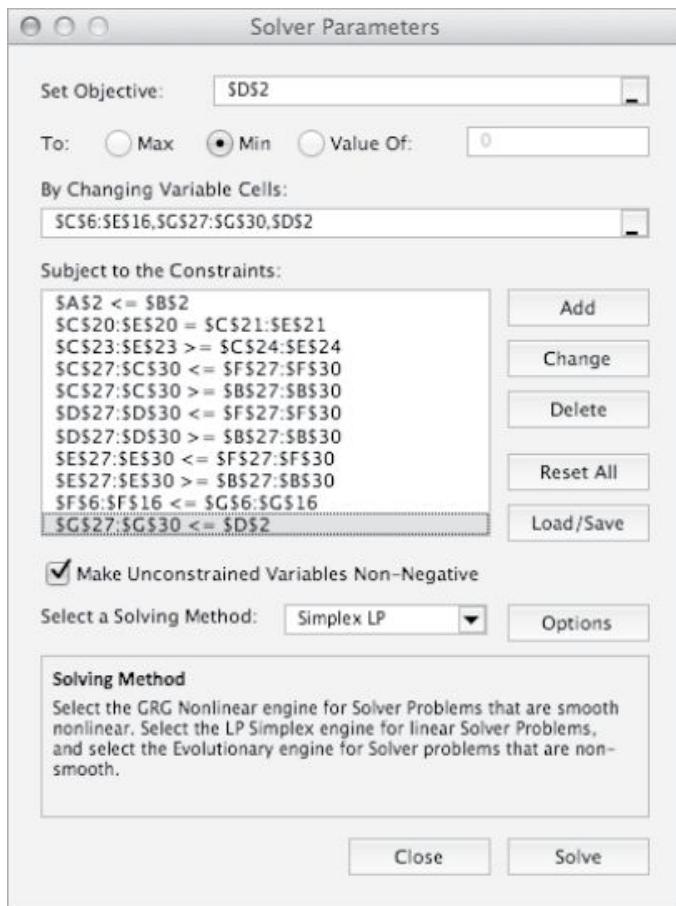
1. Defina o objetivo, D2, como uma variável de decisão, para que o algoritmo possa mover-la quando necessário. Lembre-se de que o modelo foi configurado para ser uma minimização e Simplex tentará enviar essa célula o mais para baixo possível.
2. Configure G27:G30 para ser menor ou igual a D2 usando a janela Add Constraint. D2 deve ir para o lado direito do diálogo Add Constraint para que o Excel permita um número de células desigual (4 células em um intervalo do lado esquerdo e 1 limite superior no lado direito). Diferente de qualquer outro lugar neste capítulo em que não se podia usar duas séries de tamanhos diferentes em uma restrição, isso acontece porque o Excel foi projetado para entender o caso em que o lado direito de uma restrição é uma única célula.

Tudo bem, então o que você acabou de fazer?

Bem, como a função objetiva do modelo, o simplex tentará forçar D2 a reduzir para 0, ao mesmo tempo em que as restrições de sabor e cor o forçarão a manter uma mistura funcional. Onde a célula D2 cairá? O mais baixo que ela pode ir será o máximo das quatro porcentagens de relaxamento em G27 até G30.

Uma vez que o objetivo chega naquele máximo, a única maneira que o Solver consegue progredir é forçando o máximo a diminuir. Assim como o esquilo, as restrições são a pá sob o esquilo e o objetivo da minimização é o cabo pressionando-o para baixo. Assim, obtém-se o termo “minimax”. Muito legal, não é? Ou nojento... depende de como você se sente em relação a esquilos mortos.

Agora que arrumou a fórmula em D2, a implementação no Solver (tornando D2 uma variável e adicionando  $G27:G30 \leq D2$ ) se parece com a Figura 4-29.



**Figura 4-29:** A configuração do Solver para a redução de qualidade minimax

Essa configuração no Solver produz uma redução de qualidade de 58,7%, que, embora aumente a média em 34,8% do modelo anterior, é uma grande melhora do pior caso de relaxamento de cor de 84%.

## Se-Então e a Restrição “Big M”

Agora que você tem um grande sentimento por modelagem linear vanilla, é possível acrescentar alguns inteiros. O Sr. Landingsly III eventualmente se desliga do seu plano de compras original, mas quando você o entrega à equipe de cadeia de suprimentos, os olhos dele começam a tremer incontrolavelmente.

E eles se recusam a comprar suco em qualquer mês de mais de 4 fornecedores. Muita burocracia, aparentemente.

Certo, então como controlar isso dentro do modelo?

Tire um minuto e pense sobre quais modificações de modelo podem ser necessárias antes de prosseguir.

Comece copiando a planilha Optimization Model original em uma nova aba chamada Optimization Model (Limit 4).

Agora, independente de quanto suco você compra de um fornecedor, sejam 1.000 ou 1.000.000 de galões, isso conta como um pedido de um fornecedor. Em outras palavras, é necessário descobrir uma

maneira de fazer essa troca (**switch**) rapidamente na hora de encomendar uma gota de suco de um fornecedor.

Em programação de inteiros, um “switch” é uma variável de decisão binária, que é apenas uma célula que o Solver pode definir em 0 ou 1.

Então o que você quer fazer é definir um intervalo do mesmo tamanho das suas variáveis de decisão, que armazenará apenas 0s e 1s, em que 1 é definido quando o pedido é feito.

Você pode colocar essas variáveis no intervalo C34:E44. Agora, presumindo que elas serão definidas em 1 quando fizer o pedido ao fornecedor, você pode somar cada coluna na linha 45 e certificar-se de que a soma é menor que o limite 4, que você pode jogar na linha 46. A planilha resultante está exibida na Figura 4-30.

No entanto, aqui está a pegadinha. Você não pode usar a fórmula IF que define o indicador em 1 quando a quantidade do pedido acima não é zero. Isso seria não-linear, que o forçaria a usar um algoritmo evolucionário muito mais lento. Para problemas realmente maiores com restrições se-então, os algoritmos não-lineares mais lentos tornam-se inúteis. Portanto, é necessário “ativar” o indicador usando restrições lineares.

Mas digamos que você adicione uma restrição para ativar o indicador Brazilian Hamlin ao fazer um pedido usando a restrição  $C34 \geq C6$ .

Se C34 deve ser binária, então isso limitará C6 para um máximo de 1 (isto é, 1.000 galões encomendados).

Desta maneira, você tem que modelar essa declaração se-então, “se nós encomendarmos, então ative a variável binária”, usando algo coloquialmente chamado de restrição “Big M”. Ele é apenas um número, um número grande, chamado de M. No caso de C34, M deveria ser grande o bastante para você nunca encomendar mais Brazilian Hamlin do que M. Bem, você nunca encomendará mais suco do que o disponível, certo? Para Hamlin, a quantidade disponível é de 672 mil galões. Então produza aquele M.

The screenshot shows a Microsoft Excel spreadsheet titled "OrangeJuiceBlending.xlsx". The data is organized into several sections:

- Row 1:** TOTAL COST (OBJECTIVE):
- Row 2:** \$ -
- Row 32:** INDICATORS
- Row 33:** Varietal      Region      January      February      March
- Rows 34 to 44:** Supplier names and their corresponding Region and monthly costs (all values are 0.0).
- Row 45:** Total Suppliers Used      0.0      0.0      0.0
- Row 46:** Limit 4      4      4      4

The ribbon at the top shows tabs for Home, Layout, Tables, Charts, SmartArt, Formulas, Data, Review, and other options. The formula bar at the top right shows the formula `=SUM(C34:C44)`. The status bar at the bottom indicates "Frontier Minimax Relaxed Quality Optimization Model (Limit 4)".

Figura 4-30: Adicionando os indicadores de variáveis na planilha

Em seguida, você pode definir uma restrição em que  $672 \cdot C34 \geq C6$ . Quando  $C6$  é 0,  $C34$  é *autorizado* a ser zero. E quando  $C6$  for maior do que zero,  $C34$  é *forçado* a mudar para 1 para aumentar o limite superior de 0 para 672.

Para implementar isso na planilha, é necessário definir um novo intervalo de células F34:H44 nas quais você multiplicará os indicadores à esquerda vezes suas respectivas quantidades disponíveis no intervalo G6:G16. Os resultados estão na Figura 4-31.

No Solver, é necessário adicionar C34:E44 à série de variáveis de decisão. Também é necessário torná-las binárias, o que é possível ao adicionar uma restrição bin no intervalo.

Para colocar a restrição “Big M” em vigor, você define C6:E16 ≤ F34:H44. E então é possível verificar as contagens dos fornecedores e certificar-se de que elas estão abaixo de quatro ao definir C45:E45 ≤ C46:E46. A planilha resultante está representada na Figura 4-32.

OrangeJuiceBlending.xlsx

	A	B	C	D	E	F	G	H
1	TOTAL COST (OBJECTIVE):							
2	\$	-						
32	INDICATORS				BIG M			
33	Varietal	Region	January	February	March	January	February	March
34	Hamlin	Brazil	0.0	0.0	0.0	=C34*\$G6	0.0	0.0
35	Mosambi	India	0.0	0.0	0.0	0.0	0.0	0.0
36	Valencia	Florida	0.0	0.0	0.0	0.0	0.0	0.0
37	Hamlin	California	0.0	0.0	0.0	0.0	0.0	0.0
38	Gardner	Arizona	0.0	0.0	0.0	0.0	0.0	0.0
39	Sunstar	Texas	0.0	0.0	0.0	0.0	0.0	0.0
40	Jincheng	China	0.0	0.0	0.0	0.0	0.0	0.0
41	Berna	Spain	0.0	0.0	0.0	0.0	0.0	0.0
42	Verna	Mexico	0.0	0.0	0.0	0.0	0.0	0.0
43	Blondo Commune	Egypt	0.0	0.0	0.0	0.0	0.0	0.0
44	Belladonna	Italy	0.0	0.0	0.0	0.0	0.0	0.0

Figura 4-31: Configurando os valores da nossa restrição “Big M”

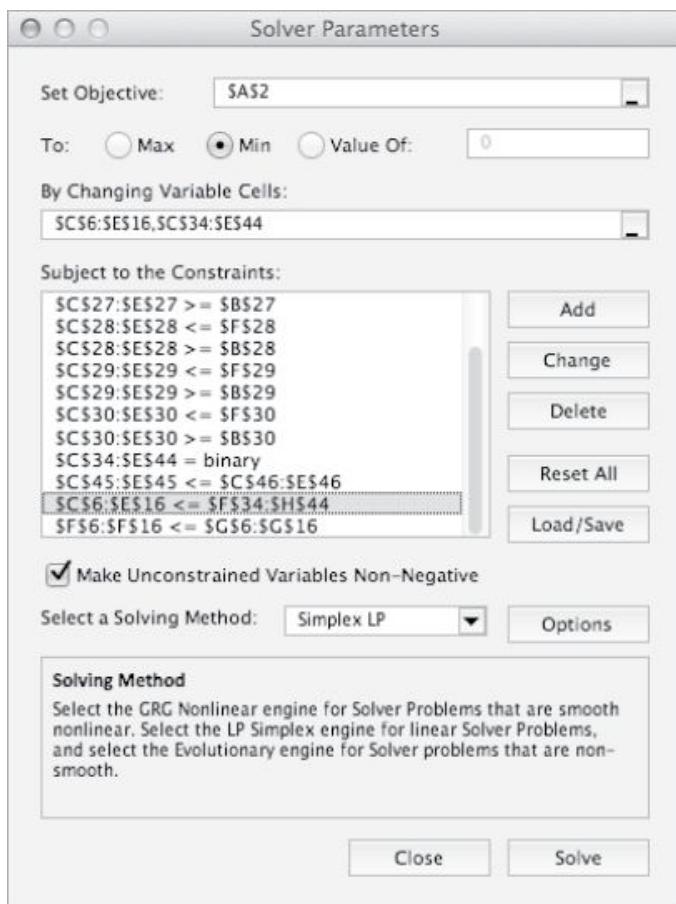


Figura 4-32: Iniciando o Solver

Pressione Solve. Você notará que o problema demora mais para resolver com o acréscimo das variáveis binárias. Ao usar variáveis binárias e inteiras em sua fórmula, o Solver exibirá a melhor solução “vigente” que ele encontrar na barra de status. Se por alguma razão o Solver demorar demais, é possível pressionar a tecla Escape e manter a melhor solução vigente encontrada até o momento.

Como mostra a Figura 4-33, a solução ideal do modelo restrito aos quatro fornecedores por mês é 1,24 milhões, aproximadamente 16.000 a mais do que o original ideal. Com este plano, você pode retornar à equipe de cadeia de suprimentos e perguntar a eles se sua burocracia reduzida vale os 16.000.

Quantificar a introdução das novas regras de negócios e restrições dessa maneira é um dos marcos de utilização de modelagem de otimização em negócios. Você pode colocar um cífrão em uma prática de negócios e tomar uma decisão em relação à questão, “Vale a pena?”

TOTAL COST (OBJECTIVE):							
\$ 1,243,658							
PURCHASE DECISIONS							
Varietal	Region	January	February	March	Total Order Available	Brix / Acid Ratio	
Hamlin	Brazil	0.0	0.0	0.0	672	10.5	
Mosambi	India	0.0	0.0	0.0	400	6.5	
Valencia	Florida	259.7	253.3	280.0	793.1	12.0	
Hamlin	California	0.0	0.0	0.0	168	11	
Gardner	Arizona	0.0	84.0	0.0	84.0	12	
Sunstar	Texas	75.4	0.0	134.6	210.0	10	
Jincheng	China	0.0	0.0	0.0	588	9	
Berna	Spain	0.0	0.0	156.2	168	15	
Verna	Mexico	0.0	137.6	129.2	266.8	8	
Biondo Commune	Egypt	210.0	0.0	0.0	210.0	13	
Belladonna	Italy	54.9	125.1	0.0	180.0	14	
Monthly Cost Totals:		\$ 366,233	\$ 344,840	\$ 426,596			
Shipping		\$ 37,379	\$ 33,071	\$ 35,538			
INDICATORS							
BIG M							
Varietal	Region	January	February	March	January	February	March
Hamlin	Brazil	0.0	0.0	0.0	0.0	0.0	0.0
Mosambi	India	0.0	0.0	0.0	0.0	0.0	0.0
Valencia	Florida	1.0	1.0	1.0	1200.0	1200.0	1200.0
Hamlin	California	0.0	0.0	0.0	0.0	0.0	0.0
Gardner	Arizona	0.0	1.0	0.0	0.0	84.0	0.0
Sunstar	Texas	1.0	0.0	1.0	210.0	0.0	210.0
Jincheng	China	0.0	0.0	0.0	0.0	0.0	0.0
Berna	Spain	0.0	0.0	1.0	0.0	0.0	168.0
Verna	Mexico	0.0	1.0	1.0	0.0	300.0	300.0
Biondo Commune	Egypt	1.0	0.0	0.0	210.0	0.0	0.0
Belladonna	Italy	1.0	1.0	0.0	180.0	180.0	0.0
Total Suppliers Used		4.0	4.0	4.0			
Limit 4		4	4	4			

Figura 4-33: A solução ideal limitada a quatro fornecedores por período

É assim que as restrições “Big M” são configuradas; elas aparecerão novamente no gráfico do problema de agrupamento no Capítulo 5.

## Multiplicando Variáveis: Aumentando a Quantidade para 11

### O OPEN Solver É NECESSÁRIO PARA O EXCEL 2010 E 2013

Essa última parte foi difícil, mas foi brincadeira de criança comparada à próxima regra de negócios que você modelará.

Para este próximo problema, por favor, mantenha a planilha trabalhada disponível para download consigo para referência. Este é difícil mas vale a pena aprender se seu negócio for confrontado por problemas

complexos de otimização. Além disso, nada no livro é dependente do conteúdo desta seção, então, se ficar difícil demais, apenas pule. Dito isso, eu o encorajo a mergulhar e experimentar.

Se você está trabalhando no Excel 2010 ou 2013, precisará do OpenSolver instalado e carregado (veja o Capítulo 1 para uma explicação). Se você não usar o OpenSolver para resolver o problema nessas versões do Excel, receberá um erro dizendo que o modelo de otimização é grande demais. Para usar o OpenSolver neste capítulo, configure o problema normalmente como mostrado nesta seção, mas quando chegar a hora de solucionar, use o botão Solver da Faixa de Opções.

Antes de implementar o plano limitado de fornecedor, você é informado que novos “redutores de acidez” chegaram na fábrica de mistura. Usando um permutador de íons com uma base de citrato de cálcio, a tecnologia consegue neutralizar 20% da acidez no suco que passa por ela. Isso não apenas reduz a porcentagem de ácido em 20 mas também aumenta o índice Brix/Acid em 25%.

Mas a energia e matérias primas necessárias para executar o redutor custam \$20 por 1.000 galões de suco colocados nele. Nem todas as encomendas de fornecedores precisam passar pelo processo de desacidificação; entretanto, se um pedido é processado pelo permutador de íons, todo o pedido precisa ser tratado.

É possível criar um plano ideal que tente usar o permutador de íons para reduzir o custo ideal? Pense em como isso poderia ser configurado. Agora você tem que fazer um novo conjunto de decisões relacionadas a quando reduzir e quando não reduzir a acidez. Como essas decisões podem interagir com quantidades de encomendas?

Comece copiando a aba Optimization Model (Limit 4) para uma nova aba. Chame-a de **Optimization Model Integer Acid**.

O problema com esta regra de negócios é que a maneira natural de modelá-la é não-linear, e isso levaria ao uso de um algoritmo de otimização lento. Você poderia ter uma variável binária que “ativa” quando quer desacidificar um pedido, porém, isso significa que o custo dessa desacidificação é de:

Indicador de desacidificação \* Quantidade comprada \* 20

Não se pode multiplicar duas variáveis a não ser que você queira mudar para o uso do solver não-linear, contudo, ele nunca descobrirá as complexidades desse modelo. Deve haver um caminho melhor para fazer isso. Lembre-se disso quando fizer programação linear: existem poucas coisas que não podem ser linearizadas pelo uso criterioso de variáveis manipuladas por restrições adicionais e a função objetiva como um par de pinças de salada.

A primeira coisa a ser feita é configurar novas variáveis binárias que estejam “ativas” quando você escolher desacidificar um lote de suco. Você pode inserir um novo pedaço delas em um retângulo entre as encomendas Valência e as restrições de qualidade (células C26:E36).

Além disso, não é obrigatório usar o produto de Indicador de desacidificação \* Quantidade comprada, em vez disso você criará uma nova tabela de variáveis abaixo dos indicadores que forçará a igualar essa quantidade sem expressamente tocá-las (à la esquilo morto). Insira essas células vazias em C38:E48.

A planilha agora possui duas tabelas de variáveis vazias — os indicadores e a quantidade total de suco sendo nutrido por uma redução de acidez — como mostra a Figura 4-34.

Agora, se você quiser multiplicar uma variável binária de desacidificação pela quantidade de suco que você encomendou, quais são os valores que esse produto pode assumir? Há infinitas possibilidades distintas:

- Se tanto o indicador quanto a quantidade de produto comprado são 0, seu produto é 0.

- Se você encomendar algum suco mas decidir não reduzir a acidez, o produto ainda é 0.
- Se escolher reduzir a acidez, o produto é apenas a quantidade de suco encomendada.

Em todo caso, o total possível de suco que pode ser desacidificado é limitado pelo indicador de variável de desacidificação multiplicado pelo total de suco disponível para compra. Se você não quer reduzir a acidez, esse limite superior vai para zero. Se escolher por reduzir, o limite superior pula para o máximo disponível para compra. Essa é uma restrição “Big M” como na seção anterior.

Então, para Brazilian Hamlin, a restrição “Big M” poderia ser calculada como o indicador na célula C26 vezes a quantidade disponível para compra, 672.000 galões, na célula G6. Ao acrescentar esse cálculo próximo ao indicador de variáveis na célula G26, é possível copiar isso para os meses seguintes e varietais.

Isso gera a planilha representada na Figura 4-35.

The screenshot shows a Microsoft Excel spreadsheet titled "OrangeJuiceBlending.xlsx". The spreadsheet contains several rows of data and formulas. The first few rows show product names and their origin: Verna (Mexico), Biondo Commune (Egypt), and Belladonna (Italy). Below these are rows for "Monthly Cost Totals" and "Shipping" costs. Rows 20 and 21 show "Total Ordered" and "Total Required" respectively. Row 23 shows "Valencia Ordered" and row 24 shows "Valencia Required". The last two rows, 26 and 27, are labeled "Acid Reduction Indicator" and contain large empty rectangular boxes. The next few rows (28-36) are blank. The final two rows, 38 and 39, are labeled "Total Reduced" and also contain large empty rectangular boxes. The bottom of the screen shows the Excel ribbon with tabs like Home, Layout, Tables, Charts, SmartArt, etc., and a status bar indicating "Optimization Model Integer Acid".

	A	B	C	D	E
14	Verna	Mexico	0.0	0.0	0.0
15	Biondo Commune	Egypt	0.0	0.0	0.0
16	Belladonna	Italy	0.0	0.0	0.0
17	Monthly Cost Totals:	Price	\$ -	\$ -	\$ -
18		Shipping	\$ -	\$ -	\$ -
19					
20	Total Ordered		0.0	0.0	0.0
21	Total Required		600	600	700
22					
23	Valencia Ordered		0.0	0.0	0.0
24	Valencia Required		240	240	280
25					
26	Acid Reduction	Indicator			
27					
28					
29					
30					
31					
32					
33					
34					
35					
36					
37					
38		Total Reduced			
39					
40					
41					
42					
43					
44					
45					
46					
47					
48					
49					

Figura 4-34: O indicador e a quantidade de variáveis adicionadas à decisão de desacidificação

Por outro lado, o total de suco possível que pode ser desacidificado é limitado pela quantidade que você decide comprar, como mostra C6:E16. Então agora você tem dois limites superiores nesse produto:

- Indicador de desacidificação \* Quantidade disponível para compra
- Quantidade comprada

Esse é um limite superior por variável no produto não-linear original.

Mas você não pode parar aí. Se decidir desacidificar um lote, precisa processar o lote **todo**. Isso significa que você deve acrescentar um limite inferior aos dois limites superiores para ajudar a “recolher” a quantidade desacidificada em C38:E48.

Que tal usar apenas a quantidade comprada como limite inferior? No caso em que você decide desacidificar, isso funciona perfeitamente. Você terá um limite inferior e um superior para a quantidade de compra, e um limite superior para a quantidade total disponível para compra multiplicada por um indicador de desacidificação definido em 1. Esses limites superiores e inferiores forçam a quantidade que passa pela desacidificação a ser o carregamento todo, que é o que você quer.

The screenshot shows a Microsoft Excel spreadsheet titled "OrangeJuiceBlending.xlsx". The data is organized into several sections:

- TOTAL COST (OBJECTIVE):** Row 1, columns A, B, C, D, E, F, G, H, I.
- PURCHASE DECISIONS:** Rows 4-16. It includes columns for **Varietal**, **Region**, **January**, **February**, **March**, **Total Ordered**, and **SPECS** (Qty Available, Brix / Acid Ratio, Acid (%)).
- Monthly Cost Totals:** Rows 17-18. It includes columns for **Price**, **Shipping**, and **Total**.
- Total Ordered:** Row 20. It includes columns for **0.0**, **0.0**, and **0.0**.
- Total Required:** Row 21. It includes columns for **600**, **600**, and **700**.
- Valencia Ordered:** Row 23. It includes columns for **0.0**, **0.0**, and **0.0**.
- Valencia Required:** Row 24. It includes columns for **240**, **240**, and **280**.
- Acid Reduction:** Rows 26-36. It includes a column for **Indicator** (containing values like 0, 1, 0, 0, etc.) and a formula column with the formula  $=C26*\$G6$ .

Figura 4-35: Cálculo adicionado para o limite superior de quanto suco pode ser desacidificado

Mas se você escolher não desacidificar um lote? Então um dos limites superiores se torna um indicador de 0 multiplicado pela quantidade disponível para compra, enquanto o limite inferior ainda é a quantidade comprada. Nesse caso, uma quantidade de compra diferente de zero que não é desacidificada se torna impossível.

Hmmm.

Então você precisa de uma maneira de “desligar” esse limite inferior na situação em que você escolhe não desacidificar o suco.

Em vez de fazer do limite inferior a quantidade que você encomendou, por que não fazer o seguinte:

Quantidade comprada - Quantidade disponível para compra \* (1 - indicador de desacidificação)

No caso de escolher desacidificar, esse limite inferior salta para a quantidade que você comprou. No caso em que você não desacidifica, esse valor se torna menor ou igual a 0. A restrição ainda existe, mas é, para todos os efeitos, inútil.

É um pouco bagunçado, eu sei.

Tente trabalhar com um exemplo. Você compra 40.000 galões de suco Brazilian Hamlin. Mais adiante, você decide desacidificar.

Os limites superiores da quantidade que você está desacidificando são a quantidade comprada de 40 e o indicador de desacidificação vezes a quantidade disponível de 672.

O limite inferior da quantidade que você está desacidificando é  $40 - 672 * (1-1) = 40$ . Em outras palavras, você tem limite superior e inferior de 40, então você comprimiu a quantidade que está desacidificando em ~~Indicador de desacidificação \* Quantidade comprada~~ sem calcular essa quantidade.

Se eu escolher não desacidificar o Hamlin, o indicador é definido para 0. Nesse caso, tem-se limites superiores de 40 e  $672 * 0 = 0$ . Tem-se um limite inferior de  $40 - 672 * (1-0) = -632$ . E como a caixa de verificação foi marcada tornando todas as variáveis não negativas, isso significa que a quantidade de Hamlin que você está desacidificando está entre 0 e 0.

Perfeito!

Então vamos colocar esse limite inferior em uma tabela à direita do cálculo do limite superior. Na célula K26, digite:

=C6 - \$G6 \* (1-C26)

E você pode copiar aquela fórmula para cada varietal e mês, gerando a planilha na Figura 4-36.

Próximo à seção Total Reduced, subtraia o valor do total de compras em C6:E16 para obter as quantidades de suco remanescentes de Not Reduced. Por exemplo, na célula G38, coloque:

=C6 - C38

Você pode arrastar isso até as outras células na tabela (veja a Figura 4-37).

Para concluir a formulação, você precisa alterar os cálculos de custo, Brix/Acid e Acid %. Para custo, deve-se adicionar 20 vezes a soma dos valores Total Reduced do mês na célula Price. Por exemplo, o cálculo de Price de January seria:

=SUMPRODUCT(C6:C16,\$L6:\$L16)+20\*SUM(C38:C48)

que você pode arrastar até February e March.

The screenshot shows a Microsoft Excel spreadsheet titled "OrangeJuiceBlending.xlsx". The data is organized into several sections:

- Header Rows:** Row 4 contains column headers like "SPECs", "Total", "City", "Brix / Acid", "Astringency", "Color (1-10 Scale)", "Price (per 1K Gallons)", and "Shipping".
- Data Rows:** Rows 5 through 16 contain monthly data for January, February, and March. Each row includes columns for "Total Ordered", "Available Ratio", "Acid (%)", and "Stringency".
- Constraint Rows:** Rows 17 and 18 show the sum of "Total Ordered" for each month.
- Total Possible Rows:** Rows 26 through 36 show the total possible values for each column, calculated as the sum of the first 16 rows minus the sum of rows 17 and 18.
- Formulas:** A formula is visible in cell C26:  $=C6-\$G6*(1-C26)$ .
- Solver Status:** The status bar at the bottom indicates "Optimization Model (Limit 4) Optimization Model Integer Acid" and "Sum = -672".

Figura 4-36: Acrescentando um limite inferior à desacidificação

Os cálculos Brix/Acid e Acid % agora serão calculados fora das quantidades divididas nas seções Total Reduced e Not Reduced da planilha. Os valores Not Reduced terão SUMPRODUCT com suas especificações originais, enquanto o mesmo SUMPRODUCT usando o suco de ácido reduzido será escalado por 1,25 e 0,8, respectivamente, para BAR e Acid e adicionado ao total nas médias mensais.

Por exemplo, Brix/Acid para January em C51 pode ser calculado como:

$$= (\text{SUMPRODUCT}(\text{G38 : G48}, \$\text{H6 : \$H16}) + \text{SUMPRODUCT}(\text{C38 : C48}, \$\text{H6 : \$H16}) * 1.25) / \text{C21}$$

Agora é preciso modificar o modelo no Solver. A função objetiva permanece a mesma (soma do preço e envio), mas as variáveis de decisão agora incluem os indicadores de desacidificação e quantidades a serem reduzidas localizadas em C26:E36 e C38:E48.

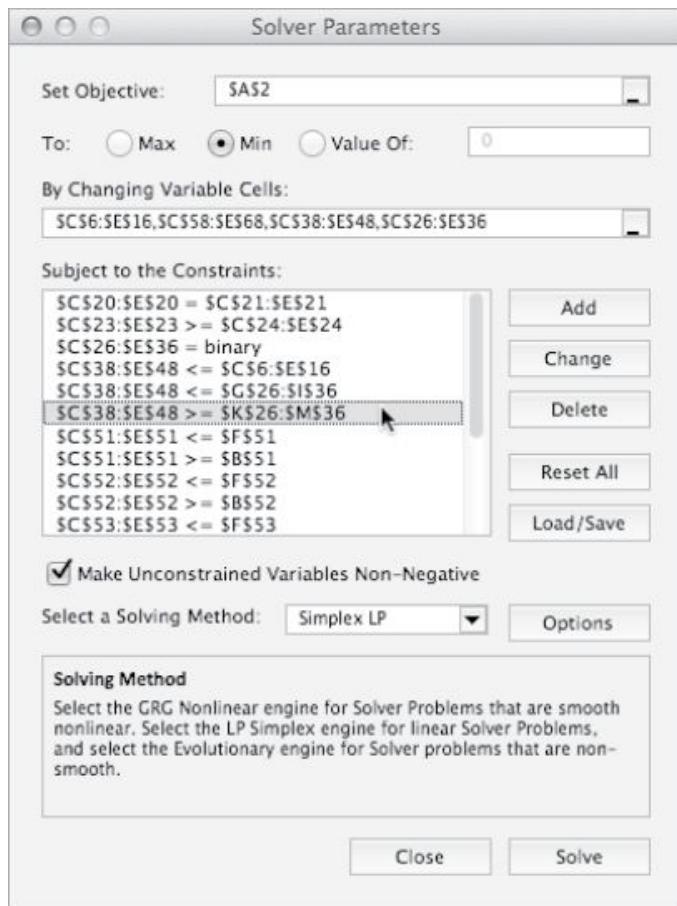
Em relação às restrições, é necessário indicar que C26:E36 é bin. Ainda, C38:C48 é menor ou igual aos dois limites superiores em C6:E16 e G6:I36. Além disso, é preciso uma restrição de limite inferior em que C38:E48 seja maior ou igual a K26:M36.

Tudo isso resulta em um novo modelo representado na Figura 4-38.

OrangeJuiceBlending.xlsx

5 Varietal	Region	January	February	March	Total Ordered	Qty Available	Brix / Acid Ratio	Acid (%)
		6 Hamlin	Brazil	0.0	0.0	0.0	672	10.5 0.60%
7 Mosambi	India	0.0	0.0	0.0	0.0	400	6.5 1.40%	
8 Valencia	Florida	0.0	0.0	0.0	0.0	1200	12 0.95%	
9 Hamlin	California	0.0	0.0	0.0	0.0	168	11 1.00%	
10 Gardner	Arizona	0.0	0.0	0.0	0.0	84	12 0.70%	
11 Sunstar	Texas	0.0	0.0	0.0	0.0	210	10 0.70%	
12 Jincheng	China	0.0	0.0	0.0	0.0	588	9 1.35%	
13 Berna	Spain	0.0	0.0	0.0	0.0	168	15 1.10%	
14 Verna	Mexico	0.0	0.0	0.0	0.0	300	8 1.30%	
15 Biondo Comune	Egypt	0.0	0.0	0.0	0.0	210	13 1.30%	
16 Belladonna	Italy	0.0	0.0	0.0	0.0	180	14 0.50%	
<b>25</b>								
26 Acid Reduction	Indicator	0	0	0		0	0	0
27		0	0	0		0	0	0
28		0	0	0		0	0	0
29		0	0	0		0	0	0
30		0	0	0		0	0	0
31		0	0	0		0	0	0
32		0	0	0		0	0	0
33		0	0	0		0	0	0
34		0	0	0		0	0	0
35		0	0	0		0	0	0
36		0	0	0		0	0	0
37								
38	Total Reduced	0	0	0	Not Reduced	=C6-C38	0	0
39		0	0	0			0	0
40		0	0	0			0	0
41		0	0	0			0	0
42		0	0	0			0	0
43		0	0	0			0	0
44		0	0	0			0	0
45		0	0	0			0	0
46		0	0	0			0	0
47		0	0	0			0	0
48		0	0	0			0	0

Figura 4-37: Adicionando um cálculo “Not-Reduced”



**Figura 4-38:** Formulação do Solver para o problema de desacidificação

Pressione Solve e deixe Branch e Bound fazerem seus trabalhos. Você terminará com uma solução ideal que é aproximadamente \$4.000 menor do que na formulação anterior. Ao examinar as novas variáveis de decisão, descobre-se que dois lotes — um do Arizona e um do Texas — estão passando pelo processo de desacidificação. Os limites inferiores e superiores para esses dois lotes correspondem com precisão a fim de posicionar o produto das variáveis (veja a Figura 4-39).

## Modelagem de Risco

Aquela última regra de negócios foi difícil, mas ela ilustra como um modelador consegue linearizar a maioria dos problemas de negócios acrescentando mais restrições e variáveis. Entretanto, independente de quão fácil ou difícil os problemas anteriores foram, todos tinham algo em comum — eles trataram cada dado de entrada como verdadeiros.

Isso nem sempre confirma a realidade na qual muitos negócios se encontram. Partes não são todas para especificações, envios nem sempre chegam a tempo, demanda não corresponde a previsão, e por aí vai. Em outras palavras, existe uma variabilidade e **risco** nos dados.

Então como pegar tal risco e modelá-lo com um modelo de otimização?

Figura 4-39: O modelo de desacidificação resolvido

## Dados Distribuídos Normalmente

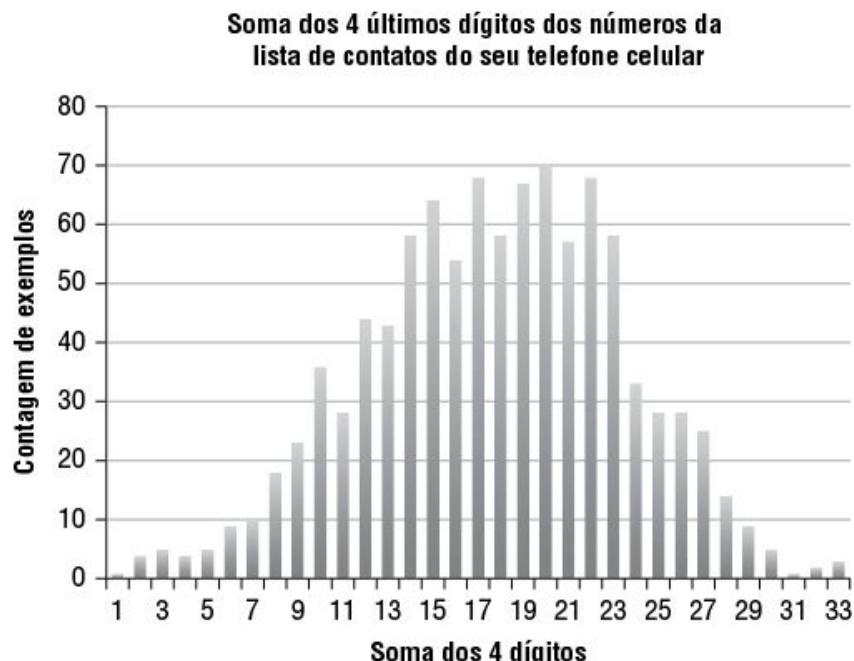
No problema do suco de laranja, você está tentando misturar sucos para tirar variabilidade. Logo, seria aceitável esperar que o produto recebido dos seus fornecedores não tenham especificações variáveis?

Provavelmente o envio do suco de laranja Biondo Commune que você está recebendo do Egito não terá um índice Brix/Acid exato de 13. Pode ser o número esperado, mas provavelmente há alguma resiliência nele. E, quase sempre, essa flexibilidade pode ser caracterizada usando uma **distribuição de probabilidade**.

Uma distribuição de probabilidade, falando livremente, gera uma probabilidade para cada resultado possível de uma situação, e todas as probabilidades somam 1. Talvez a distribuição mais famosa e mais usada seja a distribuição normal, também conhecida como **curva de sino**. A razão pela qual a curva de sino ocorre bastante é porque quando se tem vários fatores independentes, complexos e realistas somados que produzem dados aleatoriamente, esses dados serão **frequentemente** distribuídos de uma maneira normal ou em forma de sino. Isso é chamado de **teorema do limite central**.

Para ver isso, faremos um pequeno experimento. Pegue seu celular e separe os últimos quatro dígitos de cada número de telefone dos contatos salvos. O dígito um provavelmente será **distribuído uniformemente** entre 0 e 9, significando que cada um desses dígitos exibirá aproximadamente a mesma quantidade. O mesmo vale para os dígitos 2, 3 e 4.

Agora, pegue essas quatro “variáveis aleatórias” e some. O menor número que você pode obter é 0 ( $0 + 0 + 0 + 0$ ) e o maior é 36 ( $9 + 9 + 9 + 9$ ). Há apenas uma forma de obter 0 e 36. Há quatro formas de obter 1 e quatro para obter 20, mas há uma tonelada de formas de obter 20. Então se você fez isso para números de telefone o suficiente e fez um gráfico de barras das várias somas, você teria uma curva de sino parecida com a Figura 4-40 (eu usei 1.000 números de telefone para obter a figura, porque eu sou bem popular).

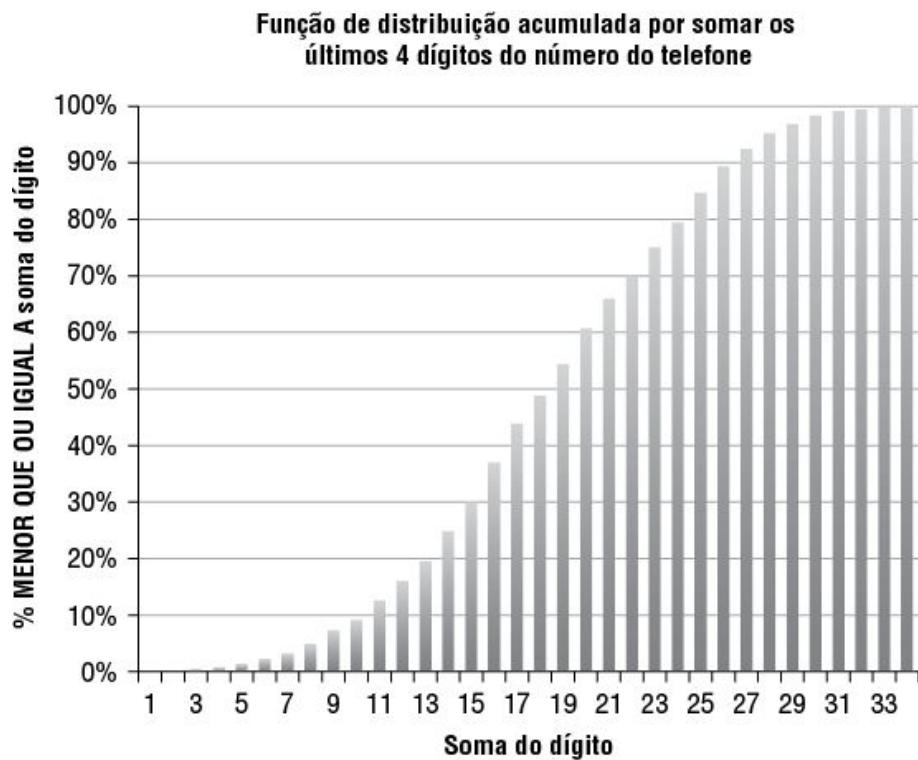


**Figura 4-40:** Combinando variáveis aleatórias independentes para ilustrar como elas se reúnem em uma curva de sino

### A Função de Distribuição Acumulada

Há outra maneira de desenhar essa distribuição que será muito útil, e é chamada de **função de distribuição acumulada (FDA)**. A função de distribuição acumulada entrega a probabilidade de um resultado que é *menor ou igual* a um valor específico.

No caso dos dados do telefone celular, somente 12% dos casos são menores ou iguais a 10, enquanto que 100% dos casos são menores ou iguais a 36 (já que esse é o maior valor possível). Essa distribuição acumulada está representada na Figura 4-41.



**Figura 4-41:** A função de distribuição acumulada para as somas de contatos do telefone celular

Eis o legal sobre a função de distribuição acumulada — *você pode lê-la em modo inverso para gerar amostras da distribuição.*

Por exemplo, se você quisesse gerar um valor aleatório dessa distribuição de soma de quatro dígitos da lista de contatos, poderia gerar um valor aleatório entre 0 e 100%. Digamos que encontra 61% como seu valor aleatório. Procurando por isso no eixo vertical da FDA, 61% alinha-se com 19 no eixo horizontal. E isso poderia ser feito várias vezes para gerar muitas amostras da distribuição.

Agora, uma FDA normal pode ser caracterizada completamente por dois números: uma *média* e um *desvio padrão*. A média nada mais é do que o centro de distribuição. E o desvio padrão mede a variabilidade ou a dispersão da curva de sino em torno da média.

Digamos que no caso do suco encomendado do Egito, tenha uma média Brix/Acid de 13 e um desvio padrão de 0,9. Isso significa que 13 é o centro da distribuição de probabilidade e 68% dos pedidos estarão dentro de  $+/-0,9$  de 13,95% que estarão dentro de dois desvios padrões ( $+/-1,8$ ), e 99,7% estarão dentro de três desvios padrões ( $+/-2,7$ ). Isso às vezes é chamado de regra “68-95-99,7.”

Em outras palavras, é muito provável que você receba um lote 13,5 Brix/Acid do Egito, mas é muito improvável que receba um lote 10 Brix/Acid.

#### CALCULANDO A AMOSTRA DA MÉDIA E O DESVIO PADRÃO

Para aqueles que nunca calcularam o desvio padrão e estão interessados em saber como é feito, é muito fácil.

A Figura 4-42 mostra os últimos 11 pedidos do suco de laranja Biondo Commune do Egito e suas respectivas medidas Brix/Acid na coluna B. A amostra da média dessas medidas é 13, como indicado na planilha de especificações original.

A estimativa do desvio padrão de amostra é a raiz quadrada do erro da média ao quadrado. Por “erro”, eu quero dizer o desvio de cada pedido a partir do valor esperado de 13.

Na coluna C da Figura 4-42, você pode ver o cálculo do erro, e o cálculo do erro quadrado está na coluna D. A média do erro quadrado é `AVERAGE(D2:D12)`, que resulta em 0,77. A raiz quadrada da média do erro quadrado então é 0,88. Muito fácil!

Na prática, entretanto, ao calcular o desvio padrão de amostra para uma pequena quantidade de pedidos, você recebe uma estimativa melhor se somar os erros quadrados e dividir por 1 a menos do que seus pedidos totais (nesse caso 10 em vez de 11).

Se fizer esse ajuste, o desvio padrão torna-se 0,92, como mostra a Figura 4-42.

Order	BAR	Error	Squared Error	MEAN	
2	1	14	1	1	<b>13</b>
3	2	13	0	0	
4	3	13	0	0	Mean Squared Error
5	4	13,5	0,25	0,25	<b>0,77</b>
6	5	14	1	1	Standard Deviation
7	5	13	0	0	<b>0,88</b>
8	7	12,5	-0,25	0,25	
9	8	11	-2	4	Sum Squared Error / N-1
10	9	13	0	0	<b>0,85</b>
11	10	12	1	1	Adjusted Standard Deviation
12	11	14	1	1	<b>0,92</b>

Figura 4-42: Um exemplo do cálculo do desvio padrão da amostra

## Gerando Cenários a partir dos Desvios Padrões no Problema de Mistura

### NOTA

Como na seção anterior, aqueles usando Excel 2010 e Excel 2013 precisarão utilizar o OpenSolver. Apenas configure o problema normalmente e use o botão Solve do OpenSolver na faixa de seleção quando for a hora. Veja o Capítulo 1 para mais detalhes sobre o OpenSolver.

Imagine que no lugar de receber a aba Specs, você recebeu desvios padrões junto com suas especificações em uma aba chamada Specs Variability, conforme a Figura 4-43. O objetivo é encontrar um plano de mistura que custe menos que \$1,25 milhões de dólares a fim de melhor atender as expectativas de qualidade à luz de variabilidade de fornecedor.

Você pode criar uma cópia da aba original Minimax Relaxed Quality chamada Robust Optimization Model. Os novos desvios padrões irão para N6:Q16 adjacentes às antigas especificações.

Uma vez lá, o que fazer com eles?

Você usará a média e o desvio padrão para as especificações para aplicar a abordagem **simulação de Monte Carlo** para resolver esse problema. O método Monte Carlo diz que em vez de incorporar a distribuição diretamente no modelo de alguma forma, é possível experimentar a distribuição, criar cenários ou instâncias de cada conjunto de amostras e, então, incluir aquelas amostras no modelo.

Um cenário é uma possível resposta para a questão, “Se essas são as distribuições para as minhas estatísticas, como seria um pedido real?” Para criar um cenário, lê-se a FDA normal — caracterizada pela média e pelo desvio padrão — de trás para frente, conforme abordado anteriormente com a Figura 4-41.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1			SPECS										
2	Varietal	Region	Qty Available (1,000 Gallons)	Brix / Acid Ratio	Acid (%)	Astringency	Color (1-10 Scale)	Price (per 1K Gallons)	Shipping				
3	Hamlin	Brazil	672	10.5	0.60%	3	3	\$ 500.00	\$100.00	2	0.12%	0.7	1
4	Mosambi	India	400	6.5	1.40%	7	1	\$ 310.00	\$150.00	1.1	0.09%	0.05	1.3
5	Valencia	Florida	1200	12	0.95%	3	3	\$ 750.00	\$ -	0.2	0.19%	0.7	1.4
6	Hamlin	California	168	11	1.00%	3	5	\$ 600.00	\$ 60.00	1	0.18%	0.9	0.9
7	Gardner	Arizona	84	12	0.70%	1	5	\$ 600.00	\$ 75.00	1.3	0.13%	0.6	0.3
8	Sunstar	Texas	210	10	0.70%	1	5	\$ 625.00	\$ 50.00	1.4	0.09%	0.4	1
9	Jincheng	China	588	9	1.35%	7	3	\$ 440.00	\$120.00	0.3	0.19%	0.2	0.3
10	Berna	Spain	168	15	1.10%	4	8	\$ 600.00	\$110.00	0.8	0.12%	0.4	0.9
11	Verna	Mexico	300	8	1.30%	8	3	\$ 300.00	\$ 90.00	1	0.17%	0.5	0.2
12	Biondo Com	Egypt	210	13	1.30%	3	5	\$ 460.00	\$130.00	0.9	0.17%	0.7	0.1
13	Belladonna	Italy	180	14	0.50%	3	9	\$ 505.00	\$115.00	0.6	0.07%	0.9	0.1

Figura 4-43: Especificações com o desvio padrão adicionado

A fórmula no Excel para ler a FDA normal de trás para frente (ou “invertida” se preferir) é NORMINV.

Então gere um cenário na coluna B, começando na linha 33 abaixo de tudo que já está na planilha. Você pode chamar de Scenario 1.

Em B34:B44, você produzirá um cenário real com valores Brix/Acid para todos os fornecedores. Em B34, gere um valor aleatório para Brazilian Hamlin em que sua média Brix/Acid seja 10,5 (H6) e seu desvio padrão, 2 (N6) usando a fórmula NORMINV:

```
=NORMINV(RAND(), $H6, $N6)
```

Você está alimentando um número aleatório entre 0 e 100% em uma NORMINV junto com a média e o desvio padrão, e exibindo um valor Brix/Acid aleatório. Vamos arrastar essa fórmula para baixo até B44.

Começando em B45, pode-se fazer a mesma coisa para Acid, então Astringency, e Color. Agora o intervalo B34:B77 contém um único cenário, aleatoriamente retirado das distribuições. Ao arrastar esse cenário pelas colunas até CW (repare a referência absoluta que permite isso), é possível gerar 100 cenários de especificações aleatórias. O Solver não consegue entendê-los se eles permanecerem em fórmulas não-lineares. Portanto, siga em frente e copie e cole os cenários por cima deles mesmos como **apenas valores**. Agora, os cenários são dados fixos.

Esse monte de cenário de dados em B34:CW77 está representado na Figura 4-44.

OrangeJuiceBlending\_.xlsx

	A	B	C	D	E	F	G	H	I	J	K
33	SCENARIO	1	2	3	4	5	6	7	8	9	10
34	BAR	7.7	11.8	6.5	13.7	12.2	8.0	8.8	11.9	11.9	12.7
35		6.8	6.6	6.8	6.2	4.2	7.8	8.0	5.3	6.4	4.6
36		12.1	12.3	11.5	12.1	12.1	11.8	11.8	12.0	12.1	12.0
37		8.8	12.5	11.7	10.4	9.6	10.8	10.4	11.0	9.0	10.5
38		12.1	12.7	13.2	12.2	10.2	13.1	11.2	11.9	12.5	10.6
39		9.2	9.9	8.6	10.5	9.1	7.3	9.7	9.4	8.5	8.4
40		8.9	9.0	8.8	9.1	8.9	9.3	8.6	9.4	9.1	9.2
41		16.2	16.3	15.5	16.1	16.0	16.3	15.2	13.2	14.6	15.5
42		7.8	8.2	6.9	7.2	7.3	6.8	7.9	8.3	8.8	8.3
43		13.7	13.4	13.7	14.5	12.3	13.4	11.2	14.2	12.3	12.9
44		13.1	13.2	13.7	13.6	13.9	13.6	14.1	13.9	14.1	14.3
45	Acid	0.56%	0.64%	0.68%	0.71%	0.46%	0.33%	0.49%	0.65%	0.43%	0.62%
46		1.38%	1.44%	1.33%	1.40%	1.33%	1.46%	1.37%	1.34%	1.44%	1.33%
47		0.93%	0.78%	1.07%	1.15%	0.89%	0.89%	0.60%	1.12%	0.75%	0.71%
48		1.23%	1.14%	0.77%	0.82%	0.88%	1.31%	0.93%	1.28%	1.11%	0.66%
49		0.76%	0.67%	0.79%	0.61%	0.74%	0.84%	0.69%	0.64%	0.57%	0.66%
50		0.60%	0.83%	0.76%	0.71%	0.71%	0.73%	0.82%	0.72%	0.71%	0.64%
51		1.58%	1.27%	1.46%	1.55%	1.45%	1.56%	1.42%	1.47%	1.44%	1.52%
52		1.23%	1.04%	1.20%	1.05%	1.23%	1.22%	1.14%	0.97%	1.10%	1.21%
53		1.24%	1.34%	1.21%	1.34%	1.19%	1.66%	1.38%	1.14%	1.35%	1.53%
54		1.50%	1.31%	1.41%	1.04%	1.57%	1.66%	1.20%	1.19%	1.60%	1.37%
55		0.55%	0.51%	0.52%	0.50%	0.48%	0.53%	0.49%	0.60%	0.57%	0.55%
56	Astringency	3.5	2.0	2.9	4.3	2.8	3.3	3.1	2.1	3.3	3.1
57		7.0	6.8	7.0	6.9	6.9	7.0	7.1	6.9	7.0	7.0
58		2.8	3.5	2.6	2.9	1.7	3.8	3.9	3.1	2.9	3.0
59		2.5	2.6	0.5	2.2	1.8	1.2	1.8	3.9	3.5	2.2
60		0.7	0.9	1.2	1.4	1.3	1.1	2.1	0.0	1.3	1.2
61		0.8	1.1	0.7	0.6	0.8	1.1	0.7	1.3	0.9	0.4
62		7.0	6.9	7.2	6.9	7.0	7.1	6.9	6.8	7.2	6.5
63		3.5	3.4	3.7	4.2	3.9	4.6	4.0	4.5	4.3	4.5
64		7.7	7.7	8.2	8.4	7.7	8.4	7.7	7.4	7.9	8.7
65		2.6	3.3	3.0	3.1	2.9	2.0	2.8	3.3	3.2	2.6
66		2.7	2.8	3.1	3.2	2.7	3.4	3.3	4.1	2.5	2.6
67	Color	1.7	2.2	3.5	3.4	3.2	3.9	3.0	1.9	5.0	3.5
68		1.0	1.1	-0.7	-2.0	0.1	1.2	-1.7	0.0	0.5	1.5
69											

Figura 4-44: 100 cenários de especificações de suco gerados

### Configurando as Restrições de Cenários

Certo, então o que você quer fazer é encontrar uma solução que simplifique os limites de qualidade ao **mínimo** para atendê-los em todo cenário gerado. Apenas encontre uma solução que proteja o produto.

Portanto, abaixo do primeiro cenário na célula B79, calcule BAR para January desta forma:

```
=SUMPRODUCT($C$6:$C$16,B34:B44)/$C$21
```

Pode-se fazer o mesmo para February e March nas linhas 80 e 81 e então arrastar todo o cálculo para a direita até a coluna CW para obter Brix/Acid para cada cenário.

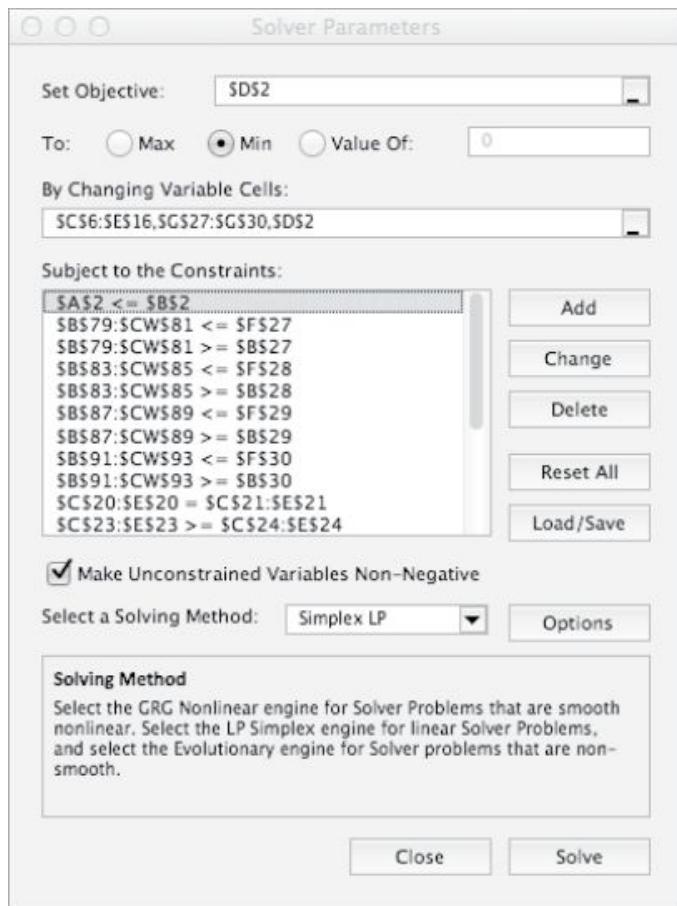
Ao fazer o mesmo com as outras especificações, obtém-se cálculos em cada cenário, como exibido na Figura 4-45.

	A	B	C	D	E	F	G
75		3.0	3.0	3.3	3.1	2.9	3
76		5.0	5.2	5.0	5.1	5.0	5
77		8.9	9.1	9.0	9.0	9.0	9
78							
79	BAR January	11.76340	11.94695	11.44005	11.82407	11.45364	11.5354
80	BAR February	11.67134	11.99191	11.00551	11.77203	11.50634	11.0080
81	BAR March	11.52702	12.18763	11.73502	11.97544	11.40177	11.6557
82							
83	ACID January	0.00989	0.00953	0.01065	0.01077	0.00985	0.0107
84	ACID February	0.01020	0.00939	0.01056	0.01045	0.00992	0.0112
85	ACID March	0.01065	0.00964	0.01025	0.01004	0.00994	0.0113
86							
87	ASTRINGENCY January	3.58260	4.09735	3.89476	4.12443	3.35370	4.3885
88	ASTRINGENCY February	3.45779	3.78433	3.51102	3.76686	3.12396	4.3338
89	ASTRINGENCY March	3.16257	3.63475	2.97732	3.41668	2.68769	3.4501
90							
91	COLOR January	4.14428	3.63160	5.63969	4.71924	4.07250	4.8937
92	COLOR February	3.88211	3.82494	6.30259	4.72132	3.96255	5.1515
93	COLOR March	4.26506	3.63309	5.49265	4.50521	3.88635	5.0760

Figura 4-45: Cálculos de especificações para cada cenário

Configurar o modelo não é tão difícil. Coloca-se um limite superior de custo de 1,25 milhões em B2. Você ainda está minimizando D2, o relaxamento de qualidade, em uma configuração minimax. Tudo que precisa fazer é colocar os limites de qualidade por todos os cenários em vez de apenas os valores de qualidade esperados.

Desta forma, para BAR, adicione B79:CW81 ≥ B27 e ≤ F27 similarmente para Acid, Astringency e Color, produzindo a formulação exibida na Figura 4-46.



**Figura 4-46:** A configuração do Solver para otimização robusta

Pressione Solve. Você receberá uma solução rapidamente. Agora, se você gerou os cenários aleatórios por si só em vez de manter aqueles oferecidos na planilha disponível para download, a solução encontrada será diferente. Para os meus 100 cenários, a melhor qualidade que puder obter é um relaxamento de 133% mantendo o custo abaixo de 1,25 milhões.

Só por diversão, você pode aumentar o limite superior de custo para R\$1,5 milhões e solucionar novamente. Você recebe um relaxamento de 114% sem o custo subir ao limite superior e ficando em aproximadamente 1,3 milhões. Parece que elevar o custo mais do que isso não lhe oferece qualquer margem para aumentar a qualidade (veja a solução na Figura 4-47).

E é isso! Agora você tem um balanço para custo e qualidade que atende as restrições mesmo em situações reais e aleatórias.

The screenshot shows a Microsoft Excel spreadsheet titled "OrangeJuiceBlending\_.xlsx". The spreadsheet contains several tables and formulas related to orange juice procurement and quality constraints.

**Purchase Decisions:**

Varietal	Region	January			February			March			SPECs		
		Qty	Brix	Acid (%)	Qty	Brix	Acid (%)	Qty	Brix	Acid (%)	Total Order	Available Acid	Acid (%)
Hamilin	Brazil	0.0	0.0	0.0	0.0	672	10.5	0.60%					
Mosambi	India	0.0	0.0	0.0	0.0	400	6.5	1.40%					
Valencia	Florida	240.0	240.0	280.0	760.0	1200	12	0.95%					
Hamilin	California	50.6	117.4	0.0	168.0	158	11	1.00%					
Gardner	Arizona	57.3	26.7	0.0	84.0	84	12	0.70%					
Sunstar	Texas	0.0	25.5	184.5	210.0	210	10	0.70%					
Jincheng	China	46.5	4.2	69.3	120.0	588	9	1.35%					
Berna	Spain	0.0	55.6	112.4	168.0	158	15	1.10%					
Verna	Mexico	0.0	0.0	0.0	0.0	300	8	1.30%					
Biondo Comune	Egypt	122.1	87.9	0.0	210.0	210	13	1.30%					
Belladonna	Italy	83.6	42.6	53.8	180.0	180	14	0.50%					

**Quality Constraints:**

Constraint	Minimum			Maximum			% Relaxed		Minimum		Maximum	
	BAR	ACID	ASTRINGENCY	COLOR	BAR	ACID	ASTRINGENCY	COLOR	BAR	ACID	ASTRINGENCY	COLOR
BAR	11.1609716	12.16546	12.26475	11.81125	12.839	0.33903	11.5	12.5	0.0075	0.01	0	4
ACID	0.00546195	0.009698	0.009741	0.009132	0.01204	0.81522	0.0075	0.01	0.0075	0.01	0	4
ASTRINGENCY	-0.1134007	3.118925	2.946547	3.029595	4.1134	0.02835	0	4	0	4	0	4
COLOR	3.35277953	4.602398	4.747972	4.791111	5.64722	1.14722	4.5	5.5	4.5	5.5	0	4

Figura 4-47: Solução para o modelo de otimização robusto

#### UM EXERCÍCIO PARA O LEITOR

Se você está sedento por dor, eu gostaria de lhe oferecer mais uma formulação para trabalhar.

No problema anterior, você minimizou a porcentagem que deveria ser reduzida e aumentou os limites de qualidade para que cada restrição fosse atendida. Mas e se você quisesse apenas 95% dos cenários cumpridos?

Você ainda minimizaria a porcentagem de relaxamento de qualidade, mas precisaria colocar um indicador de variável em cada cenário e usar restrições para defini-lo em 1 quando as restrições de qualidade do cenário fossem violadas. A soma desses indicadores poderia ser definida  $\leq 5$  como uma restrição.

Experimente. Veja se pode fazer isso.

## Resumindo

Se você ficou comigo nos últimos modelos, então **parabéns**. Aqueles otários não eram problemas infantis. Na verdade, esse talvez seja o capítulo mais difícil neste livro. A partir de agora tudo será mais fácil!

Esta é uma pequena recapitulação do que você acabou de aprender:

- Programação linear simples
- A fórmula minimax
- Adicionar variáveis e restrições inteiras
- Modelar a lógica se-então usando uma restrição “Big M”

- Modelar o produto de variáveis de decisão de uma forma linear
- Distribuição normal, teorema de limite central, funções de distribuição acumulada e o método Monte Carlo
- Usar o método Monte Carlo para modelar o risco dentro de um programa linear

Sua cabeça provavelmente está girando com todos os tipos de aplicações dessas coisas no seu negócio agora. Ou você simplesmente se embebedou com uma bebida forte e nunca mais quer lidar com programação linear. Eu espero que seja a primeira opção, porque a verdade é, você pode ficar arbitrariamente criativo e complexo com programação linear. Em muitos contextos de negócios, você geralmente encontrará modelos com dezenas de milhões de variáveis de decisão.

#### PRATIQUE, PRATIQUE, PRATIQUE! E LEIA MAIS UM POUCO

Programas de modelagem linear, especialmente quando você tem que executar truques de “remoção de esquilo”, podem ser não-intuitivos. A melhor forma de se especializar nisso é encontrar oportunidades em sua própria linha de trabalho em que possa usar modelagem e se dedicar.

Não é possível memorizar tudo isso; você adquire uma ideia de trabalhar com certas peculiaridades de modelagem com a prática.

Se quiser alguma literatura adicional de programação linear para complementar sua prática, estes são alguns recursos online gratuitos que recomendo:

- O livro de modelagem de otimização AIMMS disponível em <http://www.aimms.com/downloads/manuals/optimization-modeling> em inglês, é um recurso incrível. Não pule os dois capítulos Tip and Tricks; eles são geniais.
- “Formulating Integer Linear Programs: A Rogue’s Gallery” de Brown e Dell da Naval Postgraduate School:  
[http://faculty.nps.edu/gbrown/docs/Brown\\_Dell\\_INFORMS\\_Transactions\\_on\\_Education\\_January2007.pdf](http://faculty.nps.edu/gbrown/docs/Brown_Dell_INFORMS_Transactions_on_Education_January2007.pdf) em inglês.

# 5

## Análise↑de↑Grupo↑Parte↑II: Gráficos↑de↑Rede↑e↑Detecção↑de Comunidade

**E**ste capítulo continua a discussão sobre a identificação e análise de grupo usando o conjunto de dados de venda atacadista de vinho do Capítulo 2. Apesar de ser perfeitamente normal saltar pelo livro, neste caso eu recomendo pelo menos dar uma olhada no Capítulo 2 antes de ler este capítulo, pois não repito os passos de preparação de dados, e você usará a similaridade do cosseno, que foi abordada no final do Capítulo 2.

Além disso, as técnicas usadas aqui dependem das técnicas de otimização de restrição “Big M” apresentadas no Capítulo 4, logo alguma familiaridade com isso seria útil.

Este capítulo continua a abordar o problema de detecção de grupos de clientes interessantes baseados em suas compras, mas ele aborda o problema a partir de uma direção fundamentalmente diferente.

Em vez de pensar em clientes se movimentando ao redor de sinalizadores colocados na pista de dança para atribuí-los a grupos, como fez com agrupamento k-means (Capítulo 2), você verá seus clientes de uma maneira mais relacional. Os clientes compram coisas similares, e dessa forma eles estão relacionados entre si. Alguns são mais “amigáveis” do que outros, pois estavam interessados na mesma coisa. Então, pensando em quão relacionado ou não relacionado os clientes estão entre si, você pode identificar comunidades de clientes sem precisar sinalizar nos dados que são movimentados até que as pessoas se sintam à vontade.

O conceito-chave que permite que você aborde agrupamento de cliente desta forma relacional é chamado de **gráfico de rede**. Um gráfico de rede, como você verá na próxima seção, é uma maneira simples de

armazenar e visualizar entidades (como os clientes) que estão conectadas (por exemplo, por dados de compra).

Atualmente, visualização e análise de rede está na moda, e as técnicas usadas para explorar conhecimentos de gráficos de rede geralmente funcionam melhor do que técnicas tradicionais (como agrupamento k-means no Capítulo 2), então é importante que um analista moderno entenda e consiga alavancar gráficos de rede em seu trabalho.

Ao fazer análise de grupo em uma rede, as pessoas geralmente usam o termo **detecção de comunidade**, o que faz sentido porque muitos gráficos de redes são sociais por natureza e seus grupos realmente criam comunidades. Este capítulo foca em um algoritmo de detecção de comunidade particular chamado **maximização de modularidade**.

Em um alto nível, a maximização de modularidade recompensa todas as vezes que você coloca dois bons amigos juntos em um grupo e penaliza todas as vezes que alguns estranhos estão juntos. Ao coletar todas as recompensas que puder e evita penalidades, a técnica leva a um agrupamento natural de clientes. Você verá a parte legal em instantes — diferente da abordagem do agrupamento k-means, não é necessário escolher  $k$ . O algoritmo faz isso para você! Dessa forma, a técnica de agrupamento usada leva o aprendizado de máquina **não supervisionado** a um novo nível de descoberta de conhecimento.

Além disso, a partir de uma perspectiva matemática, o agrupamento k-means, enquanto radiano, existe por mais de meio século. As técnicas utilizadas neste capítulo foram desenvolvidas nos últimos anos e são de última tecnologia.

## O↑que↑É↑um↑Gráfico↑de↑Rede?

Um gráfico de rede é uma coleção de coisas chamadas **nós** que são conectadas por relacionamentos chamados **ligações (arestas)**. Redes sociais como o Facebook fornecem muitos dados de rede diagramáveis,

tal como amigos que estão conectados a você e possivelmente um ao outro. Assim, o termo “o gráfico social” cresceu muito ao longo dos anos.

Os nós em um gráfico de rede não precisam ser pessoas, e as ligações que representam relacionamentos não precisam ser relacionamentos interpessoais. Por exemplo, você poderia ter nós que são usuários do Facebook e outros que são páginas de produtos que eles curtiram. Tais “curtidas” alteram as ligações dos gráficos. Da mesma forma, você poderia criar um gráfico de rede de todas as paradas do sistema de transporte da sua cidade. Ou todos os destinos e rotas de um mapa de voo da Delta (na verdade, se procurar pela rota de voo no web site de qualquer companhia, verá que é um gráfico de rede canônico).

Ou você poderia ser um espião e representar em gráfico qualquer um que tenha ligado para alguém por telefone via satélite dentro da al-Qaeda no Magrebe Islâmico. Com a divulgação de material dos serviços de espionagem da NSA por Edward Snowden, o último tipo de gráfico de rede recebeu muita atenção na mídia. Um exemplo é a discussão congregacional a respeito da habilidade da NSA de executar um questionário “três-pulos” — ou seja, entrar no gráfico de rede de dados de ligações e encontrar pessoas que estão a três pulos de um terrorista conhecido (nós estamos conectados a um terrorista por um caminho de três ligações no gráfico).

Qualquer que seja o seu negócio, eu garanto que você tem um gráfico oculto em seus dados. Um dos meus projetos de gráfico de rede favoritos é chamado DocGraph (<http://www.docgraph.com> em inglês). Algumas pessoas intrépidas usaram um pedido de Ato de Liberdade de Expressão para criar um gráfico de todos os tipos de dados referenciais do Medicare (sistema de seguro de saúde do governo americano). Médicos se conectam a outros médicos por referências, e o gráfico pode ser usado para identificar comunidades, formadores de opinião (o médico a que todos vão para a opinião final em um diagnóstico complicado), e até casos de fraude e abuso.

Os gráficos de rede são uma contradição rara no mundo analítico. Eles são esteticamente bonitos e ainda extremamente úteis na forma como armazenam e habilitam certas análises. Tais gráficos permitem que os analistas façam todo o tipo de descobertas de forma visual e algorítmica, como grupos, valores atípicos, influentes locais, e pontes entre grupos diferentes.

Na próxima seção, você visualizará alguns dados de rede para ter uma ideia de como essas coisas funcionam.

## Visualizando um Gráfico Simples

A série de TV *Friends* foi uma das séries mais populares dos anos 90 e início de 2000. A série era centrada em seis amigos: Ross, Rachel, Joey, Chandler, Monica e Phoebe. Se você nunca ouviu falar do programa ou desses personagens, ou você é muito jovem ou está preso em uma caverna.

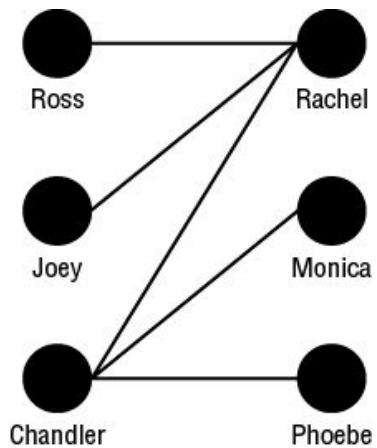
Esses seis personagens se envolveram uns com os outros em vários romances de vários tipos: romances verdadeiros, romances de fantasia que nunca deram em nada, romances de brincadeira baseados em alguma competição ou aposta, e assim por diante.

Pense nesses personagens como seis nós ou vértices no gráfico. Os relacionamentos entre eles são as ligações. De improviso, posso pensar nestas ligações:

- Ross e Rachel, obviamente.
- Monica e Chandler terminaram casados.
- Joey e Rachel tiveram um pequeno romance, mas finalmente decidiram que era muito esquisito.
- Chandler e Rachel se encontram em um episódio de flashback em um incidente na mesa de sinuca, e a Rachel imagina como seria estar com o Chandler.

- Chandler e Phoebe brincam de relacionamento e precisam se beijar, porque o Chandler se recusa a admitir que está com a Monica.

Esses seis personagens e suas cinco ligações podem ser visualizados como mostra a Figura 5-1.



**Figura 5-1: Diagrama de (falsos) romances de Friends**

Muito simples, não é? Ligações e nós. Isso é tudo o que um gráfico de rede representa. E repare como gráficos de rede não têm nada a ver com os gráficos que você está familiarizado, como o gráfico de pontos, de linha e de barra. Não, esses gráficos são um animal completamente diferente.

#### NODEXL

Se você está no Excel 2007 ou 2010, a Social Media Research Foundation lançou um modelo que permite visualização de rede no Excel chamado NodeXL. Não é abordado neste livro porque o software ainda é recente, e os usuários de LibreOffice e Excel 2011 para Mac não conseguiram acompanhar. Se está interessado, você pode conferir NodeXL em <http://smrfoundation.org/nodexl/> — em inglês.

A Figura 5-1 é o que chamamos de gráfico de rede *não direcional*, porque os relacionamentos são mútuos por definição. Algo como dados

de Twitter, por outro lado, é *direcionado*, isto é, eu posso seguir você, mas você não tem que me seguir. Quando visualizamos um gráfico direcionado, as ligações geralmente são setas direcionais.

Agora, uma das desvantagens em usar o Excel para trabalhar com gráficos de rede é que, diferente de outras capacidades gráficas, o Excel não fornece ferramentas para sua visualização.

Então para este capítulo, eu quebrarei minhas próprias regras neste livro e usarei uma ferramenta externa chamada de Gephi para visualização e computação, abordadas na próxima seção. Dito isso, *você pode ignorar todos os aspectos Gephi deste capítulo se quiser*; essa parte é só para diversão.

Mas visualização à parte, se você quer trabalhar nesse tipo de gráfico, precisa de uma representação numérica dos dados. Uma representação intuitiva é chamada de *matriz de adjacência*. Uma matriz de adjacência é apenas uma tabela nó por nó de 0s e 1s, na qual um 1 em uma célula específica significa “colocar uma ligação aqui” e um 0 significa “esses nós estão desconectados”.

Você pode criar uma matriz de adjacência a partir dos dados de *Friends*, conforme a Figura 5-2 (a matriz parece um pouco com uma lagosta na minha opinião). Os nomes dos amigos alinham as colunas e linhas, e os relacionamentos entre eles são exibidos com 1s. Repare como o gráfico é *simétrico* na diagonal, porque o gráfico é não direcional. Se Joey tem um vértice com Rachel, então o inverso é verdadeiro, e a matriz de adjacência mostra isso. Se relacionamentos fossem unilaterais, você poderia ter uma matriz sem essa simetria.

Embora os vértices estejam representados com 1s, não é obrigatório. Você pode adicionar pesos aos vértices, como capacidade — pense em diferentes aviões com diferentes rotas de voos ou larguras de bandas variáveis disponíveis em diferentes links de uma rede de tecnologia. Uma matriz de adjacência ponderada também é chamada de *matriz de afinidade*.

	A	B	C	D	E	F	G
1	Ross	Rachel	Chandler	Monica	Joey	Phoebe	
2	ROSS		1				
3	Rachel		1	1		1	
4	Chandler		1	1	1		1
5	Monica			1			
6	Joey			1			
7	Phoebe				1		

**Figura 5-2:** Uma matriz de adjacência para os dados de Friends

## Breve Introdução à Gephi

Vamos prosseguir e fazer Gephi executar para que você possa importar e visualizar o conjunto de dados *Friends*. Depois, você saberá orientar-se quando as coisas se tornarem reais aqui.

Gephi é uma ferramenta de visualização de rede de código aberto escrito em Java, e é a principal responsável por trás de muitos gráficos de visualização de rede que você vê na mídia hoje em dia. É fácil produzir imagens impressionantes, e as pessoas aproveitaram-na para criar gráficos como coelhos e cenouras e tweetá-los.

A razão pela qual eu exibi a minha hesitação usual em permanecer no Excel é porque Gephi preenche a lacuna de visualização de rede no Excel, é gratuito e funciona no Windows, Mac, e Linux, então não importa qual computador você esteja usando, você consegue acompanhar.

Você não precisa seguir estes passos de visualização. Se só quiser acompanhar nos cálculos sinta-se livre, mas eu recomendo colocar as

mãos na massa. É divertido. No entanto, lembre-se de que esse livro não é sobre Gephi. Se quiser enlouquecer com essa ferramenta, dê uma olhada nos recursos em [wiki.gephi.org](http://wiki.gephi.org) — em inglês, para instruções mais especializadas.

## Instalação do Gephi e Preparação de Arquivo

Para fazer download do Gephi, vá até [www.gephi.org](http://www.gephi.org) em inglês, no seu navegador de internet, e então baixe e instale o pacote seguindo as instruções para o seu sistema operacional em <http://gephi.org/users/install/> em inglês.

Se você deseja um tutorial geral sobre Gephi, verifique o guia de iniciação rápida em <https://gephi.org/users/quick-start/> em inglês. Além disso, dentro da aplicação, Gephi possui uma seleção Help na barra de menus se você precisar.

Uma vez que Gephi é instalado, é preciso preparar a matriz de adjacência para importar para a ferramenta de visualização.

Agora, eu acho que importar uma matriz de adjacência para o Gephi demora um pouco mais do que deveria. Por quê? Porque Gephi não aceita matrizes de adjacência separadas por vírgulas. Cada valor deve ser separado por um ponto e vírgula.

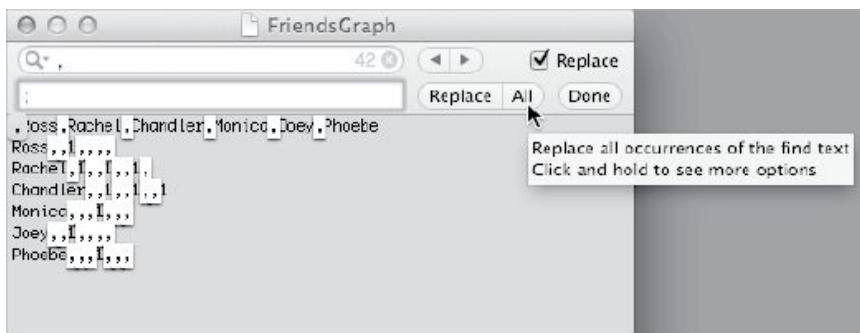
Embora Kurt Vonnegut tenha dito em *Um Homem Sem Pátria*, “Não use ponto e vírgula. Eles são travestis hermafroditas que não representam absolutamente nada. Tudo o que fazem é mostrar que você se deu bem na faculdade”, Gephi ignorou o conselho sensato dele. Peço desculpas. Então acompanhe comigo e mostrarei o processo de importação.

Tornei a planilha FriendsGraph.xlsx disponível com o livro (faça download na página da editora em [www.altabooks.com.br](http://www.altabooks.com.br), procurando pelo nome do livro) ou, se preferir, pode apenas usar o pequeno conjunto de dados da matriz de adjacência exibida na Figura 5-2.

A primeira coisa que você fará é importar esse gráfico para o Gephi e salvá-lo como um CSV (um texto simples) em formato de arquivo separado por vírgula. Para fazer isso, vá até Save As no Excel e escolha

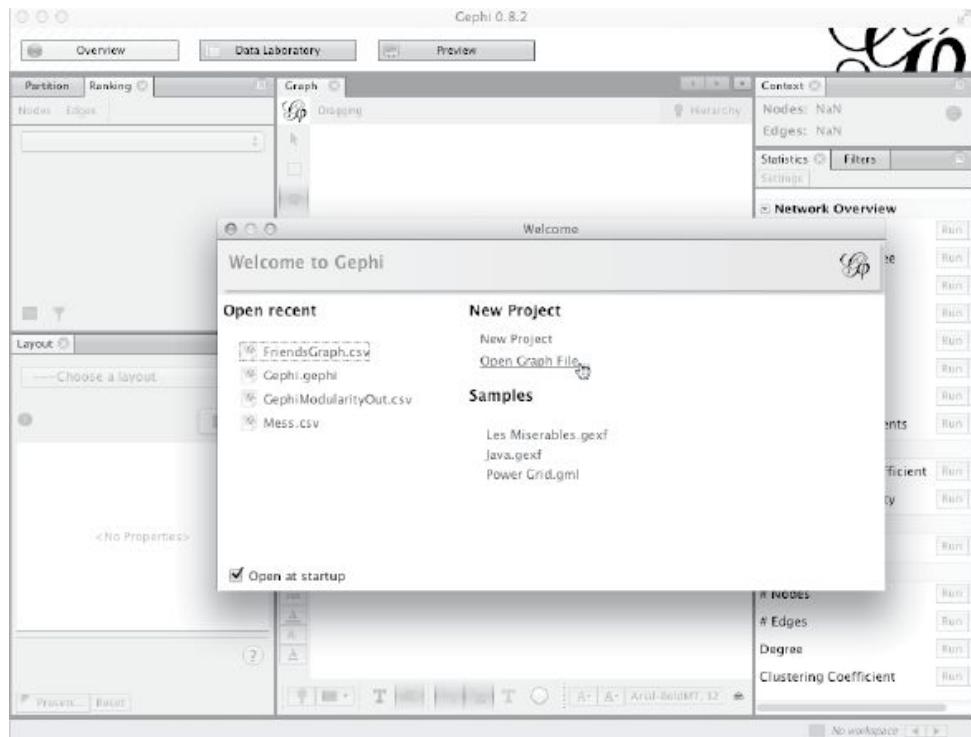
CSV na lista de formatos. O nome do arquivo será FriendsGraph.csv e, quando salvá-lo, o Excel pode exibir alguns avisos, que eu lhe dou permissão para ignorar.

Uma vez que tenha exportado o arquivo, você precisa substituir todas as vírgulas por pontos e vírgulas. Para tal, abra o arquivo em um editor de texto (como o Notepad no Windows ou oTextEdit no Mac OS) e encontre e substitua as vírgulas por ponto e vírgula. No Excel em português o CSV já é salvo com ponto e vírgula. Salve o arquivo. A Figura 5-3 mostra esse processo noTextEdit do Mac.



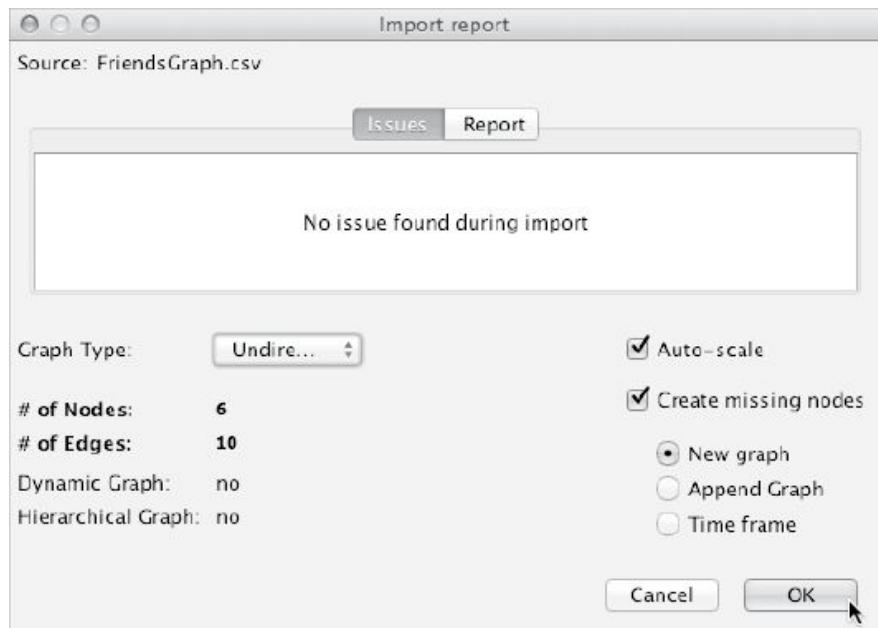
**Figura 5-3:** Substituindo vírgulas por ponto e vírgulas no gráfico CSV Friends

No fim desse processo, abra sua cópia recém-instalada de Gephi e, usando a opção Open Graph File na tela Welcome (veja a Figura 5-4), selecione o arquivo FriendsGraph.csv que acabou de editar.



**Figura 5-4:**↑Abra↑o↑arquivo↑FriendsGraph.csv↑no↑Gephi

Quando tentar abrir o arquivo, uma janela Import Report aparecerá. Note que seis nós e dez vértices foram detectados. A razão das dez ligações (arestas) estarem listadas é porque a matriz de adjacência é simétrica, então cada relacionamento é duplicado. Para resolver essa duplicação, mude o Graph Type de direcional para não direcional na janela Import (veja a Figura 5-5). Pressione OK.



**Figura 5-5:** Importando o gráfico Friends

## Expondo o Gráfico

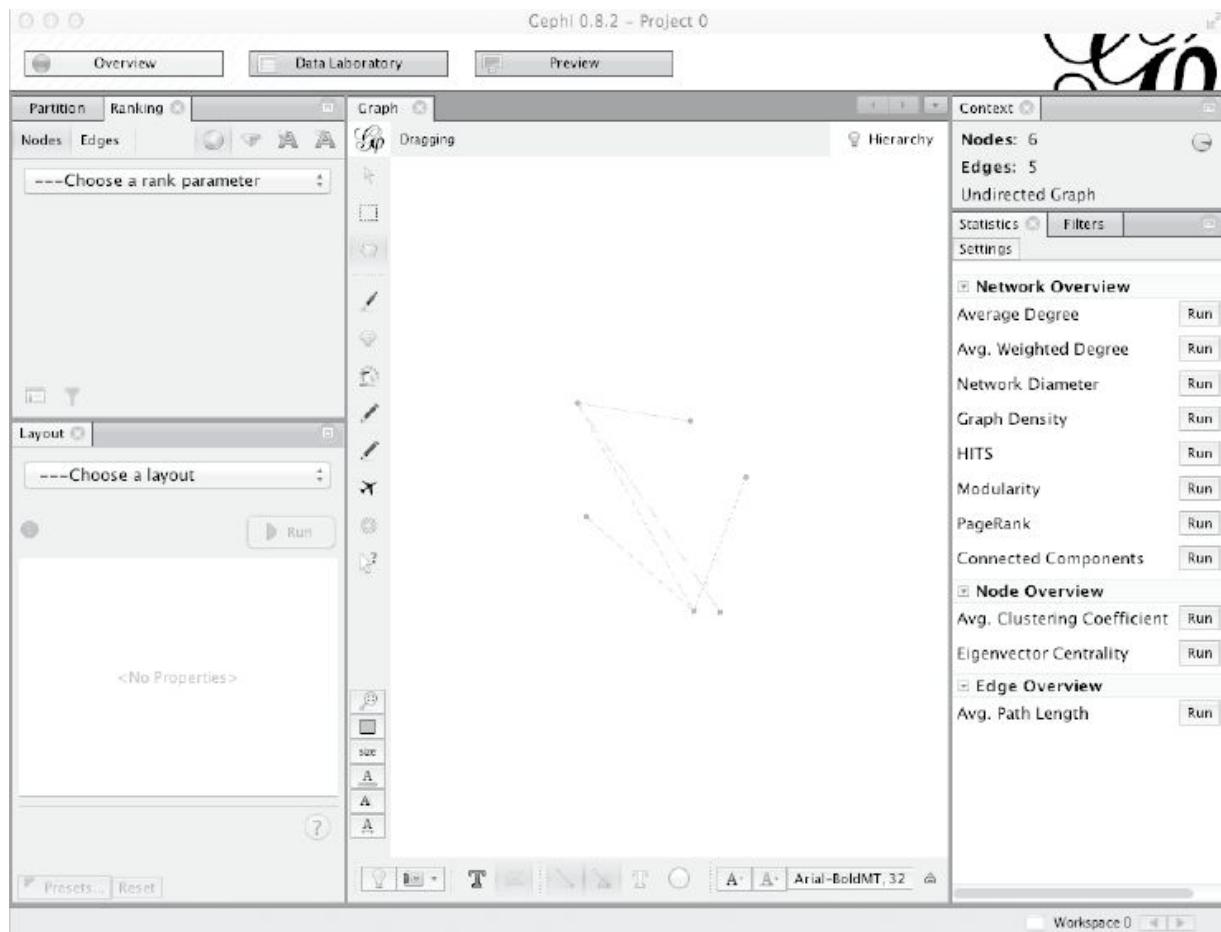
Certifique-se de que a aba Overview esteja selecionada no canto superior esquerdo na janela do Gephi. Se está selecionada, sua janela do Gephi deve parecer com a Figura 5-6. Os nós e as arestas estão expostos aleatoriamente no espaço. O zoom está desajustado, então o gráfico é pouco visível. Seu layout inicial provavelmente parecerá diferente.

Vamos deixar esse gráfico um pouco mais bonito. Você deve saber sobre dois itens navegacionais — pode-se aumentar o zoom com o botão de rolar do mouse e movimentar a tela clicando com o botão direito no espaço e arrastando o gráfico até ele estar centrado.

Clicando no botão T no rodapé da janela overview, você consegue adicionar rótulos aos gráficos de nós para que saiba qual personagem é qual nó. Após aumentar o zoom, ajustar, e adicionar rótulos, o gráfico ficará como mostra a Figura 5-7.

Você precisa expor esse gráfico de uma forma melhor. E, por sorte, Gephi tem vários algoritmos para automatizar esse processo. Muitos deles usam forças como gravidade entre nós e repulsão entre nós desconectados para colocar tudo no lugar. A seção layout do Gephi está

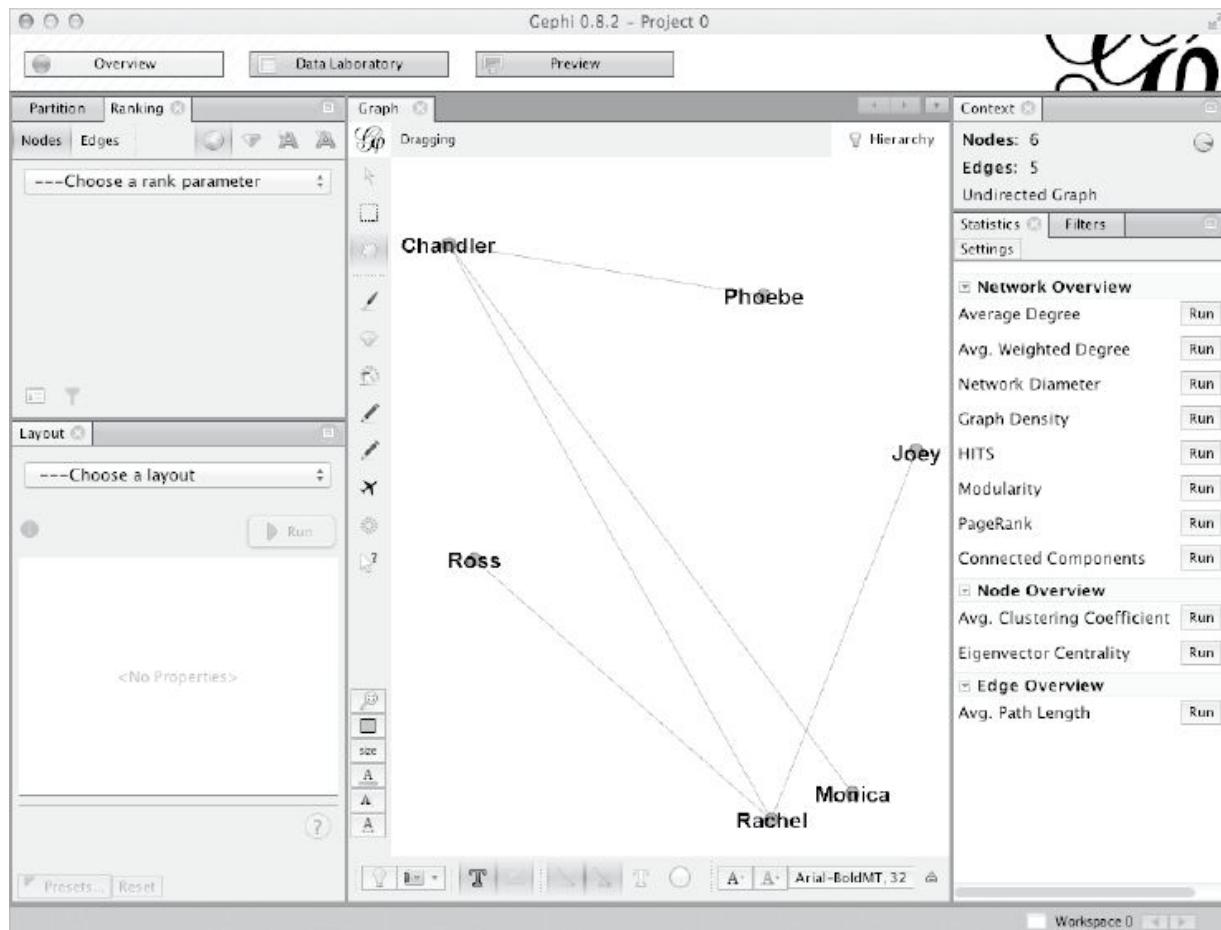
no canto inferior esquerdo da janela do painel overview. Fique à vontade para selecionar algo aleatório do menu e testá-las.



**Figura 5-6:** Layout inicial do gráfico Friends

### NOTA

Fique ciente de que alguns dos algoritmos de layout diminuirão ou expandirão tanto o gráfico que você precisará aumentar e diminuir o zoom para ver o gráfico novamente. Além disso, os tamanhos dos seus rótulos ficarão desajustados, mas há uma seleção Label Adjust abaixo do menu suspenso Layout para corrigir isso.



**Figura 5-7:** O gráfico Friends está decifrável mas bagunçado

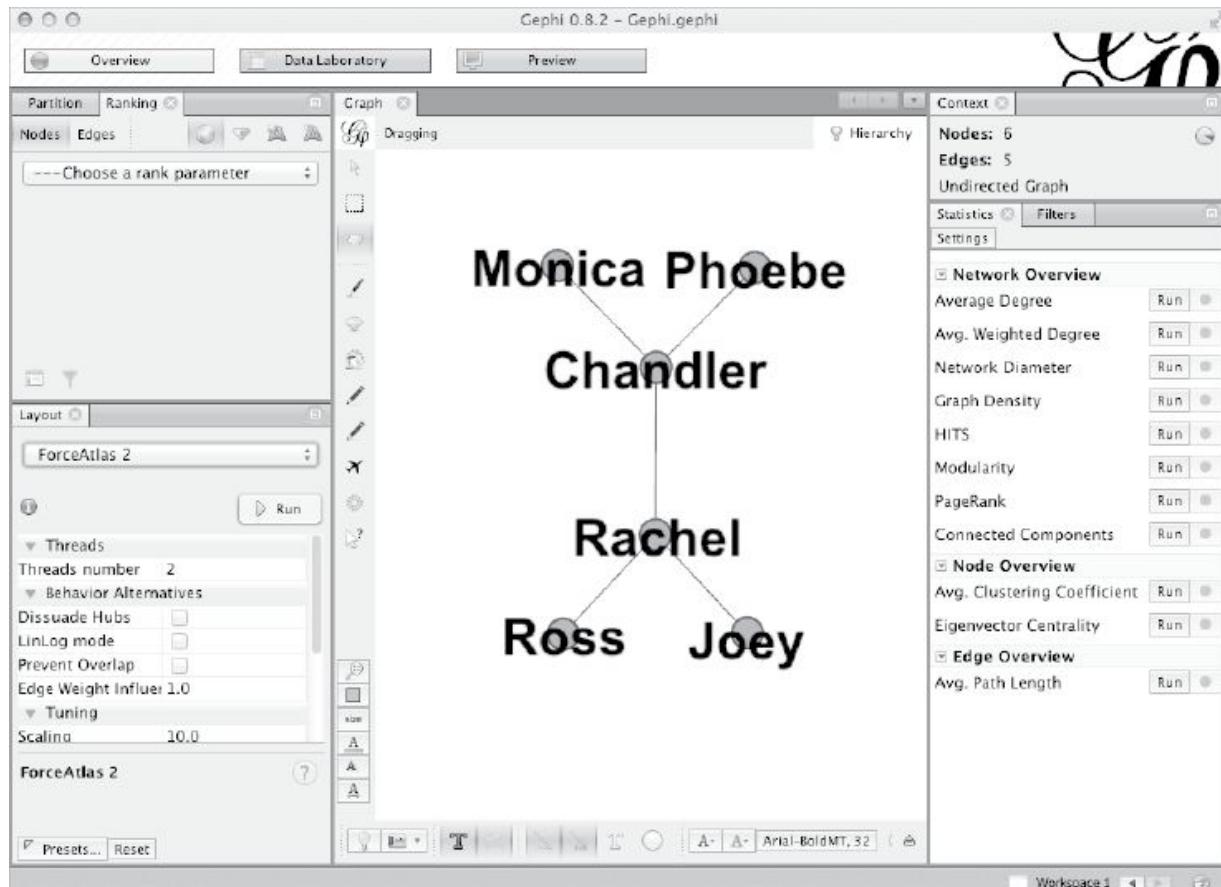
Para chegar no meu layout favorito, a primeira coisa que farei é selecionar ForceAtlas 2 a partir do menu layout e pressionar o botão Run. Isso movimentará meus nós para melhorar as posições. Mas agora os rótulos estão imensos (veja a Figura 5-8).

Selecione Label Adjust no menu e pressione Run. Haverá um resultado bem melhor. Está visível que Rachel e Chandler são os mais bem conectados no gráfico. Obviamente, Monica e Ross estão distantes porque são irmãos, e assim por diante.

## O Grau De Um Nô

Um conceito em gráfico de rede que será importante nesse capítulo é o *grau*. O grau de um nó é simplesmente a contagem dos vértices

conectados a ele. Então Chandler tem grau 3, enquanto Phoebe tem grau 1. Você pode usar esses graus em Gephi para redimensionar os nós.



**Figura 5-8:** Após a execução de ForceAtlas2 no gráfico Friends

### GRAU DE ENTRADA, GRAU DE SAÍDA, IMPORTÂNCIA E MAU COMPORTAMENTO

Em um gráfico direcionado, a contagem das arestas (ou ligações) que entram em um nó é chamada *grau de entrada (indegree)*. A contagem das arestas de saída é *grau de saída (outdegree)*. Grau de entrada em uma rede social é uma simples forma de medir o prestígio de um nó. Esse é frequentemente o primeiro valor que as pessoas veem no Facebook ou Twitter para medir a importância. “Ah, eles têm muitos seguidores... devem ser importantes.”

Agora, essa métrica pode ser caçada. Quem exatamente são esses seguidores cujas arestas entram no seu nó? Talvez eles sejam falsos usuários que você criou para elevar seu próprio prestígio.

O Google usa grau de entrada (em linguagem de mecanismo de busca isso é uma contagem *backlink*) em seu algoritmo PageRank. Quando alguém falsifica links de entrada em seu web site para aumentar o prestígio e subir nos resultados de busca, isso é chamado de *link spam*. Em contextos de uma busca de Internet em que classificações significam grandes negócios, medidas mais complexas de prestígio, influência e centralidade evoluíram para explicar tal comportamento deturpado.

Como você verá no Capítulo 9, esses conceitos de gráfico de rede são úteis em *deteção de valores atípicos*. Em vez de encontrar quem é central em um gráfico, pode-se usar o grau de entrada para encontrar quem está na periferia.

Para ter uma ideia do grau médio do gráfico e quem tem esse grau, pressione o botão Average Degree no lado do Gephi na seção Statistics. Uma janela como a exibida na Figura 5-9 abrirá, na qual o grau médio do gráfico é 1,6667 com quatro nós de grau 1 e dois nós de grau 3 (Rachel e Chandler).

Feche essa janela e navegue até a seção Ranking da janela Overview na caixa de diálogo do canto superior esquerdo. Selecione Nodes e alterne os tamanhos mínimo e máximo para nodes. Ao pressionar Apply, Gephi redimensionará os nós usando grau como um proxy para importância. Eu chamei atenção para essa seção na janela Overview na Figura 5-10.

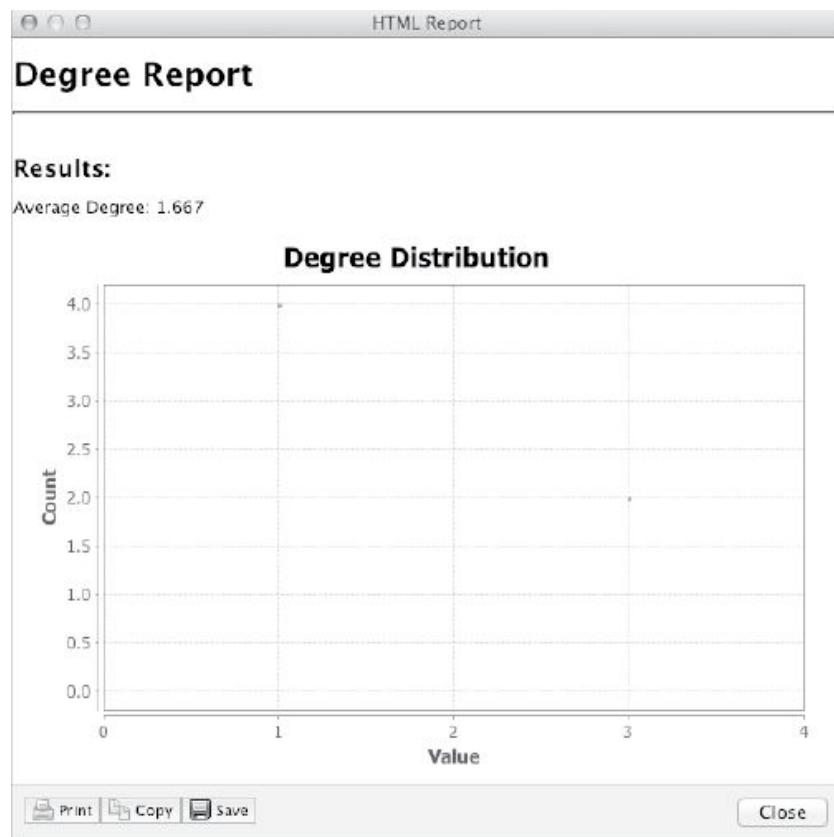
## Uma↑Bela↑Impressão

Embora essas figuras pareçam interessantes, você não as pendurará na sua parede. Para preparar o gráfico para imprimir uma imagem, clique no painel Preview na parte superior do Gephi.

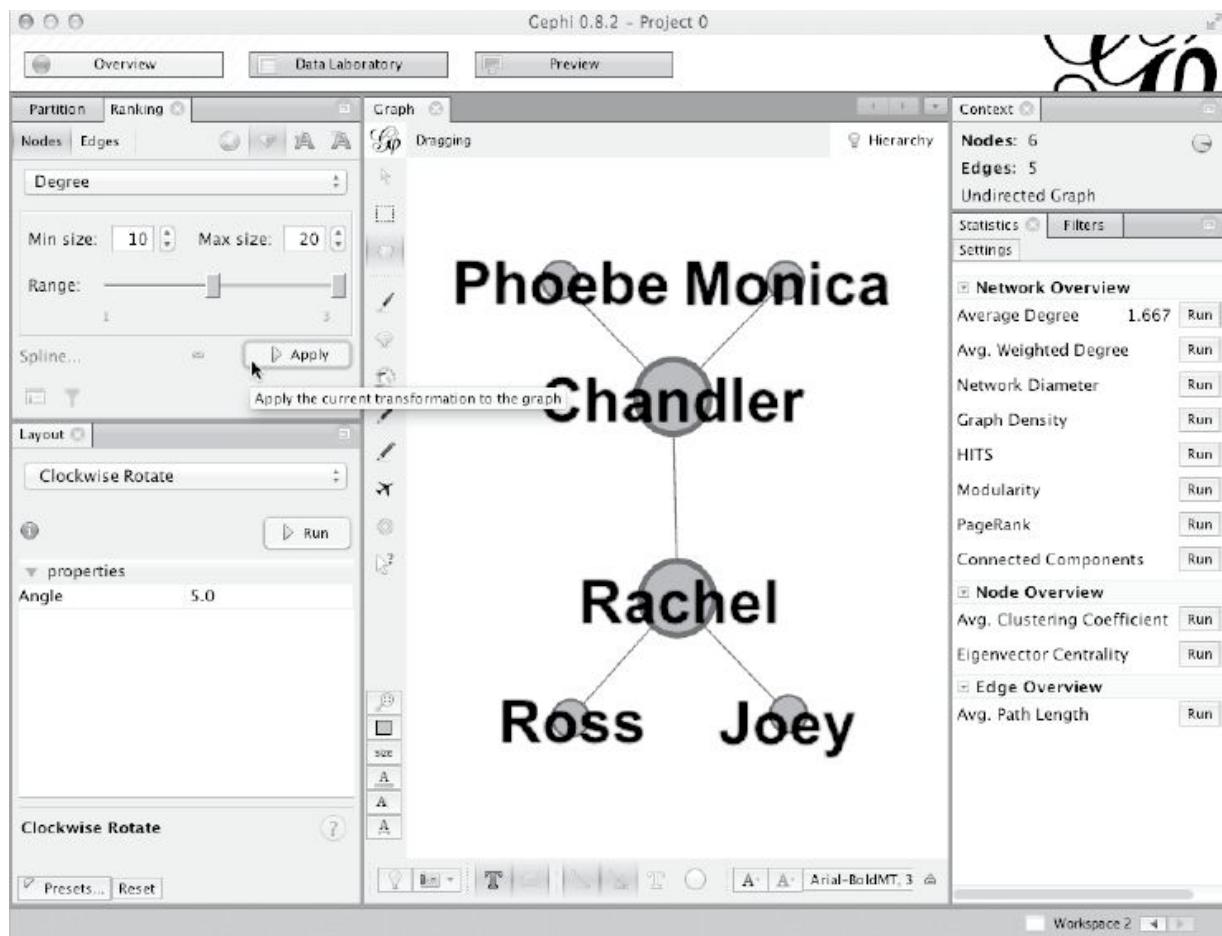
Abaixo da aba Preview Settings, selecione o ajuste Black Background do menu suspenso Presets (porque você tem momentos de hacker) e clique no botão Refresh no canto inferior esquerdo da janela.

Gephi pintará o gráfico com uma beleza sinuosa incomparável (veja a Figura 5-11). Repare como os rótulos são redimensionados com os nós, o que é excelente. Achei os vértices desse gráfico um tanto estreitos, então

alterei a grossura dos vértices de 1 a 3 no painel de configuração do lado esquerdo.



**Figura 5-9:** Calculando o grau médio de um gráfico



**Figura 5-10:** Redimensionando o gráfico de acordo com o grau do nó

Se você quiser exportar essa imagem para um arquivo gráfico (por exemplo, um arquivo .png), pressione o botão Export no canto inferior esquerdo da seção preview settings. É possível distribuir o gráfico em um web site, em uma apresentação de Power Point, ou, até mesmo, em um livro de data science.

## Tocando os Dados do Gráfico

Antes de você voltar ao Excel e confrontar o problema de venda atacadista de vinho do Capítulo 2, daremos uma olhada na seção Data Laboratory do Gephi. Clique em data Laboratory no topo do Gephi para ver os dados ocultos importados para o gráfico.

Repare que existem duas seções de dados: Nodes e Edges. Na seção Nodes, os seis personagens estão visíveis. E como você passou pelo

cálculo Average Degree antes, uma coluna para Degree foi adicionada ao conjunto de dados do nó. Se quiser, é possível exportar essa coluna de volta para o Excel pressionando o botão Export Table na barra de menu. Veja a Figura 5-12.



**Figura 5-11:** Um gráfico de Friends mais bonito

Data Table			
Nodes	Id	Label	Degree
ROSS	ROSS	ROSS	1
Rachel	Rachel	Rachel	3
Chandler	Chandler	Chandler	3
Joey	joey	Joey	1
Monica	Monica	Monica	1
Phoebe	Phoebe	Phoebe	1

**Figura 5-12:** A informação do nó com contagem de grau no Data Laboratory

Clique na seção de arestas, as cinco arestas com suas extremidades são exibidas. Cada aresta tem peso 1, porque a matriz de adjacência foi toda importada com 1s. Se você tivesse modificado alguns desses valores para serem maiores no caso de, digamos, um casamento real, então os pesos maiores estariam refletidos nessa coluna (eles também teriam afetado o layout ForceAtlas 2).

Certo! Então aqui está o seu tour ligeiro no Gephi. Vamos voltar para o agrupamento dos dados de venda atacadista de vinho, e retornaremos ao Gephi depois para fazer mais algumas visualizações e computações.

## Construindo um Gráfico do Dado de Venda Atacadista de Vinho

## NOTA

A pasta de trabalho do Excel usada nesse capítulo, “WineNetwork.xlsx”, está disponível para download na página da editora, em [www.altabooks.com.br](http://www.altabooks.com.br), procurando pelo nome do livro. Essa pasta de trabalho inclui todos os dados iniciais se você trabalhar a partir dela. Ou você pode apenas acompanhar usando as planilhas que já disponibilizei com a pasta de trabalho.

Neste capítulo, quero demonstrar como detectar grupos dentro dos seus dados de compra de cliente representando tais dados como um gráfico. Alguns negócios possuem dados que já são gráficos, como o dado referencial Medicare abordado anteriormente.

Mas, nesse caso, a matriz de compra de vinho do Capítulo 2 não representa relacionamentos cliente a cliente adequados.

Para começar, você deveria descobrir como fazer o gráfico do conjunto de dados de venda atacadista de vinho como uma rede. E isso significa construir uma matriz de adjacência similar à matriz de adjacência *Friends* exibida na Figura 5-2. A partir daí, você conseguirá visualizar e computar o que quiser no gráfico.

Eu continuarei a análise usando a aba Matrix na pasta de trabalho WineNetwork.xlsx (disponível para download na página da editora). Se você lembra, essa é a mesma aba Matrix que criou no início do Capítulo 2 a partir dos dados transacionais de venda de vinho e do metadado de oferta de venda atacadista.

Representadas na Figura 5-13, as linhas da aba Matrix fornecem detalhes de 32 ofertas de vinho oferecidas por Joey Bag O’ Donuts no ano passado. Nas colunas da planilha estão os nomes dos clientes, e cada célula (oferta, cliente) tem um valor 1 se aquele cliente adquiriu aquela oferta.

	B	C	D	E	F	G	H	I	J	K	L	
1	Campaign	Varietal	Minimum Qty (kg)	Discount (%)	Origin	Past Peak	Adams	Allen	Anderson	Bailey	Baker	Bar
2	January	Malbec	72	56	France	FALSE						
3	January	Pinot Noir	72	17	France	FALSE						
4	February	Espumante	144	32	Oregon	TRUE						
5	February	Champagne	72	48	France	TRUE						
6	February	Cabernet Sauvignon	144	44	New Zealand	TRUE						
7	March	Prosecco	144	86	Chile	FALSE						
8	March	Prosecco	6	40	Australia	TRUE			1	1		
9	March	Espumante	6	45	South Africa	FALSE						
10	April	Chardonnay	144	57	Chile	FALSE	1					
11	April	Prosecco	72	52	California	FALSE				1		
12	May	Champagne	72	85	France	FALSE						
13	May	Prosecco	72	83	Australia	FALSE						
14	May	Merlot	6	43	Chile	FALSE						
15	June	Merlot	72	64	Chile	FALSE						
16	June	Cabernet Sauvignon	144	19	Italy	FALSE						
17	June	Merlot	72	88	California	FALSE						
18	July	Pinot Noir	12	47	Germany	FALSE						
19	July	Espumante	6	50	Oregon	FALSE	1					
20	July	Champagne	12	66	Germany	FALSE				1		
21	August	Cabernet Sauvignon	72	82	Italy	FALSE						
22	August	Champagne	12	50	California	FALSE						
23	August	Champagne	72	63	France	FALSE						
24	September	Chardonnay	144	39	South Africa	FALSE						
25	September	Pinot Noir	6	34	Italy	FALSE			1			
26	October	Cabernet Sauvignon	72	59	Oregon	TRUE						
27	October	Pinot Noir	144	83	Australia	FALSE			1			
28	October	Champagne	72	88	New Zealand	FALSE	1					
29	November	Cabernet Sauvignon	12	56	France	TRUE						
30	November	Pinot Grigio	6	87	France	FALSE	1					
31	December	Malbec	6	54	France	FALSE	1			1		
32	December	Champagne	72	89	France	FALSE						
33	December	Cabernet Sauvignon	72	45	Germany	TRUE				1		

**Figura 5-13:** A aba Matrix mostrando quem comprou o quê

Então é preciso transformar esses dados do Capítulo 2 em algo parecido com a matriz de adjacência *Friends*, mas como você faz isso?

Se você criou a matriz Distances para a silhueta k-means no Capítulo 2, já viu algo similar. Para aquele cálculo, você criou uma matriz de distâncias entre cada cliente baseada nas ofertas que eles aceitaram (apresentado na Figura 5-14).

		0	1	2	3	4	5	6	7	8	9
		Adams	Allen	Anderson	Bailey	Baker	Barnes	Bell	Bennett	Brooks	Brown
3	0	Adams	0.000	2.236	2.236	1.732	2.646	2.646	2.646	2.646	2.646
4	1	Allen	2.236	0.000	2.000	2.000	2.449	2.449	2.449	2.449	2.449
5	2	Anderson	2.236	2.000	0.000	2.000	2.449	2.449	2.449	2.449	2.449
6	3	Bailey	1.732	2.000	2.000	0.000	2.000	2.449	2.449	2.449	2.449
7	4	Baker	2.646	2.449	2.449	2.000	0.000	2.000	2.449	2.449	2.449
8	5	Barnes	2.646	2.449	2.449	2.449	2.000	0.000	2.449	2.449	2.449
9	6	Bell	2.646	2.449	1.414	2.449	2.828	2.828	0.000	2.449	2.449
10	7	Bennett	1.732	2.000	2.000	2.000	2.449	2.449	2.449	2.449	2.449
11	8	Brooks	2.646	2.449	2.449	2.449	2.828	2.828	2.449	2.449	2.449
12	9	Brown	1.414	2.236	2.236	1.000	2.236	2.646	2.646	2.646	2.646

Figura 5-14: A tabela customer distances do Capítulo 2

Esse conjunto de dados foi orientado de forma cliente a cliente assim como o conjunto de dados *Friends*. Conexões entre clientes foram caracterizadas por como suas compras se alinharam.

Mas há dois problemas com essa matriz de distância cliente a cliente criada no Capítulo 2:

- No final do Capítulo 2 você descobriu que similaridade assimétrica e medidas de distância entre clientes funcionam muito melhor do que distância Euclidiana no caso do dado de compra. Você se importa com compras, não com “não compras”.
- Se você quer extrair ligações entre dois clientes, quer fazê-lo porque os dois clientes são similares, não porque estão distantes, então esse cálculo precisa ser revertido. Essa proximidade de compras é capturada pela *similaridade do cosseno*, então é preciso criar uma matriz de similaridade para contrastar com a matriz de distância do Capítulo 2.

## Criando uma Matriz de Similaridade do Cosseno

Nesta seção, você pegará a aba Matrix no seu notebook e a partir dela construirá um gráfico cliente a cliente usando similaridade do cosseno. O processo para fazer isso no Excel, usando linhas numeradas e colunas com a fórmula `OFFSET`, é idêntica à usada no Capítulo 2 para a planilha de distância Euclidiana. Para mais informações sobre `OFFSET`, veja o Capítulo 1.

Você começará criando uma aba chamada Similarity na qual colará uma tabela cliente a cliente, em que cada cliente é numerado em uma direção. Lembre-se de que copiar e colar clientes da aba Matrix para as linhas requer o uso do recurso Paste Special no Excel com a caixa Transpose marcada.

A tabela vazia está exibida na Figura 5-15.

The screenshot shows a Microsoft Excel spreadsheet titled "WineNetwork.xlsx". The ribbon menu is visible at the top, showing tabs for Home, Layout, Tables, Charts, SmartArt, Formulas, etc. The main area displays a table with columns labeled A through H and rows labeled 1 through 14. Row 1 contains numerical values 0, 1, 2, 3, 4. Row 2 contains names: Adams, Allen, Anderson, Bailey, Baker, Barne. Rows 3 through 14 contain names: Adams, Allen, Anderson, Bailey, Baker, Barnes, Bell, Bennett, Brooks, Brown, Butler, Campbell. The bottom of the screen shows the status bar with "Normal View" and "Ready".

	A	B	C	D	E	F	G	H
1		0	1	2	3	4		
2		Adams	Allen	Anderson	Bailey	Baker	Barne	
3	0	Adams						
4	1	Allen						
5	2	Anderson						
6	3	Bailey						
7	4	Baker						
8	5	Barnes						
9	6	Bell						
10	7	Bennett						
11	8	Brooks						
12	9	Brown						
13	10	Butler						
14	11	Campbell						

Figura 5-15: A tabela vazia para a matriz de similaridade do cosseno

Comece computando a similaridade do cosseno entre Adams e ele mesmo (que deveria ser 1). Como uma revisão, lembre-se da definição de similaridade do cosseno entre dois vetores de compra binários de clientes que você leu no Capítulo 2.

*A contagem de compras correspondentes nos dois vetores dividida pelo produto da raiz quadrada do número de compras no primeiro vetor vezes a raiz quadrada do número de compras no segundo vetor.*

O vetor de compra de Adams é `Matrix!$H$2:$H$33`; então, para computar a similaridade do cosseno de Adams para ele mesmo, use a seguinte fórmula na célula C3:

```
=SUMPRODUCT(Matrix!$H$2:$H$33, Matrix!$H$2:$H$33) /  
(SQRT(SUM(Matrix!$H$2:$H$33)) * SQRT(SUM(Matrix!$H$2:$H$33))))
```

No topo da fórmula tire `SUMPRODUCT` dos vetores de compra que importam para contar as compras correspondentes. No denominador, tire a raiz quadrada do número de compras para cada cliente e os multiplique.

Agora, essa computação funciona para Adams, mas você quer arrastá-la pela planilha para que não precise digitar cada fórmula individualmente. E, para fazer isso acontecer, use a fórmula `OFFSET`. Substituindo `Matrix!$H$2:$H$33` por `OFFSET(Matrix!$H$2:$H$33, 0, Similarity!C$1)` para as colunas e, similarmente usando

`OFFSET(Matrix!$H$2:$H$33, 0, Similarity!$A3)` para as linhas, você recebe uma fórmula que usa os números de clientes na coluna A e linha 1 para alternar os vetores de compra sendo usados no cálculo de similaridade.

Isso leva a uma fórmula ligeiramente mais feia (desculpe!) para a célula C3:

```
=SUMPRODUCT(OFFSET(Matrix!$H$2:$H$33, 0, Similarity!C$1),  
OFFSET(Matrix!$H$2:$H$33, 0, Similarity!$A3)) /  
(SQRT(SUM(OFFSET(Matrix!$H$2:$H$33, 0, Similarity!C$1))))  
*SQRT(SUM(OFFSET(Matrix!$H$2:$H$33, 0, Similarity!$A3))))
```

Essa fórmula trava `Matrix!$H$2:$H$33` por referências absolutas, então enquanto você arrasta a fórmula pela planilha, ela permanece igual. `Similarity!C$1` modificará as colunas mas ficará na linha 1 onde você quer, e `Similarity!$A3` ficará na coluna A.

Mas ainda não acabou. Você está interessado em criar um gráfico para os clientes que são similares uns aos outros, mas, honestamente, não se importa com a diagonal da matriz. Sim, Adams é idêntico a ele mesmo e tem uma similaridade do cosseno de 1, mas você não está interessado em

desenhar um gráfico com vértices que voltem a onde você começou, então precisa tornar todas essas entradas 0.

Isso apenas significa envolver o cálculo da similaridade do cosseno em uma declaração IF para verificar se o cliente na linha é igual ao na coluna. Assim, você recebe a fórmula final de:

```
IF(C$1=$A3, 0, SUMPRODUCT(OFFSET(Matrix!$H$2:$H$33, 0, Similarity!C$1)  
,  
    OFFSET(Matrix!$H$2:$H$33, 0, Similarity!$A3)) /  
    (SQRT(SUM(OFFSET(Matrix!$H$2:$H$33, 0, Similarity!C$1))))  
    *SQRT(SUM(OFFSET(Matrix!$H$2:$H$33, 0, Similarity!$A3)))))
```

Agora que tem uma fórmula que consegue arrastar, segure o canto inferior direito de C3, arraste-o para o outro lado da planilha até CX3, e para baixo até CX102.

Você tem uma matriz de similaridade do cosseno que mostra quais clientes correspondem uns aos outros. Colocando alguma formatação condicional na tabela, você obtém o que está na Figura 5-16.

The screenshot shows an Excel spreadsheet titled "WineNetwork.xlsx". The formula bar contains the formula for cell C3: `=IF(C$1=$A3, 0, SUMPRODUCT(OFFSET(Matrix!$H$2:$H$33, 0, Similarity!C$1), OFFSET(Matrix!$H$2:$H$33, 0, Similarity!$A3)) / (SQRT(SUM(OFFSET(Matrix!$H$2:$H$33, 0, Similarity!C$1)))) * SQRT(SUM(OFFSET(Matrix!$H$2:$H$33, 0, Similarity!$A3)))))`. The table below has columns labeled A through G and rows labeled 1 through 10. The first row and column are headers. Rows 3 through 10 contain client names: Adams, Allen, Anderson, Bailey, Baker, Barnes, Bell, Davis, Evans, and Franklin. The matrix values are as follows:

		0	1	2	3	4
0	Adams	0.000000	0.000000	0.000000	0.408248	0.000000
1	Allen	0.000000	0.000000	0.000000	0.000000	0.000000
2	Anderson	0.000000	0.000000	0.000000	0.000000	0.000000
3	Bailey	0.408248	0.000000	0.000000	0.000000	0.353553
4	Baker	0.000000	0.000000	0.000000	0.353553	0.000000
5	Barnes	0.000000	0.000000	0.000000	0.000000	0.500000
6	Bell	0.000000	0.000000	0.707107	0.000000	0.000000
7	Davis	0.000000	0.000000	0.000000	0.000000	0.000000
8	Evans	0.000000	0.000000	0.000000	0.000000	0.000000
9	Franklin	0.000000	0.000000	0.000000	0.000000	0.000000

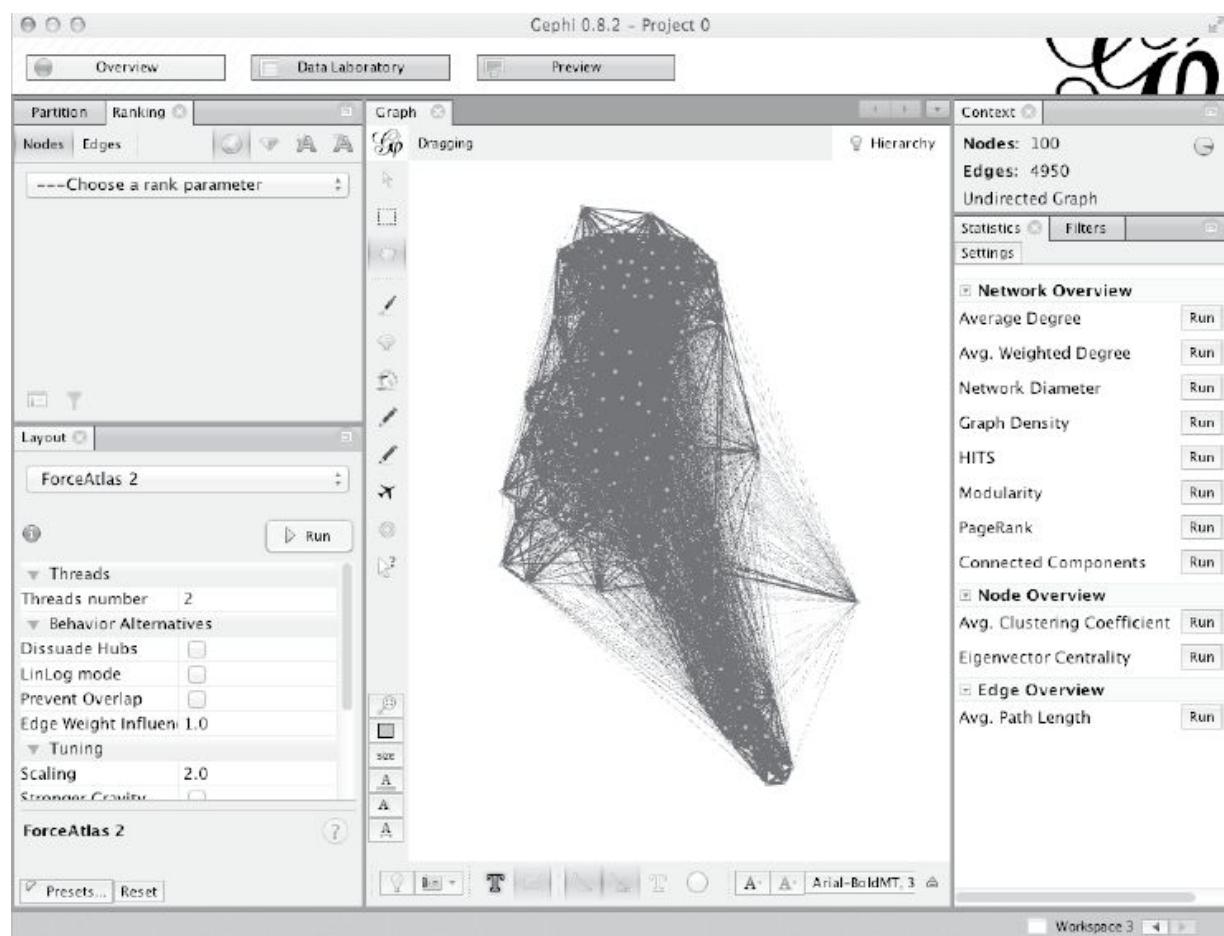
Figura 5-16: A matriz de similaridade do cosseno de cliente completa

## Produzindo um gráfico de vizinhança R

A aba Similarity é um gráfico ponderado. Cada par de clientes tem um 0 entre eles ou algum valor de similaridade do cosseno que mostra quanto forte seu vértice deveria ser. Como está, essa matriz de similaridade é uma matriz de afinidade.

Então por que não apenas dispensar essa matriz de afinidade e espiá-la no Gephi? Talvez você esteja pronto para fazer a análise no gráfico.

Claro, exportar o CSV e importá-lo para o Gephi é possível nesse momento. Mas deixe-me poupará-lo da dor de cabeça e apenas lançar uma imagem (Figura 5-17) do gráfico após ter sido exposto no Gephi. É uma confusão de arestas indo para todo lugar. Conexões demais impedem que os algoritmos de layout apropriadamente afastem nós uns dos outros, então no final você tem um pedaço oval de ruído.



**Figura 5-17:** A confusão de um gráfico de similaridade do cosseno cliente a cliente

Você pegou aproximadamente 30 compras e transformou-as em milhares de arestas no gráfico. Provavelmente seja possível conseguir alguns desses vértices aleatoriamente. Sim, talvez eu e você alinhemos em 1 de nossas 10 compras de vinho, e você tem uma pequenina similaridade do cosseno, mas essa é a aresta que vale a pena desenhar no gráfico?

Para dar sentido aos dados, é melhor se você *tirar* do gráfico as arestas que não importam tanto assim, e manter apenas os relacionamentos mais fortes nele — os relacionamentos que não vêm apenas de uma compra compartilhada por sorte.

Certo, então quais arestas eu deveria remover?

Existem duas técnicas populares para remover arestas de gráficos de rede. Você pode tirar a matriz de afinidade e construir uma das seguintes:

- **Um gráfico de vizinhança  $r$ :** Em um gráfico de vizinhança  $r$ , você mantém apenas os vértices que são de uma certa força. Por exemplo, na matriz de afinidade, os vértices ponderados podem variar de 0 a 1. Talvez você devesse tirar todos os vértices abaixo de 0,5. Isso seria um exemplo de gráfico de vizinhança  $r$  em que  $r$  é 0,5.
- **Um gráfico do vizinho mais próximo (kNN)  $k$ :** Em um gráfico kNN, mantém-se um número determinado de vértices ( $k$ ) saindo de cada nó. Por exemplo, se você fixar  $k$  em 5, manteria os cinco vértices saindo de cada nó que tem as maiores afinidades.

Nenhum dos gráficos é superior ao outro. Depende da situação.

Este capítulo foca na primeira opção, um gráfico de vizinhança  $r$ . Eu o deixo como um exercício para você voltar e trabalhar o problema com um gráfico kNN. É muito simples implementar no Excel usando a fórmula `LARGE` (veja o Capítulo 1 para mais informações sobre `LARGE`). No

Capítulo 9, nós usaremos um gráfico kNN para detecção de valores extremos.

Tudo bem. Então como você pega a aba Similarity e a transforma em uma matriz de adjacência de vizinhança  $r$ ? Bom, primeiro você precisa definir o que  $r$  deveria ser.

No espaço em branco abaixo da matriz de similaridade, conte quantas arestas (valores de similaridade não zero) você tem na matriz de afinidade usando a fórmula na célula C104:

```
=COUNTIF(C3:CX102, ">0")
```

Isso retorna 2.950 arestas feitas a partir das 324 vendas. E se você mantiver apenas as melhores 20% delas? Qual seria o valor de  $r$  para fazer isso acontecer? Bom, como você tem 2.950 arestas, o 80º percentil de valor de similaridade seria o que o 590º vértice tiver. Então, abaixo da contagem de vértice em C105, use a fórmula LARGE para obter 590º, a maior aresta ponderada (veja a Figura 5-18):

```
=LARGE(C3:CX102, 590)
```

Isso retorna um valor de 0,5. Então é possível manter os melhores 20% de arestas jogando fora tudo com similaridade do cosseno menor que 0,5.

The screenshot shows a Microsoft Excel spreadsheet titled "WineNetwork.xlsx". The spreadsheet contains a network matrix and some calculations. The matrix starts at row 2 and column C, listing nodes from 91 to 102. The columns are labeled with node numbers 0 through 6. The matrix entries are cosine similarity values. Row 104 contains the formula =LARGE(C3:CX102, 590) in cell C105, which returns 0.5. Row 105 contains the value 2950 in cell C104 and 0.5 in cell C105.

		A	B	C	D	E	F	G	H	I	J
1				0	1	2	3	4	5	6	
2				Adams	Allen	Anderson	Bailey	Baker	Barnes	Bell	Benne
94	91	Walker		0.816	0.000	0.000	0.000	0.000	0.000	0.000	0.5
95	92	Ward		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.0
96	93	Watson		0.408	0.000	0.000	0.500	0.354	0.000	0.000	0.5
97	94	White		0.289	0.000	0.000	0.354	0.000	0.250	0.000	0.0
98	95	Williams		0.333	0.000	0.000	0.000	0.289	0.577	0.000	0.0
99	96	Wilson		0.408	0.000	0.000	0.500	0.000	0.000	0.000	0.5
100	97	Wood		0.000	0.000	0.000	0.000	0.500	0.500	0.000	0.0
101	98	Wright		0.000	0.354	0.000	0.000	0.000	0.250	0.000	0.0
102	99	Young		0.000	0.289	0.000	0.000	0.204	0.408	0.000	0.0
103											
104		Edge Count		2950							
105		80th Ptile		0.5							

### **Figura 5-18:** Calculando o 80º percentil de vértices ponderados

Agora que você tem o corte para o gráfico de vizinhança  $r$ , a construção da matriz de adjacência é muito fácil. Primeiro crie uma nova aba na pasta de trabalho chamada **r-NeighborhoodAdj**, e cole os nomes dos clientes na coluna A e linha 1 para criar uma tabela.

Em qualquer célula na tabela, você coloca 1 se o valor de similaridade na aba Similarity anterior for maior do que 0,5. Então, por exemplo, na célula B2, pode-se usar a seguinte fórmula:

```
=IF(Similarity!C3>=Similarity!$C$105, 1, 0)
```

A fórmula IF simplesmente verifica o valor de similaridade apropriado contra o corte em `Similarity!$C$105` (0.5) e atribui 1 se for grande o bastante. Como `Similarity!$C$105` está travado com referências absolutas, pode-se arrastar essa fórmula pelas colunas e linhas para preencher toda a matriz de adjacência, como mostra a Figura 5-19 (eu usei um pouco de formatação condicional para beneficiar a figura).

Agora você tem um gráfico de vizinhança  $r$  dos dados de compras dos clientes. Você transformou os dados de compras em relacionamentos de clientes e então os reduziu para um conjunto significante.

Se exportasse a matriz de adjacência de vizinhança  $r$  para Gephi e expusesse isso, você obteria algo muito melhor do que a Figura 5-17. Exporte o gráfico você mesmo, faça o passo do ponto e vírgula, e dê uma olhada comigo.

Como exibido na Figura 5-20, existem, pelo menos, duas comunidades fortemente unidas no gráfico que parecem com tumores. Uma delas está bem separada do resto da horda, o que é ótimo, porque isso significa que seus interesses as separaram do resto dos clientes.

WineNetwork.xlsx

Home Layout Tables Charts SmartArt Formulas Data

B2 : fx =IF(Similarity!C3>=Similarity!\$C\$105,1,0)

	A	B	C	D	E	F	G	H	
1		Adams	Allen	Anderson	Bailey	Baker	Barnes	Bell	Benn
2	Adams		0	0	0	0	0	0	0
3	Allen		0	0	0	0	0	0	0
4	Anderson		0	0	0	0	0	1	
5	Bailey		0	0	0	0	0	0	
6	Baker		0	0	0	0	0	1	0
7	Barnes		0	0	0	0	1	0	0
8	Bell		0	0	1	0	0	0	0
9	Bennett		0	0	0	0	0	0	0
10	Brooks		0	0	0	0	0	0	0
11	Brown		1	0	0	0	1	0	0

Matrix Similarity r-NeighborhoodAdj

Normal View Ready

Figura 5-19: A matriz de adjacência de vizinhança 0,5

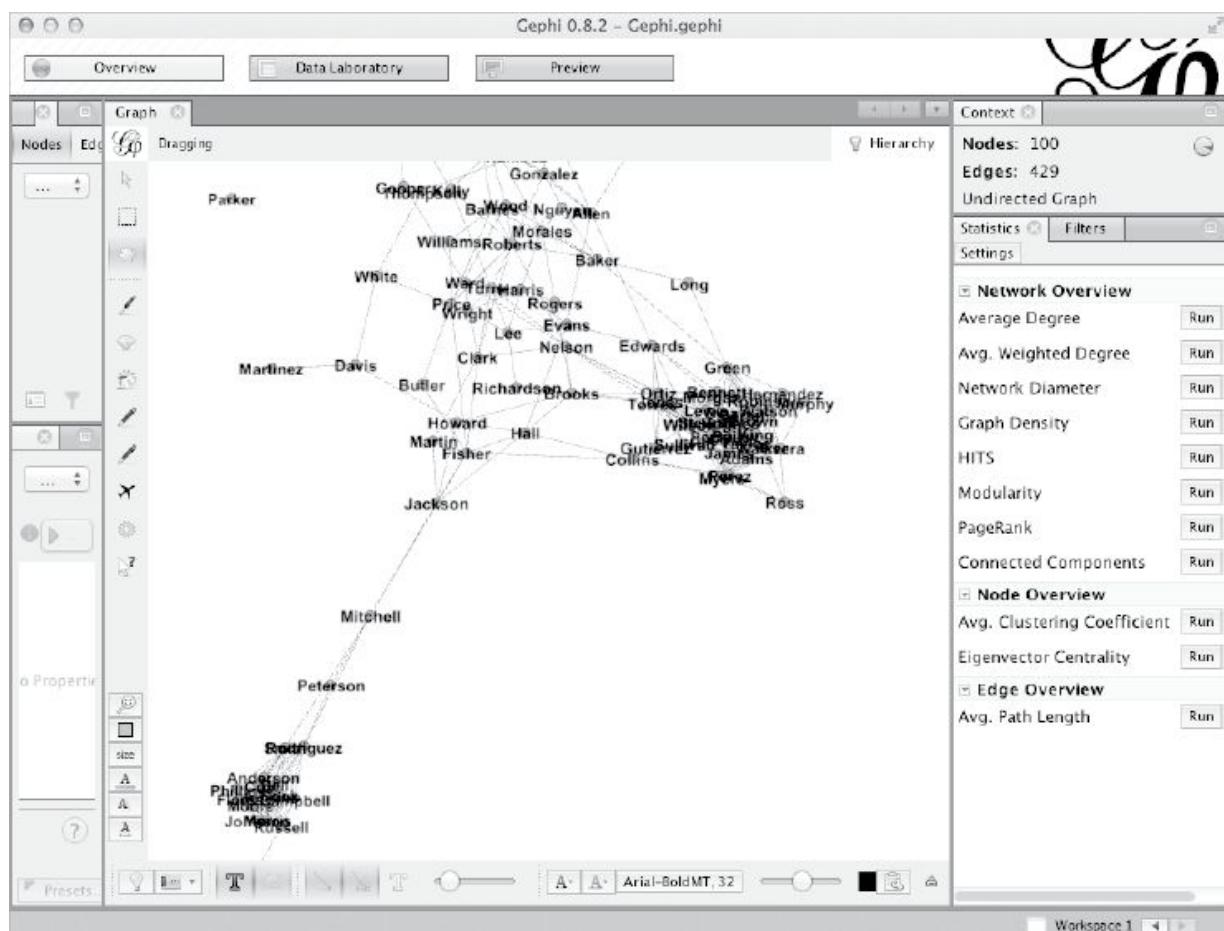


Figura 5-20: Visualização Gephi do gráfico de vizinhança

E então há o pobre velho Parker, o cliente que não ficou com nenhum vértice maior ou igual à similaridade do cosseno 0,5. Portanto, ele está sozinho, chorando em seu chá. Eu honestamente me sinto mal pelo cara, porque os algoritmos de layout tentarão jogá-lo o mais distante possível da parte conectada do gráfico.

Tudo certo! Então agora você tem um gráfico que pode ver. E, de fato, apenas expondo um gráfico e olhando para ele — separando-o em comunidades por inspeção — não é tão ruim. Você pegou dados de alta dimensão e os destilou em algo plano como a pista de dança da escola de ensino médio do Capítulo 2. Mas se tivesse milhares de clientes em vez de uma centena, seus olhos não seriam muito úteis. De fato, até mesmo agora, há uma trama de clientes no gráfico que são difíceis de agrupar. Eles estão em uma comunidade ou em várias?

É aqui que entra a maximização da modularidade. O algoritmo usa esses relacionamentos no gráfico para tomar decisões de agrupamento de comunidade mesmo quando seus olhos têm problemas.

## Quanto Vale um Vértice? Pontuações e Penalidades em Modularidade de Gráfico

Suponha que eu sou um cliente no meu gráfico e quero saber quem pertence a uma comunidade na qual eu estou.

Que tal aquela moça que está conectada a mim por uma aresta? Talvez. Provavelmente. Estamos todos conectados afinal.

E aquele cara do outro lado do gráfico que não se conecta a mim por nenhuma aresta? Hmm, é bem menos provável.

A modularidade de gráfico quantifica essa sensação de que **comunidades são definidas por conexões**. A técnica atribui pontuações para cada par de nós. Se dois nós não estão conectados, eu preciso ser penalizado por colocá-los em uma comunidade. Se dois nós

estão conectados, eu preciso ser recompensado. Qualquer atribuição de comunidade que eu faço, a modularidade do gráfico é conduzida pela soma dessas pontuações para cada par de nós que terminam juntos em uma comunidade.

Usando um algoritmo de otimização (você sabia que o Solver estava chegando!), você pode “testar” diferentes atribuições de comunidade no gráfico e ver qual acumula mais pontos com menos penalidades. Isso lhe dará uma pontuação de modularidade vencedora.

## O que É um Ponto e o que É uma Penalidade?

Em maximização de modularidade você dá a si mesmo um ponto toda vez que agrupa dois nós que compartilham uma aresta na matriz de adjacência. Você recebe zero pontos todas as vezes que agrupa aquelas que não compartilham.

Fácil.

E as penalidades?

É aqui que o algoritmo de maximização de modularidade realmente fica criativo. Considere novamente o gráfico *Friends*, originalmente representado na Figura 5-1.

A maximização de modularidade baseia suas penalidades por colocar dois nós juntos em uma questão:

*Se você tivesse esse gráfico e apagasse o meio de cada aresta e o “reconectasse” algumas vezes aleatoriamente, qual é o número de arestas esperado que receberia entre dois nós?*

Esse número esperado de arestas é a penalidade.

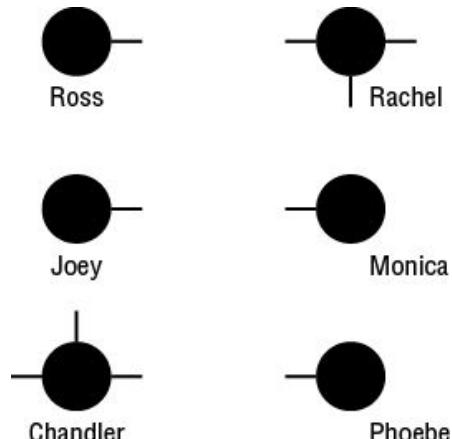
Por que o número esperado de arestas entre dois nós é a penalidade?

Bem, você não quer recompensar o modelo por agrupar pessoas com base em um relacionamento que era provável de acontecer de qualquer forma porque ambas as partes são extremamente sociais.

Eu quero saber quanto desse gráfico é relacionamento *intencional* e conexão, e quanto é apenas porque, “Sim, bem, Chandler está conectado a muitas pessoas, então há uma chance de que Phoebe seja uma delas”.

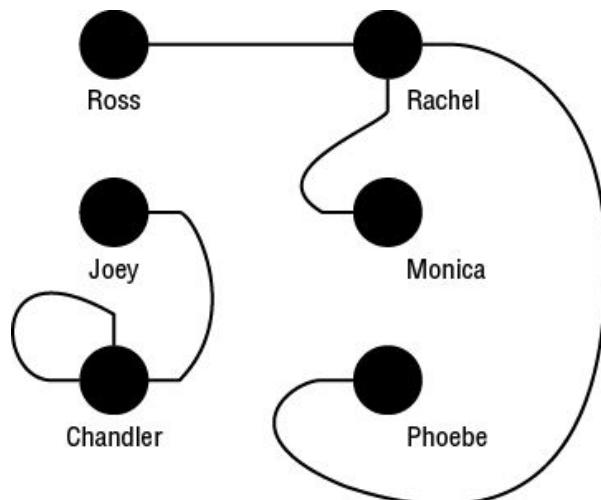
Isso significa que arestas entre dois indivíduos altamente seletivos são “menos aleatórios” e valem mais do que arestas entre duas socialites.

Para entender isso mais claramente, veja a versão do gráfico *Friends* na qual eu apaguei o meio de cada aresta. Essas meio-arestas são chamadas de **stubs**. Veja a Figura 5-21.



**Figura 5-21:** Gráfico de stubs Friends

Agora, pense sobre conectar o gráfico aleatoriamente. Na Figura 5-22, eu desenhei uma religação aleatória feia. E sim, em uma religação aleatória é totalmente possível conectar alguém a ele mesmo se essa pessoa possui múltiplos stubs saindo de si. Que viagem.

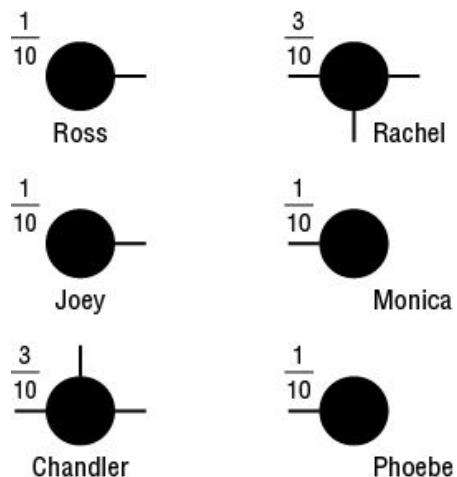


**Figura 5-22:** Uma religação para o gráfico Friends

A Figura 5-22 é apenas uma maneira de ligar, certo? Existem milhares de possibilidades até com um gráfico de apenas cinco vértices. Note que Ross e Rachel foram escolhidos. Quais foram as probabilidades disso acontecer? Com base naquela probabilidade, qual é o número de vértices esperado entre os dois se você religasse o gráfico aleatoriamente várias vezes?

Bem, ao desenhar um vértice aleatório, você precisa selecionar dois stubs aleatoriamente. Então qual a probabilidade de os stubs dos nós serem selecionados?

No caso da Rachel, ela tem três stubs de um total de dez (duas vezes o número de arestas) no gráfico. Ross possui um stub. Então a probabilidade de você ter selecionado Rachel para qualquer aresta é 30%, e a probabilidade de ter selecionado o stub do Ross para qualquer aresta é 10%. As probabilidades de seleção de nós estão representadas na Figura 5-23.



**Figura 5-23:** Probabilidades de seleção de nós no gráfico Friends

Então se você selecionasse nós aleatoriamente para ligar, você poderia selecionar Ross e então Rachel ou Rachel e em seguida Ross. Isso é grosseiramente 10% **vezes** 30% ou 30% **vezes** 10%, que é 2 vezes 0,3 vezes 0,1. Que resulta em 6%.

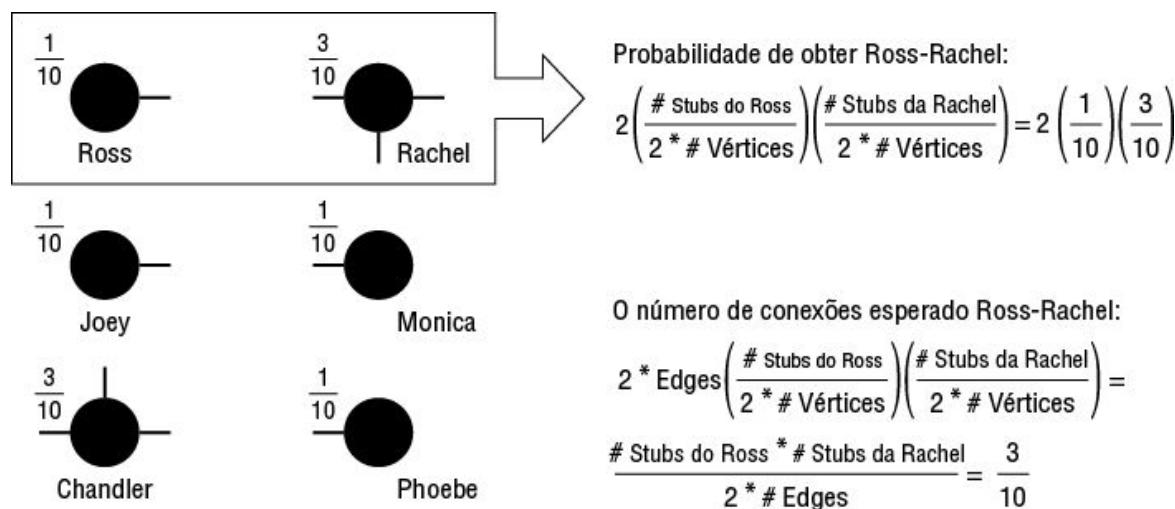
Mas você não está desenhando apenas uma aresta, está? Você precisa desenhar um gráfico aleatoriamente com cinco arestas, para receber cinco

tentativas de combinação. O número de arestas esperado entre Ross e Rachel é 6% vezes 5, ou 0,3 arestas. Sim, isso está certo, arestas esperadas podem ser fracionadas.

Eu lhe surpreendi no estilo *A Origem*? Pense assim. Se eu jogo uma moeda de dólar Sacagawea para cima, com a qual você fica se ela cair em cara e não em coroa, então 50% das vezes você receberá um dólar e 50% das vezes nada. Sua recompensa **esperada** é  $0,5 * \$1 = \$0,50$ , ainda que você nunca ganhe cinquenta centavos em um jogo.

Similarmente, aqui, você apenas encontrará gráficos em que Ross e Rachel estão ou não estão conectados, mas a aresta esperada deles é, todavia, 0,3.

A Figura 5-24 exibe, em detalhes, esses cálculos.



**Figura 5-24:** O número esperado de vértices entre Ross e Rachel

Reunindo esses pontos e penalidades, tudo deve se tornar mais claro.

Se você colocar Ross e Rachel juntos em uma comunidade, você não recebe um ponto inteiro de 1. Isso porque é penalizado em 0,3 pontos já que é o número esperado de vértices que um gráfico aleatório teria de qualquer maneira. Isso o deixa com uma pontuação de 0,7.

Se você não agrupasse Ross e Rachel, então receberia 0 em vez de 0,7 pontos.

Por outro lado, Rachel e Phoebe *não* estão conectadas. Elas têm o mesmo valor de aresta esperada de 0,3 no entanto. Isso significa que, se colocá-las juntas em uma comunidade, você ainda teria uma penalidade mas nenhum ponto, então a pontuação seria ajustada por -0,3.

Por quê? Porque o fato de não existir aresta entre Rachel e Phoebe significa alguma coisa! O número esperado de arestas era 0,3 e ainda assim esse gráfico não tem um, então a pontuação deveria contar para uma possível separação intencional.

Se você não colocasse Rachel e Phoebe juntas em uma comunidade, elas não receberiam nenhuma pontuação, então, nas mesmas condições, seria melhor separá-las em dois grupos diferentes.

**Para somá-las**, os pontos e penalidades capturam a quantidade que a estrutura do gráfico desvia da estrutura **esperada** do gráfico. Você precisa atribuir comunidades que respondam por esses desvios.

A modularidade de uma atribuição de comunidade é a soma desses pontos e penalidades de pares de nós colocados juntos em uma comunidade, dividida pelo **número total de stubs** no gráfico. Você divide pelo número de stubs, então qualquer que seja o tamanho do gráfico, a pontuação máxima de modularidade é 1, o que facilita comparações pelos gráficos.

## Configurando a Planilha de Pontuação

Chega de conversa! Vamos de fato calcular essas pontuações para cada par de clientes no gráfico.

Para começar, vamos contar quantos stubs saem de cada cliente e quanto stubs totais estão no gráfico. Note que a contagem de stub de um cliente é apenas o grau do nó.

Então, na aba r- NeighborhodAdj você pode contar o grau de um nó simplesmente somando uma coluna inteira ou uma linha. Se há um 1, há um vértice, logo um stub, logo é contada. Então, por exemplo, quantos stubs Adams têm? Na célula B102, você pode colocar a seguinte fórmula para contá-los:

=SUM(B2:B101)

Você obtém 14. Similarmente, poderia somar pela linha 2 colocando CX2 na fórmula:

=SUM(B2:CW2)

Você recebe 14 nesse caso também, que é o que esperaria já que o gráfico é indireto.

Copiando essas fórmulas pela planilha, colunas e linhas, você pode contar stubs para cada nó. E simplesmente somando CX na linha 102, você recebe o número total de stubs para o gráfico. Como exibido na Figura 5-25, o gráfico tem um total de 858 stubs.

Agora que você tem as contagens de stubs, é possível criar a aba Scores em sua pasta de trabalho, na qual coloca os nomes dos clientes a partir da linha 1 para baixo e a partir da coluna A, assim como na aba r-NeighborhoodAdj.

Considere a célula B2, que é a pontuação por Adams conectar-se a si mesmo. Isso recebe um ponto ou nenhum? Bem, você pode ler o valor a partir da matriz de adjacência, 'r-NeighborhoodAdj'!B2, e acabou. Se a matriz de adjacência é um 1, é copiada. Simples.

Para o cálculo do vértice esperado que você precisa receber como penalidade, pode-se fazê-lo da mesma maneira como foi mostrado na Figura 5-24:

**# stubs customer A \* # stubs customer B / Total stubs**

Reunindo esses pontos e penalidades na célula B2, você fica com esta fórmula:

```
= 'r-NeighborhoodAdj'!B2 -  
( ('r-NeighborhoodAdj'!$CX2*'r-NeighborhoodAdj'!B$102) /  
'r-NeighborhoodAdj'!$CX$102)
```

WineNetwork.xlsx

	A	CS	CT	CU	CV	CW	CX	
1		Williams	Wilson	Wood	Wright	Young	DEGREE	
91	Torres	0	1	0	0	0	15	
92	Turner	0	0	0	1	0	8	
93	Walker	0	0	0	0	0	14	
94	Ward	0	0	0	1	0	5	
95	Watson	0	0	0	0	0	16	
96	White	0	0	0	0	0	3	
97	Williams	0	0	0	0	0	3	
98	Wilson	0	0	0	0	0	18	
99	Wood	0	0	0	0	0	5	
100	Wright	0	0	0	0	0	4	
101	Young	0	0	0	0	0	4	
102	DEGREE	3	18	5	4	4	858	

**Figura 5-25:** Contando vértices stubs no gráfico r-NeighborhoodAdj

Você tem a pontuação de 0/1 adjacência menos a contagem esperada.

Note que a fórmula usa referências absolutas de células nos valores stub para que, quando arrastar a fórmula, tudo mude apropriadamente. Assim, ao arrastar a fórmula pela aba Scores, você recebe os valores exibidos na Figura 5-26.

`=r-NeighborhoodAdj'!B2-((r-NeighborhoodAdj'!$CX2*r-NeighborhoodAdj'!B$102)/r-NeighborhoodAdj'!$CX$102)`

	A	B	C	D	E	F	G	H	I
1		Adams	Allen	Anderson	Bailey	Baker	Barnes	Bell	Bennett
2	Adams	-0.228	-0.1	-0.21212	-0.28	-0.11	-0.098	-0.2	-0.245
3	Allen	-0.098	-0	-0.09091	-0.12	-0.05	-0.042	-0.1	-0.105
4	Anderson	-0.212	-0.1	-0.19697	-0.26	-0.11	-0.091	0.8	-0.227
5	Bailey	-0.277	-0.1	-0.25758	-0.34	-0.14	-0.119	-0.3	-0.297
6	Baker	-0.114	-0	-0.10605	-0.14	-0.06	0.951	-0.1	-0.122
7	Barnes	-0.098	-0	-0.09091	-0.12	0.95	-0.042	-0.1	-0.105
8	Bell	-0.212	-0.1	0.80303	-0.26	-0.11	-0.091	-0.2	-0.227
9	Bennett	-0.245	-0.1	-0.22727	-0.3	-0.12	-0.105	-0.2	-0.262
10	Brooks	0.131	0.1	0.12121	0.16	0.07	0.056	0.1	0.14

**Figura 5-26:** Atribuições

Para levar esse placar para casa, verifique a célula K2. Esse é o placar para um agrupamento Adams/Brown. É 0,755.

Adams e Brown compartilham uma aresta na matriz de adjacência, então você recebe 1 ponto por agrupá-los (`'r-NeighborhoodAdj' !K2` na fórmula), mas Adams tem uma contagem de stubs de 14 e Brown 15, então a contagem esperada de vértice deles é  $14 * 15 / 858$ . Essa segunda parte da fórmula parece com isto:

$$((r-NeighborhoodAdj' !$CX2 * r-NeighborhoodAdj' !K$102) / \\ r-NeighborhoodAdj' !$CX$102)$$

que resulta em 0,245. Juntando tudo, você obtém  $1 - 0,245 = 0,755$  para a pontuação.

## Vamos Agrupar!

Agora você tem as pontuações que precisa. Tudo o que você deve saber fazer é configurar um modelo de otimização para encontrar as atribuições de comunidades ideais.

Agora, serei honesto com você. Encontrar comunidades ideais usando modularidade de gráfico é uma configuração de otimização mais intensa do que o que você encontrou no Capítulo 2. Esse problema é frequentemente resolvido com heurísticas complexas como o popular método “Louvain” (para mais informações veja <http://perso.uclouvain.be/vincent.blondel/research/louvain.html> — em inglês), mas essa é uma área sem código, então você vai se virar com o Solver.

Para tornar isso possível, você atacará o problema usando uma abordagem chamada *agrupamento divisional* ou *particionamento hierárquico*. Isso significa que você configurará o problema para encontrar a melhor forma de dividir o gráfico em duas comunidades. Então, dividirá essas duas em quatro, e assim por diante até o Solver decidir que a melhor forma de maximizar modularidade é parando de dividir as comunidades.

#### NOTA

O agrupamento divisional é o oposto de outra abordagem frequentemente usada chamada *agrupamento aglomerativo*. Em agrupamento aglomerativo, cada cliente começa em seu próprio grupo, e você recursivamente agrupa junto os dois grupos mais próximos até chegar no ponto de interrupção.

## Divisão↑Número↑1

Tudo certo. Então você começa esse processo de agrupamento divisional dividindo o gráfico em duas comunidades para que a pontuação de modularidade seja maximizada.

Primeiro, crie uma nova planilha chamada **Split1** e cole os clientes pela coluna A. Cada atribuição de comunidade de cliente irá para a coluna B, que você nomeará **Community**. Como você está dividindo o gráfico pela metade, faça com que a coluna **Community** seja uma variável de decisão binária no Solver, onde o valor 0/1 denotará se você está na comunidade

0 ou na comunidade 1. Nenhuma comunidade é melhor do que a outra. Não há vergonha em ser um 0.

### **Pontuando Atribuição de Comunidade de Cada Cliente**

Na coluna C, você calculará as pontuações que recebe ao colocar cada cliente em sua respectiva comunidade. Com isso, eu quero dizer que se colocar Adams na comunidade 1, você calculará o pedaço dele do placar total de modularidade somando todos os valores da linha dele na aba Scores cujas colunas de clientes também foram para a comunidade 1.

Considere como você adicionaria essas pontuações em uma fórmula. Se Adams está na comunidade 1, precisa somar todos os valores da aba Scores na linha 2 onde o cliente correspondente no modelo de otimização também foi atribuído a 1. Como os valores de atribuição são 0/1, pode usar SUMPRODUCT para multiplicar o vetor de comunidade pelo vetor de placar e então somar o resultado.

Embora os valores de pontuação atravessem a aba Scores, no modelo de otimização as atribuições vão de cima a baixo, então você precisa utilizar TRANSPOSE nos valores de pontuação para fazer isso funcionar (e usar TRANSPOSE significa tornar isso uma fórmula de array):

```
{=SUMPRODUCT(B$2:B$101, TRANSPOSE(Scores!B2:CW2))}
```

A fórmula simplesmente multiplica os valores Scores de Adams pelas atribuições de comunidade. Apenas pontuações que correspondam à atribuição de comunidade 1 ficam, enquanto as outras fixam em 0. SUMPRODUCT apenas soma tudo.

Mas e se Adams fosse atribuído à comunidade 0? Você precisa apenas trocar as atribuições de comunidade subtraindo 1 delas para fazer a soma das pontuações funcionar.

```
{=SUMPRODUCT(1 - (B$2:B$101), TRANSPOSE(Scores!B2:CW2))}
```

Em um mundo ideal, você poderia juntar esses dois com uma fórmula IF que verifica a atribuição de comunidade de Adams e então usa uma dessas duas fórmulas para somar as pontuações corretas de vizinhos. Mas para usar uma fórmula IF, é preciso usar o solver não linear (veja o

Capítulo 4 para detalhes), e, nesse caso em particular, maximizar modularidade é muito difícil para o solver não linear lidar com eficiência. Você precisa tornar o problema linear.

### ***Tornando o Cálculo de Pontuação um Modelo Linear***

Se você leu o Capítulo 4, lembra de um método para modelar a fórmula  $I_F$  usando restrições lineares, chamado de restrição “Big M”. Você usará essa ferramenta aqui.

Ambas as fórmulas anteriores são lineares; então e se você apenas fixasse uma variável de placar para Adams para ser menor do que ambas? Você está tentando maximizar as pontuações totais de modularidade para que a pontuação de Adams suba até bater na menor dessas duas fórmulas de restrição.

Mas como saber qual cálculo de pontuação correspondente à atribuição de comunidade atual é o menor? Você não sabe.

Para corrigir isso, é necessário **desativar** quaisquer fórmulas que não estejam participando. Se Adams está atribuído a 1, a primeira fórmula torna-se um limite superior e a segunda fórmula é **desligada**. Se Adams é um zero, você tem o oposto.

Como desligar um dos dois limites superiores? Adicione um “Big M” nele — grande o suficiente apenas para que seu limite seja insignificante, porque o limite legítimo é menor.

Considere esta modificação na primeira fórmula:

```
{=SUMPRODUCT(B$2:B$101, TRANSPOSE(Scores!B2:CW2)) +  
(1-B2)*SUM(ABS(Scores!B2:CW2))}
```

Se Adams está atribuído à comunidade 1, a adição que você fez no final da fórmula vira 0 (porque está multiplicando por 1-B2). Dessa forma, a fórmula se torna idêntica à primeira que você examinou. Mas se Adams for atribuído à comunidade 0, essa fórmula não se aplica mais e precisa ser desligada. Então o pedaço da fórmula  $(1-B2)*SUM(ABS(Scores!B2:CW2))$  adiciona um multiplicado pela soma de todos os valores absolutos das

pontuações que Adams poderia obter, que garante que a fórmula é maior do que sua versão invertida que está em jogo agora:

```
{=SUMPRODUCT(1-(B$2:B$101),TRANSPOSE(Scores!B2:CW2))+  
B2*SUM(ABS(Scores!B2:CW2))}
```

Tudo o que você está fazendo é ajustando a pontuação de Adams para ser menor ou igual ao cálculo correto e removendo a outra fórmula de consideração tornando-a maior. É uma declaração IF hackeada estilo gueto.

Assim, na coluna C, você pode criar uma coluna de pontuação que será uma variável de decisão, enquanto nas colunas D e E na planilha você pode colocar essas duas fórmulas como limites superiores da pontuação (veja a Figura 5-27).

The screenshot shows a Microsoft Excel spreadsheet titled "WineNetwork.xlsx". The formula bar at the top contains the following array formula:

```
{=SUMPRODUCT(B$2:B$101,TRANSPOSE(Scores!B2:CW2))+  
(1-B2)*SUM(ABS(Scores!B2:CW2))}
```

Below the formula bar is a table with columns A through F. Column A lists names: Adams, Allen, Anderson, Bailey, Baker, and Parker. Column B is labeled "Community". Column C is labeled "Score". Columns D and E are labeled "UB1" and "UB2" respectively. The data for Adams is as follows:

	Community	Score	UB1	UB2
2 Adams		20.1351981	-1.249E-15	
3 Allen		11.6083916	-8.604E-16	
4 Anderson		21.6060606	3.1919E-16	
5 Bailey		23.8554779	-1.11E-16	
6 Baker		13.2494172	6.8001E-16	
7 Parker		11.4545455	5.551E-16	

The formula in the formula bar is highlighted with a yellow box, and an arrow points from the formula to the formula bar. The status bar at the bottom of the Excel window shows "Normal View" and "Ready".

**Figura 5-27:** Adicionando dois limites superiores à variável de pontuação de clientes

Repare que, na fórmula, as referências absolutas são usadas na série de atribuição de comunidade, então quando você as arrasta fórmula abaixo, nada muda.

Somando as pontuações na célula G2 para cada eventual atribuição de comunidade na coluna C, obtém-se a pontuação total, que pode ser

normalizada pela conta total em 'r-NeighborhoodAdj' !CX102 para conseguir o cálculo de modularidade:

=SUM(C2:C101) / 'r-NeighborhoodAdj' !CX102

Isso gera a planilha exibida na Figura 5-28.

	A	B	C	D	E	F	G
1		Community	Score	UB1	UB2		Total Score
2	Adams			20.1351981	-1.249E-15		0.000
3	Allen			11.6083916	-8.604E-16		
4	Anderson			21.6060606	3.1919E-16		
5	Bailey			23.8554779	-1.11E-16		
6	Baker			13.2494172	6.8001E-16		

Figura 5-28: Abaixo Split1 preenchida, pronta para otimização

### Configurando o Programa Linear

Agora tudo está pronto para a otimização. Abra a janela do Solver e especifique que você está maximizando o gráfico modularidade de pontuação na célula G2. As variáveis de decisão são as atribuições de comunidade em B2:B101 e suas pontuações de modularidade estão em C2:C101.

É preciso adicionar uma restrição forçando as atribuições de modularidade em B2:B101 a serem binárias. Além disso, é necessário tornar as variáveis de pontuação de cliente na coluna C menores do que os limites superiores nas colunas D e E.

Como mostra a Figura 5-29, é possível fixar todas as variáveis como não negativas com a caixa de verificação e selecionar Simplex LP como o algoritmo de otimização.

Mas espere. Tem mais!

Um dos problemas com o uso da restrição “Big M” é que o Solver frequentemente tem problemas para confirmar se a solução ideal realmente foi encontrada. Então, ele apenas senta lá e gira suas roletas

mesmo tendo uma excelente solução no bolso. Para evitar que isso aconteça, pressione o botão Options no Solver e ajuste o valor de Max Subproblems em 15.000. Isso garante que o Solver finalize depois de aproximadamente 20 minutos no meu laptop.

Prossiga e pressione Solve — independente de você estar usando Solver ou OpenSolver (veja perto da barra de ferramentas), quando o algoritmo termina devido a um limite definido por usuário, ele pode lhe dizer que, embora tenha encontrado uma solução viável, ele não resolveu adequadamente. Isso apenas significa que o algoritmo não garante otimização (similar a como Solvers não lineares não conseguem garantir otimização), mas, nesse caso, sua solução deveria ser forte apesar disso.



**Figura 5-29:** A formulação LP para o primeiro split

Uma vez que você tenha uma solução, a aba Split1 deveria parecer como na Figura 5-30.

Se você está no Excel 2010 ou 2013 no Windows, esse problema é difícil demais para o Solver disponível, e você precisará usar OpenSolver, como abordado nos Capítulos 1 e 4.

Se usar o OpenSolver, configure o problema com o Solver normal, mas antes de solucionar, abra o plug-in do OpenSolver para reforçar seu sistema. O OpenSolver tem a mesma dificuldade com restrições “Big M”, então antes de executar o modelo, clique no botão Options do OpenSolver e ajuste o limite de tempo em 300 segundos. Se não fizer isso, o tempo de execução padrão no OpenSolver é realmente alto, e ele pode girar algumas roletas, forçando-o a fechar o Excel.

Se você está no Excel 2007 ou Excel 2011 para Mac, está pronto para usar o Solver, mas se quiser usar o OpenSolver com o Excel 2007, você pode. Se está no LibreOffice, você deve ficar bem.

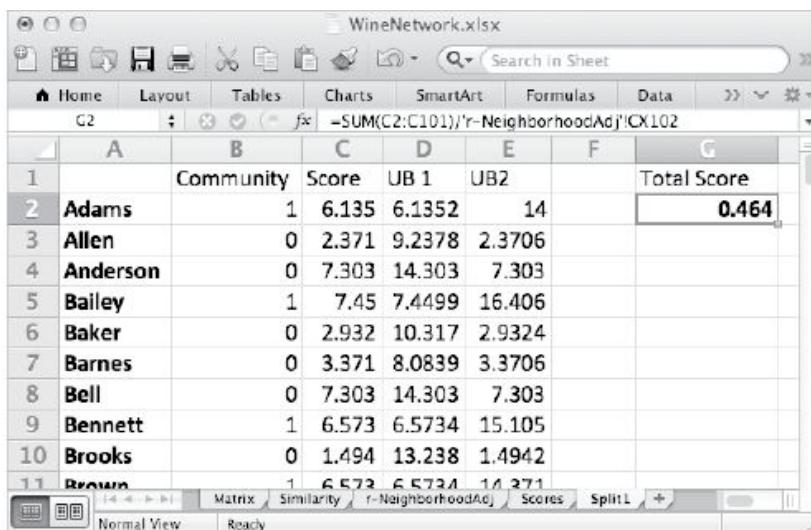


Figura 5-30: Solução ideal para a primeira divisão

Minha execução do Solver encontrou 0,464 para a modularidade; sua solução pode ser melhor se você usa o OpenSolver. Descendo pela coluna B, é possível ver quem ficou na comunidade 0 e quem está na comunidade 1. A questão então é, acabou? Existem apenas duas comunidades ou existe mais alguma?

Para responder essa questão, é necessário tentar dividir essas duas comunidades. Se você terminou, o Solver não terá nenhuma. Mas se fazer

três ou quatro comunidades a partir dessas duas melhora a modularidade, bom, então o Solver fará isso.

## Divisão<sup>2</sup>: Boogaloo<sup>Elétrico</sup>

Tudo certo. Divida essas comunidades como se estivesse fazendo divisão de células. Você começa fazendo uma cópia da aba Split1 e chamando-a de **Split2**.

A primeira coisa que deve-se fazer é inserir uma nova coluna após os valores de comunidade na coluna B. Nomeie essa nova coluna C como **Last Run** e copie os valores de B para C. Isso gera a planilha exibida na Figura 5-31.

Nesse modelo, as decisões são as mesmas — os clientes recebem 1 ou 0. Mas você precisa lembrar que se dois clientes recebem 1s, desta vez eles não necessariamente estão na mesma comunidade. Se um deles estava na comunidade 0 na primeira execução e o outro estava na comunidade 1, eles estão em duas comunidades diferentes.

Em outras palavras, as únicas pontuações que Adams pode obter por estar em, digamos, comunidade 1-0, são daqueles clientes que também estavam na comunidade 0 na primeira divisão e na comunidade 1 na segunda. Assim, é preciso modificar os limites superiores do cálculo de pontuação. O calculo de pontuação para a coluna E (aqui você exibe E2) então requer uma verificação contra a execução anterior na coluna C:

```
{=SUMPRODUCT(B$2:B$101, IF(C$2:C$101=C2, 1, 0), TRANSPOSE(Scores!B2:CW2))}
```

	A	B	C	D	E	F	G	H
1		Community	Last Run	Score	UB 1	UB2		Total Score
2	Adams		1	1	6.135	6.135	14	0.464
3	Allen		0	0	2.371	7.35	2.371	
4	Anderson		0	0	7.303	14.3	7.303	
5	Bailey		1	1	7.45	7.45	16.41	
6	Baker		0	0	2.932	8.529	2.932	
7	Barnes		0	0	3.371	8.084	3.371	
8	Bell		0	0	7.303	14.3	7.303	
9	Bennett		1	1	6.573	6.573	15.1	

**Figura 5-31:** A tabela Split2 com valores de execução anteriores

A declaração IF, IF(C\$2:C\$101=C2, 1, 0) evita que Adams receba pontos a não ser que seus vizinhos estejam com ele na primeira divisão.

Pode-se usar uma declaração IF aqui, porque a coluna C não é uma variável de decisão dessa vez. Aquela divisão foi fixada na última execução, então não há nada não linear sobre isso. Pode-se adicionar a mesma declaração IF em uma parte “Big M” da fórmula para fazer o cálculo final na coluna E:

```
=SUMPRODUCT(B$2:B$101, IF(C$2:C$101=C2, 1, 0), TRANSPOSE(Scores!B2:CW2))
) +
(1 -
B2)*SUMPRODUCT(IF(C$2:C$101=C2, 1, 0), TRANSPOSE(ABS(Scores!B2:CW2)))
```

Igualmente, pode-se adicionar as mesmas declarações IF no segundo limite superior na coluna F:

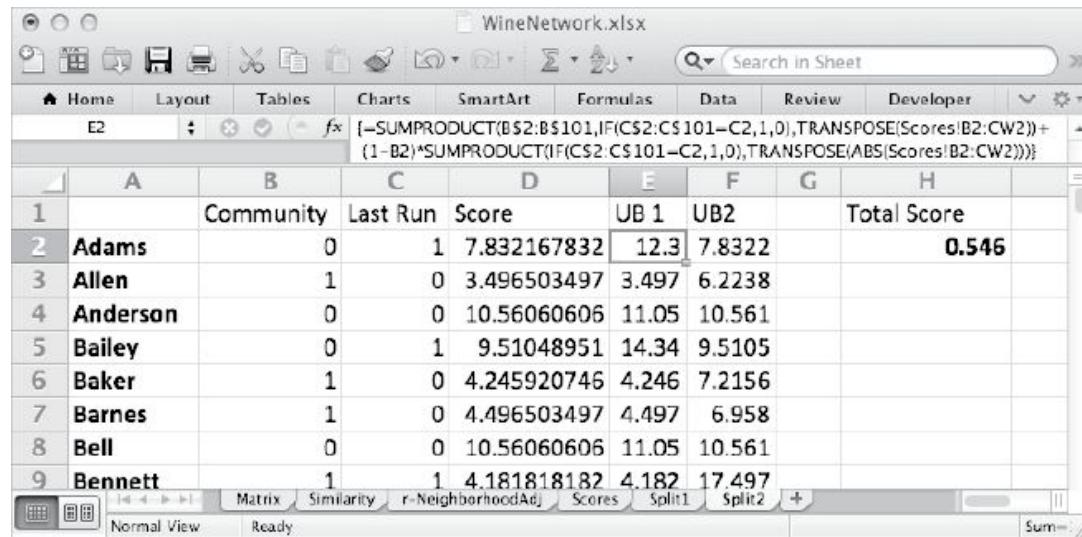
```
=SUMPRODUCT(1-
(B$2:B$101), IF(C$2:C$101=C2, 1, 0), TRANSPOSE(Scores!B2:CW2))
+B2*SUMPRODUCT(IF(C$2:C$101=C2, 1, 0), TRANSPOSE(ABS(Scores!B2:CW2)))
```

Tudo o que você fez foi armazenar o problema — aqueles que foram divididos para a comunidade 0 na primeira vez têm seu próprio mundo de pontuações para brincar e o mesmo vai para aqueles que terminaram em 1 da primeira vez.

E esta é a parte legal — não é preciso mudar a formulação do Solver em nada! Mesma formulação, mesmas opções! Se você está usando o

OpenSolver, ele pode não economizar suas opções de limite de tempo máximo da aba anterior. Redefina as opções para trezentos segundos. Solucione novamente.

Na minha execução em Split2, eu terminei com a modularidade final de 0,546 (veja a Figura 5-32), que é uma melhora substancial sobre 0,464. Isso significa que dividir foi uma boa ideia. (Sua solução pode terminar diferente e possivelmente melhor.)



**Figura 5-32:** A solução ideal para Split2

## E...↑Divisão↑3:↑Divisão↑com↑Vingança

Certo, então você deveria parar por aqui ou deveria continuar? A melhor opção é dividir novamente, e se o Solver não consegue fazer melhor do que 0,546, você acabou.

Comece criando uma aba Split3, renomeando Last Run para **Last Run 2**, e então inserindo uma nova Last Run na coluna C. Então, copie os valores da coluna B para C.

Adicione mais declarações IF ao limites superiores para verificar atribuições de comunidade na execução anterior. Por exemplo, F2 torna-se:

```
=SUMPRODUCT(B$2:B$101,
IF(D$2:D$101=D2,1,0),IF(C$2:C$101=C2,1,0),
```

```

TRANSPOSE (Scores!B2:CW2) ) +
(1-B2)*SUMPRODUCT (
IF(C$2:C$101=C2,1,0), IF(D$2:D$101=D2,1,0),
TRANSPOSE (ABS (Scores!B2:CW2)) )

```

Novamente, a formulação do Solver não altera. Redefina seu tempo máximo de solução se necessário, pressione Solve, e deixe o modelo avançar. No caso do meu modelo, eu não vi melhorias em modularidade (veja a Figura 5-33).

Dividir mais uma vez não acrescentou em nada, então isso significa que a modularidade foi efetivamente maximizada em Split2. Vamos fazer as atribuições de grupo a partir daquela aba e investigar.

## Codificando e Analisando as Comunidades

Para investigar essas atribuições de comunidade, a primeira coisa a ser feita é pegar essa **árvore binária** que foi criada pelas divisões sucessivas e transformar aquelas colunas em rótulos de grupos.

Crie uma aba chamada **Communities** e cole nomes de usuário, comunidade e valores de última execução da aba Split2. Você pode renomear as duas colunas binárias **Split2** e **Split1**. Para transformar seus valores binários em números individuais, o Excel proporciona uma elegante fórmula binária para decimal chamada **BIN2DEC**. Então, na coluna D, começando em D2, você pode adicionar:

```
=BIN2DEC (CONCATENATE (B2, C2))
```

WineNetwork.xlsx

F2:  $=\text{SUMPRODUCT}(\text{B2:B\$101}, \text{IF}(\text{D2:D\$101}=D2, 1, 0), \text{IF}(\text{C2:C\$101}=C2, 1, 0), \text{TRANSPOSE}(\text{Scores!B2:CW2}) + (1-B2)*\text{SUMPRODUCT}(\text{IF}(\text{C2:C\$101}=C2, 1, 0), \text{IF}(\text{D2:D\$101}=D2, 1, 0), \text{TRANSPOSE}(\text{ABS}(\text{Scores!B2:CW2})))}$

	A	B	C	D	E	F	G	H	I	J
1		Community	Last Run	Last Run 2	Score	UB 1	UB2		Total Score	
2	Adams	0	0	1	7.832	12.3	7.8322		0.536	
3	Allen	0	1	0	4.231	5.49	4.2308			
4	Anderson	0	0	0	10.56	11.05	10.561			
5	Bailey	0	0	1	9.51	14.34	9.5105			
6	Baker	1	1	0	4.143	4.143	7.3182			
7	Barnes	1	1	0	4.266	4.266	7.1888			
8	Bell	0	0	0	10.56	11.05	10.561			

Figura 5-33: Nenhuma melhora de modularidade em Split3

Ao copiar essa fórmula para baixo, obtém-se as atribuições de comunidade exibidas na Figura 5-34 (suas atribuições podem variar dependendo do Solver).

WineNetwork.xlsx

D2:  $=\text{BIN2DEC}(\text{CONCATENATE}(\text{B2}, \text{C2}))$

	A	B	C	D	Community
1		Split 2	Split 1		Community
2	Adams	0	0	0	0
3	Allen	0	1	1	1
4	Anderson	1	1	3	3
5	Bailey	0	0	0	0
6	Baker	0	1	1	1
7	Barnes	0	1	1	1
8	Bell	1	1	3	3

Figura 5-34: Rótulos finais de comunidade para maximização de modularidade

Você recebe quatro grupos com rótulos de 0 a 3 a partir da codificação decimal. Então quais são esses grupos ideais? Bom, você pode descobrir da mesma forma que analisou os grupos no Capítulo 2 — investigando as compras mais populares de seus membros.

Para começar, assim como no Capítulo 2, crie uma nova aba chamada **TopDealsByCluster** e cole as informações de ofertas das colunas de A a G na aba Matrix. Próximo a Matrix, coloque os rótulos de grupos 0, 1, 2

e 3 nas colunas H até K. Isso lhe dá a planilha apresentada na Figura 5-35.

	A	B	C	D	E	F	G	H	I	J	K
1	Offer #	Campaign	Varietal	Minimum Qt Discount (%)	Origin	Past Peak	0	1	2	3	
2	1	January	Malbec	72	56 France	FALSE					
3	2	January	Pinot Noir	72	17 France	FALSE					
4	3	February	Espumante	144	32 Oregon	TRUE					
5	4	February	Champagne	72	48 France	TRUE					
6	5	February	Cabernet Sauvign	144	44 New Zealand	TRUE					
7	6	March	Prosecco	144	96 Chile	FALSE					

**Figura 5-35:** A aba inicial TopDealsByCluster

Para o rótulo 0 na coluna H, você quer procurar todos os clientes na aba Communities que foram atribuídos à comunidade 0 e somar quantos deles aceitaram cada oferta. Assim como no Capítulo 2 e nas abas Split anteriores, use SUMPRODUCT com uma declaração IF para obter isto:

```
{=SUMPRODUCT(IF(Communities!$D$2:$D$101=TopDealsByCluster!H$1, 1, 0)
,
TRANSPOSE(Matrix!$H2:$DC2))}
```

Nesta fórmula você verifica quais clientes correspondem ao 0 no rótulo da coluna em H1, e, quando eles combinam, soma se eles aceitaram a primeira oferta ou não verificando H2:DC2 na aba Matrix. Repare que você usa TRANSPOSE para orientar tudo verticalmente. Isso significa que precisa fazer o cálculo em uma fórmula array.

Note que você usou referências absolutas nas atribuições de comunidade de clientes, as linhas de cabeçalho, e as colunas matrizes de compra. Isso possibilita que arraste a fórmula para a direita e para baixo, lhe entregando uma imagem cheia das compras populares para cada grupo (veja a Figura 5-36).

Assim como no Capítulo 2, é preciso aplicar um filtro na planilha e ordenar por contagem de oferta decrescente em comunidade 0 na coluna H. Isso lhe dá a Figura 5-37, a comunidade de cliente de pouca

quantidade (seus grupos podem variar em ordem e composição dependendo da solução com a qual o Solver terminou em cada passo).

	A	B	C	D	E	F	G	H	I	J	K
1	Offer #	Campaign	Varietal	Minimum	Discount	Origin	Past Peak	0	1	2	3
2	1	January	Malbec	72	56	France	FALSE	0	9	0	1
3	2	January	Pinot Noir	72	17	France	FALSE	0	4	0	6
4	3	February	Espumante	144	32	Oregon	TRUE	0	6	0	0
5	4	February	Champagne	72	48	France	TRUE	0	12	0	0
6	5	February	Cabernet Sauvi	144	44	New Zealand	TRUE	0	4	0	0
7	6	March	Prosecco	144	86	Chile	FALSE	0	11	1	0
8	7	March	Prosecco	6	40	Australia	TRUE	14	5	0	0
9	8	March	Espumante	6	45	South Africa	FALSE	8	4	8	0
10	9	April	Chardonnay	144	57	Chile	FALSE	0	10	0	0
11	10	April	Prosecco	72	52	California	FALSE	0	5	1	1

Figura 5-36: TopDealsByCluster com contagens de compras completas

	A	B	C	D	E	F	G	H	I	J	K
1	Offer #	Campaign	Varietal	Minimum	Discount	Origin	Past Peak	0	1	2	3
2	30	December	Malbec	6	54	France	FALSE	17	5	0	0
3	7	March	Prosecco	6	40	Australia	TRUE	14	5	0	0
4	29	November	Pinot Grigio	6	87	France	FALSE	13	1	3	0
5	18	July	Espumante	6	50	Oregon	FALSE	11	1	2	0
6	8	March	Espumante	6	45	South Africa	FALSE	8	4	8	0
7	13	May	Merlot	6	43	Chile	FALSE	5	0	1	0
8	11	May	Champagne	72	85	France	FALSE	1	12	0	0
9	12	May	Prosecco	72	83	Australia	FALSE	1	3	0	1
10	21	August	Champagne	12	50	California	FALSE	1	3	0	0
11	1	January	Malbec	72	56	France	FALSE	0	9	0	1
12	2	January	Prosecco	72	47	France	FALSE	0	4	0	0

Figura 5-37: Melhores ofertas para comunidade 0

Ordenando pela comunidade 1, você recebe o que parece ser o grupo de grande volume de champanhe francês (veja a Figura 5-38). Fascinante.

	A	B	C	D	E	F	G	H	I	J	K
1	Offer #	Campaign	Varietal	Minimum	Discount	Origin	Past Peak	0	1	2	3
2	22	August	Champagne	72	63	France	FALSE	0	21	0	0
3	31	December	Champagne	72	89	France	FALSE	0	17	0	0
4	11	May	Champagne	72	85	France	FALSE	1	12	0	0
5	4	February	Champagne	72	48	France	TRUE	0	12	0	0
6	6	March	Prosecco	144	86	Chile	FALSE	0	11	1	0
7	9	April	Chardonnay	144	57	Chile	FALSE	0	10	0	0
8	1	January	Malbec	72	56	France	FALSE	0	9	0	1
9	14	June	Merlot	72	64	Chile	FALSE	0	9	0	0
10	27	October	Champagne	72	88	New Zealand	FALSE	0	7	1	1
11	3	February	Espumante	144	32	Oregon	TRUE	0	6	0	0
12	15	June	Cabernet Sauvi	144	19	Italy	FALSE	0	6	0	0
13	20	August	Cabernet Sauvi	72	82	Italy	FALSE	0	6	0	0
14	25	October	Cabernet Sauvi	72	59	Oregon	TRUE	0	6	0	0

Figura 5-38: Estourando champanhe na comunidade 1

Em relação à comunidade 2, ela parece com a comunidade 0, exceto pelo principal impulsionador que é o Espumante de março (veja a Figura 5-39).

	B	C	D	E	F	G	H	I	J	K
1	Campaign	Varietal	Minimum	Discount	Origin	Past Peak	0	1	2	3
2	March	Espumante	6	45	South Africa	FALSE	8	4	8	0
3	November	Pinot Grigio	6	87	France	FALSE	13	1	3	0
4	July	Espumante	6	50	Oregon	FALSE	11	1	2	0
5	March	Prosecco	144	86	Chile	FALSE	0	11	1	0
6	October	Champagne	72	88	New Zealand	FALSE	0	7	1	1
7	April	Prosecco	72	52	California	FALSE	0	5	1	1
8	November	Cabernet Sauvi	12	56	France	TRUE	0	5	1	0
9	May	Merlot	6	43	Chile	FALSE	5	0	1	0
10	August	Champagne	72	63	France	FALSE	0	21	0	0

Figura 5-39: Pessoas que gostaram da oferta do Espumante de março

E a comunidade 3 são as pessoas do Pinot Noir. Vocês não ouviram falar de Cabernet Sauvignon, pessoal? Eu admito, tenho um péssimo paladar para vinho. Veja a Figura 5-40.

É isso aí! Você tem quatro grupos, e, honestamente, três deles fazem sentido completamente, embora eu acredite ser possível você ter um

grupo de pessoas que realmente gostam de Espumante em março. E você pode receber isso no seu trabalho — alguns agrupamentos atípicos indecifráveis.

	B	C	D	E	F	G	H	I	J	K
1	Campaign	Varietal	Minimum	Discount	Origin	Past Peak	0	1	2	3
2	October	Pinot Noir	144	83	Australia	FALSE	0	3	0	12
3	September	Pinot Noir	6	34	Italy	FALSE	0	0	0	12
4	July	Pinot Noir	12	47	Germany	FALSE	0	0	0	7
5	January	Pinot Noir	72	17	France	FALSE	0	4	0	6
6	October	Champagne	72	88	New Zealand	FALSE	0	7	1	1
7	April	Prosecco	72	52	California	FALSE	0	5	1	1
8	January	Malbec	72	56	France	FALSE	0	9	0	1
9	June	Merlot	72	88	California	FALSE	0	4	0	1
10	September	Chardonnay	144	39	South Africa	FALSE	0	4	0	1

Figura 5-40: Espiadas no Pinot

Contudo, observe o quanto similar essa solução é aos grupos encontrados no Capítulo 2. No Capítulo 2, você usou uma metodologia completamente diferente mantendo o vetor de oferta de cada cliente na mistura e usando isso para medir suas distâncias do centro do grupo. Aqui, não há conceito de centro e nem mesmo quais ofertas um cliente adquiriu foram ofuscadas. O que importa é a distância até outros clientes.

## Lá e de Volta Outra Vez: um Conto Gephi

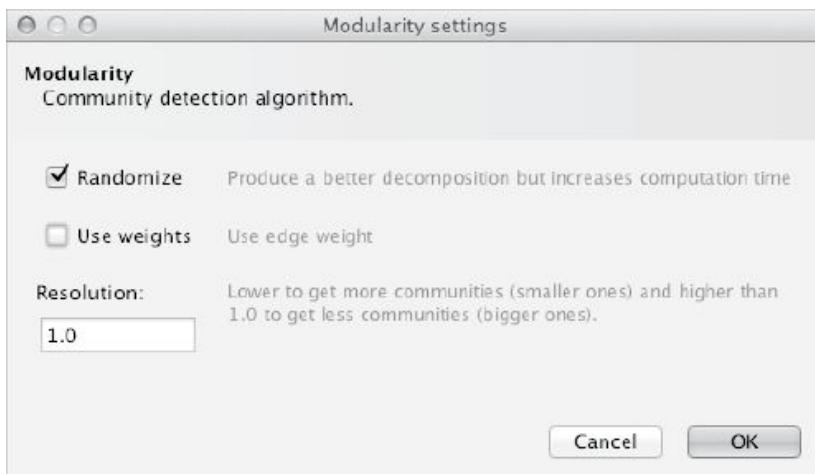
Agora que você passou por todo o processo de agrupamento, eu gostaria de mostrar o mesmo processo no Gephi. Na Figura 5-20, você examinou uma exportação de um gráfico r-Neighborhood para o Gephi, sobre o qual eu volto a falar nesta seção.

O próximo passo o deixará com inveja, mas aqui vai. No Excel você teve que solucionar para o gráfico de modularidade ideal usando agrupamento divisional. No Gephi, há um botão Modularity. Ele pode

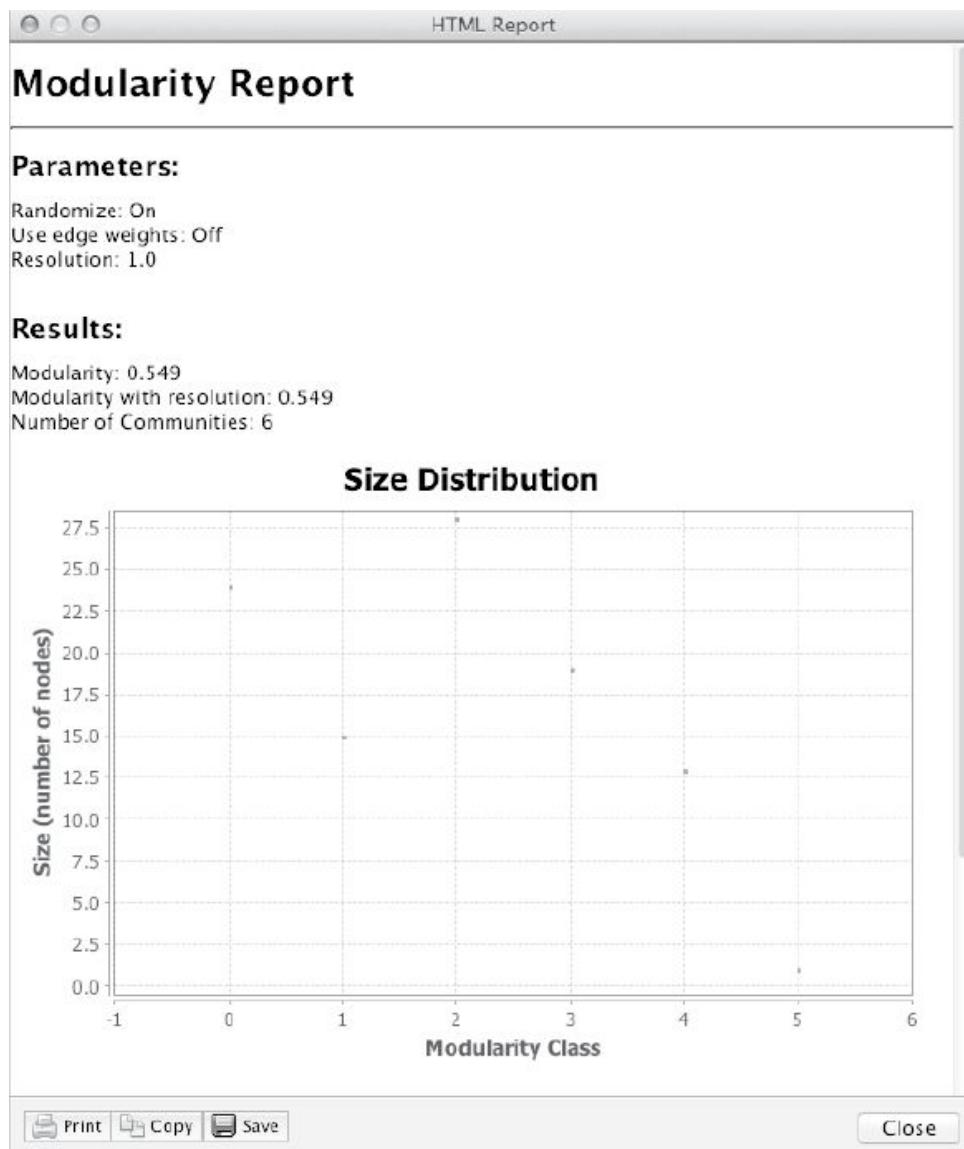
ser encontrado do lado direito da janela na seção Network Overview da aba Statistics.

Quando você pressiona o botão Modularity, uma janela de configuração abre. Não é preciso usar vértices ponderados já que você exportou a matriz de adjacência (veja a Figura 5-41 para a janela de configurações de modularidade do Gephi).

Pressione OK. A otimização de modularidade executará usando um algoritmo de aproximação que é muito rápido. Um relatório então é exibido com uma pontuação total de modularidade de 0,549 como o tamanho de cada grupo detectado (veja a Figura 5-42). Note que, se você executar isso no Gephi, a solução pode sair diferente já que o cálculo é aleatório.



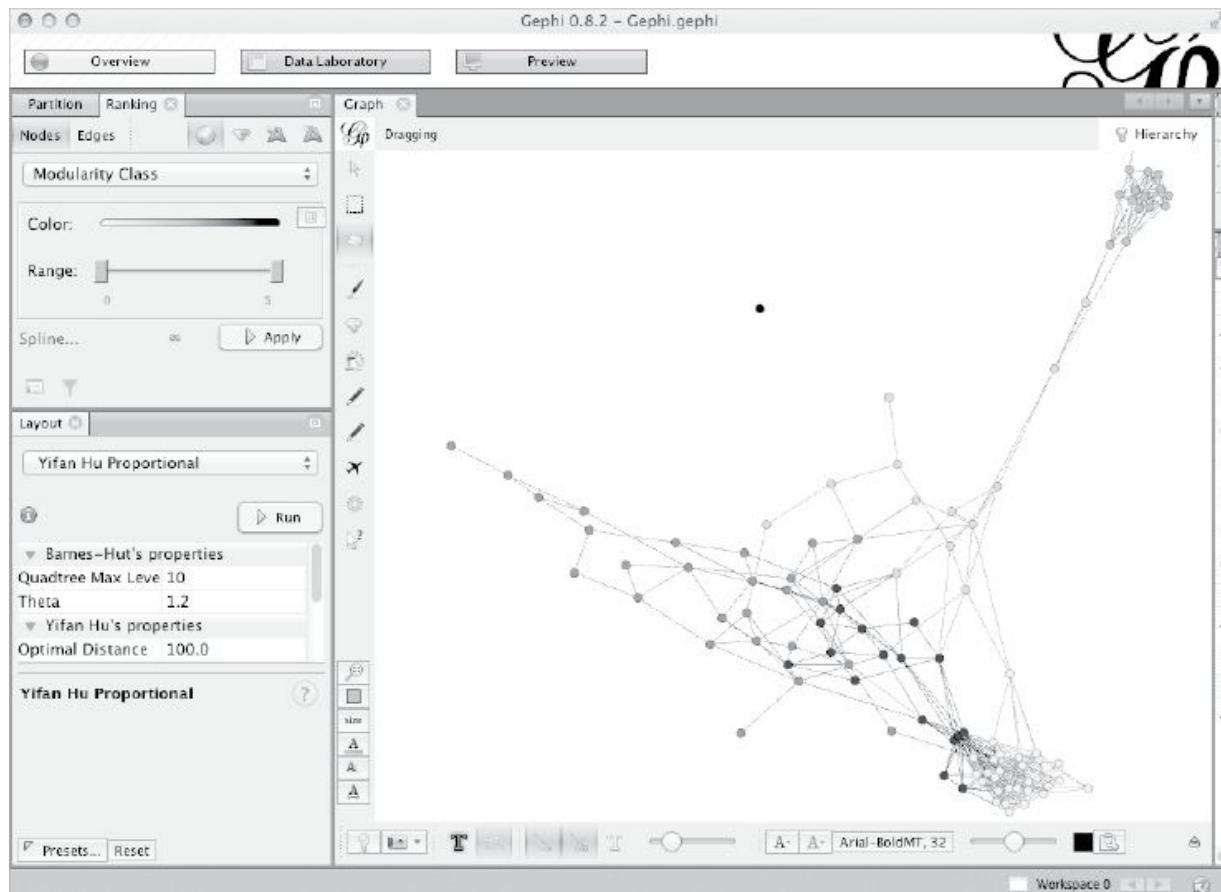
**Figura 5-41:** Configurações de modularidade do Gephi



**Figura 5-42:** Pontuação de modularidade a partir do Gephi

Uma vez que tenha os grupos a partir do Gephi, você pode fazer algumas coisas com eles.

Primeiro, é possível recolorir o gráfico usando a modularidade. Assim como você redimensionou a gráfico *Friends* usando grau de nó, pode navegar até a janela Ranking na parte superior esquerda da janela no Gephi e ir para a seção Nodes. A partir de lá, você pode selecionar Modularity Class no menu suspenso, escolher qualquer esquema de cor que quiser, e pressionar Apply para recolorir o gráfico (veja a Figura 5-43).



**Figura 5-43:** O gráfico customer recolorido para mostrar modularidade de grupos

Legal! Agora você consegue ver que as duas partes “esquema de tumor” do gráfico são de fato comunidades. A seção espalhada no meio do gráfico foi dividida em três grupos. E o pobre Parker foi colocado em seu próprio grupo, desconectado de qualquer pessoa. Que solitário e triste.

A segunda coisa que você pode fazer com a informação de modularidade é exportar de volta para o Excel e examiná-la, assim como fez com seus próprios grupos. Para realizar isso, vá para a aba Data Laboratory, que você visitou antes no Gephi. Você notará que as classes de modularidade já foram populadas como uma coluna na tabela de dados Nodes. Pressionando o botão Export Table, pode-se selecionar o rótulo e as colunas de classe de modularidade a serem jogadas em um arquivo CSV (veja a Figura 5-44).

Gephi 0.8.2 – Gephi.gephi

Nodes	Id	Label	Modularity Class
Adams	Adams	Adams	2
Brown	Brown	Brown	2
Carter	Carter	Carter	2
Cruz	Cruz	Cruz	2
Diaz	Diaz	Diaz	2
Hill	Hill	Hill	2
Hughes	Hughes	Hughes	2
James	James	James	2
King	King	King	2
Myers	Myers	Myers	2
Perez	Perez	Perez	2
Perry	Perry	Perry	2
Stewart	Stewart	Stewart	2
Taylor	Taylor	Taylor	2
Walker	Walker	Walker	2
Allen	Allen	Allen	0
Edwards	Edwards	Edwards	3
Evans	Evans	Evans	3
Gonzalez	Gonzalez	Gonzalez	0
Lopez	Lopez	Lopez	0
Ramirez	Ramirez	Ramirez	0
Reaves	Reaves	Reaves	0

Figura 5-44: Exportando classes de modularidade de volta ao Excel

Pressione OK na janela export para exportar as suas classes de modularidade para um CSV onde quiser e então abrir aquele arquivo no Excel. De lá, pode-se criar uma aba na pasta de trabalho principal chamada **CommunitiesGephi**, onde é possível colar as classes que o Gephi encontrou para você (veja Figura 5-45). Será preciso usar o recurso filter no Excel para ordenar os clientes por nome como eles estão no restante da pasta de trabalho.

Apenas por diversão, vamos confirmar que esse agrupamento ganha a pontuação original na coluna C. Você não está mais ligado por restrições de modelagem linear, então consegue totalizar as pontuações de modularidade de cliente usando a seguinte fórmula (exibida aqui usando nosso cliente favorito, Adams, na célula C2):

$$\{=\text{SUMPRODUCT}(\text{IF}(\$B\$2:\$B\$101=B2, 1, 0), \text{TRANSPOSE}(Scores!B2:CW2))\}$$

A fórmula somente procura por clientes no mesmo grupo usando uma declaração IF, atribui 1s a esses clientes e 0s a todos o resto, e então usa SUMPRODUCT para somar suas pontuações de modularidade.

Pode-se clicar duas vezes na fórmula para enviá-la para baixo pela coluna C. Somando a coluna na célula E2 e dividindo pela contagem total de stub de 'r-NeighborhoosAdj' !CX102, obtém-se uma pontuação de modularidade de 0,549 (veja a Figura 5-46). Então a heurística do Gephi ultrapassou a heurística de agrupamento divisional em 0,003. Ah, bem! Muito perto. (Se você usou o OpenSolver, pode realmente ganhar o Gephi.)

	A	B	C
1	Id	Modularity Class	
2	Adams	0	
3	Allen	4	
4	Anderson	3	
5	Bailey	0	
6	Baker	4	
7	Barnes	4	
8	Bell	3	
9	Bennett	0	
10	Brooks	1	
11	Brown	0	
12	Butler	2	
13	Campbell	3	

Figura 5-45: Classes de modularidade do Gephi de volta no Excel

	A	B	C	D	E	F
1	Id	Modularity Class	Score		Modularity	
2	Adams	0	7.24475524		0.549	
3	Allen	4	3.24475524			
4	Anderson	3	10.5909091			
5	Bailey	0	8.7972028			
6	Baker	4	4.11888112			
7	Barnes	4	5.24475524			
8	Bell	3	10.5909091			

**Figura 5-46:** Reproduzindo a pontuação de modularidade para as comunidades detectadas pelo Gephi

Vejamos quais grupos o Gephi realmente criou. Para começar, vamos fazer uma cópia da aba **TopDealsByCluster**, que você deve renomear para **TopDealsByClusterGephi**. Uma vez feita a cópia, ordene as ofertas de volta na coluna A e desista do filtro colocado na sua tabela. Agora, no agrupamento do Gephi, você tem seis grupos com rótulos de 0 a 5 (seus resultados podem ser diferentes já que o Gephi utiliza um algoritmo aleatório), então vamos adicionar 4 e 5 à mistura nas colunas L e M.

A fórmula na célula H2 precisa apenas ser modificada para referenciar a coluna B na aba **CommunitiesGephi** em vez da coluna D na aba **Communities**. Você pode arrastar essa fórmula pelo restante da planilha, produzindo a Figura 5-47.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Offer #	Campaign	Varietal	Minimum	Discount	Origin	Past Peak	0	1	2	3	4	5
2	1	January	Malbec	72	56	France	FALSE	0	0	5	0	5	0
3	2	January	Pinot Noir	72	17	France	FALSE	0	0	5	5	0	0
4	3	February	Espumante	144	32	Oregon	TRUE	0	5	0	0	1	0
5	4	February	Champagne	72	48	France	TRUE	0	6	2	0	4	0
6	5	February	Cabernet Sauvignon	144	44	New Zealand	TRUE	0	0	0	0	4	0
7	6	March	Prosecco	144	86	Chile	FALSE	0	7	0	0	5	0
8	7	March	Prosecco	6	40	Australia	TRUE	14	3	0	0	2	0
9	8	March	Espumante	6	45	South Africa	FALSE	10	10	0	0	0	0
10	9	April	Chardonnay	144	57	Chile	FALSE	0	0	1	0	0	0

**Figura 5-47:** Melhores compras por grupo a partir do Gephi

Se você ordenar novamente por coluna, verá todos os grupos familiares — pouca quantidade, vinho espumante, Francófilos, Pinot, grande quantidade, e, em último lugar, mas não menos importante, o próprio Parker.

## Resumindo

No Capítulo 2, você viu agrupamento k-means. Usando os mesmos dados neste capítulo, você abordou gráficos de rede e agrupamento via maximização de modularidade. Em mais detalhes, estes foram alguns itens sobre os quais aprendeu:

- Como gráficos de rede são visualmente representados e como eles são representados numericamente usando matrizes de adjacência e afinidade.
- Como carregar um gráfico de rede no Gephi para aumentar as deficiências de visualização do Excel.
- Como reduzir arestas de gráficos de rede com gráfico de vizinhança r. Você também aprendeu o conceito de um gráfico kNN, com o qual eu recomendei que volte e brinque.
- As definições de grau de nó e modularidade de gráfico e como calcular pontuações de modularidade para agrupar dois nós juntos.
- Como maximizar modularidade de gráfico usando um modelo de otimização linear e agrupamento divisional.
- Como maximizar modularidade de gráfico no Gephi e exportar os resultados.

Agora, você pode estar pensando, “John, por que você me fez passar por todo aquele processo de maximização de modularidade de gráfico quando o Gephi pode fazer isso por mim?”

Lembre-se, o propósito deste livro não é pressionar botões cegamente, sem entender o que eles fazem. Agora você sabe como construir e preparar um gráfico de dados para detecção de grupo. E você sabe como a detecção de comunidade em gráfico de dados funciona. Você fez isso. Então da próxima vez que o fizer, mesmo se estiver apenas pressionando um botão, saberá o que está acontecendo nos bastidores, e esse nível de entendimento e confiança no processo é inestimável.

Embora o Gephi seja um dos melhores lugares para fazer essa análise, se você está procurando um lugar para codificar com gráfico de dados, a biblioteca igraph, que tem ganchos em R e Python, é excelente para trabalhar com gráficos de rede.

Também vale a pena mencionar os bancos de dados gráficos Neo4J e Titan. Esses bancos de dados são feitos para armazenar dados de gráficos para consultas posteriores, seja essa consulta simples como “encontre os filmes favoritos dos amigos do John” ou complexa como “encontre o menor caminho no Facebook entre John e Kevin Bacon”.

Então é isso. Siga em frente, gere o gráfico, e encontre comunidades!

## 6

# O↑Avô↑da↑Inteligência↑Artificial Supervisionada↑—↑Regressão

## Espere,↑o↑Quê?↑Você↑Está↑Grávida?

**E**m um recente artigo da revista *Forbes*, foi noticiado que a Target havia criado um modelo de inteligência artificial (IA) que poderia prever quando uma cliente estaria grávida e usar essa informação para direcioná-la a ofertas e marketing direcionado relacionado à gravidez. Novos pais gastam muito dinheiro em produtos para criação dos filhos, e que melhor momento para torná-los clientes fiéis senão antes de o bebê aparecer? Eles comprarão fraldas da marca da loja durante anos!

Essa história sobre a Target é apenas uma das muitas que apareceram na mídia recentemente. Watson ganhou *Jeopardy!*. O Netflix ofereceu um prêmio de um milhão de dólares para aprimorar seu sistema de recomendações. A campanha de reeleição de Obama usou inteligência artificial para ajudar no direcionamento, on-line e na mídia ao vivo, e em operações de arrecadação de fundos. E há o Kaggle.com, no qual surgem competições para prever de tudo, desde um motorista ficando com sono a quanto um comprador gastará em compras de mantimentos.

Mas essas são apenas as aplicações com manchetes chamativas. A IA é útil em qualquer indústria na qual possa pensar. Sua empresa de cartão de crédito usa a IA para identificar transações estranhas na sua conta. O inimigo no seu jogo de tiro do Xbox roda em IA. Há filtro de e-mail spam, detecção de fraude fiscal, autocorreção ortográfica e recomendação de amigos em redes sociais.

Com simplicidade, um bom modelo IA pode ajudar um negócio a tomar melhores decisões, comercializar melhor, aumentar a renda e reduzir

custos. A IA pode ajudar a prever quais ofertas trarão um cliente de volta à sua loja. A IA pode identificar candidatos que mentem em seus perfis de namoro on-line ou farão uma cirurgia cardíaca no ano que vem. Você escolhe e, se houver um bom histórico de dados, um modelo IA treinado pode ajudar.

## Não se Engane

Pessoas que não sabem como modelos IA funcionam, frequentemente enfrentam alguma combinação de medo e respeito quando escutam falar em como esses modelos podem prever o futuro. Mas para parafrapear o grande filme de 1992 *Quebra de Sigilo*, “Não se engane. Não é tão [inteligente].”

Por quê? Porque modelos IA não são mais espertos do que a soma de suas partes. Em um nível simplista, você alimenta um algoritmo IA **supervisionado** com algum histórico de dados, como compras na Target, por exemplo, e diz ao algoritmo, “Ei, essas compras foram de mulheres grávidas, e essas outras compras foram de mulheres não-tão-grávidas”. O algoritmo mastiga esses dados e então apresenta um modelo. No futuro, você fornece o modelo às compras de uma cliente e pergunta, “Esta pessoa está grávida?” e o modelo responde, “Não, esse é um cara de 26 anos que mora no porão de sua mãe.”

Isso é extremamente útil, mas o modelo não é um mágico. Ele apenas espertamente transforma dados passados em uma fórmula ou estabelece regras que usa para prever um caso futuro. Como vimos no caso de naïve Bayes no Capítulo 3, é a habilidade de um modelo IA recordar esses dados e associar regras de decisão, probabilidades, ou coeficientes que o deixa tão eficaz.

Nós fazemos isso o tempo todo em nossas próprias vidas não inteligentes artificialmente. Por exemplo, usando dados históricos pessoais, meu cérebro sabe que quando eu como um sanduíche com brotos de alfafa amarronzados, há uma boa chance de eu passar mal em algumas horas. Eu usei dados passados (eu passei mal) e *treinei* meu

cérebro com ele, então agora eu tenho uma regra, fórmula, modelo, o que você quiser chamar: brotos amarronzados = pesadelo gastrointestinal.

Neste capítulo, implementaremos dois *modelos de regressão* diferentes apenas para ver o quanto direta IA pode ser. Regressão é o avô de modelagem de previsão supervisionada com pesquisa sendo feita nele tão rápida quanto a virada do século 19. É velho, mas sua linhagem contribui para seu poder — regressão teve tempo de construir todo tipo de rigor ao seu redor de formas que algumas técnicas IA não fizeram. Em contraste com a sensação MacGyver de naïve Bayes no Capítulo 3, você sente o peso do rigor estatístico de regressão neste capítulo, particularmente quando investigamos teste de significância.

Da mesma forma como usamos o modelo naïve Bayes no Capítulo 3, usaremos esses modelos para classificação. No entanto, como verá, o problema em mãos é bem diferente daquele problema de classificação de documento saco de palavras que encontramos previamente.

## Prevendo↑Clientes↑Grávidas↑na RetailMart↑Usando↑Regressão↑Linear

### NOTA

A pasta de trabalho do Excel usada neste capítulo, “RetailMart.xlsx”, está disponível para download na página da editora, em [www.altabooks.com.br](http://www.altabooks.com.br), procurando pelo título do livro. Essa pasta de trabalho inclui todos os dados iniciais se você quiser trabalhar a partir dela. Ou você pode apenas acompanhar usando as planilhas

Suponha que você seja um gerente de marketing na sede da empresa RetailMart responsável por mercadorias infantis. Seu trabalho é ajudar a vender mais fraldas, kits, pijamas, berços, carrinhos, chupetas, etc., para novos pais, mas você tem um problema.

Você sabe de grupos especializados que novos pais adquirem hábitos com produtos para bebês. Eles encontram marcas de fraldas que gostam bem cedo e lojas que têm os melhores preços dessas marcas. Eles encontram a chupeta que funciona com seu bebê e sabem onde ir para comprar pacotes baratos com duas. Você quer que RetailMart seja a primeira loja onde esses novos pais comprem fraldas. Você quer maximizar as chances de RetailMart ser o lugar no qual os pais vão para fazer compras para seus bebês.

Mas para fazer isso, você precisa comercializar para esses pais antes de comprarem o primeiro pacote de fraldas em outro lugar e do bebê aparecer. Dessa forma, quando o bebê chegar, os pais já decidiram e possivelmente já usaram aquele cupom que receberam por e-mail para fraldas e pomadas.

Muito simples: você precisa de um modelo preditivo para ajudar a identificar potenciais clientes grávidas para marketing direcionado.

## O↑Conjunto↑de↑Características

Você tem uma arma secreta à sua disposição para construir esse modelo: dados de contas de clientes. Você não possui esses dados para todo cliente; não, você terá problemas com o cara que vive no meio do nada e paga apenas em dinheiro. Mas para aqueles que usam um cartão de crédito da loja ou possuem uma conta on-line vinculada ao principal cartão de crédito, pode-se associar compras não necessariamente a um indivíduo mas, pelo menos, a uma família.

No entanto, você não pode apenas alimentar um histórico completo de compras, desestruturado, para um modelo IA e esperar que as coisas aconteçam. É preciso ser esperto em tirar prognosticadores relevantes do conjunto de dados. Então a pergunta que você deveria se fazer é quais compras passadas são preditivas a favor ou contra uma pessoa da família estar grávida?

A primeira compra que vem em mente é um teste de gravidez. Se uma cliente compra um teste de gravidez, é mais provável que ela esteja

grávida do que uma outra cliente. Esses prognosticadores são geralmente chamados de modelo de **recursos, características** ou **variáveis independentes**, enquanto o que estamos tentando prever “Grávida (sim/não)?”, seria a **variável dependente** no sentido que seu valor é dependente dos dados variáveis independentes que estamos colocando no modelo.

Pause por um momento e anote seus pensamentos em possíveis características para um modelo IA. Qual histórico de compras RetailMart deveria considerar?

Eis uma lista de exemplos de características que poderiam ser geradas a partir de registros de compras de uma cliente e informações de conta associada:

- Titular da conta é Masculino/Feminino/Desconhecido combinando sobrenome com dados de censo.
- Endereço do titular da conta é uma casa, apartamento ou caixa postal.
- Recentemente comprou um teste de gravidez.
- Recentemente comprou anticoncepcionais.
- Recentemente comprou produtos de higiene feminina.
- Recentemente comprou suplementos de ácido fólico.
- Recentemente comprou vitaminas pré-natais.
- Recentemente comprou DVD de ioga pré-natal.
- Recentemente comprou body pillow (travesseiro de corpo).
- Recentemente comprou ginger ale (refrigerante de gengibre).
- Recentemente comprou Sea-Bands (uma pulseira elástica que evita enjoos fazendo pressão em alguns pontos corporais, segundo a acupuntura)
- Comprou cigarros regularmente até recentemente, depois parou.
- Recentemente comprou cigarros.

- Recentemente comprou produtos para parar de fumar (chiclete, adesivos, etc.).
- Recentemente comprou vinho.
- Recentemente comprou roupas para maternidade.

Nenhum desses prognosticadores são perfeitos. Clientes não compram tudo na RetailMart; um cliente pode escolher comprar o teste de gravidez em uma farmácia local em vez de na RetailMart ou os suplementos pré-natais podem ser prescritos. Mesmo se o cliente comprou tudo na RetailMart, **famílias** grávidas ainda podem ter um fumante ou um consumidor de bebida alcoólica. Roupas de maternidade são frequentemente usadas por pessoas não grávidas, especialmente quando a cintura Empire está na moda — graças a Deus não existe RetailMart em romances da Jane Austen. O gengibre ajuda na náusea, mas também é ótimo com uísque. Você entendeu.

Nenhum desses prognosticadores farão cortes, mas a esperança é que, com seus poderes combinados no estilo Capitão Planeta, o modelo conseguirá classificar os clientes razoavelmente bem.

## Reunindo↑os↑Dados↑em↑Treinamento

Pesquisas conduzidas pela RetailMar apontaram que 6% das famílias de clientes estão grávidas em qualquer período. É preciso pegar alguns exemplos desses grupos do banco de dados da RetailMart e reunir suas características de modelagem em simples históricos de compras antes de elas darem à luz. Da mesma forma, é preciso reunir essas características para uma amostra de clientes que não estão grávidas.

Uma vez que reúna essas características para um monte de famílias grávidas e não-grávidas, pode-se usar esses exemplos conhecidos para treinar um modelo IA.

Mas como você poderia identificar famílias grávidas anteriores nos dados? Entrevistar clientes para construir um conjunto em treinamento sempre é uma opção. Você está apenas construindo um protótipo, então

talvez aproximar famílias que acabaram de ter um bebê vendo hábitos de compras é bom o bastante. Para clientes que subitamente começaram a comprar fraldas para recém-nascidos e continuaram a comprar fraldas aumentando os tamanhos por, pelo menos, um ano, pode-se presumir, razoavelmente, que a família do cliente tem um novo bebê.

Então olhando o histórico de compras do cliente antes do evento de compra de fralda, você reúne as características listadas para uma família grávida. Imagine que você tire 500 exemplos de famílias grávidas e reúna seus dados relativos do banco de dados da RetailMart.

E para clientes não-grávidas, você pode reunir históricos de compras de uma seleção aleatória de clientes no banco de dados da RetailMart que não atendam ao critério de “compra de fralda contínua”. Claro, uma ou duas pessoas grávidas podem cair na categoria não grávida, mas como famílias grávidas apenas constituem uma pequena porcentagem da população RetailMart (e isso é antes de remover compradores de fraldas), essa amostra aleatória deve ser clara o suficiente. Imagine que você pegue outros 500 exemplos dessas clientes não-grávidas.

Se você colocasse as 1.000 linhas (500 grávidas, 500 não) em uma planilha ela pareceria como a Figura 6-1.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	PREGNANT
1	Home/ Implied Gender	Apt/ PO Box	Pregnancy Test	Birth Control	Feminine Hygiene	Folic Acid	Prenatal Vitamins	Prenatal Yoga	Body Pillow	Ginger Ale	Sea Bands	Stopped buying cigglies	Cigar ettes	Smoking cessation	Wine	Wine	Maternity Clothes			
2	M	A	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	
3	M	H	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	
4	M	H	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	
5	U	H	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	
6	F	A	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	1	
7	F	H	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	
8	M	H	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	1	
9	F	H	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	
10	F	H	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	
11	F	H	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	

**Figura 6-1:** Dado bruto em treinamento

## RESOLVENDO O DESEQUILÍBRIO DE CLASSE

Agora, você sabe que apenas 6% da nossa população de clientes está grávida em

qualquer período de tempo, mas o conjunto em treinamento que você montou é 50/50. Isso é chamado de **superamostragem**. Gravidez seria a “minoria” ou classe rara nos dados, e, equilibrando a amostra, o classificador que você treinará não ficará sobre carregado com clientes não-grávidas. Afinal, se você deixou a amostra em uma divisão natural 6/94, então apenas nomeando todas como não-grávidas leva a um índice de precisão de 94%. Isso é perigoso já que gravidez, enquanto em minoria, é, de fato, a classe para a qual você se importa em comercializar.

Esse reequilíbrio dos dados em treinamento apresentará uma polarização ao modelo — ele pensará que gravidez é mais comum do que realmente é. Mas está tudo bem, porque não é necessário obter as probabilidades reais de estar grávida do modelo. Como verá neste capítulo, você precisa apenas encontrar o ponto ideal para pontuações de gravidez saindo do modelo que equilibre **verdadeiros positivos** e **falsos positivos**.

Nas primeiras duas colunas do conjunto de dados em treinamento, você tem dados categóricos para gênero e tipo de endereço. O restante das características são binários onde um 1 significa TRUE. Então, por exemplo, se olhar para a primeira linha na coluna, é possível ver que essa cliente foi confirmada grávida (coluna S). Essa é a coluna que você treinará o modelo para prever. E se olhar o histórico de compras dessa cliente, verá que ela comprou um teste de gravidez e algumas vitaminas pré-natais. Além disso, ela **não** comprou nem cigarros nem vinho recentemente.

Ao passear pelos dados, verá todos os tipos de clientes, alguns com muitos indicadores e outros com poucos. Como esperado, a família de grávidas de vez em quando comprará cigarros e vinho, enquanto que a família das não-grávidas comprará produtos associados à gravidez.

## Criando↑Variáveis↑Dummy

Você pode pensar em um modelo IA como nada além do que uma fórmula que admite números, os mastiga um pouco e cospe uma previsão que deve parecer com os 1s (grávida) e 0s (não) na coluna S da planilha.

Mas o problema com esses dados é que as primeiras duas colunas não são números, são? Elas são letras representando **categorias**, como feminino e masculino.

Esse problema, controlar **dados categóricos**, isto é, dados que são agrupados por um número infinito de rótulos sem equivalentes numéricos inerentes, constantemente é uma pedra no sapato de mineradores de dados. Se você envia um questionário para seus clientes e eles têm que reportar no que eles trabalham, estado civil, o país onde moram, a raça do cachorro deles e até mesmo o episódio favorito de Gilmore Girls, você acabaria preso com um monte de dados categóricos.

Isso contrasta com **dados quantitativos**, que já são numéricos e estão prontos para serem devorados por técnicas de mineração de dados.

Então como você lida com dados categóricos? Bem, para resumir, é preciso transformá-los em dados quantitativos.

Algumas vezes, seus dados categóricos podem ter uma ordenação natural que pode-se usar para atribuir um valor a cada categoria. Por exemplo, se você tivesse uma variável no seu conjunto de dados onde as pessoas relatavam se dirigiam um Scion, um Toyota ou um Lexus, talvez você pudesse transformar essas respostas em 1, 2 e 3. *Voila*, números.

Mas, com maior frequência, não há ordenação, como em gênero. Por exemplo, masculino, feminino e desconhecido são rótulos distintos sem uma noção de ordenação. Nesse caso, é comum usar uma técnica chamada **código dummy** para converter seus dados categóricos em dados quantitativos.

O código dummy funciona pegando uma única coluna categórica (considere a coluna Implied Gender) e transformando-a em múltiplas colunas binárias. Você poderia usar a coluna Implied Gender e, no lugar dela, ter uma coluna para masculino, outra para feminino e outra para gênero desconhecido. Se um valor na coluna original era “M”, isso poderia ser codificado com um 1 na coluna male, um 0 na coluna female e um 0 na coluna unknown gender.

Isso é, na realidade, excesso, porque se as colunas male e female eram ambas 0, então a coluna unknown gender já estava subentendida. Você não precisa de uma terceira coluna.

Desta forma, quando usar o código dummy em uma variável categórica, você sempre precisa de uma coluna a menos do que tem em valores de categoria — a última categoria sempre é subentendida por outros valores. Em linguagem estatística, você diria que a variável categórica de gênero tem apenas dois *graus de liberdade*, porque os graus de liberdade sempre são um a menos dos valores possíveis que a variável pode aceitar.

Neste exemplo em particular, comece criando uma cópia da planilha Training Data chamada *Training Data w Dummy Vars*. Você dividirá os dois primeiros preditores em *duas colunas cada*, então prossiga e esvazie as colunas A e B e insira duas colunas em branco à esquerda da coluna A.

Nomeie essas quatro colunas vazias como *Male*, *Female*, *Home* e *Apt* (unknown gender e PO box tornam-se subentendidos). Como exibido na Figura 6-2, você deve ter quatro colunas em branco para abrigar o código dummy das duas variáveis categóricas.

				Pregnancy Test	Birth Control	Feminine Hygiene	Folic Acid	Prenatal Vitamins	Prenatal Yoga	Body Pillow	Ginger Ale	Sea Bands	Stopped buying cigarettes	Cigarettes	Smoking cessation	Stopped buying wine	Wine	Maternity Clothes	PREGNANT
1	Male	Female	Home	Apt	1	0	0	0	1	0	0	0	0	0	0	0	0	0	1
2					1	0	0	0	1	0	0	0	0	0	0	0	0	0	1
3					1	0	0	0	1	0	0	0	0	0	0	0	0	0	1
4					1	0	0	0	0	0	0	0	1	0	0	0	0	0	1
5					0	0	0	0	0	0	0	1	0	0	0	0	0	0	1
6					0	0	0	0	0	1	0	0	0	0	0	0	1	0	1
7					0	0	0	0	1	0	0	0	0	1	0	0	0	0	1
8					0	0	0	1	1	0	0	0	0	0	0	0	0	0	1
9					0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
10					0	0	0	0	0	0	1	0	0	0	0	0	0	0	1
11					0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
12					0	0	0	0	0	0	0	0	0	0	0	1	0	0	1

**Figura 6-2:** A tabela Training Data com novas colunas para as variáveis dummy

Considere a primeira linha dos dados em treinamento. Para transformar o “M” na coluna gender em dado codificado dummy, você coloca um 1

na coluna Male e um 0 na coluna Female. (O 1 na coluna Male já sugere que o gênero não é Unknown.)

Na célula A2 na aba Training Data w Dummy Vars, verifique a antiga categoria na aba Training Data e atribua um 1 se a categoria foi configurada em “M”:

```
=IF('Training Data'!A2="M",1,0)
```

O mesmo serve para valores “F” na coluna Female, “H” na coluna Home e “A” na coluna Apt. Para copiar essas quatro fórmulas para todas as linhas dos dados em treinamento, você pode arrastá-las, ou, ainda melhor, como explicado no Capítulo 1, realçar todas as quatro fórmulas e então dar um clique duplo no canto inferior direito de D2. Isso preencherá a planilha com valores convertidos até D1001. Uma vez que tenha convertido essas duas colunas categóricas em quatro variáveis dummy binárias (veja a Figura 6-3), estará pronto para modelar.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	PREG NANT	
1	Male	Female	Home	Apt	Test	Pregnancy Control	Birth Hygiene	Feminine Acid	Prenatal Vitamins	Prenatal Yoga	Body Pillow	Ginger Ale	Sea Bands	Stopped cigges	Cigar ettes	Smoking cessation	Stopped buying wine					Maternity Clothes	
2	1	0	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0		
3	1	0	1	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0		
4	1	0	1	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0		
5	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0		
6	0	1	0	1	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0		
7	0	1	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0		
8	1	0	1	0	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0		
9	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0		
10	0	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0		
11	0	1	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1		

Figura 6-3:↑Dados↑em↑treinamento↑com↑variáveis↑dummy↑populadas

## Vamos↑Cozinhar↑Nossa↑Própria↑Regressão Linear

Toda vez que eu digo isso, um estatístico perde suas asas, mas eu falarei de qualquer forma — se você alguma vez enfiou uma linha de tendência

por uma nuvem de pontos em um gráfico de dispersão, então você construiu um modelo IA.

Você provavelmente está pensando, “Mas não há como! Eu saberia se eu tivesse criado um robô que pudesse voltar no tempo e parar John Connor!”

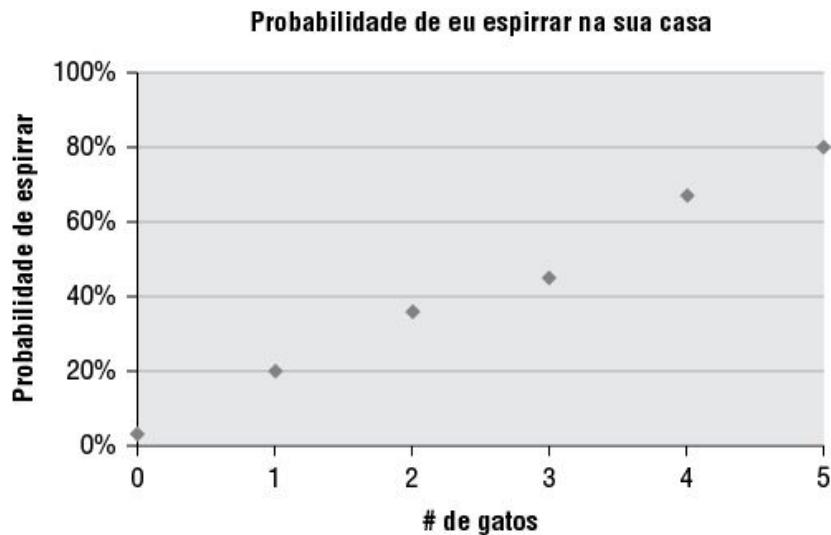
### ***Os Modelos Lineares Mais Simples***

Deixe-me explicar mostrando um exemplo na Figura 6-4.

	A	B
1	Number of cats owned	Likelihood I'll sneeze in your home
2	0	3%
3	1	20%
4	2	36%
5	3	45%
6	4	67%
7	5	80%

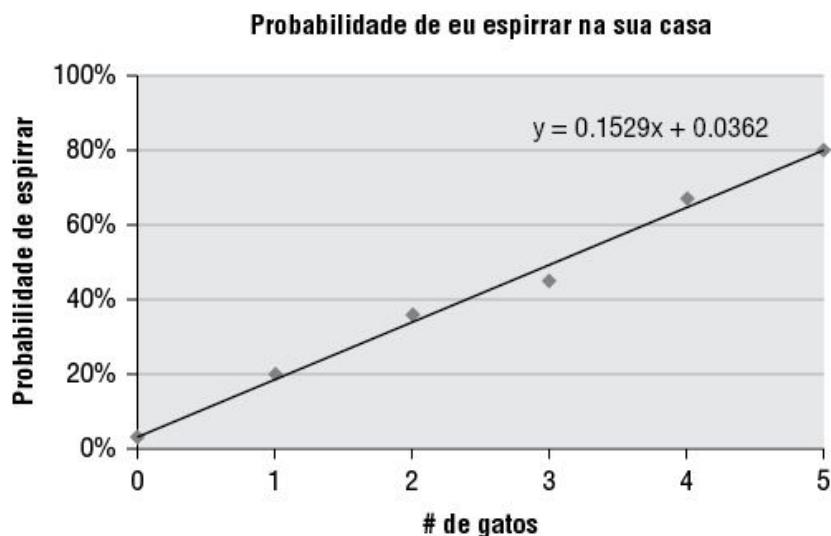
**Figura 6-4:** Proprietário de gato versus eu espirrando

Na tabela exibida, você tem o número de gatos em uma casa na primeira coluna e a probabilidade de eu espirrar dentro daquela casa na segunda coluna. Sem gatos? 3% das vezes eu espirro de qualquer forma só porque eu sei que um gato platônico existe em algum lugar. Cinco gatos? Bem, então meu espirro é simplesmente garantido. Agora, nós podemos criar um gráfico de dispersão desses dados no Excel e olhar para ele conforme está na Figura 6-5 (Para mais informações sobre como inserir dispersões e gráficos veja o Capítulo 1).



**Figura 6-5:** Gráfico de dispersão de gatos versus espirro

Clicando com o botão direito nos pontos de dados no gráfico (você tem que clicar com o botão direito em um ponto de dado real, não apenas no gráfico) e selecionando Add Trendline no menu, é possível selecionar um modelo de regressão linear para adicionar ao gráfico. Abaixo da seleção “Options” na janela “Format Trendline”, você pode selecionar “Display equation on chart”. Pressionando OK, é possível ver a linha de tendência e a fórmula para a linha (Figura 6-6).



**Figura 6-6:** Modelo linear exibido no gráfico

A linha de tendência no gráfico mostra o relacionamento entre gatos e espirros corretamente com uma fórmula de:

$$Y = 0,1529x + 0,0362$$

Em outras palavras, quando x é 0, o modelo linear pensa que eu tenho uma chance de aproximadamente 3-4% de espirrar, e o modelo me dá uma chance extra de 15% por gato.

Essa linha de base de 3-4% é chamada de **intercepto** do modelo, e os 15% por gato é chamado de **coeficiente** para a variável gatos. Fazer uma previsão com um modelo linear como esse requer nada além de pegar meu dado futuro e combiná-lo com os coeficientes e os interceptos do modelo.

Na verdade, você pode copiar a fórmula  $=0,1529x+0,0362$  do gráfico se preferir e colá-la em uma célula para fazer previsões, substituindo o x por um número. Por exemplo, se no futuro eu for em uma casa com três gatos e meio (o pobre Timmy perdeu suas patas traseiras em um acidente de barco), então eu pegaria uma “combinação linear” dos coeficientes e meus dados, adicionaria ao intercepto, e obteria minha previsão::

$$0,1529 \cdot 3,5 \text{ gatos} + 0,0362 = 0,57$$

Uma chance de 57% de espirrar! Isso é um modelo IA no sentido que eu peguei um variável independente (gatos) e uma variável dependente (espirro) e pedi para o computador descrever seu relacionamento como uma fórmula que melhor se ajusta ao nosso dado histórico.

Agora, você pode estar se perguntando como o computador descobriu essa linha de tendência a partir dos dados. Parece boa, mas como ele saberia onde colocá-la? Basicamente, o computador procurou uma linha de tendência que **melhor se ajustava** aos dados, em que por **melhor** ajuste eu quero dizer a linha de tendência que minimiza **a soma dos erros quadrados** com os dados em treinamento.

Para entender o que a soma dos erros quadrados significa, se você avaliar a linha de tendência para um gato, você obtém:

$$0,1529 \cdot 1 \text{ gato} + 0,0362 = 0,1891$$

Mas o dado em treinamento entrega uma probabilidade de 20%, não 18,91%. Então seu erro nesse momento na linha de tendência é de 1,09%. Esse valor de erro é quadrático para certificar um valor positivo, independente de a linha de tendência estar acima ou abaixo desse ponto de dados. 1,09% ao quadrado é 0,012%. Agora, se somar cada um desses valores de erro quadrado para os pontos em nossos dados em treinamento, você obteria a soma do erro quadrado (frequentemente chamada de **soma dos quadrados**). E isso é o que o Excel minimizou ao ajustar a linha de tendência para o gráfico de espirro.

Embora seus dados RetailMart tenham dimensões demais para jogar em um gráfico de dispersão, nas próximas seções você ajustará o mesmo tipo de linhas aos dados a partir do zero.

### ***De volta aos Dados RetailMart***

Certo, então está na hora de construir um modelo linear como o modelo Kitty Sneeze no conjunto de dados RetailMart. Primeiro, crie uma nova aba chamada **Linear Model**, e cole os valores da aba Training Data w Dummy Vars, mas, quando colá-los, comece na coluna B para economizar espaço para alguns rótulos de linhas na coluna A e na linha 7 para deixar espaço no topo da planilha para os coeficientes do modelo linear e outros dados avaliativos que você rastreará.

Cole a linha de cabeçalho para suas variáveis dependentes novamente na linha 1 para manter organizada. E, na coluna U, adicione o rótulo **Intercept** porque seu modelo linear precisará de uma linha de base assim como no exemplo anterior. Além disso, para incorporar o intercepto ao modelo mais facilmente, preencha a sua coluna intercept (U8:U1007) com 1s. Isso permitirá que você avalie o modelo usando **SUMPRODUCT** da linha de coeficientes com uma linha de dados que incorporará o valor do intercepto.

Todos os coeficientes para esse modelo irão para a linha 2 na planilha, então nomeie a linha 2 como **Model Coefficients** e coloque 1 em cada célula. Você também pode colocar alguma formatação condicional na

linha de coeficiente para que possa ver as diferenças nelas uma vez que elas estejam configuradas.

Seu conjunto de dados agora parece com a Figura 6-7.

**Figura 6-7:** Configuração de modelagens lineares

Uma vez que os coeficientes na linha 2 estiverem configurados, pode-se tirar uma combinação linear (fórmula `SUMPRODUCT`) dos coeficientes com uma linha de dado de clientes e obter a previsão de gravidez.

Você tem colunas demais aqui para construir um modelo linear criando um gráfico dela como fiz com os gatos, então em vez de fazer isso, você treinará o modelo sozinho. O primeiro passo é adicionar uma coluna na planilha com uma previsão em uma das linhas de dados.

Na coluna W, perto dos dados de clientes, adicione a coluna nomeada ***Linear Combination (Prediction)*** na linha 7 e abaixo dela tire uma combinação linear dos coeficientes e dados de clientes (coluna do intercepto incluída). A fórmula que você coloca na linha 8 para fazer isso para o primeiro cliente é:

```
=SUMPRODUCT(B$2:U$2, B8:U8)
```

A referência absoluta dever ser colocada na linha 2, para que você possa arrastar essa fórmula para todos os clientes sem mudar a linha de coeficiente.

## DICA

Além disso, você pode querer realçar a coluna W, clicar com o botão direito, selecionar “Format Cells...” e formatar os valores como um número com duas casas decimais, apenas para evitar que seus olhos sangrem ao ver tantos

Uma vez adicionada essa coluna, seus dados estarão parecidos com a Figura 6-8.

	N	O	P	Q	R	S	T	U	V	W
1	Sea Bands	Stopped buying ciggies	Cigarettes	Smoking Cessation	Stopped buying wine	Wine	Maternity Clothes	Intercept		
2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00		
3										
4										
5										
6										
7	Sea Bands	Stopped buying ciggies	Cigarettes	Smoking Cessation	Stopped buying wine	Wine	Maternity Clothes	Intercept	PREGNANT	Linear Combination (Prediction)
8	0	0	0	0	0	0	0	1	1	=SUMPRODUCT(B\$2:U\$2,B8:U8)
9	0	0	0	0	0	0	0	1	1	5.00
10	1	0	0	0	0	0	0	1	1	5.00
11	0	0	0	0	0	0	0	1	1	3.00
12	0	0	0	0	0	0	0	1	1	5.00

Figura 6-8: A coluna de previsão para um modelo linear

Idealmente, a coluna de previsão (coluna W) pareceria idêntica ao que nós conhecemos como verdade (coluna V), mas usando coeficientes de 1 para toda variável é fácil ver o caminho. O primeiro cliente recebe uma previsão de 5, ainda que gravidez seja indicada com um 1 e não gravidez com um 0. O que é um 5? Muito, muito grávida?

### Adicionando um Cálculo de Erro

É preciso fazer o computador configurar esses modelos de coeficientes para você, mas para que ele saiba fazer isso, é preciso fazer com que a máquina saiba quando uma previsão está certa e quando está errada.

Com esse objetivo, adicione um cálculo de erro na coluna X. Use erro quadrado, que é apenas o quadrado da distância do valor de PREGNANT (coluna V) do valor previsto (coluna W).

Elevar o erro ao quadrado permite que cada cálculo de erro seja positivo, para que você possa somá-los para saber o erro total do modelo. Você não quer que erros positivos e negativos cancelam uns aos outros. Então para o primeiro cliente na planilha, você teria a seguinte fórmula:

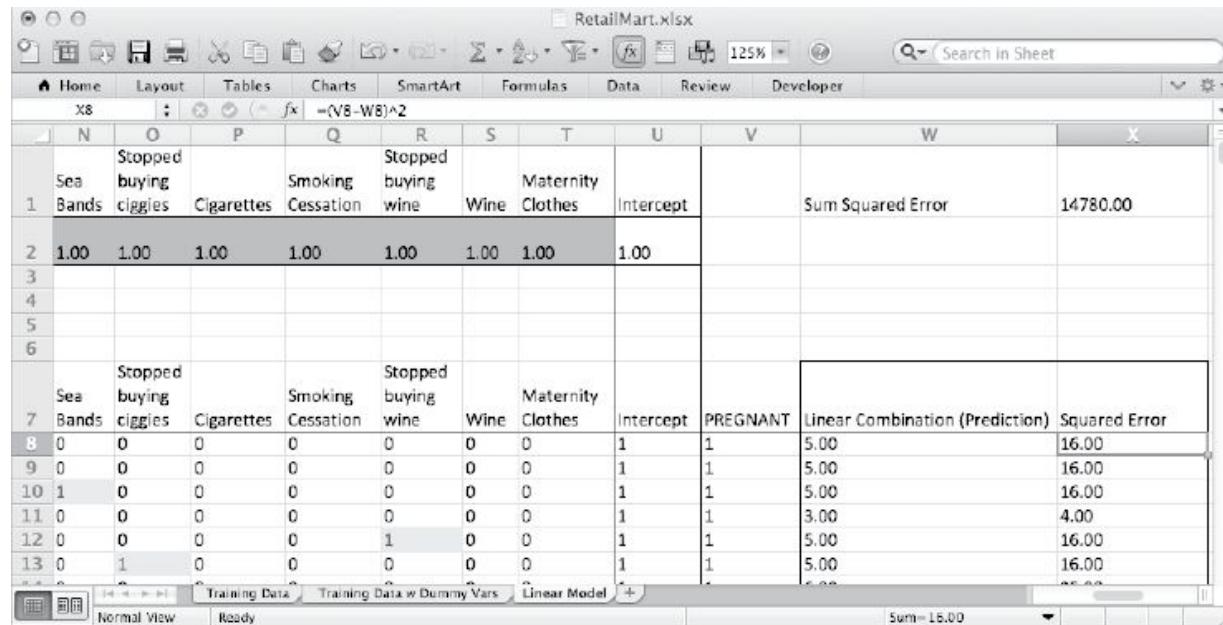
$$= (V8 - W8)^2$$

Pode-se arrastar essa célula pelo restante da coluna para fornecer seu próprio cálculo de erro a cada previsão.

Agora, adicione uma célula acima das previsões na célula X1 (nomeada em W1 como **Sum Squared Error**), onde você somará a coluna de erro quadrático usando a fórmula:

$$=\text{SUM}(X8:X1007)$$

Sua planilha parece com a Figura 6-9:



The screenshot shows an Excel spreadsheet titled "RetailMart.xlsx". The formula bar at the top displays the formula  $=\text{SUM}(X8:X1007)$ . The main table has columns labeled N through X. Row 1 contains labels: Sea Bands, Stopped buying ciggies, Cigarettes, Smoking Cessation, Stopped buying wine, Wine, Maternity Clothes, Intercept, Sum Squared Error, and an empty cell. Row 2 contains values: 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, 1.00, and 14780.00 respectively. Rows 3 through 6 are empty. Row 7 starts a new section with labels: Sea Bands, Stopped buying ciggies, Cigarettes, Smoking Cessation, Stopped buying wine, Wine, Maternity Clothes, Intercept, and PREGNANT. It also includes a header row for a secondary table below it. This secondary table has columns: Linear Combination (Prediction) and Squared Error. Data rows 8 through 13 show values: (5.00, 16.00), (5.00, 16.00), (5.00, 16.00), (3.00, 4.00), (5.00, 16.00), (5.00, 16.00), and (5.00, 16.00). The status bar at the bottom indicates "Sum=16.00".

**Figura 6-9:** Previsões e soma dos erros quadrados

## Treinando com o Solver

Agora você está pronto para treinar seu modelo linear. Você quer configurar os coeficientes para cada variável de tal forma que a soma dos erros quadrados seja a mais baixa possível. Se parece ser um trabalho para o Solver, você está certo. Assim como nos Capítulos 2, 4 e 5, você abrirá o Solver e fará com que o computador encontre os melhores coeficientes para você.

A função objetiva será o valor da soma dos erros quadrados da célula X1, a qual você vai querer minimizar “ao mudar as células variáveis” B2 pela U2, que são seus coeficientes modelos.

Agora, o quadrado do erro é uma função quadrática das suas variáveis de decisão, os coeficientes, então você não pode simplesmente usar Simplex-LP como método de resolução como usou ao longo de todo o Capítulo 4. Simplex é muito rápido e garante encontrar a melhor resposta, mas requer que o modelo considere somente combinações lineares das decisões. Você precisará usar o algoritmo evolucionário no Solver.

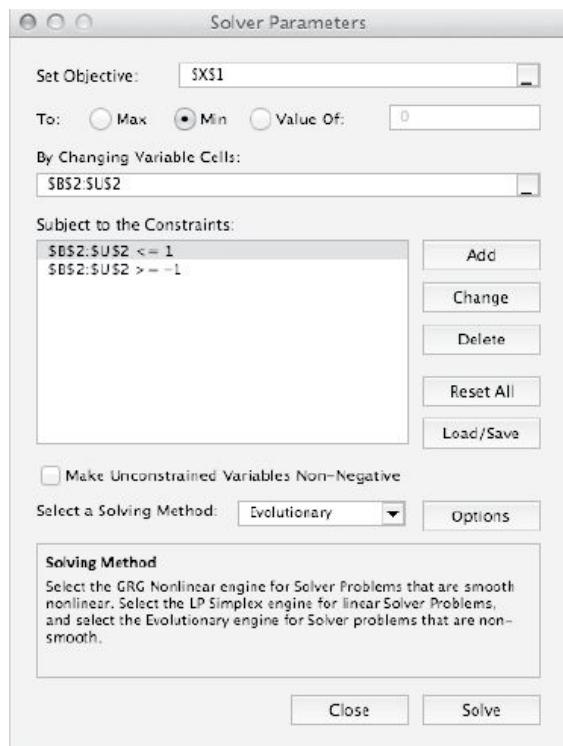
## REFERÊNCIA

Para mais modelos de otimização não-linear e trabalhos internos do algoritmo evolucionário de otimização, veja o Capítulo 4. Se gostar, você também pode brincar com outros algoritmos de otimização não-lineares no Excel chamado GRG (Gradiente Reduzido Generalizado).

Basicamente, o Solver investigará os valores de coeficiente que fazem a soma das raízes caírem até ele achar que encontrou uma boa solução. Mas, para usar o algoritmo evolucionário efetivamente, você precisará traçar limites superiores e inferiores para cada um dos coeficientes que está tentando configurar.

Peço que brinque bastante com esses limites. Quanto mais apertados eles estiverem (nem tanto assim), melhor os algoritmos funcionam. Para este modelo, eu os configurei para entre -1 e 1.

Uma vez que tenha completado esses itens, a configuração do Solver deve se parecer com a Figura 6-10.



**Figura 6-10:** A configuração do Solver para modelos lineares

Pressione o botão Solve e aguarde! Como o Evolutionary Solver experimenta vários coeficientes para o modelo, você verá os valores mudarem. A formatação condicional nas células lhe dará uma sensação de magnitude. Além do mais, a soma dos erros quadrados deveria saltar, mas geralmente ela diminui com o tempo. Quando o Solver termina , ele mostra que o problema está otimizado. Clique em OK e terá seu modelo de volta.

Na Figura 6-11, você verá que o Solver executado terminou com uma soma dos erros quadrados de 135,52. Se você está acompanhando e quiser executar o Solver sozinho, esteja ciente de que duas execuções do algoritmo evolucionário não precisam ser iguais — sua soma dos quadrados pode ser mais alta ou mais baixa que a do livro, com uma pequena diferença nos coeficientes de modelo final. O modelo linear otimizado está exibido na Figura 6-11.

The screenshot shows a Microsoft Excel spreadsheet titled "RetailMart.xlsx". The active sheet is named "X1". The formula bar displays the formula: =SUMX(X8:X1007). The data is organized into several columns:

- Row 1:** Contains labels "N", "O", "P", "Q", "R", "S", "T", "U", "V", "W", and "X". Below these, under column "N", are "Sea Bands" and "Stopped buying ciggies". Under column "Q", are "Cigarettes" and "Smoking Cessation". Under column "R", are "Stopped buying wine" and "Maternity Clothes". Under column "U", is "Intercept". To the right of "W" is "Sum Squared Error" with the value "135.52".
- Row 2:** Contains numerical values: 0.15, 0.16, -0.16, 0.16, 0.19, -0.21, 0.24, and 0.48.
- Row 7:** Contains labels "Sea Bands" and "Stopped buying ciggies" under "N", "Cigarettes" and "Smoking Cessation" under "Q", "Stopped buying wine" and "Maternity Clothes" under "R", and "Intercept" under "U". It also includes "PREGNANT" and "Linear Combination (Prediction)" under "V", and "Squared Error" under "X".
- Rows 8-18:** Data points for each row, showing values for Sea Bands, Stopped buying ciggies, Cigarettes, Smoking Cessation, Stopped buying wine, Maternity Clothes, Intercept, PREGNANT, Linear Combination (Prediction), and Squared Error.

The status bar at the bottom indicates "Normal View" and "Sum = 135.52".

**Figura 6-11:** Modelo linear otimizado

## USANDO A FÓRMULA LINEST() EM UMA REGRESSÃO LINEAR

Alguns leitores talvez estejam cientes de que o Excel possui sua própria fórmula de regressão linear chamada LINEST(). Em alguns casos, a fórmula pode, de fato, fazer o que você fez manualmente. Ela admite 64 características, porém, para regressões grandes, você precisará executar a sua própria de qualquer forma.

Sinta-se à vontade para experimentar essa fórmula neste conjunto de dados. Mas tenha cuidado! Leia a documentação de ajuda do Excel sobre a fórmula primeiro. Para retirar todos os seus coeficientes em ordem reversa, é preciso usá-la como uma fórmula de array (veja o Capítulo 1). E, também, ela coloca os coeficientes para fora na ordem reversa (Male será o último coeficiente antes da interrupção), o que é bem perturbador.

LINEST() é bem prática pois ela computa automaticamente muitos dos valores necessários para desempenhar um teste estatístico em seu modelo linear, como o temido cálculo do coeficiente do erro padrão que você verá na próxima seção.

Mas, neste capítulo, você fará tudo manualmente para que conheça bem o que LINEST() (e outros pacotes de software de funções de modelo linear) está fazendo e fique confortável ao depender dela no futuro. Ao fazer o processo manualmente, você também ajudará a transição para regressão logística, que o Excel *não suporta*.

---

## USANDO A REGRESSÃO MEDIANA PARA LIDAR MELHOR COM VALORES ATÍPICOS

Na *regressão mediana*, você minimiza a soma dos *valores absolutos* dos erros em vez de a soma dos erros quadrados. Essa é a única diferença da regressão linear.

O que você entende disso?

Na regressão linear, os *valores atípicos (outliers)*, valores marcados distantes do restante dos dados) em seu conjunto de dados possuem maior influência e podem prejudicar o processo de ajuste do modelo. Quando os erros do valor atípico são altos, a regressão linear irá cada vez mais atrás deles, alcançando um equilíbrio entre o erro alto e muitos outros erros menores nos pontos normais do que o equilíbrio que é alcançado na regressão mediana. Nela, a linha que é ajustada ao dado ficará mais perto dos pontos de dados internos e usuais do que perto dos valores atípicos.

Enquanto não trabalhamos com regressão mediana neste capítulo; seria bom você tentar sozinho. Apenas troque o termo do quadrado do erro para o valor absoluto (Excel possui a função ABS) e está pronto.

Dito isso, se você usa o Windows e tiver o OpenSolver instalado (veja o Capítulo 1), então este é um grande problema bônus!

Já que na regressão mediana você está minimizando o erro, e já que um valor absoluto também pode ser uma função máxima (o máximo de um valor e  $-1$  vezes aquele valor), tente linearizar a regressão à média como um modelo de otimização estilo minimax (veja o Capítulo 4 para mais sobre modelos de otimização minimax). Dica: você precisará criar uma variável por linha de dado em treinamento e é por isso que precisa do OpenSolver — o Solver comum não consegue lidar com milhares de decisões e duas mil limitações.

Boa sorte!

# Estatísticas da Regressão Linear: R-quadrado, Testes F e Testes t

## NOTA

Esta próxima seção possui o conteúdo mais pesado sobre estatística do livro inteiro. Na verdade, esta seção contém os cálculos mais complexos do livro inteiro — o cálculo do modelo do erro padrão do coeficiente. Tentei descrever tudo o mais claramente possível, mas alguns cálculos trazem explicações a um nível complicado para texto. Não quero me desvirtuar do assunto e ensinar álgebra linear aqui.

Tente entender esses conceitos da melhor forma que puder. Pratique-os. E, se quiser saber mais, pegue um livro didático de estatística de nível iniciante (por exemplo, *Statistics in Plain English* de Timothy C. Urdan [Routledge, 2010]).

Agora você tem um modelo linear que ajustou minimizando a soma dos quadrados. Dando uma olhada nas previsões da Coluna Y, todas parecem certas. Por exemplo, a cliente grávida na linha 27 que comprou um teste de gravidez, vitaminas para o pré-natal e roupas para grávidas obteve um escore de 1,07 enquanto que a cliente na lina 996 que comprou apenas um vinho obteve 0,15. Dito isso, as questões permanecem:

- Quão bem a regressão realmente *ajusta o dado* a partir de uma perspectiva quantitativa sem ser pela olhada?
- Esse ajuste geral é por acaso ou é estatisticamente significativo?
- Quão útil é cada uma das características para o modelo?

Para responder a essas questões para uma regressão linear, você pode computar *R-quadrado*, um *teste F geral* e *testes t* para cada um dos coeficientes.

## *R-quadrado — Acessando o Benefício do Ajuste*

Se você não soubesse nada sobre um cliente dos conjuntos em treinamento (da coluna B até a T estavam faltando) mas foi forçado a fazer previsões sobre gravidez de qualquer forma — a melhor maneira de minimizar a soma dos erros quadrados seria colocar a média da coluna V na planilha para cada previsão. Neste caso, a média 0,5 é dada com divisão de 500/500 nos dados em treinamento. E, já que cada valor atual é 0 ou 1, cada erro seria de 0,5, tornando cada erro quadrado em 0,25. Após 1000 previsões teríamos, então, a soma dos quadrados de 250.

Esse valor é chamado de a **soma total dos quadrados**. É a soma dos quadrados dos desvios para cada valor na coluna V a partir da média da coluna V. O Excel oferece uma fórmula elegante para esse cálculo em um passo, `DEVSQ`.

Em X2, você pode calcular a soma total dos quadrados como:

=DEVSQ(V8 : V1007)

Mas, enquanto colocar a soma para cada previsão produziria uma soma dos erros quadrados de 250, a soma dos erros quadrados dada pelo modelo linear que você ajustou antes é bem menor do que isso. Apenas 135,52.

Isso significa que 135,52 dos 250 da soma total dos quadrados permanece **sem explicação** após o ajuste da regressão (nesse contexto, a soma dos erros quadrados é chamada de **soma de quadrados residual**).

Dando uma volta por esse valor, a **soma explicada dos quadrados** (e é exatamente isso — a quantidade que pode ser explicada em seu modelo) é  $250 - 135,52$ . Coloque isso em X3 como:

=X2-X1

Isso resulta em 114,48 na soma dos quadrados explicada (se você não obteve a soma dos erros quadrados de 135,52 ao ajustar a regressão, seus resultados vão variar um pouco).

Então, quão bom é esse ajuste?

Geralmente, obtemos essa resposta ao olhar a relação da soma dos quadrados explicada com a soma total dos quadrados. Esse valor é chamado de **R-quadrado**. Podemos calcular essa relação em X4:

$$=X3/X2$$

Como mostra a Figura 6-12, 0,46 é o valor do R-quadrado. Se o modelo se ajustasse perfeitamente, você teria 0 erros ao quadrado, a soma dos quadrados explicada seria igual ao total e o R-quadrado seria 1. Se o modelo não se ajustasse, o R-quadrado seria próximo a 0. Então, no caso desse modelo, dadas as entradas dos dados em treinamento, o modelo pode fazer um trabalho ok-não-tão-perfeito de replicar a variável dependente dos dados em treinamento (a coluna Pregnancy, gravidez).

1	Sea Bands	Stopped buying ciggies	Cigarettes	Smoking Cessation	Stopped buying wine	Wine	Maternity Clothes	Intercept		Sum Squared Error	135.52
2	0.15	0.16	-0.16	0.16	0.19	-0.21	0.24	0.48		Total Sum of Squares	250.00
3										Explained Sum of Squares	114.48
4										R squared	0.46
5											
6											
7	Sea Bands	Stopped buying ciggies	Cigarettes	Smoking Cessation	Stopped buying wine	Wine	Maternity Clothes	Intercept	PREGNANT	Linear Combination (Prediction)	Squared Error
8	0	0	0	0	0	0	0	1	1	0.88	0.01
9	0	0	0	0	0	0	0	1	1	0.87	0.02
10	1	0	0	0	0	0	0	1	1	0.72	0.08
11	0	0	0	0	0	0	0	1	1	0.69	0.10
12	0	0	0	1	0	0	0	1	1	0.96	0.00
13	0	1	0	0	0	0	0	1	1	0.88	0.01
14	0	0	0	0	0	0	0	1	1	0.72	0.08
15	0	0	0	0	0	0	1	1	1	0.67	0.11
16	0	0	0	0	0	0	0	1	1	0.66	0.12
17	0	0	0	0	0	0	1	1	1	0.96	0.00
18	0	0	0	1	0	0	0	1	1	0.62	0.15

**Figura 6-12:** R-quadrado de 0,46 para a regressão linear

Agora, lembre-se que o cálculo do R-quadrado somente funciona para encontrar relações lineares entre dados. Se você tem uma relação não-linear (seja em forma de V ou U) entre uma variável dependente e independente em um modelo, o valor R-quadrado não poderia capturar a relação.

## **O Teste F — O Ajuste É Estatisticamente significativo?**

Frequentemente, as pessoas param no R-quadrado ao analisar o ajuste de uma regressão.

“Olha, o ajuste parece bom! Terminei.”

Não faça isso.

O R-quadrado apenas mostra se o modelo se ajustou aos dados corretamente. O que ele não mostra é se o ajuste é **estatisticamente significativo**.

É fácil, especialmente com os conjuntos de dados rarefeitos (com algumas observações), obter um modelo bem ajustado, mas sendo **estatisticamente insignificante**, indicando que a relação entre as características e a variável dependente talvez não seja real.

O ajuste do seu modelo é por acaso? Um pouco de sorte? Para que um modelo seja estatisticamente significativo, você deve rejeitar a **hipótese** ajuste-por-acaso. Então, **presuma** por um momento que seu modelo está ajustado por acaso. Que o ajuste se deu totalmente por sorte da obtenção das 1.000 observações aleatórias que você obteve do banco de dados RetailMart. Esse pressuposto com muitas controvérsias é chamado de **hipótese nula**.

A regra padrão é rejeitar a hipótese nula se ela for verdade, a probabilidade de obter um ajuste no mínimo tão bom quanto esse é menor do que 5%. Essa probabilidade é chamada de **valor p**.

Para calcular a probabilidade, realiza-se um teste F. Um **teste F** pega três partes da informação sobre o nosso modelo e os executa usando uma distribuição de probabilidade chamada de distribuição F (para uma explicação do termo distribuição de probabilidade, veja a discussão sobre distribuição normal no Capítulo 4). Essas três partes são:

- **Quantidade de coeficientes de modelo** — É 20 no nosso caso (19 características mais o intercepto).

- *Graus de liberdade* — Esta é a quantidade de observações nos dados em treinamento menos o número de coeficientes de modelo.
- *A estatística F* — A estatística F é a relação entre os erros quadrados explicados e inexplicados ( $X3/X1$  na planilha) vezes a razão entre os graus de liberdade e as variáveis dependentes.

Quando maior a estatística F, menor é a probabilidade da hipótese nula. Dada a explicação anterior sobre a estatística F, como aumentá-la? Aumente uma das duas relações no cálculo. Você também pode explicar mais dos dados (isto é, ter um melhor ajuste) ou obter mais dados da mesma quantidade de variáveis (isto é, tenha certeza de que seu ajuste adere a uma amostra maior).

Ao retornar para a planilha, precisamos contar o número de observações e de coeficientes de modelo que temos.

Nomeie Y1 como *Observation Count* e em Z1 conte todos os valores de gravidez na coluna V:

=COUNT (V8 : V1007)

Você deveria, como era de se esperar, obter 1.000 observações.

Em Z2, obtenha a Contagem do Coeficiente de Modelo contando as observações na linha 2:

=COUNT (B2 : U2)

Você deveria obter 20 incluindo o intercepto. Calcule os Graus de Liberdade em Z3 ao subtrair a contagem do coeficiente de modelo da contagem de observação:

=Z1 - Z2

Você obterá um valor de 980 graus de liberdade.

Vamos para a estatística F em Z4. Como mostrado anteriormente, essa é apenas a relação dos erros quadrados explicados com os inexplicados

$(X_3/X_1)$  vezes a relação dos graus de liberdade com as variáveis dependentes ( $Z_3/(Z_2-Z_1)$ ):

$$= (X_3/X_1) * (Z_3 / (Z_2 - 1))$$

Portanto, podemos posicionar esses valores dentro da distribuição F em Z5 usando a função FDIST no Excel. Nomeie a célula como **F Test P Value**. FDIST leva a estatística F, a quantidade de variáveis dependentes no modelo e os graus de liberdade:

$$=\text{FDIST}(Z4, Z2-1, Z3)$$

Como a Figura 6-13 mostra, a probabilidade de obter um ajuste como este, dada a hipótese nula, é efetivamente 0. Então, você pode rejeitar a hipótese nula e concluir que o ajuste é estatisticamente significativo.

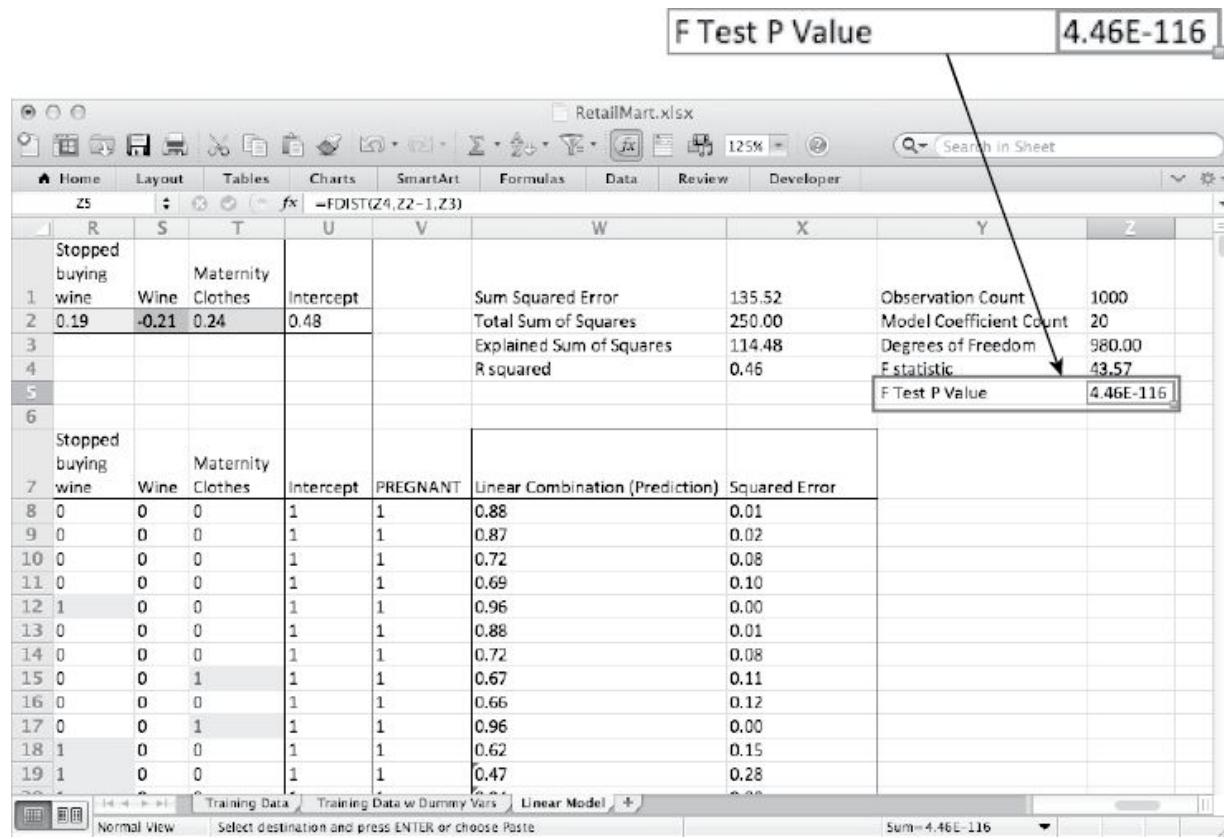


Figura 6-13: O resultado do teste F

## Coeficiente do Teste T — Quais Variáveis São Significativas?

### AVISO: MATRIZ DE MATEMÁTICA À FRENTE!

Enquanto as duas estatísticas anteriores não foram difíceis de computar, realizar um teste t em uma regressão linear múltipla requer uma multiplicação e inversão de matriz. Se você não se lembra de como essas operações funcionam desde o ensino médio ou introdução à matemática na faculdade, dê uma olhada em um livro de álgebra linear ou cálculo. Ou leia a Wikipédia. Use a pasta de trabalho que está disponível para download com este capítulo para certificar-se de que sua matemática está correta.

No Excel, a multiplicação de matriz usa a função `MMULT` enquanto que a inversão usa a função `MINVERSE`. Já que a matriz nada mais é do que um conjunto de números em um retângulo, essas fórmulas são conjuntos de fórmulas (veja o Capítulo 1 sobre como usar conjuntos de fórmulas no Excel).

---

O teste F verificou que a regressão inteira foi significativa, e você também pode verificar a significância das variáveis individuais. Ao testar a significância de características individuais, pode enxergar o que está impulsionando os resultados do seu modelo. As variáveis estatisticamente insignificantes talvez possam ser eliminadas ou, se você tiver completa certeza de que a variável insignificante **deveria ser importante**, então talvez deva investigar se há algum problema com a limpeza dos dados em seu conjunto em treinamento.

O teste para a significância do coeficiente de modelo é chamado de **teste t**. Quando você executa um teste t, parecido com o teste F, você presume que o coeficiente de modelo que está testando não vale nada e deveria ser 0. Com essa premissa, o teste t calcula a probabilidade de obter um coeficiente tão longe de 0 quanto o que obteria da sua amostra.

Quando executar um teste t em uma variável dependente, o primeiro valor que deve calcular é a **previsão de erro padrão**. Esse é o desvio padrão da previsão de erro da amostra (veja o Capítulo 4 para saber mais sobre desvio padrão), significando que é uma medida de variabilidade nos erros das previsões do modelo.

Pode-se calcular a previsão de erro padrão em X5 como a raiz quadrada da soma dos erros quadrados (X1) dividido pelos graus de liberdade

(Z3):

$$=\text{SQRT}(X1/Z3)$$

Isso nos dá a planilha mostrada na Figura 6-14.

Ao usar esse valor, pode-se calcular os **erros padrões de coeficientes do modelo**. Pense no erro padrão de um coeficiente como o desvio padrão se você continuar a desenhar mil novas amostras de clientes do banco de dados da RetailMart e ajustar regressões lineares novas para os conjuntos em treinamento. Você não obteria os mesmos coeficientes todas as vezes; haveria uma pequena variação. O erro padrão do coeficiente determina a variabilidade que esperaria ver.

The screenshot shows a Microsoft Excel spreadsheet titled "RetailMart.xlsx". The active sheet is labeled "Linear Model". The data is organized into several sections:

- Top Left:** A table with columns R, S, T, U, V, W, X, Y, Z. Rows 1 and 2 show coefficients for "Stopped buying wine" (Intercept: 0.48, Wine: -0.21, Clothes: 0.24). Row 5 shows the prediction standard error as 0.37.
- Top Right:** Summary statistics including Sum Squared Error (135.52), Total Sum of Squares (250.00), Explained Sum of Squares (114.48), R squared (0.46), Prediction Standard Error (0.37), Observation Count (1000), Model Coefficient Count (20), Degrees of Freedom (980.00), F statistic (43.57), and F Test P Value (4.46E-116).
- Middle:** A table with columns R, S, T, U, V, W, X, Y, Z. Rows 7 through 12 show data points for "Stopped buying wine" (Wine: 0, Clothes: 0) and "PREGNANT" status (1 or 0). The "W" column contains "Linear Combination (Prediction)" values (e.g., 0.88, 0.87, 0.72, 0.69, 0.96) and the "X" column contains "Squared Error" values (e.g., 0.01, 0.02, 0.08, 0.10, 0.00).
- Bottom:** A section labeled "Training Data" with tabs for "Training Data", "Training Data w Dummy Vars", and "Linear Model". It shows a sum of 0.37.

**Figura 6-14:** A previsão de erro padrão para a regressão linear

Para começar esse cálculo, crie uma nova aba na pasta de trabalho chamada **ModelCoefficientStandardError**. Agora, o que dificulta o erro padrão é que precisamos entender como os dados em treinamento para um coeficiente variam lado a lado com outras variáveis. O próximo passo para desvendar esse mistério é multiplicar o conjunto em treinamento como uma matriz gigante (geralmente chamada de **matriz de design**).

O produto da matriz de design (B8:U1007) com ela mesma constitui no que chamamos de matriz de **soma dos quadrados e produtos vetoriais**

(SSCP, *sum of squares and cross products*). Para ver como ela é, primeiro cole as linhas de cabeçalho para os dados em treinamento na aba ModelCoefficientStandardError em B1:U1 e transponha as linhas A2:A21. Isso inclui o cabeçalho Intercept.

Para multiplicar a matriz de design por ela mesma, você a insere na função MMULT do Excel, primeiro transposta, e depois invertida:

```
{=MMULT(TRANSPOSE('Linear Model'!B8:U1007), 'Linear Model'!B8:U1007)}
```

Já que essa função retorna uma matriz de tamanho variáveis-por-variáveis, na verdade, deve-se destacar toda a extensão de B2:U21 na aba ModelCoefficientStandardError e executar a função como uma fórmula array (veja o Capítulo 1 para mais sobre fórmula array).

Isso gera a aba mostrada na Figura 6-15.

Repare nos valores da matriz de soma dos quadrados e produtos vetoriais. Por toda a diagonal, você está contando combinações de cada variável com ela mesma — o mesmo que somar todos os 1 em cada coluna da matriz de design. O intercepto recebe 1000, por exemplo, na célula U21, porque nos dados em treinamento originais aquela coluna é formada por 1000.

Nas células não-diagonais, você acaba por contar as combinações entre preditores diferentes. Enquanto que Male e Female obviamente nunca combinam por design, Pregnancy Test e Birth Control aparecem juntos em linhas de seis clientes dos dados em treinamento.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	SSCP MATRIX	Male	Female	Home	Age	Pregnancy	Birth	Feminine	Folic	Prenatal	Prenatal	Body	Ginger	Sea	Stopped	buying	Smoking	Stopped	buying	Maternity	
2	Male	401	0	196	169	27	62	67	42	45	8	8	29	14	36	45	23	46	51	50	401
3	Female	0	495	239	207	37	63	61	54	71	9	8	31	14	44	47	32	69	59	71	495
4	Home	196	239	488	0	43	57	74	54	59	10	14	44	11	45	46	29	63	62	65	488
5	Apt	169	207	0	420	59	58	39	57	8	3	19	16	38	42	20	56	51	54	420	
6	Pregnancy Test	27	37	43	26	75	6	5	13	19	3	2	8	5	9	5	18	17	2	17	75
7	Birth Control	62	63	57	59	6	140	24	5	13	0	1	5	1	3	20	5	20	22	7	140
8	Feminine Hygiene	67	61	74	58	5	24	141	7	14	4	4	6	3	5	19	3	12	25	17	141
9	Folic Acid	42	54	54	39	13	5	7	106	22	3	1	11	5	14	4	12	25	4	23	106
10	Prenatal Vitamins	45	71	59	57	19	13	14	22	128	2	4	9	10	22	9	8	24	9	22	128
11	Prenatal Yoga	8	9	10	8	3	0	4	3	2	18	1	2	1	0	0	1	3	1	5	18
12	Body Pillow	8	8	14	3	2	1	4	1	4	1	18	0	0	2	0	1	5	1	4	18
13	Ginger Ale	29	31	44	19	8	5	6	11	9	2	0	69	1	6	7	8	8	5	17	69
14	Sea Bands	14	14	12	16	5	5	3	5	10	1	0	1	30	3	3	3	3	1	5	30
15	Stopped buying cig	36	44	45	38	9	3	5	14	22	0	2	6	3	92	0	10	20	6	19	92
16	Cigarettes	45	47	46	42	5	20	19	4	9	0	0	7	3	0	97	5	7	19	11	97
17	Smoking Cessation	23	32	29	20	18	5	3	12	8	1	1	8	3	10	5	60	13	2	18	60
18	Stopped buying wi	46	69	63	56	17	10	12	25	24	3	5	8	3	20	7	13	130	0	22	130
19	Wine	51	59	62	51	2	22	25	4	9	1	5	1	6	19	2	0	123	10	123	123
20	Maternity Clothes	50	71	65	54	17	7	17	23	22	5	4	17	5	19	31	18	22	30	331	331
21	Intercept	401	495	488	420	75	140	141	306	128	18	18	69	30	92	97	60	130	123	131	1000

Figura 6-15: A matriz SSCP

A matriz SSCP mostra uma parte da magnitude de cada variável e quanto elas se sobrepõem e se movem umas com as outras.

O cálculo do erro padrão do coeficiente usa o inverso da matriz SSCP. Para obter o inverso, cole os cabeçalhos das variáveis novamente abaixo na matriz SSCP em B24:U24 e em A25:A44. O inverso da matriz SSCP em B2:U21 é então calculado destacando B25:U44 e empregando a função MINVERSE como uma fórmula array:

{=MINVERSE(B2:U21)}

Isso gera a planilha mostrada da Figura 6-16.

Os valores exigidos no cálculo do erro padrão do coeficiente são os da diagonal da matriz inversa SSCP. Cada erro padrão do coeficiente é calculado como a previsão de erro padrão para o modelo inteiro (calculado como 0,37 na aba Linear Model na célula X5) modificado pela raiz quadrada do valor adequado da diagonal inversa SSCP.

Por exemplo, o erro padrão do coeficiente para Male seria a raiz quadrada da sua entrada Male-to-Male na matriz inversa SSCP (raiz quadrada de 0,0122) vezes a previsão de erro padrão.

Para fazer esse cálculo para todas as variáveis, numere cada variável começando com 1 em B46 até 20 em U46. O valor diagonal adequado agora pode ser lido por todos os preditores usando a fórmula INDEX. Por exemplo,

`INDEX(ModelCoefficientStdError!B25:B44, ModelCoefficientStdError!B46)` retorna a entrada diagonal Male-to-Male (veja mais sobre a fórmula `INDEX` no Capítulo 1).

**Figura 6-16:** O inverso da matriz SSCP

Pegando a raiz quadrada desse valor e multiplicando pela previsão de erro padrão, o erro padrão do coeficiente Male é calculado na célula B47 como:

```
= 'Linear Model'!$X5*SQRT(INDEX(ModelCoefficientStdError!B25:B44,  
ModelCoefficientStdError!B46))
```

Isso se transforma em 0,04 para que o modelo se ajuste no livro.

Arraste essa fórmula para a coluna U e obtenha todos os valores do erro padrão de coeficiente como mostra a Figura 6-17:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
24	INVERSE	Male	Female	Home	Apt	Prez-	Birth	Femirine	Folic	Pre-	Pre-	Pre-	Body	Ginger	Sea	Stopped	
41	Stopped buying wine	0.000	0.000	0.000	0.000	-0.001	0.000	0.000	-0.001	0.000	0.000	-0.001	0.000	0.001	-0.001	0.000	
42	Wine	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.001	0.000	0.000	0.000	0.001	0.000	0.000	0.000	
43	Maternity Clothes	0.000	0.000	0.000	0.000	0.000	0.001	0.000	-0.001	0.000	-0.001	-0.001	-0.001	0.000	-0.001	0.000	
44	Intercept	-0.010	-0.010	-0.012	-0.012	-0.001	-0.003	-0.001	-0.001	-0.001	0.000	0.001	0.001	0.001	-0.002	0.001	
45		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
46	Coefficient Std Error	0.04	0.0403	0.043	0.04	0.046	0.035	0.0343	0.04	0.04	0.09	0.09	0.047	0.07	0.0417	0.04	0.0516
47	ModelCoefficientStdError																

Figura 6-17: O erro padrão de cada coeficiente de modelo

Na aba Linear Model, nomeie A3 como **Coefficient Standard Error**. Copie os erros padrões do coeficiente e cole seus valores de volta na linha 3 da aba Linear Model (B3:U3).

Ufa! Daqui para frente é moleza. Sem mais matrizes matemáticas no restante do livro. Juro.

Agora você possui tudo o que precisa para calcular a **estatística t** de cada coeficiente (parecido com a estatística F do modelo inteiro da seção anterior). Você executará o que se chama de teste t bicaudal, o que significa que você calculará a probabilidade de obter um coeficiente no mínimo bem grande na direção **positiva** ou **negativa** se, na realidade, não houver relação entre a característica e a variável dependente.

A estatística t para o teste pode ser calculada na linha 4 como o valor absoluto do coeficiente normalizado pelo erro padrão dos coeficientes. Para a funcionalidade Male isso é:

$$=ABS(B2/B3)$$

Copie isso na coluna U em todas as variáveis.

O teste t pode ser chamado ao avaliar a **distribuição t** (outra distribuição estatística como a distribuição normal apresentada no Capítulo 4) no valor da estatística t para o seu valor particular de graus de liberdade. Nomeie a linha 5 como **t Test p Value**, e em B5 use a

fórmula TDIST para calcular a probabilidade de um coeficiente ser no mínimo desse tamanho dada a hipótese nula:

=TDIST(B4, \$Z3, 2)

O 2 na fórmula indica que você está executando o teste t bicaudal. Ao copiar essa fórmula por todas as variáveis e aplicar a formatação condicional nas células acima de 0,05 (5 por cento de probabilidade), pode-se ver quais funcionalidades não são significantes estatisticamente. Enquanto seus resultados podem variar baseados no ajuste do seu modelo, na pasta de trabalho na Figura 6-18, as colunas Female, Home e Apt são dadas como insignificantes.

		Male	Female	Home	Apt	Pregnancy Test	Birth Control	Feminine Hygiene	Folic Acid	Prenatal Vitamins	Prenatal Yoga	Body Pillow	Ginger Ale	Sugar
2	Model Coefficients	-0.10	-0.03	-0.03	-0.01	0.22	-0.27	-0.24	0.35	0.29	0.33	0.19	0.23	0.00
3	Coefficient Standard Error	0.041	0.040	0.043	0.043	0.046	0.035	0.034	0.039	0.036	0.089	0.089	0.047	0.00
4	t Statistic	2.39	0.6651	0.6516	0.307	4.654338	7.8738	6.939359	8.83	8.16017	3.645	2.166	4.882	2.00
5	t Test p Value	0.02	0.51	0.51	0.76	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00

Figura 6-18: ↑Female, ↑Home e ↑Apt são previsões insignificantes de acordo com o teste

É possível remover essas colunas do seu modelo nas execuções de treinamento no futuro.

Agora que você já aprendeu como avaliar o modelo usando testes estatísticos, vamos mudar nossa visão e começar a medir o desempenho do modelo ao fazer previsões reais em um conjunto de teste.

## Fazendo Previsões em Dados Novos e Medindo o Desempenho

A última seção foi toda sobre estatística. Podemos dizer que foi um trabalho manual. Não foi o dia mais divertido da sua vida, mas validar o benefício do ajuste e da significância é uma habilidade importante para

dominar. Mas agora é hora de levar esse modelo para a pista de corrida e se divertir um pouco!

Como você sabe se seu modelo linear vai, de fato, fazer uma previsão correta no mundo real? Afinal, seu conjunto em treinamento não engloba o registro de todos os clientes possíveis, e seus coeficientes são direcionados a ajustar o conjunto em treinamento (apesar de ter feito o trabalho correto, o conjunto em treinamento quase lembra o mundo no geral).

Para ter uma melhor ideia de como o modelo desempenhará no mundo real, você deveria passar alguns clientes pelo modelo que não foram usados no processo de treinamento. Você verá um conjunto separado de exemplos usados para testar um modelo geralmente chamado de **conjunto de validação**, **conjunto de teste** ou **conjunto holdout**.

Para agrupar seu conjunto de teste, você pode retornar ao banco de dados do cliente e selecionar outro conjunto de dados de clientes aleatórios (prestando bastante atenção para não puxar os mesmos clientes usados no treinamento). Agora, como vimos anteriormente, 6% das clientes da RetailMart estão grávidas, então, se você selecionasse aleatoriamente mil clientes do banco de dados, dificilmente 60% delas estariam grávidas.

Como houve a sobreamostragem da classe pregnant ao treinar o modelo, a relação de grávidas será de 6% para que as medidas de precisão do modelo do teste sejam exatas para quando o modelo desempenhar em contextos reais.

Na planilha da RetailMart disponível para download que acompanha este capítulo, você encontrará uma aba chamada Test Set, que constitui de milhares de linhas de dados idênticos aos dados em treinamento. As primeiras 60 clientes estão grávidas, enquanto as outras 940 não estão (veja a Figura 6-19).

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	Male	Female	Home	Apt	Pregnancy Test	Birth Control	Feminine Hygiene	Folic Acid	natal Vit.	natal Yoga	Body Pillow	Ginger Ale	Sea Bands	Stopped buying ciggies	Cigarettes	Smoking Cessation	buying wine	Stopped Wine	Maternity Clothes	PREG NANT	
2	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	1	1	0	1	
3	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
4	1	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1
5	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1
6	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
7	1	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0	1
8	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	1
9	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Sum=0																					

**Figura 6-19:** Teste do conjunto de dados

Assim como fez na aba Linear Model, execute esses dados novos pelo modelo levando uma combinação linear de dados e coeficientes de clientes e adicionando ao intercepto. Posicionando essa previsão na coluna V, você tem a seguinte fórmula para o primeiro cliente na linha 2 (como o teste não tem a coluna Intercept, você a adiciona separadamente):

```
=SUMPRODUCT('Linear Model'!B$2:T$2,'Test Set'!A2:S2)+'Linear Model'!U$2
```

Copie esse cálculo em todos os clientes. A planilha resultante se parece com a Figura 6-20.

RetailMart.xlsx

	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	Prenatal Yoga	Body Pillow	Ginger Ale	Sea Bands	Stopped buying ciggies	Cigarettes	Smoking Cessation	Stopped buying wine	Wine	Maternity Clothes		PREGNANT	Linear Prediction
2	0	0	0	1	0	0	0	1	1	0		1	0.60
3	0	0	0	0	0	0	0	0	0	0		1	0.46
4	0	0	0	0	1	0	0	0	0	0		1	0.52
5	0	0	0	0	0	0	0	1	0	0		1	0.55
6	0	0	0	0	0	0	0	0	0	0		1	0.37
7	0	0	1	0	0	0	0	1	0	0		1	1.12
8	0	0	0	0	0	0	1	0	1			1	0.86
9	0	0	0	0	0	0	0	0	0	0		1	0.46
10	0	0	0	0	0	0	0	0	0	0		1	0.46
11	0	0	0	1	1	0	0	0	1			1	0.67
12	0	0	1	0	0	0	0	0	0	0		1	0.70
13	0	0	0	1	0	0	0	0	0	0		1	0.59
14	0	0	0	0	0	0	0	0	0	0		1	0.46
15	0	0	0	0	0	0	0	0	0	0		1	0.67
16	0	0	0	0	0	0	0	0	0	0		1	0.66
17	0	0	0	0	0	0	0	0	0	0		1	1.12

Figura 6-20: Previsões no conjunto de testes

É possível ver na Figura 6-20 que o modelo identificou muitos dos grupos de grávidas com previsões mais próximas a 1 do que 0. Os valores da previsão mais altos são para os grupos que compraram produtos relacionados à gravidez, como ácido fólico ou vitaminas para pré-natal.

Por outro lado, fora das 60 grávidas, há algumas que nunca compraram nada que indicassem estarem grávidas. Claro, elas *não* compraram álcool ou tabaco, mas como o escore de gravidez indica, não comprar alguma coisa *não* significa nada.

Em contrapartida, há alguns erros se você olhar as previsões para as mulheres que não estão grávidas. Por exemplo, se você está acompanhando a pasta de trabalhos, na linha 154 uma cliente não-grávida comprou roupas para maternidade e parou de comprar cigarros — o modelo atribuiu um escore de 0,76 a ela.

Fica claro que se você for usar essas previsões em campanhas reais de marketing, você precisa estabelecer um escore limite para quando presumir que alguém esteja grávida e alcance essa pessoa com estratégias comerciais. Talvez você só consiga atingir clientes com materiais de

marketing se elas tiverem um escore de 0,8 ou acima. Talvez esse corte devesse ser de 0,95 para que você tenha uma certeza extra.

Para estabelecer um limite na classificação, você precisa verificar as trocas na métrica do desempenho do modelo. A maioria das métricas de desempenho do modelo preditivo é baseada em contagens e relações de quatro valores que vêm das previsões no nosso conjunto de teste:

- Positivos verdadeiros — Nomear uma cliente grávida como grávida
- Negativos verdadeiros — Nomear uma cliente não-grávida como não-grávida
- Positivos falsos (também chamados de *erro tipo I*) — Chamar uma cliente não-grávida de grávida. Na minha experiência, esse positivo falso específico é muito agressivo. Não tente isso em casa.
- Negativos falsos (também chamados de *erro tipo II*) — Não saber como identificar uma cliente grávida. Isso não é nem um pouco agressivo na minha opinião.

Como você verá, há muitas métricas de desempenhos diferentes para um modelo preditivo. Todos eles se parecem um pouco com comida mexicana — são combinações dos mesmos quatro ingredientes listados anteriormente.

### *Estabelecendo os Valores de Corte*

Crie uma nova planilha com o nome **Performance**. O menor valor que poderia ser usado como corte entre uma grávida e uma não-grávida é o menor valor de previsão do conjunto de teste. Nomeie A1 como **Min Prediction** e em A2 você pode calcular desta forma:

```
=MIN('Test Set'!V2:V1001)
```

Da mesma maneira, o valor de corte mais alto seria a previsão máxima do conjunto de teste. Nomeie A4 como **Max Prediction** e em A5 você

pode calcular desta forma:

```
=MAX('Test Set'!V2:V1001)
```

Os valores retornados são  $-0,35$  e  $1,25$ , respectivamente. Lembre-se que sua regressão linear pode fazer previsões abaixo de 0 e acima de 1, pois ela não está, de fato, retornando probabilidades para as classes (voltaremos a esse tópico mais tarde com outro modelo).

Na coluna B, adicione o cabeçalho Probability Cutoff for Pregnant Classification e abaixo especifique um limite de valores de corte começando com  $-0,35$ . Na planilha mostrada na Figura 6-21, os valores de corte foram escolhidos para aumentarem em  $0,05$  os incrementos até chegar a  $1,25$  (apenas insira os três primeiros, destaque-os e arraste-os para preencher o restante).

Uma outra opção seria especificar cada valor de previsão única do conjunto de teste como corte se você quisesse ser minucioso. Não seria necessário mais nada.

### **Precisão (Valor Preditivo Positivo)**

Vamos preencher algumas métricas de desempenho de modelo para cada valor de corte usando as previsões dos dados do Conjunto de Teste começando por **precisão**, também conhecida como **valor preditivo positivo**.

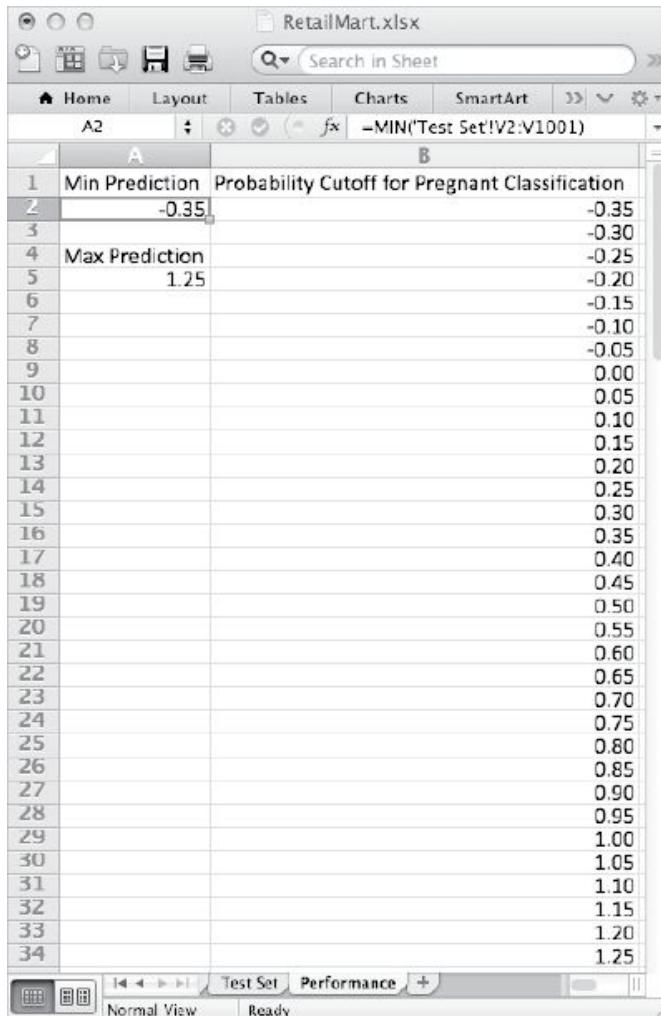
Precisão é a medida de quantos grupos de grávidas identificamos corretamente em todos os grupos em que os modelos apontam como grávidas. Em linguagem de negócios, precisão é a porcentagem de peixes na sua rede que são atuns e não golfinhos.

Nomeie a coluna C como **Precision**. Observe o escore de corte em B2 de  $-0,35$ . Qual a precisão do nosso modelo se não considerarmos nenhum escore com mínimo de  $-0,35$  estarem grávidas?

Para calcular isso, podemos ir para a aba “Test Set” e contar a quantidade de casos em que os grupos de grávidas tiveram um escore maior ou igual a  $-0,35$  dividido pelo número total de linhas com escore

maior do que -0,35. Usando a fórmula COUNTIFS para verificar os escores atuais e as previsões, a fórmula na célula C2 seria assim:

```
=COUNTIFS('Test Set'!$V$2:$V$1001,">=" & B2,  
'Test Set'!$U$2:$U$1001,"=1")/COUNTIF('Test Set'!$V$2:$V$1001,">="  
& B2)
```

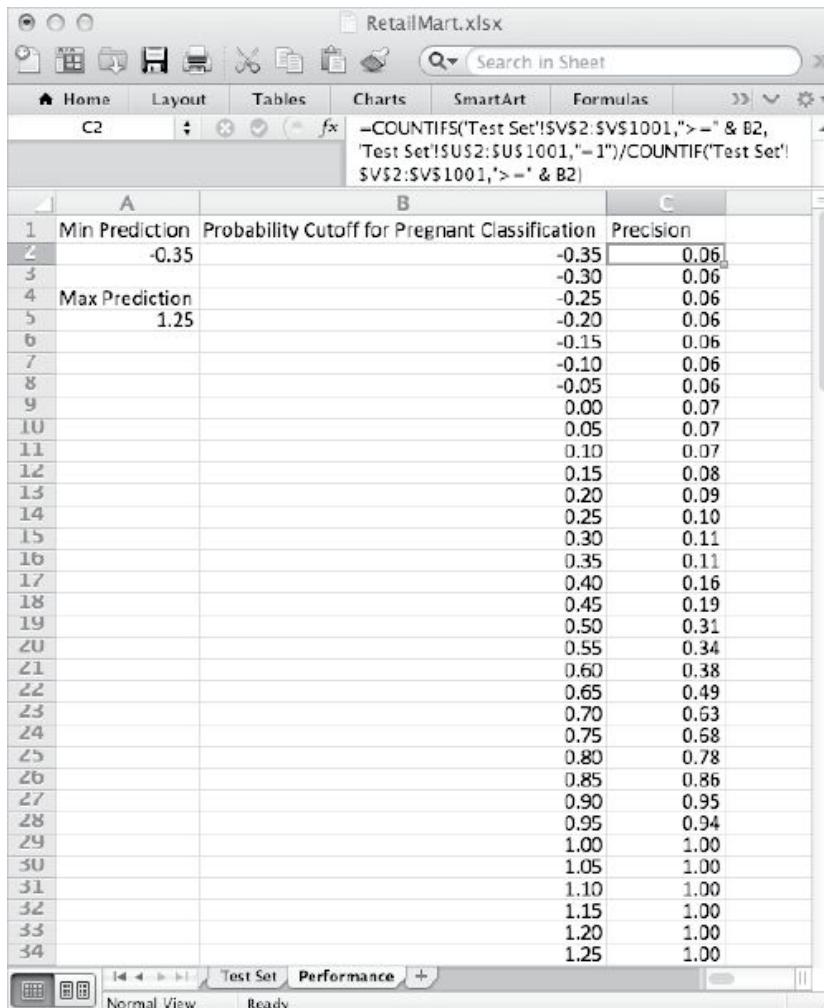


**Figura 6-21:** Valores de corte para a classificação da gravidez

A primeira declaração COUNTIFS na fórmula corresponde à gravidez atual e à previsão do modelo. O COUNTIFS no denominador apenas se refere a quem teve um escore maior do que -0,35 independente da gravidez. Você pode copiar essa fórmula em todos os grupos que avaliar.

Como visto na Figura 6-22, a precisão do modelo aumenta com o valor de corte, e no valor de corte 1, o modelo se torna completamente preciso.

Um modelo completamente preciso identifica apenas clientes grávidas como grávidas.



**Figura 6-22:** Cálculos de precisão no conjunto de dados

### Especificidade (Taxa de verdadeiro negativo)

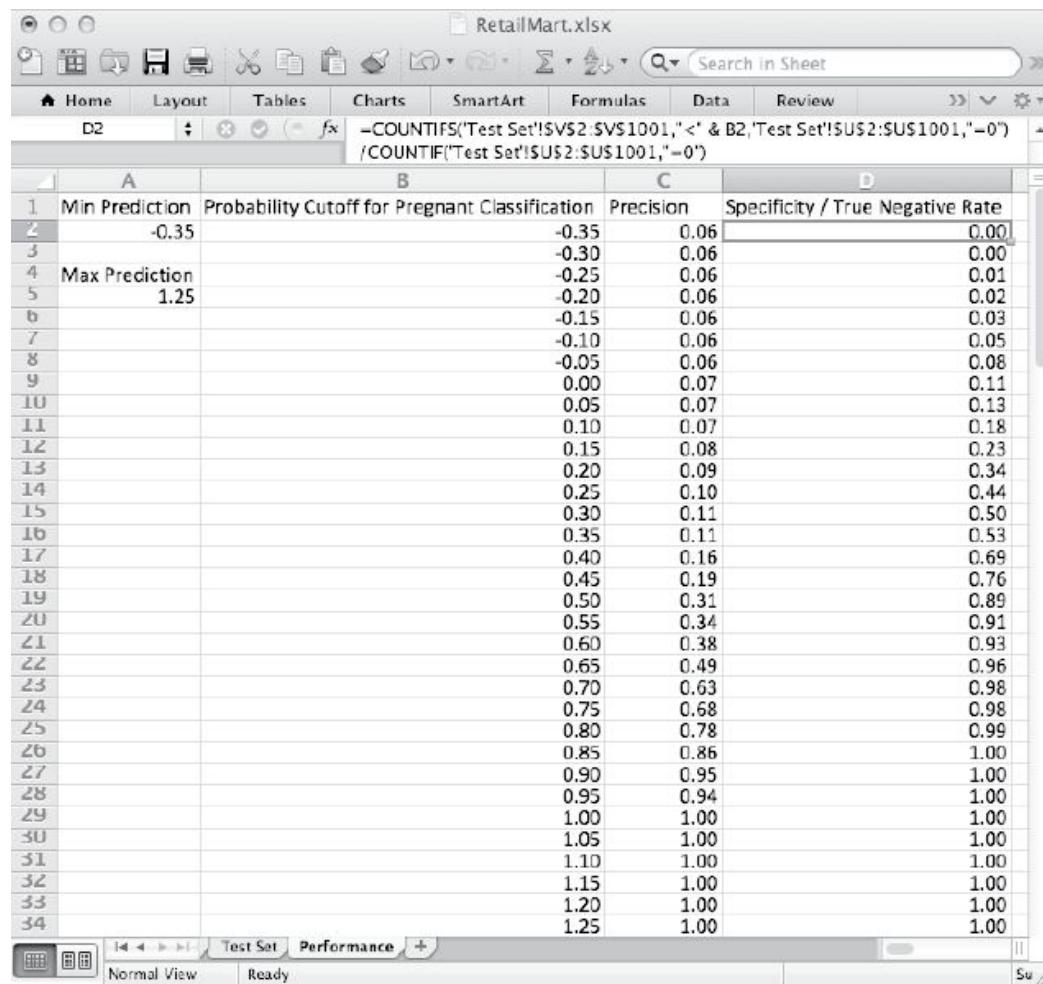
Especificidade é outro **desempenho** métrico que aumenta o valor de corte. Também chamado de Taxa de verdadeiro negativo, é uma contagem de quantas clientes não-grávidas foram previstas corretamente como tal (verdadeiros negativos) divididas pela quantidade total de casos de gravidez.

Ao nomear a coluna D como Specificity/True Negative Rate, você pode calculá-lo em D2 usando COUNTIFS no numerador para contar negativos

verdadeiros, e COUNTIFS no denominador para contar o total de clientes que não estão grávidas:

```
=COUNTIFS('Test Set'!$V$2:$V$1001,"<" & B2,
'Test Set'!$U$2:$U$1001,"=0")/COUNTIF('Test
Set'!$U$2:$U$1001,"=0")
```

Ao copiar esse cálculo para os outros valores de corte, você deverá vê-lo aumentar (veja a Figura 6-23). Uma vez que o valor de corte 0,85 é alcançado, 100% das clientes não-grávidas no conjunto de teste são devidamente previstas.



**Figura 6-23:** Cálculos de especificidade no conjunto de dados

### Taxa de Falso Positivo

A taxa de **falso positivo** é uma métrica comum destinada a entender o desempenho do modelo. E, já que você tem a taxa de verdadeiro negativo, isso pode ser facilmente calculado como um menos a taxa de verdadeiro negativo. Nomeie a coluna E ***False Positive Rate/(1 – Specificity)*** e preencha as células com um menos o valor da célula adjacente em D. Para E2, ela é escrita desta forma:

=1-D2

Ao copiar essa fórmula, percebe-se que o valor de corte aumenta e obtém-se menos positivos falsos. Em outras palavras, há uma menor quantidade de erro tipo I (classificar as clientes de grávidas quando elas não estão).

### ***Taxa de Verdadeiro Positivo/Recall/Sensibilidade***

A métrica que pode ser calculada no desempenho do seu modelo é chamada de ***taxa de verdadeiro positivo***. E ***recall***. E ***sensibilidade***. Nossa. Eles deveriam escolher apenas um nome e ficar com ele.

Uma taxa de verdadeiro positivo é a razão de mulheres grávidas corretamente identificadas dividida pelo total de mulheres grávidas atuais dentro do conjunto de testes. Nomeie a coluna F como ***True Positive Rate/Recall/Sensitivity***. Em F2, pode-se calcular a taxa verdadeira positiva com um valor de corte de -0,35 assim:

```
=COUNTIFS('Test Set'!$V$2:$V$1001,">=" & B2,  
'Test Set'!$U$2:$U$1001,"=1")/COUNTIF('Test  
Set'!$U$2:$U$1001,"=1")
```

Olhando novamente a coluna da taxa de verdadeiro negativo, esse cálculo é exatamente o mesmo exceto pelos “<” virarem “>=” e 0s virarem 1s.

Ao copiar essa fórmula, pode-se ver que, conforme o corte aumenta, algumas das mulheres grávidas param de ser identificadas como tal (esses são erros tipo II) e a Taxa de Verdadeiro Positivo cai. A Figura 6-24 mostra as taxas positivas verdadeiras e falsas nas colunas E e F:

RetailMart.xlsx

	A	B	C	D	E	F
1	Min Prediction	Probability Cutoff for Pregnant Classification	Precision	Specificity / True Negative Rate	False Positive Rate (1 - Specificity)	True Positive Rate / Recall / Sensitivity
2	-0.35	-0.35	0.06	0.00	1.00	1.00
3		-0.30	0.06	0.00	1.00	1.00
4	Max Prediction	-0.25	0.05	0.01	0.99	1.00
5		-0.20	0.06	0.02	0.98	1.00
6		-0.15	0.05	0.03	0.97	1.00
7		-0.10	0.05	0.05	0.95	1.00
8		-0.05	0.05	0.08	0.92	1.00
9		0.00	0.07	0.11	0.89	1.00
10		0.05	0.07	0.13	0.87	0.98
11		0.10	0.07	0.18	0.82	0.98
12		0.15	0.08	0.23	0.77	0.98
13		0.20	0.09	0.34	0.66	0.97
14		0.25	0.10	0.44	0.56	0.95
15		0.30	0.11	0.50	0.50	0.95
16		0.35	0.11	0.53	0.47	0.95
17		0.40	0.15	0.69	0.31	0.90
18		0.45	0.19	0.76	0.24	0.87
19		0.50	0.31	0.89	0.11	0.78
20		0.55	0.34	0.91	0.09	0.75
21		0.60	0.38	0.93	0.07	0.72
22		0.65	0.49	0.96	0.04	0.65
23		0.70	0.63	0.98	0.02	0.58
24		0.75	0.68	0.98	0.02	0.53
25		0.80	0.78	0.99	0.01	0.47
26		0.85	0.85	1.00	0.00	0.40
27		0.90	0.95	1.00	0.00	0.33
28		0.95	0.94	1.00	0.00	0.28
29		1.00	1.00	1.00	0.00	0.23
30		1.05	1.00	1.00	0.00	0.18
31		1.10	1.00	1.00	0.00	0.15
32		1.15	1.00	1.00	0.00	0.03
33		1.20	1.00	1.00	0.00	0.03
34		1.25	1.00	1.00	0.00	0.02

Figura 6-24: As taxas positivas falsas e verdadeiras

## Avaliando o Compromisso da Métrica e a Curva Característica de Operação do Receptor

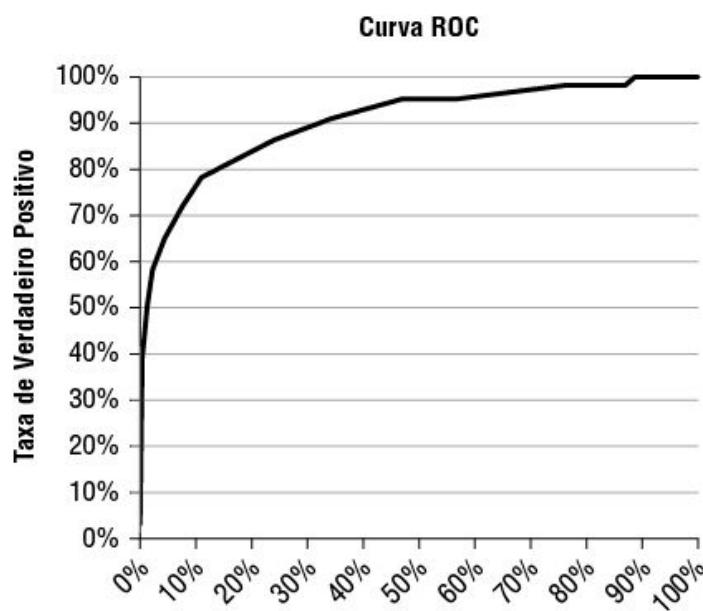
Ao escolher um valor limite para um classificador binário, é importante selecionar o melhor equilíbrio com compromisso entre essas métricas de desempenho. Quanto maior o corte, mais preciso o modelo, porém mais baixo o recall, por exemplo. Uma das visualizações mais comuns usadas para avaliar essas trocas de desempenho é a curva característica de operação do receptor (ROC, do inglês *receiver operating characteristic*). A curva ROC é apenas um gráfico da taxa de falso positivo versus a taxa de verdadeiro positivo (colunas E e F na planilha Performance).

## POR QUE ELE É CHAMADO DE CARACTERÍSTICA DE OPERAÇÃO DO RECEPTOR?

O motivo de um gráfico tão simples possuir um nome tão complexo é que ele foi desenvolvido durante a Segunda Guerra Mundial por engenheiros de sistemas em vez de publicitários premeditando quando as clientes estão grávidas.

Esse pessoal estava usando sinais para detectar os inimigos e seus equipamentos no campo de batalha, e eles queriam visualizar melhor a troca entre identificar correta e incorretamente alguém como inimigo.

Para introduzir esse gráfico, destaque os dados nas colunas E e F e selecione **straightlined scatter plot** no Excel (veja o Capítulo 1 para mais sobre inserir diagramas e gráficos). Com um pouco de formatação (ajustando os eixos entre 0 e 1 aumentando a fonte), a curva ROC se parece com a Figura 6-25.



**Figura 6-25:** A curva ROC para a regressão linear

Essa curva permite avaliar a taxa de falso positivo que está associada à Taxa de Verdadeiro Positivo a fim de entender suas opções. Por exemplo, na Figura 6-25, observa-se que o modelo é capaz de identificar 40% das

clientes grávidas usando um corte de 0,85 sem um único positivo falso. Que bom!

E se estiver tudo bem para você enviar alguns cupons com itens relacionados à gravidez para um grupo de não-grávidas, o modelo poderia alcançar 75% da taxa de positivo verdadeiro com apenas 9% da taxa de positivo falso.

Onde você decide ajustar o grupo para agir a favor do escore das grávidas é uma **decisão de negócios** e não apenas analítica. Se houver pequeno impacto negativo em prever que alguém estivesse grávida, logo uma baixa precisão talvez fosse um bom compromisso por uma alta taxa de positivos verdadeiros. Mas se estiver prevendo a probabilidade de padrão para aplicações de empréstimo, você vai querer que a especificidade e a precisão sejam **altas**, certo? Em um final alternativo, se um modelo como esse estivesse sendo usado para validar a legitimidade de ameaças estrangeiras, baseada em um corpo de inteligência, então você esperaria que o operador do modelo quisesse um nível de precisão muito alto antes de convocar um ataque de drones.

Portanto, se estamos pensando em enviar cupons pelo correio, aprovar empréstimos ou lançar bombas, o equilíbrio a ser determinado nessas métricas de desempenho é uma decisão estratégica.

### COMPARANDO UM MODELO COM O OUTRO

Como veremos mais adiante, a curva ROC é boa para escolher um modelo preditivo em vez de outro. Idealmente, a curva ROC se moveria para 1 no eixo y o mais rápido que pudesse e ficaria lá durante todo o gráfico. Portanto, o modelo que mais se parece com ele (também dizemos que possui a maior **área abaixo da curva** ou AUC, do inglês *area under the curve*) é considerado superior.

Tudo bem! Então, agora você já executou um teste de dados em um modelo, fez algumas previsões, computou seu desempenho no conjunto de teste para diferentes valores de corte e visualizou esse desempenho com a curva ROC.

Mas, para comparar o desempenho do modelo, você precisará de outro modelo para essa competição.

## Prevendo Clientes Grávidas na RetailMart Usando Regressão Logística

Se reparar nos valores previstos que saem da sua regressão linear, fica claro que enquanto seu modelo é útil para classificação, os valores de previsão certamente não são as probabilidades de classe. Não é possível engravidar com uma probabilidade de 125% ou -35%.

Mas será que existe um modelo cujas previsões são, de fato, probabilidades de classe? Uma vez que possamos construir esse modelo, o chamamos de *regressão logística*.

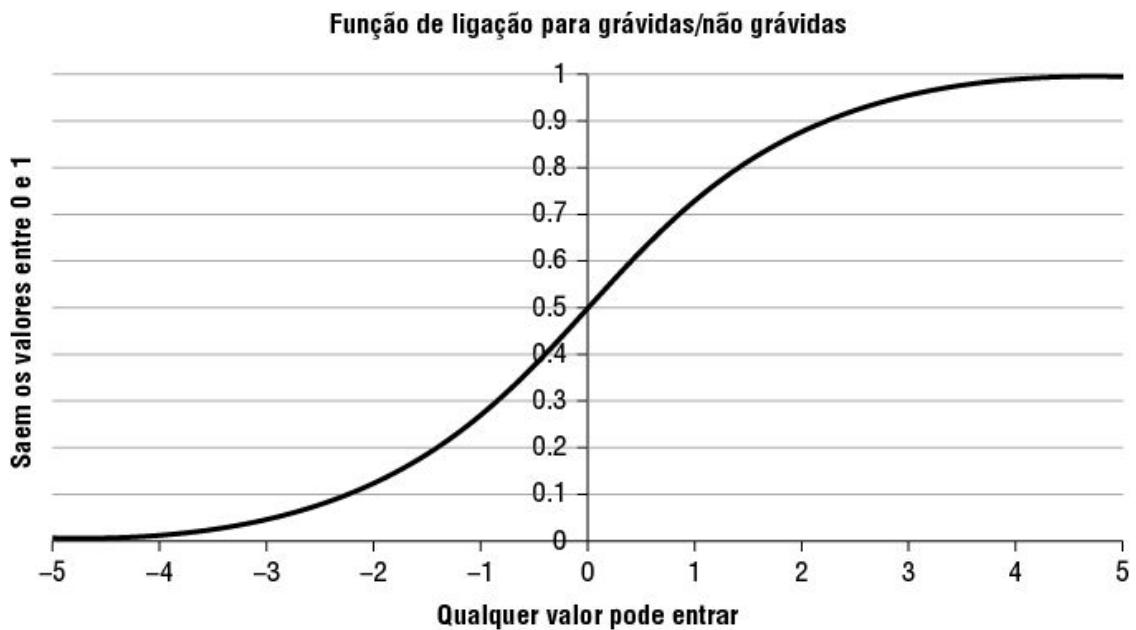
## Primeiro Você Precisa de uma Função de Ligação

Reflita sobre as previsões que seu modelo linear produz atualmente. Existe uma fórmula para compactar esses números e fazê-los ficar entre 0 e 1? Acontece que esse tipo de função é chamada de *função de ligação* e existe uma perfeita para isso:

$$\exp(x) / (1 + \exp(x))$$

Nessa fórmula,  $x$  é a combinação linear da coluna W na aba Linear Model, e  $\exp$  é a função exponencial. A função exponencial  $\exp(x)$  é, na verdade, a constante matemática  $e$  ( $2,71828\dots$ ) é igual a pi, mas um pouco menor) elevada à potência  $x$ .

Observe o gráfico dessa função na Figura 6-26.



**Figura 6-26:** A função de ligação

Essa função de ligação se parece com uma grande letra S. Qualquer valor a partir da multiplicação dos coeficientes de modelo vezes uma linha dos dados do cliente e a saída é um número entre 0 e 1. Mas por que essa função esquisita possui essa forma?

Bem, apenas arredonde e para 2,7 e imagine um caso em que a entrada para essa função seja bastante ampla, digamos 10. Logo, a função de ligação é:

$$\exp(x) / (1 + \exp(x)) = 2,7^{10} / (1 + 2,7^{10}) = 20589/20590$$

Bem, isso é praticamente 1, portanto percebemos que enquanto x fica maior, o 1 no denominador passa a não importar muito. Mas e se o x ficar negativo? Observe -10:

$$\exp(x) / (1 + \exp(x)) = 2,7^{-10} / (1 + 2,7^{-10}) = 0.00005/1.00005$$

Bem, é apenas um 0 na maior parte. Nesse caso, o 1 no denominador passa a ter mais importância e os números menores ficam em torno de 0.

Não é prático? Na verdade, essa função de ligação tem sido tão útil que passou a ser chamada de função “logística”.

## Associando a Função Logística e Reotimizando

Agora crie uma cópia da aba Linear Model na planilha e nomeie de ***Logistic Link Model***. Apague todos os dados de teste estatístico da planilha já que eles são aplicáveis somente na regressão linear. Especificamente, destaque e apague as linhas 3 até 5 e limpe todos os valores do topo das colunas W até Z, exceto pelo marcador de posição Sum Squared Error. Limpe também a coluna do erro ao quadrado e renomeie ***Prediction (after Link Function)***. Veja a Figura 6-27 e verifique como a planilha deve ficar.

	P	Q	R	S	T	U	V	W	X
1	Cigarettes	Smoking Cessation	Stopped buying wine	Maternity Wine	Maternity Clothes	Intercept	Sum Squared Error		
2	-0.16	0.16	0.19	-0.21	0.24	0.48			
3									
4	Cigarettes	Smoking Cessation	Stopped buying wine	Maternity Wine	Maternity Clothes	Intercept	PREGNANT	Linear Combination	Prediction (after Link Function)
5	0	0	0	0	0	1	1	0.88	
6	0	0	0	0	0	1	1	0.87	
7	0	0	0	0	0	1	1	0.72	
8	0	0	0	0	0	1	1	0.69	
9	0	0	1	0	0	1	1	0.96	
10	0	0	0	0	0	1	1	0.88	
11	0	0	0	0	0	1	1	0.72	
12	0	0	0	0	1	1	1	0.67	
13	0	0	0	0	0	1	1	0.66	
14	0	0	0	0	1	1	1	0.96	
15	0	0	1	0	0	1	1	0.62	

**Figura 6-27:** A planilha inicial do modelo logístico

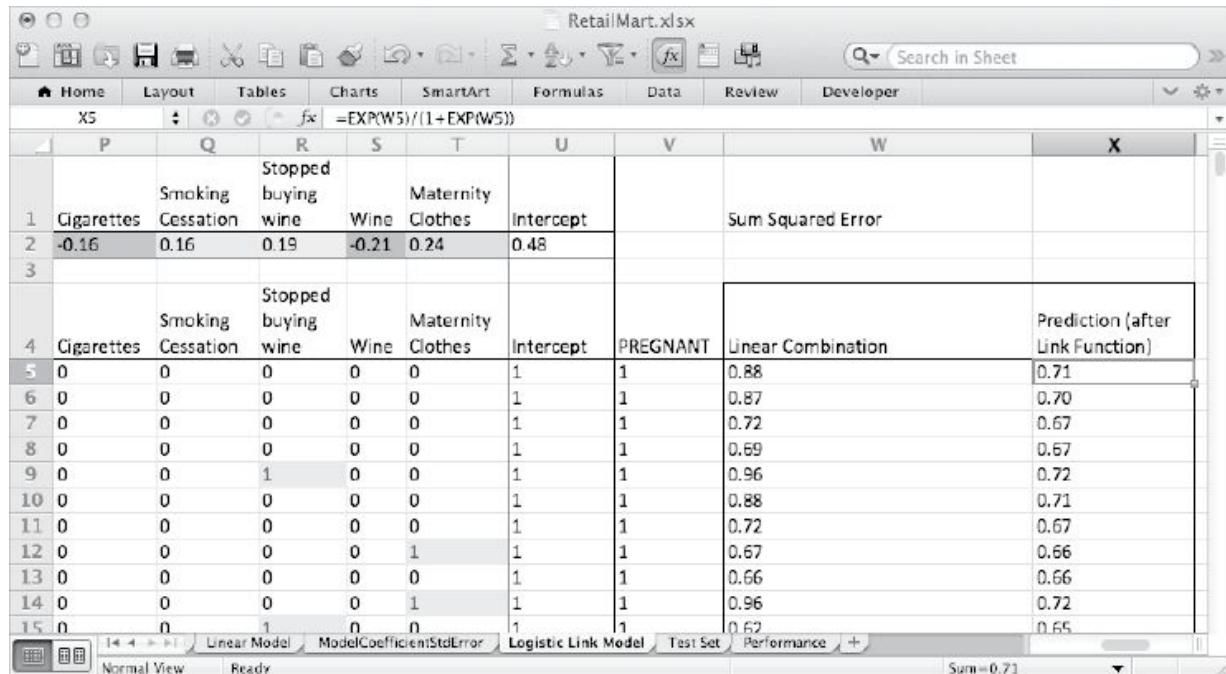
Você usará a coluna X para absorver a combinação linear dos coeficientes e dados da coluna W e colocá-los em sua função logística. Por exemplo, a primeira linha dos dados modelados do cliente seria enviada por meio da função logística ao colocar esta fórmula na célula X5:

$$=\text{EXP}(\text{W5}) / (1+\text{EXP}(\text{W5}))$$

Se copiar essa fórmula abaixo da coluna, verá que todos os valores novos estarão entre 0 e 1 (veja a Figura 6-28).

## NOTA

Sua planilha pode conter alguns valores diferentes nas colunas W e X de início, já que os coeficientes de modelo estão vindo do algoritmo evolucionário executado na aba anterior.



	P	Q	R	S	T	U	V	W	X
1	Cigarettes	Smoking Cessation	Stopped buying wine	Wine	Maternity Clothes	Intercept		Sum Squared Error	
2	-0.16	0.16	0.19	-0.21	0.24	0.48			
3									
4	Cigarettes	Smoking Cessation	Stopped buying wine	Wine	Maternity Clothes	Intercept	PREGNANT	Linear Combination	Prediction (after Link Function)
5	0	0	0	0	0	1	1	0.88	0.71
6	0	0	0	0	0	1	1	0.87	0.70
7	0	0	0	0	0	1	1	0.72	0.67
8	0	0	0	0	0	1	1	0.69	0.57
9	0	0	1	0	0	1	1	0.96	0.72
10	0	0	0	0	0	1	1	0.88	0.71
11	0	0	0	0	0	1	1	0.72	0.67
12	0	0	0	0	1	1	1	0.67	0.66
13	0	0	0	0	0	1	1	0.66	0.66
14	0	0	0	0	1	1	1	0.96	0.72
15	0	0	1	0	0	1	1	0.62	0.55

Figura 6-28: Os valores por toda a função logística

Entretanto, a maioria das previsões parecem ser regulares, entre 0,4 e 0,7. Bem, isso acontece porque não otimizamos nossos coeficientes na aba “Linear Model” para esse novo tipo de modelo. Precisamos otimizar outra vez.

Então posicione novamente a coluna do erro ao quadrado na coluna Y, sendo que, desta vez, o cálculo do erro usará as previsões geradas na função de ligação na coluna X:

$$= (V5 - X5)^2$$

Você realizará a soma tal como o modelo linear na célula X1 como:

$$=\text{SUM}(Y5:Y1004)$$

Agora pode minimizar a soma dos quadrados nesse novo modelo usando a mesma configuração do Solver (veja a Figura 6-29) no modelo linear. Exceto que, se fizer experimentos com os limites variáveis, verá que é melhor aumentá-los um pouco para um modelo logístico. Na Figura 6-29, os limites foram configurados para manter cada coeficiente entre -5 e 5.

Uma vez que tenha reotimizado para a nova função de ligação, você verá que suas previsões sobre os dados em treinamento se encontram entre 0 e 1, com muitas previsões sendo direcionadas a 0 ou a 1. Como pode ser visto na Figura 6-30, a partir de uma perspectiva estética, estas previsões agradam mais do que as de uma regressão linear.

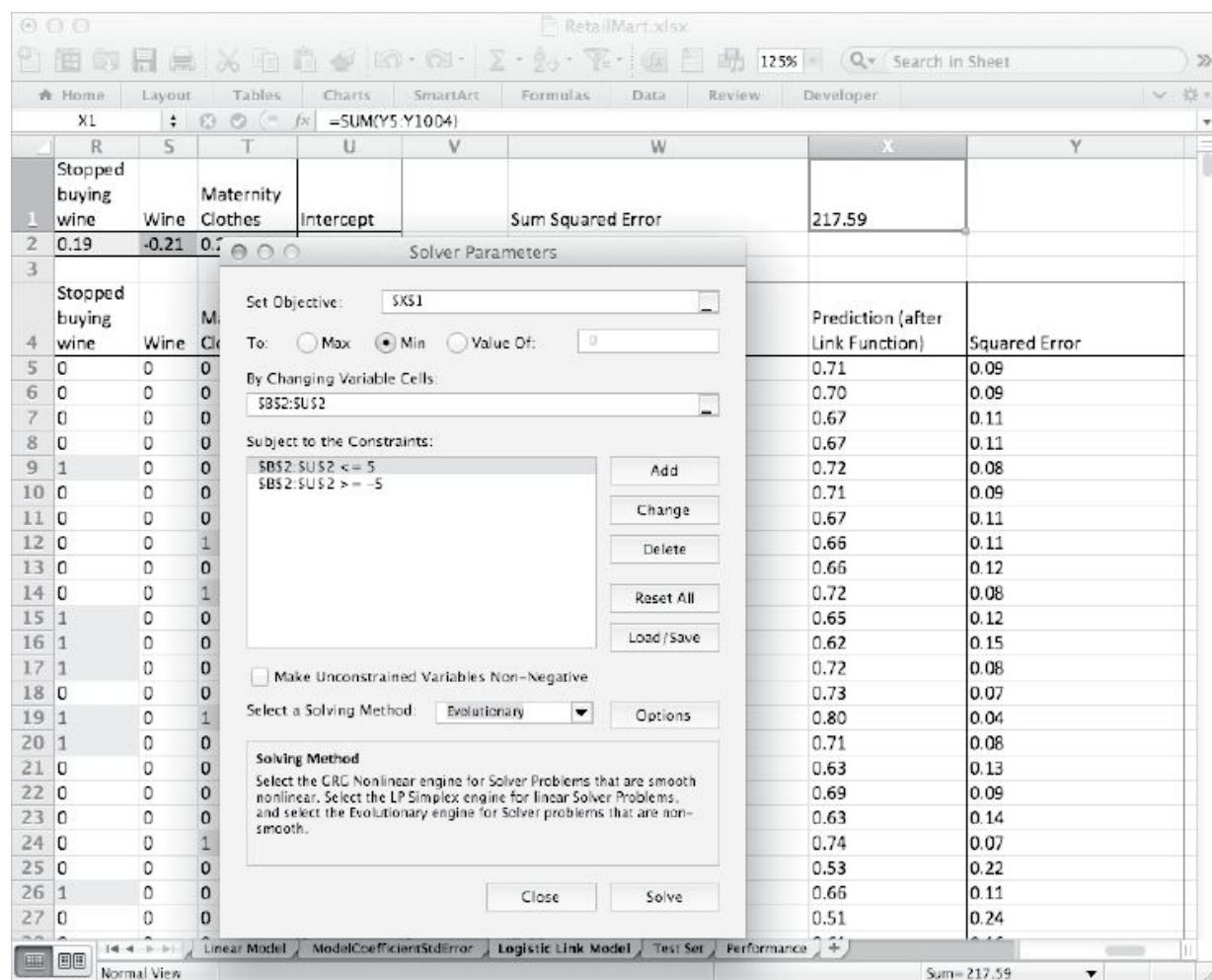


Figura 6-29: Configuração idêntica do Solver para o modelo logístico

The screenshot shows a Microsoft Excel spreadsheet titled "RetailMart.xlsx". The data is organized into several tabs at the bottom: Training Data, Training Data w/ Dummy Vars, Linear Model, Model Coefficient Std Error, Logistic Link Model, and Test Set. The "Test Set" tab is active, showing a table with 13 rows of data. The columns include "Smoking Cessation", "Stopped buying wine", "Wine", "Maternity Clothes", "Intercept", and "PREGNANT". The "Linear Combination" column contains values like 5.42, 5.08, 4.12, etc. The "Prediction (after Link Function)" column contains values like 1.00, 0.99, 0.98, etc. The "Squared Error" column contains values like 0.00, 0.00, 0.00, etc. A formula bar at the top shows =SUM(Y5:Y1004). The status bar at the bottom indicates "Sum = 116.98".

**Figura 6-30:** Modelo logístico ajustado

## Cozinhando uma Regressão Logística Real

A verdade é que, a fim de realizar uma regressão logística real que gere probabilidades precisas e imparciais, você não pode, por motivos que estão além do escopo deste livro, minimizar a soma dos erros quadrados.

Em vez disso, ajuste o modelo ao encontrar os coeficientes de modelo que maximizam a probabilidade conjunta de como você tratou esse conjunto em treinamento a partir do banco de dados da RetailMart, dado que o modelo expõe a verdade com precisão.

Portanto, qual a probabilidade de uma linha em treinamento dado um conjunto de parâmetros de modelo logístico? Para uma dada linha no conjunto em treinamento,  $p$  representa a probabilidade de classe que o seu modelo logístico está dando à coluna X. Seja  $y$  o valor atual de gravidez armazenado na coluna V. A probabilidade daquela linha em treinamento, dados os parâmetros do modelo é:

$$p^y (1-p)^{1-y}$$

Para uma cliente grávida (a coluna V é 1) com uma previsão de 1 (a coluna X possui um 1), a probabilidade desse cálculo é, também, 1. Porém, se as previsões para uma cliente grávida fossem 0, o cálculo

anterior seria 0 (insira os números e verifique). Assim, a probabilidade de cada linha é maximizada quando as previsões e os valores reais estiverem alinhados.

Presumindo que cada linha de dados é independente (veja o Capítulo 3 para mais sobre independência), como no caso de qualquer retirada aleatória de um banco de dados, pode-se calcular o registro da probabilidade conjunta dos dados ao fazer o registro de cada uma das probabilidades e as somando. O registro da equação anterior, usando as mesmas regras vistas na seção Controle de Fluxo do ponto-flutuante no Capítulo 3 é:

$$y * \ln(p) + (1-y) * \ln(1-p)$$

O registro da probabilidade é próximo a 0 quando a fórmula anterior for próxima de 1 (ou seja, quando o modelo se ajusta bem).

Em vez de minimizar a soma dos erros quadrados, pode-se calcular esse valor de registro de probabilidade para cada previsão e somá-las. Os coeficientes de modelo que *maximizarem* a probabilidade conjunta dos dados serão os melhores.

De início, faça uma cópia da aba Logistic Link Model e a nomeie de *Logistic Regression*. Na coluna Y, direcione a coluna dos erros quadrados para ler Log Likelihood. Na célula Y5, o primeiro registro de probabilidade pode ser calculado assim:

$$=IFERROR(V5 * LN(X5) + (1-V5) * LN(1-X5), 0)$$

Todo o cálculo do registro de probabilidade está acondicionado na fórmula IFERROR, porque quando os coeficientes de modelo geram uma previsão muito, muito próxima ao valor de classe real 0/1, tem-se instabilidade numérica. Nesse caso, é justo configurar o registro de probabilidade para um perfeito escore de 0.

Copie essa fórmula na coluna Y, e, em X1, some os registros das probabilidades. Ao otimizar, obtém-se um conjunto de coeficientes que se parecem com a soma dos coeficientes ao quadrado com algumas mudanças aqui e ali. Veja a Figura 6-31.

**Figura 6-31:** A planilha de Regressão Logística

Se você checar a soma dos erros quadrados associada à sua regressão logística real, a métrica é claramente otimizada.

#### TESTES ESTATÍSTICOS EM UMA REGRESSÃO LOGÍSTICA

Conceitos estatísticos análogos para os R-quadrado, teste F e teste t estão disponíveis na regressão logística. Cálculos tais como R-quadrado falso, desvio de modelo e a estatística Wald, fornecem à regressão logística muita da mesma exatidão que a regressão linear. Para mais informações, veja *Applied Logistic Regression* de David W. Hosmer, Jr., Stanley Lemeshow e Rodney X. Sturdivant (John Wiley & Sons, 2013).

## Seleção de Modelo — Comparando o Desempenho das Regressões Lineares e Logísticas

Agora que você possui um segundo modelo, pode aplicá-lo em um conjunto de teste e comparar seu desempenho ao desempenho da regressão linear. As previsões usando a regressão logística são feitas da mesma forma que são modeladas na aba Logistic Regression nas colunas W e X.

Na célula W2 da aba Test Set, pegue a combinação linear dos coeficientes de modelo e teste os dados assim:

```
=SUMPRODUCT('Logistic Regression'!B$2:T$2,'Test Set'!A2:S2) +
'Logistic Regression'!U$2
```

Em X2, execute isto por toda a função de ligação para obter as probabilidades de classe:

```
=EXP(W2)/(1+EXP(W2))
```

Copie essas células por todo o conjunto de teste para obter a planilha exibida na Figura 6-32:

	L	M	N	O	P	Q	R	S	T	U	V	W	X
1	Ginger Ale	Sea Bands	Stopped buying ciggies	Cigarettes	Smoking Cessation	Stopped buying wine	Wine	Maternity Clothes		PREGNANT	Linear Prediction	Logistic Regression -- Linear Comb.	Probability
5	0	0	0	0	0	1	0	0		1	0.55	0.41	0.60
6	0	0	0	0	0	0	0	0		1	0.37	-0.80	0.31
7	1	0	0	0	0	1	0	0		1	1.12	6.43	1.00
8	0	0	0	0	0	1	0	1		1	0.86	2.95	0.95
9	0	0	0	0	0	0	0	0		1	0.46	-0.37	0.41
10	0	0	0	0	0	0	0	0		1	0.46	-0.37	0.41
11	0	0	1	1	0	0	0	1		1	0.67	1.43	0.81
12	1	0	0	0	0	0	0	0		1	0.70	1.72	0.85
13	0	0	1	0	0	0	0	0		1	0.59	0.79	0.69
14	0	0	0	0	0	0	0	0		1	0.46	-0.39	0.40
15	0	0	0	0	0	0	0	0		1	0.67	1.97	0.88
16	0	0	0	0	0	0	0	0		1	0.66	2.02	0.88
17	0	0	0	0	0	0	0	0		1	1.12	6.32	1.00
18	0	0	0	0	0	0	0	1		1	1.04	5.77	1.00

**Figura 6-32:** Previsões de regressão logística no conjunto de dados

Para ver como as previsões se juntam, faça uma cópia da aba Performance e a nomeie como **Performance Logistic**. Direcionando as fórmulas das previsões mínimas e máximas para focalizar na coluna X da aba Test Set, os valores voltam como 0 e 1, assim como você esperaria agora que seu modelo está fornecendo probabilidades reais, em oposição à regressão linear.

## NOTA

Embora a regressão logística retorne as probabilidades de classe (previsões reais entre 0 e 1), tais probabilidades são baseadas em uma divisão de 50/50 de clientes grávidas e não grávidas no conjunto em treinamento reequilibrado.

Isso está correto se você se importar com a classificação binária em algum valor de corte em vez de usar as probabilidades reais.

Selecione valores de corte de 0 a 1 em incrementos 0,5 (na verdade, você talvez tenha que selecionar entre 1 e 0,999 para evitar que a fórmula da precisão seja divisível por 0). O que vier após a linha 22 pode ser apagado, e as métricas de desempenho precisam ser alteradas para verificar a coluna X na aba Test Set em vez de V. Esse resultado está na Figura 6-33.

Você pode configurar a curva ROC da mesma forma de antes, porém, a fim de comparar a regressão logística com a linear, adicione uma série de dados para cada métrica de performance do modelo (clique com o botão direito no gráfico e selecione Select Data para adicionar outras séries). Na Figura 6-34, fica claro que as curvas ROC para os dois modelos estão quase uma em cima da outra.

RetailMart.xlsx

The screenshot shows a Microsoft Excel spreadsheet titled "RetailMart.xlsx". The active sheet contains a table of performance metrics for a logistic regression model. The columns are labeled: Min Prediction, Classification, Precision, Specificity / True Negative Rate, False Positive Rate (1 - Specificity), and True Positive Rate / Recall / Sensitivity. The rows show values for various probability cutoffs from 0.00 to 1.00. The table includes a header row and a summary row for "Max Prediction". The "True Positive Rate / Recall / Sensitivity" column shows values ranging from 0.00 to 1.00, with the highest value being 1.00 at a 0.00 prediction threshold.

	Probability Cutoff for Pregnant	Min Prediction	Classification	Precision	Specificity / True Negative Rate	False Positive Rate (1 - Specificity)	True Positive Rate / Recall / Sensitivity
2		0.00	0.00	0.06	0.00	1.00	1.00
3			0.05	0.07	0.21	0.79	0.98
4	Max Prediction	1.00	0.10	0.09	0.41	0.59	0.97
5		0.15	0.11	0.49	0.51	0.95	
6		0.20	0.11	0.51	0.49	0.95	
7		0.25	0.11	0.54	0.46	0.93	
8		0.30	0.13	0.61	0.39	0.93	
9		0.35	0.16	0.69	0.31	0.90	
10		0.40	0.18	0.74	0.26	0.88	
11		0.45	0.30	0.88	0.12	0.80	
12		0.50	0.30	0.88	0.12	0.78	
13		0.55	0.32	0.89	0.11	0.78	
14		0.60	0.33	0.90	0.10	0.77	
15		0.65	0.36	0.92	0.08	0.75	
16		0.70	0.37	0.92	0.08	0.72	
17		0.75	0.43	0.94	0.06	0.67	
18		0.80	0.51	0.96	0.04	0.67	
19		0.85	0.62	0.98	0.02	0.60	
20		0.90	0.71	0.99	0.01	0.53	
21		0.95	0.76	0.99	0.01	0.42	
22		1.00	1.00	1.00	0.00	0.02	

Figura 6-33: A tabela de Desempenho Logístico

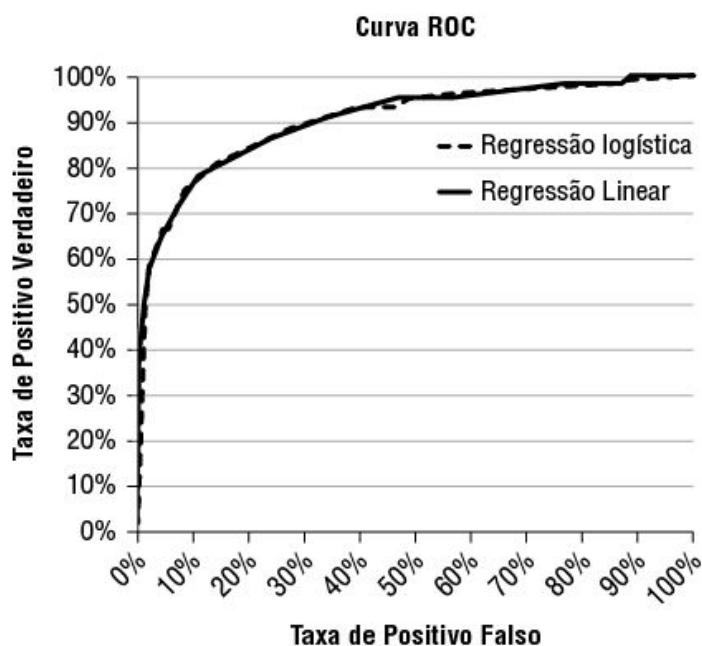


Figura 6-34: As curvas ROC das regressões linear e logística representadas

Como os desempenhos dos modelos são quase idênticos, você talvez considere usar a regressão logística por nenhum outro motivo além da

praticidade de obter probabilidades de classe reais variando entre 0 e 1 do modelo. Não há nada mais bonito.

---

### TENHA CUIDADO

Você talvez escute falar sobre seleção de modelo por aí no mundo real. Podem perguntar, “Por que você não usou máquina de vetores de suporte, redes neurais, florestas aleatórias ou árvores de decisão boosted?” Há diversos tipos de modelos IA, todos com suas vantagens e desvantagens. Eu o encorajaria a ler sobre eles e, se você usar modelo IA em seu trabalho, então deveria tentar alguns desses modelos.

Porém.

Tentar modelos IA diferentes não é a parte mais importante de um projeto de modelagem de IA. É o último passo, a cereja do bolo. Aqui é onde sites como Kaggle.com (um site de competição de modelagem de IA) abordam de forma errada.

Você faz valer mais o seu dinheiro passando o tempo selecionando dados e características boas do que modelos. Por exemplo, no problema que destaquei neste capítulo, você estaria mais informado se estivesse testando novas características possíveis como “cliente não consome mais carne por medo de listeriose” e certificando-se de que seus dados em treinamento estão melhores do que estariam se os estivesse testando em uma rede neural em seus antigos dados de treinamento.

Por quê? Porque a frase “garbage in, garbage out”, nunca foi tão bem expressa em nenhum outro campo do que em IA. Nenhum modelo IA é milagroso; ele não pode receber dados ruins e descobrir como usá-los com mágica. Então faça um favor ao seu modelo IA e equipe-a com as melhores e mais criativas características que puder.

---

## Para Mais Informações

Se você adora IA supervisionada e este capítulo não foi suficiente, sugiro as seguintes leituras:

- *Data Mining* with R, de Luis Torgo (Chapman & Hall/CRC, 2010), é um excelente próximo passo. O livro aborda aprendizado de máquina na linguagem de programação R. Essa linguagem é idolatrada por estatísticos em toda parte e não é difícil de acompanhar para fins de modelagem IA. Na verdade, se você fosse produzir algo como o modelo deste capítulo, R seria um bom lugar para treinar e executar tal modelo de produção.
- *The Elements of Statistical Learning*, de Trevor Hastie, Robert Tibshirani e Jerome

Para um debate com outros participantes, eu utilizo o fórum CrossValidated em StackExchange ([stats.stackexchange.com](https://stats.stackexchange.com) — em inglês). Com frequência, alguém já fez a pergunta que você faria, portanto esse fórum proporciona uma vasta base de conhecimento.

## Resumindo

Parabéns! Você acabou de construir um modelo de classificação em uma planilha. Dois, na verdade. Talvez dois e meio. E, se você me acompanhou no desafio de regressão mediana, você é uma fera.

Vamos recapitular os tópicos que abordamos:

- Seleção de características e montagem de dados em treinamento, incluindo criação de variáveis dummy a partir de previsões categóricas
- Treinar um modelo de regressão linear minimizando a soma dos erros quadrados
- Calcular R-quadrado, mostrando que um modelo é significativo estatisticamente usando um teste F e coeficientes de modelo que são significantes individualmente usando um teste t
- Avaliar o desempenho de modelo em um conjunto holdout em diversas classificações de valores de corte ao calcular precisão, especificidade, taxa de falso positivo e recall

- Analisar uma curva ROC
- Adicionar uma função de ligação logística a um modelo linear geral e reotimizá-lo
- Maximizar a probabilidade em uma regressão logística
- Comparar modelos com a curva ROC

Como serei o primeiro a admitir que os dados neste capítulo são apenas imaginários, deixe-me assegurá-lo de que o poder de tal modelo logístico não é para ser banalizado. Você pode utilizar algo parecido em um suporte de decisão de produção ou em um sistema de marketing automatizado para sua organização.

Se quiser continuar a usar o IA, no próximo capítulo apresentarei uma abordagem diferente chamada modelo ensemble.

# 7

## Modelos Ensemble: É Muita Pizza Ruim Junto

**N**a versão americana da popular série de TV *The Office*, o chefe, Michael Scott, compra pizza para os seus funcionários. Todos reclamam ao descobrirem que ele comprou por engano a pizza no Pizza by Alfredo em vez de Alfredo's Pizza. Embora seja mais em conta, a pizza do Pizza by Alfredo é terrível.

Em resposta às reclamações, Michael faz uma pergunta para seus funcionários: é melhor ter pouca pizza boa ou muita pizza ruim?

Para muitas implementações práticas de inteligência artificial, a resposta é possivelmente a primeira. No capítulo anterior, você criou um modelo bom e único para prever grupos de grávidas fazendo compras no RetailMart. Mas e se você se democratizasse? E se você construísse muitos modelos fracos de propósito e deixasse eles decidirem se uma cliente está grávida? A contagem dos votos seria usada como uma previsão única.

Esse tipo de abordagem é chamado de *modelagem ensemble*, e, como verá, ele transforma simples observações em ouro.

Você examinará um tipo de modelagem ensemble chamado *tocos de decisão bagged (bootstrap aggregating)* e é próximo a um método constantemente utilizado na indústria chamado de *floresta aleatória*. Na verdade, é bem próximo ao método que eu uso diariamente na minha própria vida em MailChimp.com para prever quando um usuário está prestes a enviar um spam.

Depois do método *bagging*, você investigará outra técnica superinteressante chamada de *boosting*. Ambas as técnicas encontraram formas criativas para utilizar os dados em treinamento diversas vezes a

fim de instruir todos os classificadores ensemble. Há uma tendência intuitiva a esses métodos que é remanescente do naïve Bayes — uma bobeira mas, em agregação, é inteligente.

## Usando↑os↑Dados↑do↑Capítulo↑6

### NOTA

A pasta de trabalho do Excel usada neste capítulo, “Ensemble.xlsx”, está disponível para download na página da editora em [www.altabooks.com.br](http://www.altabooks.com.br), procurando pelo nome do livro. Nela, você pode ver todos os dados iniciais se quiser usá-los. Ou, pode acompanhar por meio das planilhas na pasta de trabalho.

Este capítulo é bem rápido já que você utilizará os dados do RetailMart do Capítulo 6. Ao usar os mesmos dados, notará as diferenças dessas duas implementações de modelos dos modelos de regressão no capítulo anterior. As técnicas de modelagem demonstradas neste capítulo foram desenvolvidas mais recentemente. Elas são um tanto intuitivas, e, ainda assim, são umas das tecnologias de inteligência artificial prontas para uso mais poderosas atualmente.

Além disso, construiremos curvas ROC idênticas às do Capítulo 6, portanto não gastarei muito tempo explicando cálculos de métricas de desempenho. Veja o Capítulo 6 se você realmente quiser entender conceitos como precisão e recall.

Para começar, a pasta de trabalho disponível para download possui uma planilha chamada TD que inclui os dados em treinamento do Capítulo 6, com as variáveis dummy já instaladas adequadamente (para mais informações veja o Capítulo 6). Além disso, as características foram numeradas de 0 a 18 na linha 2. Isso será útil junto à manutenção dos registros mais tarde (veja a Figura 7-1).

A pasta de trabalho também inclui a aba Test Set do Capítulo 6.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	Male	Female	Home Apt	Test	Pregnancy	Birth	Feminine	Folic Acid	Prenatal Vitamins	Prenatal Yoga	Body Pillow	Ginger Ale	Sea Bands	buying ciggies	Stopped Cigar ettes	Smoking cessation	Stopped buying wine	Wine	Maternity Clothes	PREGNANT	
2	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18		
3	1	0	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	
4	1	0	1	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	
5	1	0	1	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	
6	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	
7	0	1	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	1	
8	0	1	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	1	
9	1	0	1	0	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	1	
10	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	
11	0	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	

Figura 7-1: A tabela TD armazena os dados do Capítulo 6

Você tentará fazer exatamente o que fez com esses dados no Capítulo 6 — prever os valores na coluna PREGNANT usando os dados à sua esquerda. Depois, verificará sua exatidão no conjunto de grupos

## A IMPUTAÇÃO DE VALORES FALTANTES

No exemplo do RetailMart apresentado no Capítulo 6 e tendo continuidade aqui, você trabalha com um conjunto de dados que não possui furos. Para muitos modelos construídos a partir de dados de negócios transacionais, muitas vezes este é o caso. Porém, haverá situações em que os elementos estarão ausentes em algumas das linhas em um conjunto de dados.

Por exemplo, se você estivesse construindo um modelo de recomendação de IA para um site de encontros e perguntasse aos usuários em seus questionários de perfil se eles escutam a banda de rock Evanescence, talvez essa questão fosse deixada em branco em algumas ocasiões.

Portanto, como treinar um modelo se algumas pessoas em seu conjunto em treinamento deixariam a questão do Evanescence em branco?

Há toda uma problemática em torno desse assunto, mas listarei rapidamente alguns pontos para começar:

- Pule apenas as linhas com valores ausentes. Se eles forem um tanto aleatórios, perder algumas linhas do conjunto em treinamento não o matará. No exemplo do site de encontros, esses espaços em branco são mais intencionais do que aleatórios; portanto, pular as linhas

poderia fazer com que os dados em treinamento obtivessem uma visão distorcida da realidade.

- Se a coluna for numérica, preencha os valores ausentes com a média dos registros que possuem valores. Esse preenchimento recebe o nome de *imputação*. Se a coluna for categórica, utilize a categoria do valor mais comum. Mais uma vez, no caso dos fãs tímidos do Evanescence, o valor mais comum provavelmente é Não, então preencher com o valor mais comum pode ser a forma errada de agir quando as pessoas estão se abstraindo da resposta.
- Uma outra opção é adicionar outra coluna indicadora que possua um 0, a menos que você tivesse um valor ausente em sua coluna original, e um 1, caso contrário. Desta forma, você terá preenchido os valores ausentes da melhor maneira possível, embora tenha acabado de dizer ao modelo para não confiar tanto assim.
- Em vez de usar a mediana, é possível treinar um modelo como o modelo linear geral apresentado no Capítulo 6 para prever o valor ausente usando os dados das outras colunas. É um pouco trabalhoso mas vale a pena se você tiver um pequeno conjunto de dados e não aguenta perder exatidão ou jogar linhas fora.
- Infelizmente, este último método (como todos os que abordei nesta nota) é muito convencido. Ele trata os pontos dos dados imputados como se fossem cidadãos de primeira classe uma vez que foram previstos a partir da regressão linear. Para sobreviver a isso, os estatísticos com frequência usam modelos estatísticos para gerar linhas de regressão múltipla. Os dados vazios são preenchidos diversas vezes usando esses modelos de regressão, criando um novo conjunto de dados imputados a cada vez. Qualquer análise será realizada nos conjuntos de dados imputados e quaisquer resultados serão reunidos no final da análise. Isso se chama *imputação múltipla* (*multiple imputation*).
- Outro método que vale a pena tentar é a *imputação k vizinhos mais próximos*. Usando a distância (veja o Capítulo 2) ou as matrizes de afinidade (Capítulo 5), calcule os k vizinhos mais próximos para uma entrada com dados ausentes. Pegue a média ponderada pela distância (ou o valor mais comum, se preferir) dos valores dos vizinhos para imputar os dados ausentes.

## Bagging: ↑Aleatorize, ↑Treine e ↑Repita

*Bagging* é uma técnica utilizada para treinar classificadores múltiplos (um ensemble se preferir) sem que todos sejam treinados no mesmo conjunto de dados em treinamento. Visto que, se os classificadores fossem treinados nos mesmos dados, eles pareceriam idênticos; você quer uma variedade de modelos, e não um monte de cópias do mesmo. Bagging permite que você introduza alguma variedade em um conjunto de classificadores que não seria possível de outra maneira.

## Toco ↑de ↑Decisão ↑Não É ↑um ↑Termo ↑Sexy para um ↑Preditor ↑Bobo

No modelo bagging que você construirá, os classificadores individuais serão os *tocos de decisão* (*decision stumps*). Um toco de decisão nada mais é do que uma única pergunta que você faz sobre os dados.

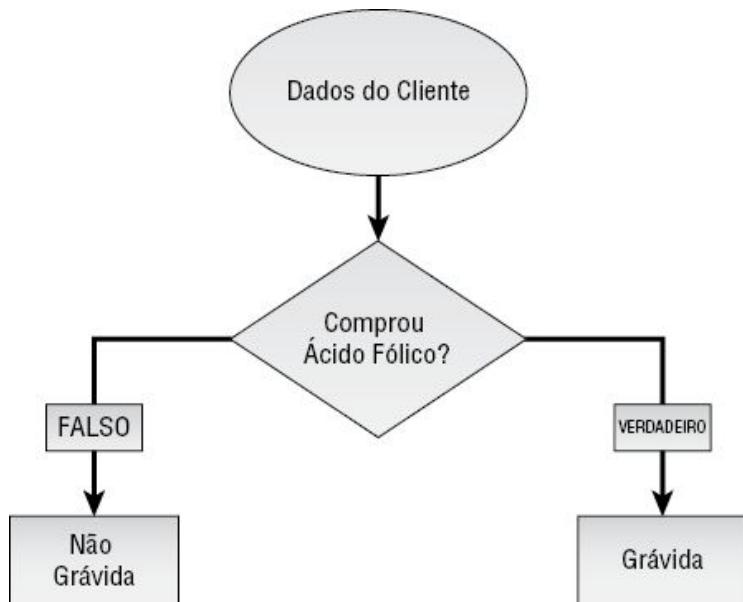
Dependendo da resposta, você diz se o grupo é de grávidas ou não. Um classificador simples como este geralmente é chamado de *weak learner* (*classificador fraco*).

Por exemplo, nos dados em treinamento, se você contar a quantidade de vezes que um grupo de grávidas comprou ácido fólico ao realçar H3:H502 e somar com a barra de resumo, encontrará 104 grupos de grávidas que fizeram a compra antes de dar à luz. Por outro lado, somente duas clientes do grupo que não estavam grávidas compraram ácido fólico.

Portanto, existe uma relação entre comprar suplementos de ácido fólico e estar grávida. Você pode usar essa simples relação para construir o seguinte weak learner:

*A cliente comprou ácido fólico? Se sim, presuma que está grávida.  
Se não, presuma que não está grávida.*

Esse preditor é visualizado na Figura 7-2:



**Figura 7-2:** O toco de decisão do ácido fólico

## Isso Não me Parece Tão Bobo Assim!

O toco na Figura 7-2 divide o conjunto de registros em treinamento em dois subconjuntos. Agora, você deve estar pensando que o toco de decisão faz todo o sentido, e está certo, faz mesmo. Mas não é tão perfeito assim. Afinal de contas, existem aproximadamente 400 grupos de grávidas nos dados em treinamento que não compraram ácido fólico e seriam classificadas erroneamente pelo toco.

Ainda assim é melhor do que não ter um modelo, certo?

Sem dúvidas. Mas a questão é *quão melhor* é o toco se não tiver um modelo. Uma forma de avaliar é por meio de uma medida chamada *impureza do nó*.

A impureza do nó mede com que frequência uma determinada cliente seria rotulada como grávida e não-grávida de forma incorreta se esse rótulo tivesse sido atribuído aleatoriamente, de acordo com a distribuição dos clientes em seu subconjunto de toco de decisão.

Por exemplo, você poderia começar a introduzir todos os 1.000 registros em treinamento dentro do mesmo subconjunto, ou seja, começar sem um modelo.

A probabilidade de retirar uma mulher grávida do amontoado é de 50%. E, se você rotulá-las aleatoriamente de acordo com a distribuição 50/50, terá 50% de chance de adivinhar o rótulo corretamente.

Assim, você tem uma chance de  $50\% * 50\% = 25\%$  de escolher uma cliente grávida e adivinhar corretamente que ela está grávida. Da mesma forma, você tem uma chance de 25% de escolher uma cliente não-grávida e adivinhar que ela não está grávida. Todos os outros casos, que não esses dois, são apenas diferentes versões de um palpite errôneo.

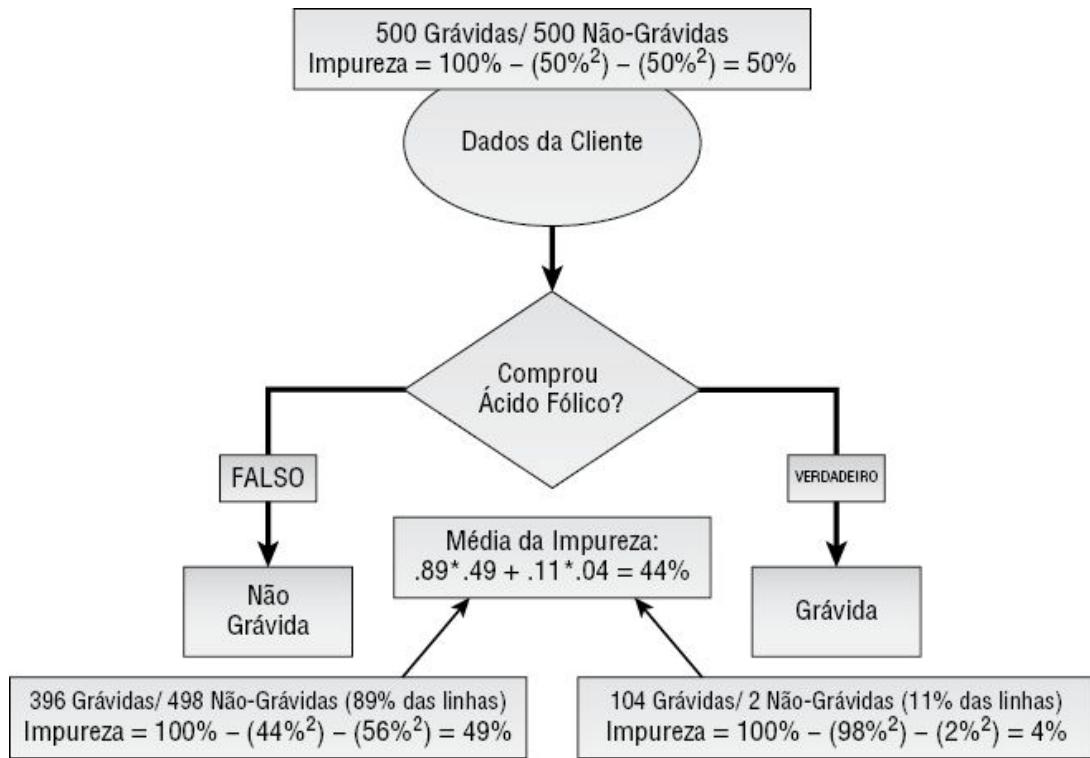
Isso significa que eu tenho uma chance de  $100\% - 25\% - 25\% = 50\%$  de rotular incorretamente uma cliente. Logo, você diria que a impureza do meu nó inicial é de 50%.

O toco de ácido fólico divide esse conjunto de 1.000 casos em dois grupos — 894 pessoas que não compraram ácido fólico e 106 que compraram. Cada um desses subconjuntos terá sua própria impureza, então, se calculá-las (levando em consideração as diferenças de tamanho), pode dizer o quanto o toco de decisão melhorou a situação.

Para aquelas 894 clientes localizadas na parte de não-grávidas, 44% delas estão grávidas e 56% não estão. Isso gera o cálculo de impureza de  $100\% - 44\%^2 - 56\%^2 = 49\%$ . Não houve uma melhoria muito grande.

Mas para as 106 clientes localizadas na categoria de grávidas, 98% delas estão grávidas e 2% não estão. Isso gera o cálculo de impureza de  $100\% - 98\%^2 - 2\%^2 = 4\%$ . Muito bem. Ao calculá-los juntos, você saberá que a impureza para o toco inteiro é de 44%. É melhor do que fazer cara ou coroa!

A Figura 7-3 mostra o cálculo da impureza.



**Figura 7-3:** Impureza do nó para o toco do ácido fólico

## DIVIDINDO UMA CARACTERÍSTICA COM MAIS DE DOIS VALORES

No exemplo do RetailMart, todas as variáveis independentes são binárias. Nunca é preciso decidir como dividir os dados em treinamento ao criar uma árvore de decisão — os 1s vão para um lado e os 0s vão para o outro. Mas, e se você tiver uma característica que possua todos os tipos de valores?

Por exemplo, no MailChimp, uma das coisas que prevemos é se um endereço de e-mail está ativo e se pode receber e-mails. Uma das métricas utilizadas para fazer isso é quantos dias se passaram desde que enviaram um e-mail para aquele endereço. (Nós enviamos sete bilhões de e-mails por mês, então possuímos dados sobre todo mundo...)

Essa característica não está nem perto de ser binária! Então, quando treinamos uma árvore de decisão que usa somente essa característica, como determinamos qual valor dividir para que uma parte dos dados em treinamento possam ir em uma direção e o restante em outra?

Na verdade, é bem fácil.

Há apenas um número finito que você pode separar. No máximo, é um valor único por registro do seu conjunto em treinamento. E provavelmente há

alguns endereços em seu conjunto em treinamento que possuem a mesma quantidade de dias desde que você enviou pela última vez.

É preciso considerar apenas esses valores. Se você tiver quatro valores únicos para serem divididos dos seus registros de treinamento (digamos 10, 20, 30 e 40 dias), dividi-los em 35 ou 30 não faz diferença. Então, apenas verifique os escores de impureza que você obtém se escolher cada valor para ser dividido, e selecione o que der a menor impureza. Feito!

---

## Você↑Precisa↑de↑Mais↑Potência!

Um único toco de decisão não é o suficiente. E se você tivesse uma grande quantidade deles, cada um treinado em partes diferentes dos dados e com impureza um pouco abaixo de 50%? Aí sim poderia permitir que eles votassem. Baseado na porcentagem de tocos que votaram em grávidas, você poderia decidir chamar uma cliente de grávida.

Mas você precisa de mais tocos.

Bem, você treinou um na coluna Folic Acid. Por que não fazer o mesmo com as outras características?

Você tem apenas 19 características, e, francamente, algumas delas, como se o endereço da cliente é de um apartamento, são muito ruins. Portanto, poderia ficar preso a 19 tocos de qualidade duvidosa.

Acontece que, por meio do bagging, você pode fazer quantos tocos quiser. Bagging será algo como:

- 1.** Primeiro, pegue um pedaço do conjunto de dados. O comum é pegar a raiz quadrada de uma contagem de características (quatro colunas aleatórias no nosso caso) e dois terços aleatórios das linhas.
- 2.** Construa um toco de decisão para cada uma das quatro características que você selecionou usando apenas os dois terços aleatórios dos dados que escolheu.
- 3.** Desses quatro tocos, selecione o mais puro. Guarde-o. Jogue tudo de volta na grande árvore e treine um toco novo.

- 4.** Uma vez que tenha uma porção de tocos, pegue todos eles, faça-os votar, e chame-os de modelo único.

# Vamos↑Praticar

Você deve ser capaz de selecionar um conjunto aleatório de linhas e colunas a partir dos dados em treinamento. O modo mais fácil de fazer isso é colocar as linhas e as colunas em ordem aleatória, como em um baralho de cartas, e então escolher o que precisa da parte de cima à esquerda da tabela.

Para começar, copie A2:U1002 da aba TD no topo da nova aba chamada **TD\_BAG** (você não precisará dos nomes das características, apenas seus valores de índice da linha 2). A maneira mais fácil de aleatorizar TD\_BAG será acrescentar uma coluna e uma linha extras próximas aos dados preenchidos com números aleatórios (usando a fórmula RAND()). Uma amostra aleatória de linhas e características é originada ao ordenar os valores aleatórios de cima para baixo e da esquerda para a direita e, depois, remover a quantidade desejada da parte de cima e da esquerda da tabela.

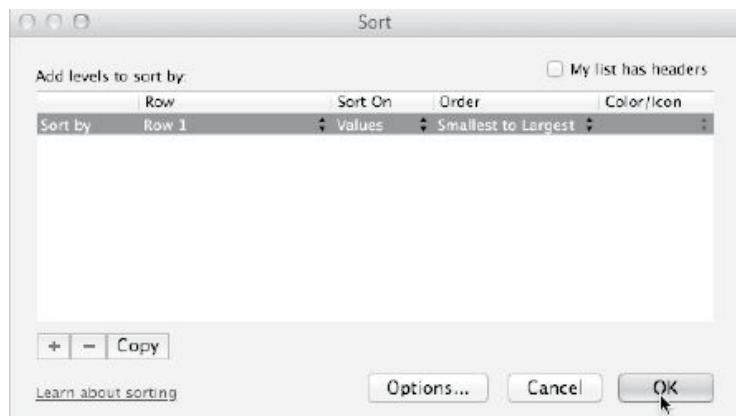
*Obtendo a Amostra Aleatória*

Insira uma linha acima dos índices de características e adicione a fórmula RAND() na linha 1 (A1:S1) e na coluna V (V3:V1002). A planilha se parecerá com a Figura 7-4. Repare que nomeei a coluna V de **RANDOM**.

**Figura 7-4:** Adicionando números aleatórios na parte de cima e na lateral dos dados

Ordene as colunas e linhas aleatoriamente. Comece pelas colunas, pois ordenar lado a lado é um pouco confuso. Para embaralhar as colunas, realce as colunas de A até S. Não realce a coluna PREGNANT; ela não é uma característica, e sim a variável dependente.

Abra a janela de ordenação (veja o Capítulo 1 para saber mais sobre esse assunto). Na janela Sort (Figura 7-5), pressione o botão Options e selecione para ordenar da esquerda para a direita a fim de ordenar as colunas. Certifique-se de que a Row 1, a linha das variáveis aleatórias, esteja selecionada como uma linha a ser ordenada. Também verifique se a caixa My List Has Headers está desmarcada, já que você não possui cabeçalhos na direção horizontal.



**Figura 7-5:** Organizando da esquerda para a direita

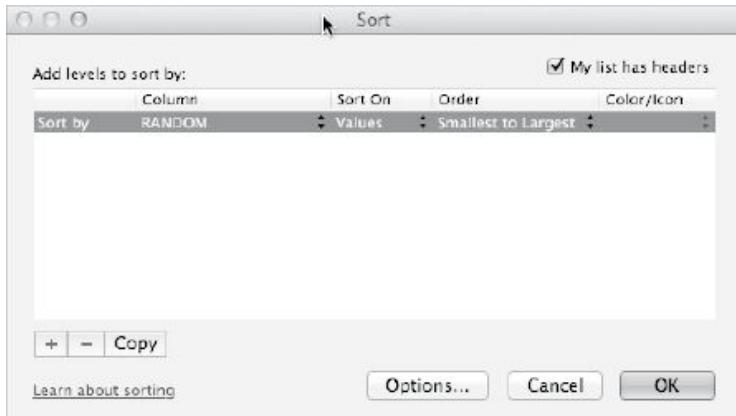
Pressione OK. Você verá as colunas se reorganizarem na planilha.

Agora precisa fazer a mesma coisa com as linhas. Desta vez, selecione a série A2:V1002, incluindo a coluna PREGNANT para que ela permaneça junto aos seus dados ao excluir os números aleatórios no topo da planilha.

Acesse a janela Custom Sort novamente e, abaixo da seção Option, selecione **sort from top to bottom** desta vez.

Certifique-se de que a caixa My List Has Headers esteja marcada desta vez, e depois selecione a coluna RANDOM da lista. A janela Sort deverá

ficar como na Figura 7-6.



**Figura 7-6:** Ordenando de cima para baixo

Agora que você ordenou seus dados em treinamento aleatoriamente, as primeiras quatro colunas e as primeiras 666 linhas formam uma amostra aleatória retangular que você pode pegar. Crie uma nova aba chamada **RandomSelection**. Para retirar a amostra aleatória, aponte a célula em A1 para o seguinte:

```
=TD_BAG!A2
```

E, então, copie a fórmula por todo D667.

Você pode obter os valores de PREGNANT próximos à amostra ao traçá-los diretamente na coluna E. E1 aponta para a célula U2 da aba anterior:

```
=TD_BAG!U2
```

Dê um clique duplo na fórmula para enviá-la para a planilha. Uma vez que faça isso, restará apenas a amostra aleatória dos dados (veja a Figura 7-7). Repare que os dados já estão ordenados aleatoriamente, provavelmente você terá quatro colunas com características diferentes.

O legal é que, se você voltar à aba TD\_BAG e ordenar novamente, essa amostra se atualizará automaticamente!

	A	B	C	D	E
1	15	6	8	11	PREGNANT
2	0	0	0	0	0
3	0	0	0	0	0
4	0	0	0	0	0
5	0	0	0	0	0
6	0	1	0	0	0
7	0	0	1	0	1
8	0	1	0	1	0
9	0	0	0	0	1
10	0	1	0	0	0
11	0	0	0	0	1
12	0	0	1	0	1

**Figura 7-7:** Quatro colunas e dois terços aleatórios das linhas

### Obtendo um Toco de Decisão a partir de uma Amostra

Ao observar uma dessas quatro características, há quatro coisas que podem acontecer entre uma única característica e a variável dependente PREGNANT:

- A característica pode ser 0 e PREGNANT pode ser 1.
- A característica pode ser 0 e PREGNANT pode ser 0.
- A característica pode ser 1 e PREGNANT pode ser 1.
- A característica pode ser 1 e PREGNANT pode ser 0.

É preciso ter uma contagem da quantidade de linhas em treinamento para cair em um desses casos a fim de construir um toco na característica similar ao exibido na Figura 7-2. Para fazer isso, enumere as quatro combinações de 0s e 1s em G2:H5. Configure I1:L1 para igualar os índices da coluna de A1:D1.

A planilha se parecerá com a Figura 7-8.

	A	B	C	D	E	F	G	H	I	J	K	L	
1	15	6	8	11	PREGNANT		PREDICTOR	PREGNANT		15	6	8	11
2	0	0	0	0		0		0	1				
3	0	0	0	0		0		0	0				
4	0	0	0	0		0		1	1				
5	0	0	0	0		0		1	0				

Normal View   Ready   Test Set   TD\_BAG   RandomSelection

Figura 7-8: Quatro possibilidades para os dados em treinamento

Uma vez configurada essa pequena tabela, você precisará preenchê-la obtendo as contagens das linhas em treinamento cujos valores correspondem à combinação dos valores do preditor e das grávidas assinaladas à esquerda. Para o canto de cima à esquerda da tabela (a primeira característica na minha amostra aleatória acabou sendo o número 15), você pode contar a quantidade de linhas em treinamento em que a característica 15 seja 0 e a coluna PREGNANT seja 1 usando a seguinte fórmula:

```
=COUNTIFS(A$2:A$667, $G2, $E$2:$E$667, $H2)
```

A fórmula COUNTIFS() permite que você conte as linhas correspondentes aos *critérios múltiplos*, por isso o S no final de IFS. O primeiro critério considera a característica de número 15 do intervalo (A2:A667) e verifica as linhas que são idênticas ao valor em G2 (0), enquanto o segundo critério considera o intervalo PREGNANT (E2:E667) e verifica as linhas que são idênticas ao valor em H2 (1).

Copie essa fórmula no restante das células na tabela para fazer as contagens para cada caso (veja a Figura 7-9).

Se você fosse tratar cada uma dessas características como um toco de decisão, qual valor para a característica indicaria gravidez? Esse seria o valor da concentração mais alta de clientes grávidas na amostra.

Na linha 6 abaixo, você pode comparar os valores de contagem com essas duas relações. Em I6 coloque a fórmula:

=IF(I2/(I2+I3)>I4/(I4+I5),0,1)

	A	B	C	D	E	F	G	H	I	J	K	L
1	15	6	8	11	PREGNANT	PREDICTOR	PREGNANT		15	6	8	11
2	0	0	0	0	0		0	1	299	315	252	293
3	0	0	0	0	0		0	0	330	254	324	325
4	0	0	0	0	0		1	1	34	18	81	40
5	0	0	0	0	0		1	0	3	79	9	8

**Figura 7-9:** Característica/resposta de pareamento para cada uma das características na amostra aleatória

Se a relação de clientes grávidas associadas ao valor 0 para a característica ( $I2/(I2+I3)$ ) é maior do que as associadas com 1 ( $I4/(I4+I5)$ ), então 0 é preditivo da gravidez neste toco. Caso contrário, o 1 é o preditivo. Copie essa fórmula por toda a coluna L. O resultado é esta planilha na Figura 7-10.

	A	B	C	D	E	F	G	H	I	J	K	L
1	15	6	8	11	PREGNANT	PREDICTOR	PREGNANT		15	6	8	11
2	0	0	0	0	0		0	1	299	315	252	293
3	0	0	0	0	0		0	0	330	254	324	325
4	0	0	0	0	0		1	1	34	18	81	40
5	0	0	0	0	0		1	0	3	79	9	8
6	0	1	0	0	0		Which value indicates pregnancy?	1	0	1	1	1

**Figura 7-10:** Calcular qual valor de característica está relacionado com a gravidez

Usando as contagens das linhas 2 até 5, você pode calcular os valores de impureza para os nós de cada toco de decisão se decidir fazer uma divisão naquela característica.

Vamos inserir os cálculos de impureza na linha 8 abaixo da contagem do caso. Assim como na Figura 7-3, é necessário calcular o valor de impureza dos casos de treinamento que tinham um valor de característica 0 e fazer uma média com os que tinham valor 1.

Se você utilizar a primeira característica (número 15 para mim), 299 mulheres grávidas e 330 não-grávidas ficaram com o nó 0, logo a impureza é de  $100\% - (299/629)^2 - (330/629)^2$ , e pode ser inserido na planilha na célula I8 desta forma:

$$=1-(I2/(I2+I3))^2-(I3/(I2+I3))^2$$

Igualmente, a impureza para o nó 1 pode ser escrita assim:

$$=1-(I4/(I4+I5))^2-(I5/(I4+I5))^2$$

Eles estão unidos em uma média ponderada ao multiplicar cada impureza pela quantidade de casos de treinamento em cada nó, somá-los e, então, dividi-los pelo número total de casos de treinamento, 666:

$$=(I8*(I2+I3)+I9*(I4+I5))/666$$

Você pode arrastar esses cálculos de impureza por todas as quatro características produzindo valores de impureza unidos para cada um dos possíveis tocos de decisão, como mostra a Figura 7-11.

The screenshot shows a Microsoft Excel spreadsheet titled "Ensemble.xlsm". The formula bar displays the formula  $=1-(I2/(I2+I3))^2-(I3/(I2+I3))^2$ . The data table consists of several columns: A through D, E, F, G, H, I, J, K, L, and M. Row 1 contains the values 15, 6, 8, 11 under columns A-D, and "PREGNANT" under column E. Rows 2 through 6 show the counts for non-pregnant (0) and pregnant (1) cases across four predictors. Row 7 contains the question "Which value indicates pregnancy?". Rows 8 through 10 show the calculated impurity values for each predictor category (0 or 1) and a combined row. The formula for row 8 is  $=1-(I2/(I2+I3))^2-(I3/(I2+I3))^2$ , and the formula for row 9 is  $=1-(I4/(I4+I5))^2-(I5/(I4+I5))^2$ . The final row 10 shows the combined impurity values.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	15	6	8	11	PREGNANT		PREDICTOR	PREGNANT	15	6	8	11	
2	0	0	0	0			0		299	315	252	293	
3	0	0	0	0			0		330	254	324	325	
4	0	0	0	0			1	1	34	18	81	40	
5	0	0	0	0			1	0	3	79	9	8	
6	0	1	0	0			0	Which value indicates pregnancy?	1	0	1	1	
7	0	0	1	0			1						
8	0	1	0	1			Impurity	0	0.499	0.494	0.492	0.499	
9	0	0	0	0			1		0.149	0.302	0.180	0.278	
10	0	1	0	0			Combined		0.479	0.466	0.450	0.483	

**Figura 7-11:** Valores de impureza combinados para quatro tocos de decisão

Passando os olhos pelos valores de impureza, para a minha pasta de trabalho (provavelmente a sua será diferente devido à organização aleatória), a característica líder é a número 8 (voltando à planilha TD, é Prenatal Vitamins) com impureza de 0,450.

### **Documentando o Vencedor**

Tudo bem, os pré-natais venceram essa amostra. Você deve ter tido um vencedor diferente e deve documentá-lo em algum lugar.

Nomeie as células N1 e N2 como **Winner** e **Pregnant Is**. O toco vencedor será salvo na coluna O. Comece salvando a coluna vencedora na célula O1. Esse seria o valor de I1:L1 pois possui a impureza mais baixa (no meu caso é 8). É possível combinar as fórmulas MATCH e INDEX para que elas façam essa verificação (veja o Capítulo 1 para saber mais sobre essas fórmulas):

```
=INDEX(I1:L1, 0, MATCH(MIN(I10:L10), I10:L10, 0))
```

MATCH(MIN(I10:L10), I10:L10, 0) encontra qual coluna possui o mínimo de impureza na linha 10 e a entrega para INDEX. INDEX localiza o nome da característica vencedora adequada.

Da mesma forma, em O2 pode-se colocar se 0 ou 1 está associado à gravidez ao encontrar o valor na linha 6 da coluna com o mínimo de impureza:

```
=INDEX(I6:L6, 0, MATCH(MIN(I10:L10), I10:L10, 0))
```

O toco de decisão vencedor e seu nó associado à gravidez são chamados, como mostra a Figura 7-12.

	G	H	I	J	K	L	M	N	O
1	PREDICTOR	PREGNANT		15	6	8	11	Winner:	8
2		0	1	299	315	252	293	Pregnant is:	1
3		0	0	330	254	324	325		
4		1	1	34	18	81	40		
5		1	0	3	79	9	8		
6	Which value indicates pregnancy?			1	0	1	1		
7									
8	Impurity	0		0.499	0.494	0.492	0.499		
9		1		0.149	0.302	0.180	0.278		
10		Combined		0.479	0.466	0.450	0.483		
11								Sum = 8	

Figura 7-12: O ambiente do vencedor para os quatro tocos de decisão

### Me Sacuda, Judy!

Uau! Eu sei que foram muitos passos pequenos para criar um toco. Mas agora que todas as fórmulas estão em seus devidos lugares, criar os próximos duzentos será muito mais fácil.

É possível criar um segundo toco rapidamente. Mas, antes de começar, salve o toco que acabou de construir. Para fazer isso, copie e cole os valores à direita de P1:P2 em O1:O2.

Para criar um toco novo, volte para a aba TD\_BAG e embaralhe as linhas e as colunas novamente.

Clique outra vez na aba RandomSelection. *Voilà!* O vencedor mudou. No meu caso, é ácido fólico, e o valor associado à gravidez é 1 (veja a Figura 7-13). O toco anterior está salvo à direita.

	G	H	I	J	K	L	M	N	O	P
1	PREDICTOR	PREGNANT	2	7	4	0		Winner:	7	8
2		0		171	262	285	222			
3		0		171	334	331	187			
4		1		161	70	47	110			
5		1		163	0	3	147			
6	Which value indicates pregnancy?			0	1	1	0			
7										
8	Impurity	0		0.500	0.493	0.497	0.496			
9		1		0.500	0.000	0.113	0.490			
10		Combined		0.500	0.441	0.468	0.494			

Figura 7-13: Reembalar os dados produz um toco novo

Para salvar o segundo toco, clique com o botão direito na coluna P e selecione Insert para deslocar o primeiro toco para a direita. Depois, coloque os valores do toco novo na coluna P. O ensemble agora se parece com a Figura 7-14.

	G	H	I	J	K	L	M	N	O	P	Q
1	PREDICTOR	PREGNANT	2	7	4	0		Winner:	7	7	8
2		0		171	262	285	222				
3		0		171	334	331	187				
4		1		161	70	47	110				
5		1		163	0	3	147				
6	Which value indicates pregnancy?			0	1	1	0				
7											
8	Impurity	0		0.500	0.493	0.497	0.496				
9		1		0.500	0.000	0.113	0.490				
10		Combined		0.500	0.441	0.468	0.494				

Figura 7-14: E tentão, teram dois

Bem, esse segundo tomou menos tempo que o primeiro. Vejamos uma coisa...

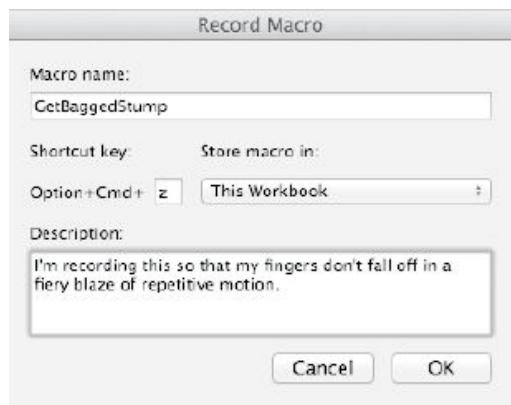
Digamos que você queira dominar os 200 tocos no modelo ensemble. Tudo o que se tem a fazer é repetir esses passos mais 198 vezes. Não é impossível, mas é cansativo.

Por que você não grava uma macro sua fazendo-o e a executa? Como você deve saber, essa operação de embaralhar é perfeita para um macro.

Para vocês que nunca gravaram um macro, nada mais é do que gravar uma série de pressionamento de botões repetitivos, para que você possa reproduzi-los em vez de desenvolver uma síndrome do túnel do carpo.

Portanto, vá em View ⇔ Macros (Tools ⇔ Macro, no Mac OS) e selecione Record NewMacro.

Ao pressionar Record, uma janela abrirá na qual você poderá nomear seu macro em algo parecido com **GetBaggedStump**. E, por mera conveniência, associaremos uma tecla de atalho a ele. Estou em um Mac, então minhas teclas de atalho começam com Option+Cmd, e vou jogar um **Z** na caixa de atalho, porque estou muito bem-humorado hoje (veja a Figura 7-15).



**Figura 7-15:** Preparando-se para gravar um macro

Pressione OK para começar a gravar. Estes são alguns passos para a gravação de um toco de decisão completo:

- 1.** Clique na aba TD\_BAG.
- 2.** Realce as colunas A até S.
- 3.** Ordene as colunas.
- 4.** Realce as linhas 2 até 1002.
- 5.** Ordene as linhas.

6. Clique na aba RandomSelection.
7. Clique com o botão direito na coluna P e insira uma nova coluna em branco.
8. Selecione e copie o toco vencedor em O1:O2.
9. Cole especial os valores em P1:P2.

Vá em View Macro Stop Recording (Tools Macro Stop Recording no Excel 2011 para Mac) para terminar a gravação.

Agora você conseguir gerar um toco de decisão novo ao pressionar uma única tecla de atalho para ativar o macro. Espere um pouco enquanto eu clico neste botão 198 vezes...

## Avaliando o Modelo Bagging

Isso é bagging! Tudo o que você faz é embaralhar os dados, retirar um subconjunto, treinar um classificador simples e fazer tudo mais uma vez. Uma vez que tenha muitos classificadores no seu ensemble, você estará pronto para fazer previsões.

Visto que executamos o macro do toco de decisão algumas centenas de vezes, a planilha RandomSelection é semelhante à Figura 7-16 (seus tocos estarão diferentes).

**Figura 7-16:** Os 200 tocos de decisão

### Previsões no Conjunto de Teste

Agora que você possui seus tocos, é hora de enviar seus dados do conjunto de teste por todo o modelo. Crie uma cópia da aba Test Set e nomeie-a de **TestBag**.

Já dentro dessa aba, insira duas linhas em branco no topo da planilha para ter espaço para seus tocos.

Cole os valores dos tocos na aba RandomSelection (P1:HG2 se você tiver 200 deles) na aba TestBag começando pela coluna W. A planilha fica como a Figura 7-17.

W1	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AE	AC	AD	AE	AF	AG	AI	AI	A	
1								Winner:	8	16	8	16	18	6	7	17	4	7	17	4	5	1
2								Pregnant is:	1	1	1	1	1	0	1	0	1	1	0	1	0	
3	Smoking Cessation	Stopped buying wine		Wine	Maternity Clothes		PREGNANT															
4	0	1		1	0				1													
5	0	0		0	0				1													
6	0	0		0	0				1													
7	0	1		0	0				1													
8	0	0		0	0				1													
9	0	1		0	0				1													
10	0	1		0	1				1													

Figura 7-17: Tocos adicionados à aba TestBag

Pode-se executar cada linha do Test Set por todos os tocos. Comece executando a primeira linha dos dados (linha 4) pelo primeiro toco na coluna W. Você pode usar a fórmula OFFSET para procurar pelo valor da coluna do toco listado em W1, e se esse valor for igual ao do W2, então o toco prevê uma cliente grávida. Caso contrário, o toco prevê não-grávida. A fórmula é assim:

```
=IF(OFFSET($A4, 0, W$1)=W$2, 1, 0)
```

Essa fórmula pode ser copiada por todos os tocos e pela planilha (repare nas referências absolutas). A planilha fica como a Figura 7-18:

Ensemble.xlsxm

	R	S	T	U	V	W	X	Y	Z	AA	AA	AA	AD	AD	AD	AFA														
2					Pregnant is:	1	1	1	1	1	0	1	0	1	1	0	1	0	1	1	1	0	0	1	1	1	1			
3		Maternity			PREGNANT																									
4	1	0			1					0	1	0	1	0	1	0	0	0	0	0	1	0	0	0	1	1	0	0		
5	0	0			1					0	0	0	0	1	0	1	0	0	1	0	1	0	0	0	1	1	0	0		
6	0	0			1					0	0	0	0	1	0	1	0	0	1	0	1	0	1	0	1	1	0	0		
7	0	0			1					0	1	0	1	0	1	0	1	0	0	1	0	1	0	0	0	1	1	0	0	
8	0	0			1					0	0	0	0	1	0	1	0	0	1	0	1	0	0	0	1	1	0	0		
9	0	0			1					0	1	0	1	0	1	1	0	1	0	1	0	0	0	1	1	1	1	1		
10	0	1			1					0	1	0	1	1	1	0	1	0	0	1	0	1	1	0	0	0	1	1	0	0
11	0	0			1					0	0	0	0	0	1	0	1	0	0	1	0	1	0	0	0	1	1	0	0	

Figura 7-18: Tocos avaliados no conjunto TestBag

Na coluna V, tire a média das linhas da esquerda a fim de obter uma probabilidade de classe para gravidez. Por exemplo, se você tiver 200 tocos em V4, usaria:

=AVERAGE (W4 : HN4)

Copie isso na coluna V para ter previsões para cada linha no conjunto de teste como mostra a Figura 7-19.

Ensemble.xlsxm

	R	S	T	U	V	W	X	Y	Z	AA	AA	AA	AD	AD	AD	AFA														
1					Winner:	8	16	8	16	18	6	7	17	4	7	17	4	5	18	13	4	14	5	7	7	7				
2					Pregnant is:	1	1	1	1	1	0	1	0	1	1	0	1	0	1	1	1	0	0	1	1	1				
3		Maternity			PREGNANT																									
4	1	0			1					0.315	0	1	0	1	0	1	0	0	0	0	0	1	0	0	0	1	1	0	0	
5	0	0			1					0.305	0	0	0	0	0	1	0	1	0	0	1	0	0	0	1	1	0	0		
6	0	0			1					0.345	0	0	0	0	0	1	0	1	0	0	1	0	1	0	1	1	0	0		
7	0	0			1					0.385	0	1	0	1	0	1	0	1	0	0	1	0	1	0	0	1	1	0	0	
8	0	0			1					0.3	0	0	0	0	0	1	0	1	0	0	1	0	1	0	0	0	1	1	0	0
9	0	0			1					0.63	0	1	0	1	0	1	1	0	1	1	0	1	0	0	0	1	1	1	1	

Figura 7-19: Previsões para cada linha

## *Desempenho*

Pode-se avaliar essas previsões usando as mesmas medidas de desempenho utilizadas no Capítulo 6. Eu não ficarei muito tempo nesses cálculos já que a técnica é exatamente a mesma do Capítulo 6. Primeiro, crie uma nova aba chamada **PerformanceBag**. Na primeira coluna, assim como no Capítulo 6, calcule as previsões máxima e mínima. Para os meus 200 tocos, a série resulta de 0,02 para 0,75.

Na coluna B, coloque uma série de valores de corte do mínimo ao máximo (no meu caso, incrementei em 0,02). Precisão, especificidade, taxa de falso positivo e sensibilidade podem ser calculadas da mesma forma que o Capítulo 6 (volte ao Capítulo 6 para mais detalhes).

Isso tem como resultado a planilha da Figura 7-20.

Repare que para uma previsão de corte de 0,5, ou seja, com metade dos tocos votando em grávidas, pode-se identificar 33% das clientes grávidas com apenas 1% de taxa de falso positivo (seus valores podem variar devido à natureza aleatória do algoritmo). Ótimo para tocos tão pequenos!

Pode-se também inserir uma curva ROC usando uma taxa de falso positivo e uma de verdadeiro positivo (colunas E e F) assim como você fez no Capítulo 6. Para os meus 200 tocos, obtive a Figura 7-21.

Ensemble.xlsm

		Cutoff for Pregnant Classification	Precision	True Negative Rate	Specificity / False Positive Rate (1 - Specificity)	True Positive Rate / Recall / Sensitivity
1	Min Prediction	0.02	0.02	0.06	0.00	1.00
2		0.04	0.06	0.06	1.00	1.00
4	Max Prediction	0.06	0.06	0.06	0.99	1.00
5	0.75	0.08	0.06	0.04	0.96	1.00
6		0.1	0.06	0.06	0.94	1.00
7		0.12	0.06	0.07	0.93	1.00
8		0.14	0.07	0.09	0.91	1.00
9		0.16	0.07	0.20	0.80	0.98

Figura 7-20: A métrica do desempenho para bagging

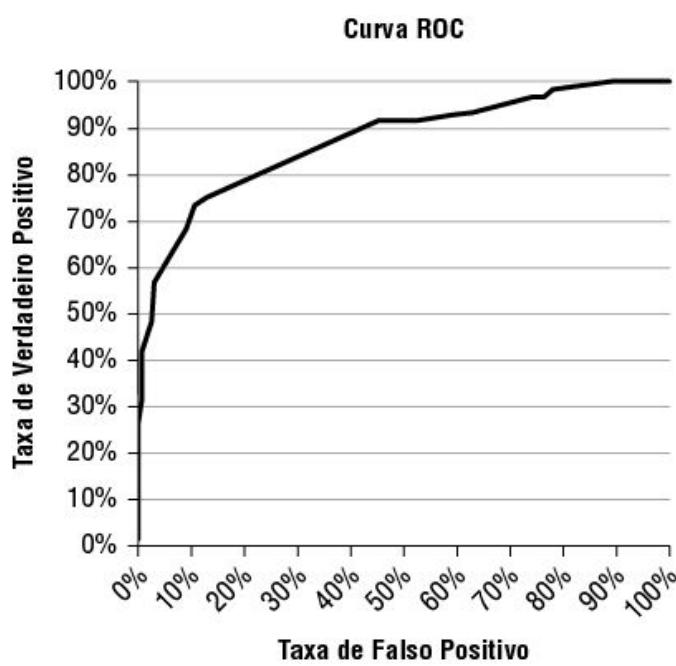


Figura 7-21: A Curva ROC para Tocos Bagged

Além do Desempenho

Enquanto esse modelo de tocos bagged é suportado pelos pacotes padrões da indústria como o pacote `randomForest` de R, é importante ressaltar as duas diferenças entre ele e as configurações de modelagem de uma típica modelagem de floresta aleatória:

- As florestas aleatórias mais suaves geralmente testam *com substituição*, o que significa que a mesma linha dos dados em treinamento podem ser retiradas da mesma amostra mais de uma vez. Ao testar com substituição, pode-se testar com o mesmo número de registros que o conjunto em treinamento em vez de limitar o teste a apenas dois terços. Na prática, embora testar com substituição possua probabilidades estatísticas melhores, se você estiver trabalhando com um conjunto de dados consideravelmente grande, não há diferença entre os dois métodos de teste.
- Por padrão, as florestas aleatórias criam árvores de classificação inteiras em vez de tocos. Uma árvore inteira é a que você divide os dados em dois nós, escolhe características novas para dividir esses nós, e assim por diante, até deparar-se com algum critério de interrupção. As árvores de classificação inteiras são melhores do que os tocos quando há interação entre os recursos que podem ser modelados.

Indo com essa conversa além da exatidão dos modelos, estas são algumas vantagens para o método bagging:

- Bagging é resistente aos valores atípicos (*outliers*) e tende a não *sobreajustar* os dados. O sobreajuste ocorre quando o modelo se ajusta a mais do que somente o sinal em seus dados e ajusta o ruído também.
- O processo de treinamento pode ser paralelizado contanto que o treinamento de um indivíduo weak learner não seja dependente

do treinamento de um weak learner anterior.

- Esse tipo de modelo pode lidar com toneladas de variáveis de decisão.

Os modelos que usamos no MailChimp para prever spam e abuso são modelos de floresta aleatória, os quais treinamos em paralelo usando aproximadamente 10 bilhões de linhas de dados crus. Isso não caberá no Excel, e eu, com certeza, não usaria um macro para fazer isso!

Não, eu uso a linguagem de programação R com o pacote `randomForest`, o qual eu recomendaria que você aprendesse em seguida se quiser conduzir um desses modelos para produção em sua organização. De fato, o modelo neste Capítulo pode ser alcançado pelo pacote `randomForest` somente desligando o teste com substituição e configurando o número máximo de nós nas árvores de decisão para 2 (veja o Capítulo 10).

## Boosting: ↑Se ↑Fizer ↑Errado, ↑Reinicie ↑e Tente ↑Novamente

Qual foi o motivo por trás de fazer bagging mesmo?

Se você treinou um amontoado de tocos de decisão sobre todo um conjunto de dados várias vezes, eles seriam idênticos. Ao escolher seleções aleatórias em um conjunto de dados, você apresenta uma certa variedade aos seus tocos e acaba capturando nuances nos dados em treinamento que um único toco nunca conseguiria.

Bem, o que o bagging faz com seleções aleatórias, o *boosting* faz com pesos. O boosting não pega partes aleatórias do conjunto de dados. Ele usa o conjunto inteiro de dados em cada iteração em treinamento. E, com cada iteração, o boosting focaliza no treinamento de um toco de decisão que conserta alguns dos pecados cometidos pelos tocos de decisão anteriores. Funciona desta forma:

- Em primeiro lugar, cada linha dos dados em treinamento é contada exatamente da mesma forma. Todas elas possuem o mesmo peso. No seu caso, você possui 1000 linhas de dados em treinamento, logo, todas elas começam com um peso de 0,001. Isso significa que os pesos somam 1.
- Avalie cada característica do conjunto inteiro de dados para escolher o melhor toco de decisão. Exceto quando se trata de boosting em vez de bagging, o toco vencedor será aquele que possuir o *erro ponderado* (*weighted error*) mais baixo. Uma penalidade igual ao peso da linha é dada para cada previsão errada para um possível toco. A soma dessas punições é o erro ponderado. Escolha o toco de decisão que apresenta o erro ponderado mais baixo.
- Os pesos são adaptados. Se o toco de decisão escolhido prever uma linha com exatidão, então o peso dessa linha diminui. Se o toco de decisão escolhido estragar uma linha, o peso dela aumenta.
- Um toco novo é treinado usando esses pesos novos. Desta forma, conforme o algoritmo segue, ele se concentra mais nas linhas nos dados em treinamento que os tocos anteriores não acertaram. Os tocos são treinados até que o erro ponderado exceda um limite.

Parte disso pode ser um pouco vago, mas o processo se tornará bastante claro em uma planilha. Vamos aos dados!

## Treinando o Modelo — Todas as Características Têm uma Chance

No boosting, cada característica é um possível toco em todas as iterações. Desta vez você não terá que selecionar a partir de quatro características.

Para começar, crie uma aba chamada **BoostStumps**. E nela, cole as combinações dos valores possíveis de característica/resposta do G1:H5 da aba RandomSelection.

Próximo a esses valores, cole os valores de índice da característica (0–18) na linha 1. Isso resulta na planilha da Figura 7-22.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	PREDICTOR	PREGNANT	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
2		0		1																	
3		0			0																
4		1			1																
5		1			0																

Figura 7-22: As partes iniciais da aba BoostStumps

Abaixo de cada índice, assim como no processo de bagging, você deve somar a quantidade de linhas do conjunto em treinamento que caem em cada uma das quatro combinações do valor da característica e das variáveis independentes listadas nas colunas A e B.

Comece na célula C2 (índice de característica 0) somando a quantidade de linhas em treinamento que possuem 0 para o valor da característica e, também, as que estão grávidas. Isso pode ser contado usando a fórmula COUNTIFS:

```
=COUNTIFS(TD!A$3:A$1002,$A2,TD!$U$3:$U$1002,$B2)
```

O uso de referências absolutas permite que você use uma cópia dessa fórmula até a U5. A planilha resultante está na Figura 7-23.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	
1	PREDICTOR	PREGNANT	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	
2		0	1	327	231	254	293	431	481	472	396	388	483	486	444	476	425	478	447	394	480	389
3		0	0	272	274	258	287	494	379	387	498	484	499	496	487	494	483	425	493	476	397	480
4		1	1	173	269	246	207	69	19	28	104	112	17	14	56	24	75	22	53	106	20	111
5		1	0	228	226	242	213	6	121	113	2	16	1	4	13	6	17	75	7	24	103	20

Figura 7-23: Contando como cada característica divide os dados em treinamento

Assim como no exemplo do bagging, em C6 você pode encontrar o valor associado à gravidez para o índice da característica 0 ao olhar as relações de gravidez associadas ao valor da característica 0 e o valor da característica 1:

=IF(C2/(C2+C3)>C4/(C4+C5), 0, 1)

Isso também deve ser copiado até a coluna U.

Agora, na coluna B, insira os pesos para cada ponto de dados. Comece em B9 com o nome Current Weights, e abaixo dela, por todo o B1009, coloque um 0,001 para cada uma das milhares de linhas em treinamento. Na linha 9, cole os nomes das características da planilha TD, para acompanhar cada característica.

Isso tem como resultado a planilha da Figura 7-24.

Para cada um desses possíveis tocos de decisão, é preciso calcular a taxa de erro ponderado. Isso é feito localizando as linhas em treinamento que estão sem categoria e penalizando-as de acordo com seu peso.

Em C10, por exemplo, você pode comparar os primeiros dados da linha em treinamento para o índice da característica 0 (A3 na aba TD), e, se não combinar com o índice de gravidez em C6, você levará uma punição (o peso na célula B10) se a linha for *não-grávida*. Se o valor da característica não combinar com C6, você levará uma punição se *a linha for grávida*. Isso resulta nas duas declarações IF:

=IF(AND(TD!A3=C\$6, TD!\$U3=0), \$B10, 0) + IF(AND(TD!A3<>C\$6, TD!\$U3=1), \$B10, 0)

As referências absolutas permitem que você copie essa fórmula por toda a U1009. O erro ponderado para cada possível toco de decisão pode então ser calculado na linha 7. O cálculo para o erro ponderado na célula 7 é:

=SUM(C10:C1009)

Ensemble.xlsxm

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	PREDICTOR	PREGNANT	0	1	2	3	4	5	6	7	8	9	10	11
2	0	1	327	231	254	293	431	481	472	396	388	483	486	
3	0	0	272	274	258	287	494	379	387	498	484	499	496	
4	1	1	173	269	246	207	69	19	28	104	112	17	14	
5	1	0	228	226	242	213	6	121	113	2	16	1	4	
6	Pregnant is:		0	1	1	0	1	0	0	1	1	1	1	
7														
8														
9	Current Weights	Male	Female	Home	Apt	Pregnancy Test	Birth Control	Feminine Hygiene	Folic Acid	Prenatal Vitamins	Prenatal Yoga	Body Pillow	Ging	
10	0.001													
11	0.001													
12	0.001													
13	0.001													
14	0.001													
15	0.001													
16	0.001													

Figura 7-24: Os pesos para cada linha dos dados em treinamento.

Copie isso pela linha 7 para obter o erro ponderado de cada toco de decisão (veja a Figura 7-25).

Ensemble.xlsxm

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	PREDICTOR	PREGNANT	0	1	2	3	4	5	6	7	8	9		
2	0	1	327	231	254	293	431	481	472	396	388			
3	0	0	272	274	258	287	494	379	387	498	484			
4	1	1	173	269	246	207	69	19	28	104	112			
5	1	0	228	226	242	213	6	121	113	2	16			
6	Pregnant is:		0	1	1	0	1	0	0	1	1			
7	Weighted error:	0.45	0.457	0.496	0.5	0.437	0.398	0.415	0.4	0.404	0.401			
8														
9	Current Weights	Male	Female	Home	Apt	Pregnancy Test	Birth Control	Feminine Hygiene	Folic Acid	Prenatal Vitamins	Prenatal Yoga			
10	0.001	0	0.001	0.001	0	0	0	0	0	0	0	0	0	
11	0.001	0	0.001	0	0	0	0	0	0	0	0	0	0	
12	0.001	0	0.001	0	0	0	0	0	0	0	0	0.001	0	
13	0.001	0	0.001	0	0	0.001	0	0	0	0	0.001	0	0	

Figura 7-25: O cálculo do erro ponderado para cada toco

## *Acrescentando o Vencedor*

Nomeie a célula 1 como **Winning Error**, e, em X1, encontre o menor valor dos erros ponderados:

```
=MIN(C7:U7)
```

Assim como na seção bagging, combine as fórmulas INDEX e MATCH em X2 para obter o índice da característica do toco vencedor:

```
=INDEX(C1:U1, 0, MATCH(X1, C7:U7, 0))
```

E em X3 você também pode obter o valor associado à gravidez para o toco usando INDEX e MATCH:

```
=INDEX(C6:U6, 0, MATCH(X1, C7:U7, 0))
```

Isso tem como resultado a planilha da Figura 7-26. Começando pelos pesos iguais para cada ponto de dado, o índice da característica 5 com um valor 0 indicando gravidez é escolhido como o melhor toco.

Voltando à aba TD, você pode ver que essa é a característica Birth Control.

The screenshot shows a Microsoft Excel spreadsheet titled "Ensemble.xlsxm". The formula bar displays the formula `=INDEX(C6:U6,0,MATCH(X1,C7:U7,0))`. The main table has columns Q through X. Row 1 contains headers 14, 15, 16, 17, 18. Rows 2 through 7 contain numerical values. Row 8 contains the calculated result. Column W contains the results of the INDEX and MATCH functions. Column X contains the value 0.398. Row 3 contains the text "Winning Error" and "Column 5". Row 4 contains the text "Pregnant is" and "0". The ribbon at the top shows tabs for Home, Layout, Tables, Charts, SmartArt, Formulas, Data, Review, and others. The status bar at the bottom shows "Normal View" and "Ready".

	Q	R	S	T	U	V	W	X
1	14	15	16	17	18		Winning Error	0.398
2	478	447	394	480	389		Column	5
3	425	493	476	397	480		Pregnant is	0
4	22	53	106	20	111			
5	75	7	24	103	20			
6	0	1	1	0	1			
7	0.447	0.454	0.418	0.42	0.409			

Figura 7-26: O primeiro toco boosted vencedor

## *Calculando o Valor Alfa para o Toco*

Boosting funciona estabelecendo pesos para as linhas em treinamento que foram classificadas erroneamente pelos tocos anteriores. Os tocos no início do processo de boosting são geralmente mais eficazes, enquanto que os tocos no final do processo de treinamento são mais especializados

— os pesos foram alterados para concentrar nos pontos perturbadores nos dados em treinamento.

Esses tocos com pesos especializados ajudam a ajustar o modelo para os pontos estranhos no conjunto de dados. No entanto, ao fazer isso, o erro ponderado deles será maior do que os dos tocos iniciais no processo de boosting. Conforme seus erros ponderados aumentam, a melhoria geral que eles contribuem para o modelo cai. Em boosting, essa relação é contabilizada por um valor chamado **alfa**:

$$\text{alfa} = 0.5 * \ln((1 - \text{erro ponderado total para o toco})/\text{erro ponderado total para o toco})$$

Da mesma forma que o erro ponderado total do toco aumenta, a fração dentro da função do log natural diminui e chega próximo a 1. Uma vez que o logaritmo natural de 1 é 0, o valor **alfa** fica cada vez menor. Dê uma olhada nele no contexto da planilha.

Nomeie a célula W4 como **Alpha**, e em X4 envie o erro ponderado da chamada de X1 por todo cálculo do **alfa**:

```
=0.5*LN((1-X1)/X1)
```

Para este primeiro toco, você possui um valor **alfa** de 0,207 (veja a Figura 7-27).

The screenshot shows a Microsoft Excel spreadsheet titled "Ensemble.xlsxm". The formula bar displays the formula `=0.5*LN((1-X1)/X1)`. The data is organized into several columns:

	Q	R	S	T	U	V	W	X
1	14	15	16	17	18		Winning Error	0.398
2	478	447	394	480	389		Column	5
3	425	493	476	397	480		Pregnant is	0
4	22	53	106	20	111		Alpha	0.207
5	75	7	24	103	20			
6	0	1	1	0	1			
7	0.447	0.454	0.418	0.42	0.409			

The formula bar also shows the address `X4` and the formula `=0.5*LN((1-X1)/X1)`.

**Figura 7-27:** O valor **alfa** para a primeira iteração do boosting

Como exatamente esses valores alfa são usados? Em bagging, cada toco votou 0 ou 1 na previsão. Quando os seus tocos boosted fazem a previsão, cada classificador votará ***alfa*** se ele achar que a linha for grávida e ***-alfa*** se não for. Então, ao ser usado no conjunto de teste, esse primeiro toco daria 0,207 pontos para qualquer cliente que não tivesse comprado pílulas anticoncepcionais e ***-0,207*** pontos para as que compraram. A previsão final do modelo ensemble é a soma de todos os valores ***alfa*** positivos e negativos.

Como verá mais adiante, para determinar a previsão de gravidez geral vindo do modelo, um corte é programado para a soma dos escores dos tocos individuais. Já que cada toco retorna um valor ***alfa positivo*** ou ***negativo*** como contribuição para a previsão, é comum usar 0 como classificação limite para gravidez, porém ela pode ser alterada para atender às necessidades da sua precisão.

### *Reconsiderando*

Agora que você já completou um toco, é hora de reconsiderar os dados em treinamento. E, para fazer isso, você precisa saber quais linhas dos dados esse toco acertou e quais colunas ele não acertou.

Na coluna V, nomeie V9 como **Wrong**. Em V10, pode-se usar a fórmula `OFFSET` em conjunto com o índice da coluna do toco vencedor (célula X2) para procurar o erro ponderado na linha em treinamento. Se o erro não for zero, então o toco está incorreto para aquela linha, e **Wrong** é configurado para 1:

```
=IF(OFFSET($C10,0,$X$2)>0,1,0)
```

Essa fórmula pode ser copiada por todas as linhas em treinamento (repare nas referências absolutas).

Agora, os pesos originais para esse toco estão na coluna B. Para adaptar os pesos de acordo com as linhas que possuam 1 na coluna **Wrong**, o boosting multiplica o peso original vezes ***exp(alpha \* Wrong)*** (em que ***exp*** é a função exponencial que você encontrou quando aplicou a regressão logística no Capítulo 6).

Se o valor da coluna Wrong for 0, então  $\exp(\alpha * \text{Wrong})$  torna-se 1, e o peso fica estagnado.

Se Wrong for configurado 1, então  $\exp(\alpha * \text{Wrong})$  é um valor maior do que 1, logo todo o peso é ampliado. Nomeie a coluna W como **Scale by Alpha**, e em W10, você pode calcular o peso novo assim:

```
=\$B10*EXP ($V10*$X$4)
```

Copie isso por todo o conjunto de dados.

Infelizmente, esses pesos novos não somam como os antigos. Eles precisam ser **normalizados** (adaptados a fim de somarem um). Portanto, nomeie X9 como **Normalize** e, em X10, divida o novo peso dimensionado pela soma de todos os pesos novos:

```
=W10/SUM (W$10:W$1009)
```

Isso garante que seus pesos novos somam um. Copie essa fórmula. Isso resulta na planilha da Figura 7-28.

The screenshot shows a Microsoft Excel spreadsheet titled "Ensemble.xlsxm". The formula bar displays the formula `=W10/SUM(W$10:W$1009)`. The data table has columns labeled R, S, T, U, V, W, and X. Row 1 contains values 15, 16, 17, 18, and empty cells for V, W, and X. Rows 2 through 5 contain numerical values. Row 6 has values 1, 1, 0, 1. Row 7 has values 0.454, 0.418, 0.42, 0.409. Row 8 is blank. Row 9 contains category names: Smoking Cessation, Stopped buying wine, Maternity Wine Clothes, Wrong, Scale by Alpha, and Normalize. Rows 10 through 13 show data points with values 0.001, 0.001, 0, 0.001, 0, 0.0010, 0, 0.0010, 0, 0.0010, and 0.0009 respectively. The "Normalize" column is highlighted in yellow.

	R	S	T	U	V	W	X
1	15	16	17	18		Winning Error	0.398
2	447	394	480	389		Column	5
3	493	476	397	480		Pregnant is	0
4	53	106	20	111		Alpha	0.207
5	7	24	103	20			
6	1	1	0	1			
7	0.454	0.418	0.42	0.409			
8							
9	Smoking Cessation	Stopped buying wine	Maternity Wine	Clothes	Wrong	Scale by Alpha	Normalize
10	0.001	0.001	0	0.001	0	0.0010	0.0009
11	0.001	0.001	0	0.001	0	0.0010	0.0009
12	0.001	0.001	0	0.001	0	0.0010	0.0009
13	0.001	0.001	0	0.001	0	0.0010	0.0009

Figura 7-28: O novo cálculo do peso

Faça Isso de Novo... e de Novo...

Agora você está pronto para construir um segundo toco. Primeiro, copie os dados do toco vencedor da iteração anterior desde X1:X4 até Y1:Y4.

Depois, copie os novos **valores** do peso da coluna X até a coluna B. A planilha inteira atualizará para selecionar o toco que for melhor para o novo conjunto de pesos. Como exibido na Figura 7-29, o segundo toco vencedor é o índice 7 (Folic Acid) em que 1 indica gravidez.

Você pode treinar 200 desses tocos como fez no processo de bagging. Apenas grave um macro que insira uma nova coluna Y, que copie os valores de X1:X4 em Y1:Y4 e que cole os pesos da coluna X na coluna B.

Após 200 iterações, sua taxa de erro ponderado terá aumentado bem próximo a 0,5 enquanto seu valor **alfa** terá caído 0,005 (veja a Figura 7-30). Lembre-se de que seu primeiro toco possuía um valor **alfa** de 0,2. Isso significa que os tocos finais são 40 vezes menos poderosos no processo de votação do que o seu primeiro toco.

X4	Q	R	S	T	U	V	W	X	Y
1	14	15	16	17	18		Winning Error	0.36801023	0.398
2	478	447	394	480	389		Column	7	5
3	425	493	476	397	480		Pregnant is	1	0
4	22	53	106	20	111		Alpha	0.270	0.20590272
5	75	7	24	103	20				
6	0	1	1	0	1				
7	0.477767	0.420369	0.38991	0.44	0.3814578				
8			Stopped buying wine	Maternity Wine	Maternity Clothes				
9	Cigarettes	Smoking				Wrong	Scale by Alpha	Normalize	
10	0	0.000916	0.00092	0	0.0009162	1	0.0012	0.0011	
11	0	0.000916	0.00092	0	0.0009162	1	0.0012	0.0011	
12	0	0.000916	0.00092	0	0.0009162	1	0.0012	0.0011	
13	0	0.000916	0.00092	0	0.0009162	1	0.0012	0.0011	

Figura 7-29: O segundo toco

	S	T	U	V	W	X	Y	Z	AA	AB	AC
1	16	17	18		Winning Error	0.49731981	0.49711002	0.4972793	0.49713974	0.49714264	0.497114
2	394	480		389	Column	5	4	14	3	8	
3	476	397		480	Pregnant is	0	1	0	0	1	
4	106	20		111	Alpha	0.005	0.006	0.005	0.006	0.006	0.0
5	24	103		20							
6	1	0		1							
7	0.49825	0.5		0.4985357							
8	Stopped buying wine		Maternity Clothes		Wrong	Scale by Alpha	Normalize				
9					0	0.0004	0.0004				
10	0.00038	0	0.0003769		0	0.0004	0.0004				
11	0.00036	0	0.0003641		0	0.0004	0.0004				
12	0.00096	0	0.0009561		0	0.0010	0.0010				

Figura 7-30: Um modelo Boosted

## Avaliando o Modelo Boosted

É isso! Agora você já treinou um modelo inteiro de tocos de decisão boosted. Pode compará-lo ao modelo bagged observando sua métrica de desempenho. Para que isso aconteça, primeiro você deve fazer previsões usando o modelo no conjunto dos dados de teste.

### Previsões do Conjunto de Teste

Primeiramente faça uma cópia de Test Set chamada **TestBoost** e insira quatro linhas em branco em seu topo para ter espaço para os tocos de decisão vencedores. Começando pela coluna W na aba TestBoost, cole seus tocos (todos os 200 no meu caso) no topo da planilha. Isso resulta na planilha exibida na Figura 7-31.

**Figura 7-31:** Os tocos de decisão colados em TestBoost

Em W6, você pode avaliar o primeiro toco na primeira linha dos dados de teste usando `OFFSET`, assim como fez no modelo bagged. Exceto que, desta vez, a previsão de uma gravidez retorna o valor **alfa** do toco (célula W4) e uma previsão de uma não-gravidez retorna **-alfa**:

```
=IF(OFFSET($A6, 0, W$2)=W$3, W$4, -W$4)
```

Copie essa fórmula por todos os tocos e por todas as linhas de teste (veja a Figura 7-32). Para fazer uma previsão para uma linha, deve-se somar esses valores por todas as previsões dos tocos individuais.

**Figura 7-32:** As previsões para cada linha dos dados de teste de cada toco

Nomeie V5 como **Score**. O escore para V6 é a soma das previsões à direita:

=SUM(W6 : HN6)

Copie essa soma. Você terá a planilha exibida na Figura 7-33. Um escore na coluna V acima de 0 significa que mais previsões com peso **alfa** foram em direção à gravidez do que à não-gravidez (veja a Figura 7-33).

### *Calculando o Desempenho*

Para medir o desempenho de um modelo boosted no conjunto de teste, simplesmente crie uma cópia da aba PerformanceBag chamada **PerformanceBoost**, aponte as fórmulas para a coluna V na aba TestBoost, e configure os valores de corte para variar do escore mínimo ao máximo produzidos pelo modelo boosted. No meu caso, incrementei os valores de corte em 0,25 entre uma previsão de escore mínimo de -8 e máximo de 4,5. Isso resulta na aba de desempenho exibida na Figura 7-34.

	R	S	T	U	V	W	X	Y	Z	AA	AB	C
1					Winning error	0.497	0.497	0.497	0.497	0.497	0.497	0
2					Column error	5	4	14	3	8	17	
3					Pregnant is	0	1	0	0	1	0	
4					Alpha	0.005	0.006	0.005	0.006	0.006	0.006	0
5	Wine	Maternity Clothes		PREGNANT	Score							
6	1	0		1	-1.575396867	0.005	-0.01	0.005	-0.01	-0.01	-0.01	
7	0	0		1	0.26830013	0.005	-0.01	0.005	0.006	-0.01	0.006	
8	0	0		1	0.110890325	0.005	-0.01	0.005	0.006	-0.01	0.006	
9	0	0		1	0.563866525	0.005	-0.01	0.005	0.006	-0.01	0.006	
10	0	0		1	-0.330642733	0.005	-0.01	0.005	-0.01	-0.01	0.006	
11	0	0		1	3.612180019	0.005	-0.01	0.005	0.006	-0.01	0.006	
12	0	1		1	2.803633128	0.005	-0.01	0.005	0.006	-0.01	0.006	0

**Figura 7-33:** As previsões finais a partir do modelo boosted

Com esse modelo, é possível ver que um corte de escore de 0 produz uma taxa positiva verdadeira de 85% com apenas 27% de taxa de falso positivo. Não é tão ruim para 200 tocos bobos.

Adicione a curva ROC do modelo boosted à curva ROC do modelo bagged para comparar as duas como fez no Capítulo 6. Como veremos na Figura 7-35, com 200 tocos, o modelo boosted ultrapassa o modelo bagged em muitos pontos no gráfico.

			Specificity /			
		Cutoff for Pregnant Classification	True Negative Rate	False Positive Rate (1 - Specificity)		True Positive Rate / Recall / Sensitivity
1	Min Prediction					
2	-8.066887574	-8	0.06	0.00	1.00	1.00
3		-7.75	0.06	0.00	1.00	1.00
4	Max Prediction	-7.5	0.06	0.00	1.00	1.00
5	4.689217724	-7.25	0.06	0.01	0.99	1.00
6		-7	0.06	0.01	0.99	1.00
7		-6.75	0.06	0.02	0.98	1.00
8		-6.5	0.06	0.02	0.98	1.00
9		-6.25	0.06	0.02	0.98	1.00
10		-6	0.06	0.02	0.98	1.00
11		-5.75	0.06	0.02	0.98	1.00
12		-5.5	0.06	0.04	0.96	1.00
13		-5.25	0.06	0.06	0.94	1.00
14		-5	0.06	0.08	0.92	1.00

Figura 7-34: A métrica do desempenho dos tocos boosted

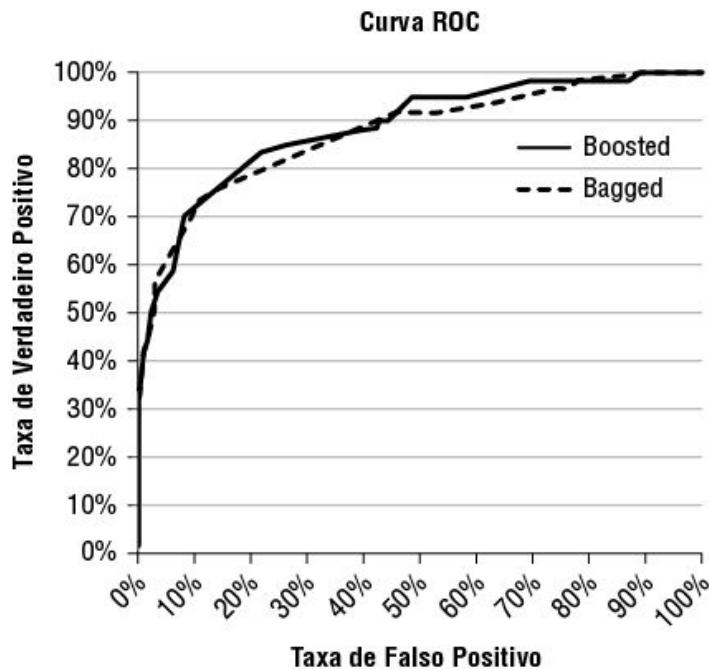


Figura 7-35: A curva ROC para os modelos boosted e bagged

### Além do Desempenho

Em geral, boosting requer menos árvores do que bagging para produzir um bom modelo. Na prática não é tão popular quanto bagging, porque há um risco bem maior de sobreajustar os dados. Uma vez que cada reconsideração dos dados em treinamento é baseada nos pontos desclassificados da iteração anterior, você pode acabar em uma situação em que esteja treinando classificadores para serem sensíveis a alguns poucos pontos nos dados.

Além disso, a redefinição iterativa dos pesos dos dados significa que o boosting, diferente de bagging, não pode ser paralelizado por todas as bases de CPU ou computadores múltiplos.

Dito isto, em um concurso cara a cara entre um modelo boosted bem ajustado e um modelo bagged bem ajustado, é difícil que o modelo bagged seja vencedor.

## Resumindo

Você acabou de ver como um monte de modelos simples podem ser combinados por meio de bagging ou boosting para formar um modelo ***ensemble***. Esses métodos eram desconhecidos até a metade de década de 1990, mas hoje eles são duas das técnicas de modelagem mais populares usadas no mercado.

Você pode fazer bag ou boost de qualquer modelo que desejar usar como um weak learner. Esses modelos não precisam ser tocos de decisão ou árvores. Por exemplo, recentemente, há muita discussão sobre boosting de modelos naïve Bayes iguais aos que você encontrou no Capítulo 3.

No Capítulo 10, você implementará um pouco do que encontrou neste capítulo usando a linguagem de programação R.

Se quiser aprender mais sobre esses algoritmos, eu recomendaria ler sobre eles em *The Elements of Statistical Learning* de Trevor Hastie, Robert Tibshirani e Jerome Friedman (Springer, 2009).

## 8

# Forecasting: ↑Respire ↑Devagar; Você ↑Não ↑Pode ↑Ganhar

**C**omo vimos nos Capítulos 3, 6 e 7, o aprendizado de máquina supervisionado é sobre prever um valor ou classificar uma observação usando um modelo treinado com dados antigos. Prever é similar. Claro, você pode prever sem dados (astrologia, alguém?). Mas, na previsão quantitativa, os dados antigos são usados para prever um resultado futuro. Na verdade, algumas dessas técnicas, tal como regressão múltipla (apresentada no Capítulo 6), são utilizadas em ambas as modalidades.

No entanto, previsão e aprendizado de máquina supervisionado diferem enormemente em seus problemas de espaço canônico. Os problemas mais típicos de previsão são sobre pegar alguns pontos de dados em alguns períodos de tempo (por exemplo vendas, demanda, fornecimento, GDP, emissão de carbono ou população) e projetar esses dados para um futuro. Na presença das tendências, ciclos e ocasionais atos de Deus, os dados do futuro podem estar extremamente fora dos limites do passado observado.

É aí que mora o problema em fazer previsões: diferentemente dos Capítulos 6 e 7 em que as mulheres grávidas adquiriam mais ou menos os mesmos produtos, fazer previsões é usado em contextos em que o futuro não se parece nem um pouco com o passado.

Quando você acha que possui uma boa projeção para a demanda imobiliária, a tal bolha explode e sua previsão vai ralo abaixo. Assim que você acha que possui uma previsão com uma boa demanda, uma enchente interrompe sua rede de fornecimento, limitando sua mercadoria, forçando a aumentar os preços e tornando suas vendas

completamente caóticas. Dados futuros de séries temporais podem e serão diferentes dos dados observados anteriormente.

*A única garantia de uma previsão é que a sua está errada.* Ouve-se muito isso no mundo da previsão. Mas isso não significa que você não tenta. Quando se trata de planejar seus negócios, às vezes a projeção é necessária. No MailChimp, talvez continuemos a crescer feito reis, ou um buraco se abrirá em Atlanta e nos engolirá. Mas nos esforçamos para prever o crescimento da melhor forma possível para que possamos planejar nossa infraestrutura e os caminhos do RH. Nem sempre você quer brincar de pega-pega.

Como verá neste capítulo, você pode tentar prever o futuro mas também pode quantificar a incerteza em torno dessa previsão.

Quantificar a incerteza da previsão criando *intervalos de previsão* é inestimável e frequentemente ignorado no mundo da previsão.

Como um sábio preditor disse, “um bom preditor não é mais esperto do que todos; eles apenas têm sua ignorância mais organizada”.

Sem mais delongas, vamos organizar um pouco de ignorância.

## O↑Mercado↑de↑Espadas↑Está↑a↑Mil

Imagine que você seja um fanático fã de *O Senhor dos Anéis*. Anos atrás, quando o primeiro filme foi lançado, você amarrou sua prótese de pé de hobbit e ficou horas na fila para a sessão de estreia da meia-noite. Em pouco tempo estava indo a convenções e discutindo em fóruns de mensagens se o Frodo poderia ter voado em uma águia direto para a Montanha da Perdição.

Um dia, decidiu colaborar com algo. Você fez um curso de especialização em trabalho com metais e começou a forjar suas próprias espadas. Sua espada favorita do livro era Anduril, a Chama do Ocidente. Você se tornou um expert em martelar suas beiradas largas em sua oficina caseira, e logo começou a vendê-las na Amazon, eBay e Mercado

Livre. Hoje em dia, suas réplicas se tornaram referência para os nerds de plantão, os negócios estão de vento em popa.

No passado, você lutou para alcançar a demanda com os materiais necessários, e então decidiu prever sua demanda futura. Para isso, você jogou suas vendas passadas em uma planilha. Mas como pegar esses dados antigos e usá-los?

Este capítulo aborda um conjunto de técnicas de previsão chamado método de ***suavização exponencial***. Elas são as técnicas mais simples e mais utilizadas no mundo dos negócios hoje em dia. Na verdade, eu tenho alguns poucos milionários na ponta da língua que usam essas técnicas para fazer previsões, porque elas se provaram como as mais precisas para os seus dados.

Essa precisão decorre da simplicidade das técnicas — elas sobrevivem ao sobreajuste dos dados históricos muitas vezes esparsos usados na previsão. Além do mais, com essas técnicas, fica relativamente fácil computar intervalos de previsão ***em torno das*** previsões de suavização exponencial, logo, você fará um pouco disso também.

## Conhecendo↑os↑Dados↑de↑Séries Temporais

### NOTA

A pasta de trabalho do Excel usada neste capítulo, “SwordForecasting.xlsx”, está disponível para download no site da editora em [www.altabooks.com](http://www.altabooks.com).  
br, procurando pelo nome do livro. Nesta pasta de trabalho estão incluídos todos os dados iniciais se você quiser trabalhar com eles. Ou, pode apenas ler acompanhando com as planilhas que já estão nela.

A pasta de trabalho para este capítulo contém as demandas das espadas dos últimos 36 meses começando em janeiro de três anos atrás. Os dados são exibidos na aba Timeseries na Figura 8-1. Como já mencionei, dados

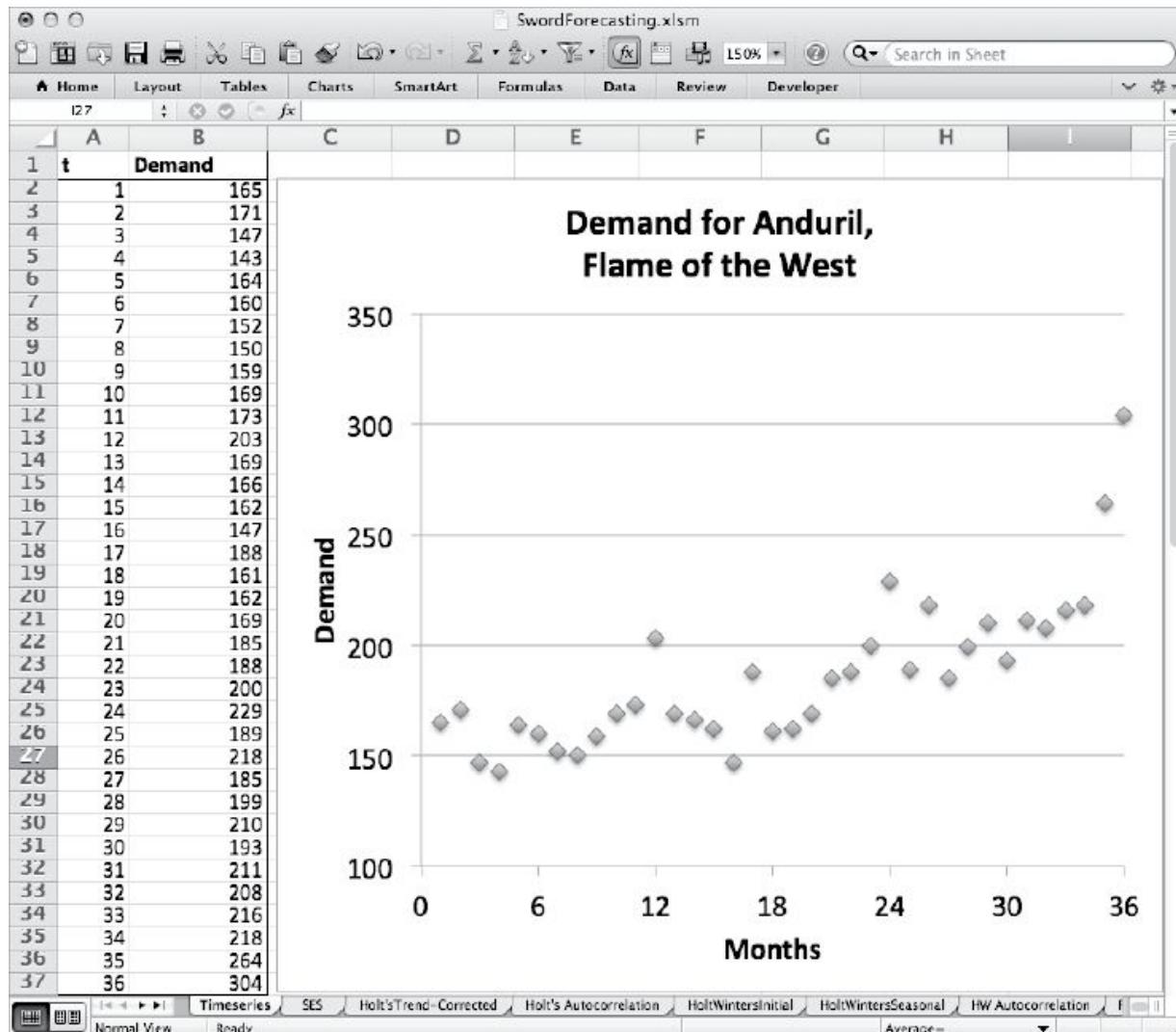
como estes — observações em intervalos de tempo regulares — são chamados de ***dados em séries temporais***. O intervalo de tempo pode ser qualquer um que seja apropriado para o problema em questão, seja sobre a população anual de habitantes ou os preços diárias da gasolina.

A	B
t	Demand
1	165
2	171
3	147
4	143
5	164
6	160
7	152
8	150
9	159
10	169
11	173
12	203
13	169
14	166
15	162
16	147
17	188
18	161
19	162
20	169
21	185
22	188
23	200
24	229
25	189
26	218
27	185
28	199
29	210
30	193
31	211
32	208
33	216
34	218
35	264
36	304

**Figura 8-1:** Dados em séries temporais

Neste caso, você possui dados de uma demanda mensal de espadas e a primeira coisa que você deveria fazer com eles é criar um gráfico, como na Figura 8-2. Para criar um gráfico como esse, destaque as colunas A e B no Excel e selecione Scatter da seção de gráficos do menu da faixa de opções do Excel (aba Charts no Mac, Insert Tab no Windows). Você pode adaptar o alcance dos seus eixos clicando com o botão direito e selecionando a opção Format.

Então, o que você vê na Figura 8-2? Os dados variam dos 140 de três anos atrás até 304 do mês passado. Isso é o dobro da demanda em três anos — então talvez ele tenha uma tendência ascendente? Voltaremos a esse pensamento em um minuto.



**Figura 8-2:** Gráfico de dispersão dos dados em séries temporais

Há alguns altos e baixos que podem indicar um padrão sazonal. Por exemplo, os meses 12, 24 e 36, todos dezembro, possuem a demanda mensal mais alta de todos os anos. No entanto, isso pode ser casual ou uma tendência. Vamos descobrir.

# Começando com Coluna com Suavização Exponencial Simples

As técnicas de suavização exponencial se baseiam em uma previsão futura a partir dos dados passados em que as observações recentes são mais ponderadas do que as antigas. Essa ponderação é feita pelas **constantes de suavização**. Esse primeiro método de suavização exponencial com o qual você lidará agora é chamado de **suavização exponencial simples** (SES, do inglês *Simple Exponential Smoothing*), e ele usa somente uma constante de suavização, como você verá.

A suavização exponencial simples presume que os seus dados em séries temporais são feitos de dois componentes: um **nível** (ou média) e algum erro em torno dele. Não há tendência nem sazonalidade, apenas um nível no qual a demanda paira com alguns errinhos aqui e ali. Ao preferir as observações mais recentes, SES pode explicar as mudanças nesse nível. Em linguagem de fórmula, temos:

*Demanda em tempo  $t$  = nível + erro aleatório próximo nível no tempo  $t$*

E a estimativa mais atual do nível serve como uma previsão para os períodos de tempo futuros. Se você estiver no mês 36, qual seria uma boa estimativa para o período 38? A estimativa mais recente do nível. E no período 40? O nível. Simples — por isso o nome suavização exponencial simples.

E como obter uma estimativa do nível?

*Se presumir que todos os seus valores históricos possuem a mesma importância, você simplesmente calcula uma média direta.*

Essa média geraria um nível e você preveria o futuro ao dizer, “a demanda no futuro é a média da demanda passada”. Há empresas que fazem isso. Já vi previsões mensais em empresas nas quais os meses futuros foram iguais à média daqueles mesmos meses ao longo dos últimos anos. Mais uma “margem de erro” para teste. Sim, previsões são

feitas tão transitoriamente que, mesmo em grande escala, palavras públicas das empresas como “margem de erro” ainda são usadas. Eca.

Porém, quando esse nível sofre mudanças ao longo do tempo, não é possível dar pesos iguais para cada ponto histórico da forma que a média dá. Os dados de 2008 a 2013 deveriam carregar o mesmo peso ao prever 2014? Talvez, mas para a maioria das empresas, provavelmente não. Portanto, você quer uma estimativa de média que dê mais peso às suas observações das demandas recentes.

Então vamos pensar sobre calcular a média, mas em vez de calcular a média diretamente, mover os pontos de dados em ordem, atualizando as médias do cálculo no caminho. Para começar, digamos que a estimativa inicial do nível é a média de alguns pontos de dados iniciais. Neste caso, escolha o valor dos dados do primeiro ano. Chame essa estimativa inicial do nível de *nível0*:

$$\text{nível0} = \text{a média da demanda do primeiro ano (meses 1 – 12)}$$

É 163 para a demanda de espadas.

Agora, o modo como a suavização exponencial trabalha é que, apesar de saber da demanda para os meses 1 ao 36, você pegará seus componentes de previsão mais recentes e os usará para prever um mês a frente ao longo da série inteira.

Desta forma, você usa o *nível0* (163) como previsão para a demanda no mês 1.

Visto que previu o período 1, você dá um passo à frente no tempo a partir do período 0 ao período 1. A demanda atual era de 165, então você errou por duas espadas. Você deveria atualizar a estimativa do nível e então explicar o erro. A suavização exponencial simples usa esta equação:

$$\text{nível1} = \text{nível0} + \text{alguma porcentagem} * (\text{demanda1} - \text{nível0})$$

Repare que (*demanda1* – *nível0*) é o erro que você obtém quando prevê o período um com a estimativa do nível inicial. Seguindo adiante:

$$\text{nível2} = \text{nível1} + \text{alguma porcentagem} * (\text{demanda2} - \text{nível1})$$

E, novamente:

$$\text{nível3} = \text{nível2} + \text{alguma porcentagem} * (\text{demanda3} - \text{nível2})$$

Agora, a porcentagem de erro que você quer aplicar ao nível é a **constante de suavização**, e, para o nível, ele é historicamente chamado de **alfa**. Pode ser qualquer valor entre 0 e cem por cento (0 e 1).

Se você configurar **alfa** para 1, explicará o erro, o que significa que o nível do período atual é a demanda do período atual.

Se configurar **alfa** para 0, você não conduz absolutamente nenhuma correção de erro naquela estimativa do primeiro nível.

Possivelmente, você vai querer algo entre esses dois extremos, mas você aprenderá a escolher o melhor valor de **alfa** mais tarde.

Você pode seguir com este cálculo através do tempo:

$$\text{nível período atual} = \text{nível período anterior} + \text{alfa} * (\text{demanda período atual} - \text{nível período anterior})$$

Eventualmente, você terá uma estimativa de nível final, **nível36**, em que as observações da última demanda contam mais porque seus ajustes de erros não foram multiplicados por **alfa** milhares de vezes:

$$\text{nível36} = \text{nível35} + \text{alfa} * (\text{demanda36} - \text{nível35})$$

A estimativa final do nível é o que você verá como a previsão dos meses futuros. A demanda pra o mês 37? Bem, é o **nível36**. E a demanda para o mês 40? **nível36**. Mês 45? **nível36**. Você entendeu. A estimativa do nível final é a melhor que você tem para o futuro, então será a que você usará.

Vamos dar uma olhada na planilha.

## Estabelecendo↑uma↑Previsão↑de↑ Suavização↑Exponencial↑Simples

A primeira coisa que fará é criar uma nova planilha nas pastas de trabalho chamada **SES**. Cole os dados em séries temporais nas colunas A e B começando na linha 4 para deixar algum espaço no topo da planilha para um valor **alfa**. Você pode colocar o número de meses que tem nos

seus dados (36) na célula A2, e uma área inicial para o valor ***alfa*** em C2. Começarei com 0,5 porque ele está entre 0 e 1, e esse é meu jeito.

Agora, na coluna C, coloque os cálculos do nível. Você precisará inserir uma nova linha 5 dentro dos dados em séries temporais no topo para a estimativa de nível inicial no tempo 0. Em C5, utilize o cálculo a seguir:

```
=AVERAGE(B6:B17)
```

Isso constitui o primeiro ano dos dados válidos para dar ao nível inicial. A planilha então se parece com a Figura 8-3.

	A	B	C
1	Total Months		Level smoothing parameter (alpha)
2	36		0.50
3			
4	t	Demand	Level Estimate
5	0	163	163
6	1	165	
7	2	171	
8	3	147	
9	4	143	
10	5	164	
11	6	160	
12	7	152	
13	8	150	
14	9	159	
15	10	169	
16	11	173	
17	12	203	
18	13	169	
19	14	166	
20	15	162	

**Figura 8-3:** Estimativa do nível inicial para suavização exponencial simples

### ***Adicionando a Previsão e o Erro de um Passo***

Agora que você adicionou o valor do primeiro nível na planilha, pode prosseguir no tempo usando a fórmula SES definida na seção anterior. Para fazer isso, precisará adicionar duas colunas: a coluna com a previsão

de um passo (D) e a coluna de previsão de erro (E). A previsão de um passo para o período de tempo 1 é *nível0* (célula C5), e o cálculo de erro é a demanda atual menos a previsão:

=B6 - D6

A estimativa do nível para o período 1 é o nível anterior adaptado pelo *alfa* vezes o erro, ou seja:

=C5 + C\$2 \* E6

Observe que coloquei um \$ na frente do valor *alfa* para que quando arrastar a fórmula pela tabela, a linha de referência absoluta deixe *alfa* em paz. Isso resulta na planilha exibida na Figura 8-4.

	A	B	C	D	E
1	Total Months		Level smoothing parameter (alpha)		
2	36		0.50		
3					
4	t	Demand	Level Estimate	One-step Forecast	Forecast Error
5	0		163		
6	1	165	164	163	2
7	2	171			
8	3	147			
9	4	143			
10	5	164			
11	6	160			
12					

**Figura 8-4:** Produzindo a previsão de um passo, o erro e o cálculo de nível para o período 1

### Jogue Tudo Para Baixo!

Tendo se divertido o suficiente, você já acabou por aqui. Arraste C6:E6 por todos os 36 meses, e *voila*, você tem o *nível36*.

Vamos adicionar os meses 37–48 à coluna A. A previsão para esses próximos 12 meses é somente o *nível36*. Então em B42, você pode adicionar:

=C\$41

como a previsão e arrastá-la pelo próximo ano.

Isso gera uma previsão de 272, como mostra a Figura 8-5.

SwordForecasting.xlsm					
Home Layout Tables Charts SmartArt Formulas Data					
B53	fx	-C\$41			
A	B	C	D	E	
1	Total Months	Level smoothing parameter (alpha)			
2	36	0.50			
3					
4	t	Demand	Level Estimate	One-step Forecast	Forecast Error
38	33	216	211.1855079	206.371016	9.62898412
39	34	218	214.592754	211.185508	6.81449206
40	35	264	239.296377	214.592754	49.407246
41	36	304	271.6481885	239.296377	64.703623
42	37	271.64819			
43	38	271.64819			
44	39	271.64819			
45	40	271.64819			
46	41	271.64819			
47	42	271.64819			
48	43	271.64819			
49	44	271.64819			
50	45	271.64819			
51	46	271.64819			
52	47	271.64819			
53	48	271.64819			

**Figura 8-5:** Previsão da suavização exponencial simples com *alfa* de 0,5

Mas, isso é o melhor que você pode fazer? Bem, a forma de otimizar esse *alfa* é configurar *alfa* — quanto maior for o *alfa*, menos importância você dará aos pontos das demandas antigas.

### Otimizando o Erro de um Passo

Semelhante à forma como você minimizou a soma dos erros quadrados quando ajustou a regressão no Capítulo 6, você pode encontrar a melhor suavização constante para a previsão ao minimizar a soma dos erros quadrados para as próximas previsões de um passo.

Vamos adicionar um cálculo do erro quadrado na coluna F, que é o valor da coluna E ao quadrado, levar esse cálculo por todos os 36 meses e somá-lo na célula E2 como a soma dos erros quadrados (SSE). Isso resulta na planilha exibida na Figura 8-6.

Além disso, você adicionará o *erro padrão* à planilha na célula F2. O erro padrão é a raiz quadrada do SSE dividido por 35 (36 meses menos a quantidade de parâmetros suavizadores no modelo, que para a suavização exponencial simples é 1).

	A	B	C	D	E	F
1	Total Months		Level smoothing parameter (alpha)		SSE	
2	36		0.50		=SUM(F6:F41)	
4	t	Demand	Level Estimate	One-step Forecast	Forecast Error	Squared Error
5	0		163			
6	1	165	164	163	2	4
7	2	171	167.5	164	7	49
8	3	147	157.25	167.5	-20.5	420.25
9	4	143	150.125	157.25	-14.25	203.0625
10	5	164	157.0625	150.125	13.875	192.515625
11	6	160	158.53125	157.0625	2.9375	8.62890625
12	7	152	155.265625	158.53125	-6.53125	42.6572266
13	8	150	152.6328125	155.265625	-5.265625	27.7268066
14	9	159	155.8164063	152.632813	6.3671875	40.5410767
15	10	169	162.4082031	155.816406	13.1835938	173.807144

Figura 8-6: A soma dos erros quadrados para a suavização exponencial simples

O erro padrão é uma estimativa do desvio padrão do erro de um passo. Você já viu o desvio padrão no Capítulo 4. Ele é apenas uma medida de propagação do erro.

Se você tiver um bom modelo de previsão ajustado, seu erro terá a média 0. Isso quer dizer que a previsão está *neutra*. Ele superestima a demanda ao mesmo tempo que subestima. O erro padrão quantifica a propagação em torno de 0 quando a previsão é neutra.

Portanto, na célula F2, você pode calcular o erro padrão desta forma:

=SQRT(E2 / (36 - 1))

Para um valor **alfa** de 0,5, ele resulta em 20,94 (veja a Figura 8-7). E, se você se lembrar da regra 68-95-99,7 da distribuição normal vista no Capítulo 4, isso quer dizer que 68% dos erros de previsão de um passo deveriam ser menores que 20,94 e maiores que -20,94.

Agora, o que você quer fazer é diminuir essa propagação o máximo que puder ao encontrar um valor **alfa** apropriado. Você poderia tentar um monte de valores diferentes para **alfa**. Mas usará o Solver pela enésima vez neste livro.

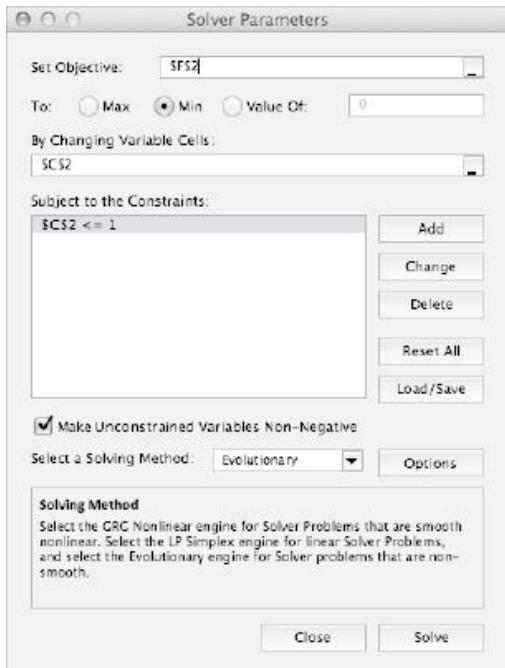
A configuração do Solver é muito fácil. Abra o Solver, configure o objetivo para o erro padrão em F2, configure a variável de decisão para **alfa** em C2, adicione uma restrição a fim de C2 ser menor do que 1, e marque a caixa de verificação para que a decisão seja não-negativa. Os cálculos de nível repetitivos que são feitos para cada erro de previsão são altamente não-lineares, portanto você precisará usar o algoritmo evolucionário para otimizar **alfa**.

The screenshot shows a Microsoft Excel spreadsheet titled "SwordForecasting.xlsm". The formula bar displays the formula =SQRT(E2/(36-1)). The data is organized into several columns: A (t), B (Demand), C (Level Estimate), D (One-step Forecast), E (Forecast Error), and F (Squared Error). Row 1 contains headers: "Total Months" (36), "Level smoothing parameter (alpha)" (0.50), "SSE" (15346.86), and "Standard Error" (20.94). Rows 2 through 12 show the actual demand values (165, 171, 147, 143, 164, 160) and their corresponding Level Estimate, One-step Forecast, Forecast Error, and Squared Error. The bottom of the screen shows the ribbon tabs: Home, Layout, Tables, Charts, SmartArt, Formulas, Data, Review, and the Solver tab.

	A	B	C	D	E	F
1	Total Months	36	Level smoothing parameter (alpha)	0.50	SSE	Standard Error
2					15346.86	20.94
3						
4	t	Demand	Level Estimate	One-step Forecast	Forecast Error	Squared Error
5		0	163			
6		1	165	164	163	2
7		2	171	167.5	164	7
8		3	147	157.25	167.5	-20.5
9		4	143	150.125	157.25	-14.25
10		5	164	157.0625	150.125	13.875
11		6	160	158.53125	157.0625	2.9375
12		7	152	155.265625	158.53125	-6.53125

Figura 8-7: O cálculo do erro padrão

A elaboração do Solver deve se parecer com o conteúdo na Figura 8-8. Ao pressionar Solve, você obtém um valor ***alfa*** de 0,73, dando-lhe um novo erro padrão de 20,39. Não foi tão melhor assim.



**Figura 8-8:** A elaboração do Solver para a otimização do ***alfa***

### ***Vamos Representar em Gráficos***

A melhor forma de checar uma previsão com absoluta certeza é representá-la em gráfico junto à sua demanda histórica e ver como a demanda prevista reaproveita a passada. Pode-se selecionar os dados de demanda histórica e a previsão e representá-los em gráfico. Eu gosto da aparência do gráfico de dispersão com linha reta do Excel. Para começar, selecione os dados históricos em A6:B41, e escolha o gráfico de dispersão linear reto na seção de gráficos do Excel.

Uma vez adicionado o gráfico, clique com o botão direito no centro dele, escolha Select Data, e adicione uma nova série ao gráfico com linha reta com apenas os valores previstos de A42:B53. Você também pode adicionar alguns nomes aos eixos se quiser, depois disso tudo deve ter algo parecido com a Figura 8-9.



**Figura 8-9:** Representando em gráfico a previsão de suavização exponencial simples final

## Talvez Você Tenha uma Tendência

Somente observando aquele gráfico, alguns pontos se destacam.

Primeiro, a suavização exponencial simples é uma linha plana — o nível. Os dados da demanda dos últimos 36 meses estão em aumento e parecem fazer parte de uma tendência ascendente, especialmente no final.

Sem denegrir a visão humana, mas como provar isso?

Prova-se ao ajustar uma regressão linear para os dados da demanda e realizando um **teste t** na inclinação da linha de tendência, como fez no Capítulo 6.

Se a inclinação da linha for não-zero e estatisticamente significante (possui um valor p menor do que 0,05 no **teste t**), você pode ficar confiante de que os dados têm uma tendência. Se a última frase não fez sentido para você, verifique a seção de teste estatístico no Capítulo 6.

Retorne à aba Timeseries nas pastas de trabalho para realizar o teste de tendência.

Bem, você provou sua coragem ao realizar o **teste F** e o teste t **manualmente** no Capítulo 6. Ninguém quer que você faça isso de novo.

Neste capítulo, você usará a função embutida `LINEST` do Excel para ajustar uma regressão linear, mexer na inclinação, no erro padrão do coeficiente de inclinação e nos graus de liberdade (veja o Capítulo 6 para entender esses termos). Depois, pode calcular sua estatística t e executá-la pela função `TDIST` assim como no Capítulo 6.

Se você não usou `LINEST` antes, a ajuda de documentação do Excel sobre a função é muito boa. Você fornece `LINEST` com os dados variáveis dependentes (a demanda na coluna B) e os dados variáveis independentes (você tem somente uma variável independente e é o tempo na coluna A).

Você também deve fornecer uma referência de `TRUE` para deixar que a função saiba ajustar um intercepto como parte da linha de regressão, e uma segunda referência `TRUE` para obter detalhes estatísticos como o erro padrão e o R-quadrado. Para os dados da aba Timeseries, uma regressão linear pode ser realizada assim:

```
=LINEST(B2:B37,A2:A37,TRUE,TRUE)
```

No entanto, essa chamada somente retornará a inclinação da linha de regressão porque `LINEST` é uma fórmula de array. `LINEST` retorna todas as estatísticas de regressão para um array. Você pode executar `LINEST` como uma fórmula de array para jogar tudo fora em uma série selecionada em uma planilha, ou executar `LINEST` pela fórmula `INDEX` e retirar apenas os valores que precisa um por um.

Por exemplo, os primeiros componentes de uma linha de regressão que `LINEST` dá são os coeficientes de regressão. Então você pode retirar a inclinação da regressão na célula B39 na aba Timeseries ao alimentar `LINEST` pelo `INDEX`:

```
=INDEX(LINEST(B2:B37,A2:A37,TRUE,TRUE),1,1)
```

Você tem como retorno uma inclinação de 2,54, o que significa que a linha de regressão está mostrando uma tendência ascendente de 2,54

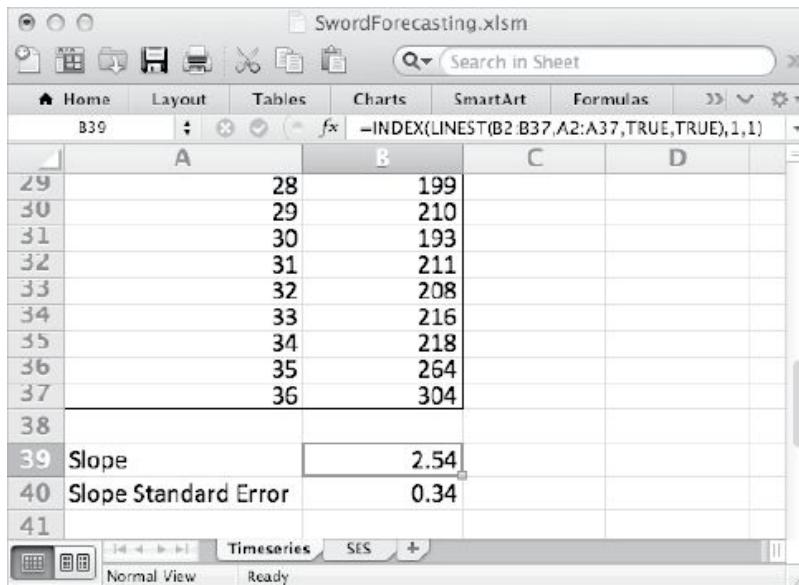
para a demanda de espadas adicionais por mês. Portanto, há uma inclinação. Mas ela é estatisticamente significante?

Para realizar um teste t na inclinação, você precisa obter o erro padrão dele e os graus de liberdade da regressão. LINEST armazena o valor do erro padrão na linha 2, coluna 1 do seu array de resultados. Então, em B40, temos:

```
=INDEX(LINEST(B2:B37,A2:A37,TRUE,TRUE),2,1)
```

A única mudança de obter a inclinação é que na fórmula INDEX você obtém a linha 2, coluna 1 para o erro padrão em vez de linha 1 coluna 1 para a inclinação.

O erro padrão da inclinação é exibido como 0,34 gerando a planilha exibida na Figura 8-10.



**Figura 8-10:** A inclinação e o erro padrão para a linha de regressão ajustada à demanda histórica

Da mesma forma, a documentação LINEST do Excel repara que os graus de liberdade para a regressão são retornados no valor da quarta linha e segunda coluna no array de resultado. Logo, em B41 pode-se arrastar desta maneira:

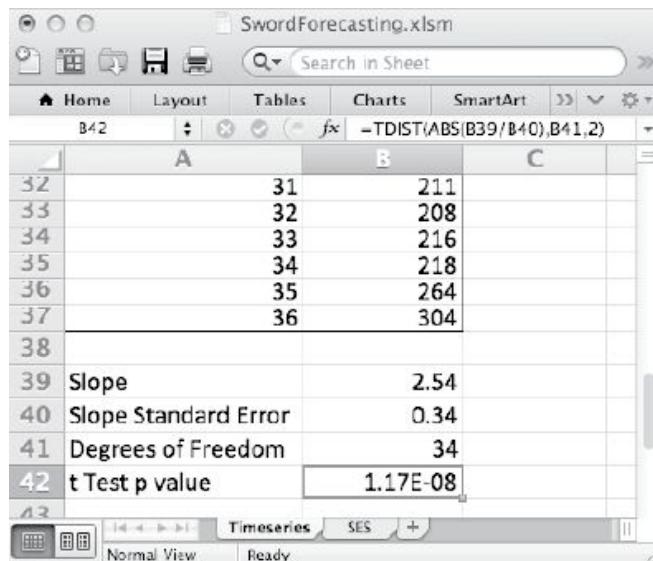
```
=INDEX(LINEST(B2:B37,A2:A37,TRUE,TRUE),4,2)
```

Você deve obter 34 para os graus de liberdade (como visto no Capítulo 6, isso é calculado como 36 pontos de dados menos 2 coeficientes vindos da regressão linear).

Você agora possui três valores necessários para realizar um teste t na significância estatística da sua tendência ajustada. Assim como no Capítulo 6, você pode calcular o teste estatístico como o valor absoluto da inclinação dividido pelo erro padrão para a inclinação. Pode-se obter o valor p para essa estatística a partir da distribuição t com 34 graus de liberdade usando a função TDIST em B42:

```
=TDIST(ABS(B39/B40), B41, 2)
```

Isso retorna um valor p próximo a 0 sugerindo que, se a tendência fosse inexistente na realidade (inclinação de 0), não haveria chance de obter uma inclinação tão extrema a partir da nossa regressão. A Figura 8-11 mostra isso.



**Figura 8-11:** Sua tendência é legítima

Muito bem! Então você possui uma tendência. Agora, precisa incorporá-la na sua previsão.

# A Suavização Exponencial com Tendência Corrigida de Holt

A *Suavização Exponencial com Tendência Corrigida de Holt* expande a suavização exponencial para criar uma previsão a partir de dados que possuem uma tendência linear. Geralmente ela é chamada de *suavização exponencial dupla*, pois diferentemente de SES que possui o parâmetro suavizador *alfa* e um componente não-erro, a suavização exponencial dupla possui dois.

Se a série temporal possui uma tendência linear, pode escrevê-la desta forma:

*Demanda no tempo t = nível + t\*tendência + erro aleatório em torno do nível no tempo t*

As estimativas mais atuais do nível e da tendência (vezes o número de períodos) servem como uma previsão para os futuros períodos de tempo. Se você estiver no mês 36, o que é uma boa estimativa de demanda no período de tempo 38? A estimativa mais atual *mais* dois meses da tendência. E o tempo 40? O nível *mais* quatro meses da tendência. Não é tão simples quanto a SES mas chega bem perto.

Agora, assim como na suavização exponencial simples, você precisa obter algumas estimativas iniciais dos valores de nível e tendência, chamado de *nível0* e *tendência0*. Uma forma comum de obtê-los é representar em gráfico a primeira metade dos seus dados da demanda e enviar uma linha de tendência por ele (assim como fez no Capítulo 6 no exemplo da alergia a gatos). A inclinação dessa linha é *tendência0* e o intercepto y é *nível0*.

A Suavização Exponencial com Tendência de Holt possui duas equações de atualização, uma para o nível enquanto você segue pelo tempo e uma para a tendência. A equação de nível ainda usa um parâmetro de suavização chamado *alfa*, enquanto que a equação de tendência usa um parâmetro frequentemente chamado de *gama*. Eles são

exatamente os mesmos — valores entre 0 e 1 que regulam quanto o erro de previsão de um passo está incorporado de volta nas estimativas.

Então, esta é a nova equação de atualização de nível:

$$\text{nível1} = \text{nível0} + \text{tendência0} + \text{alfa} * (\text{demanda1} - (\text{nível0} + \text{tendência0}))$$

Repare que  $(\text{nível0} + \text{tendência0})$  é a previsão de um passo dos valores iniciais para mês 1, então  $(\text{demanda1} - (\text{nível0} + \text{tendência0}))$  é o erro de um passo. Essa equação parece idêntica à equação de nível do SES, exceto que há a aplicação do valor da tendência para um período de tempo toda vez que você conta um intervalo. Portanto, a equação geral para a estimativa de nível é:

$$\begin{aligned} \text{nível período atual} &= \text{nível período anterior} + \text{tendência período} \\ &\quad \text{anterior} + \text{alfa} * (\text{demanda período atual} - (\text{nível período anterior} + \text{tendência período anterior})) \end{aligned}$$

Com essa nova técnica de suavização, você também precisa de uma equação de atualização de tendência. Para o primeiro intervalo é:

$$\text{tendência1} = \text{tendência0} + \text{gama} * \text{alfa} * (\text{demanda1} - (\text{nível0} + \text{tendência0}))$$

Então a equação da tendência é similar à equação de atualização de nível. Você pega a estimativa anterior da tendência e a ajusta em **gama vezes** a quantia de erro incorporada à atualização de nível correspondente (o que faz um sentido intuitivo porque somente parte dos erros que você está usando para ajustar o nível seriam atribuíveis à estimativa de tendência variável ou pobre).

Portanto, a equação geral para a estimativa da tendência é:

$$\begin{aligned} \text{tendência período atual} &= \text{tendência período anterior} + \text{gama} * \text{alfa} * \\ &\quad (\text{demanda período atual} - (\text{nível período anterior} + \text{tendência período anterior})) \end{aligned}$$

## Configurando↑Suavização↑Exponencial↑com↑Tendência↑Corrigida↑de↑Holt↑em↑uma↑Planilha

Para começar, crie uma nova aba chamada Holt's Trend-Corrected. Nessa aba, assim como a aba de suavização exponencial simples, cole os dados em séries temporais na linha 4 e insira uma linha 5 vazia para as estimativas iniciais.

A coluna C conterá novamente as estimativas iniciais, e você colocará as estimativas de tendência na coluna D. No topo dessas duas colunas você colocará os valores de *alfa* e *gama*. Você vai otimizá-los com o Solver daqui a pouco, mas, por enquanto, jogue alguns 0,5s. Isso gera a planilha exibida na Figura 8-12.

Para os valores iniciais de nível e tendência que vão em C5 e D5, vamos representar em um gráfico de dispersão os primeiros 18 meses dos dados e adicionar uma linha de tendência a ele com a equação (se você não sabe como adicionar uma linha de tendência em um gráfico de dispersão, veja o Capítulo 6 e procure um exemplo). Isso produz uma tendência inicial de 0,8369 e um nível inicial (intercepto da linha de tendência) de 155,88.

Screenshot of Microsoft Excel showing the initial setup for Holt's Trend-Corrected forecasting. The spreadsheet includes headers for parameters and actual demand data, and tabs for Timeseries, SES, and Holt'sTrend-Corrected.

	A	B	C	D
1	Total months		Level smoothing parameter (alpha)	Trend smoothing parameter (gamma)
2	36		0.5	0.5
3		Actual Demand	Level	Trend
4	t			
5	0			
6	1	165		
7	2	171		
8	3	147		
9	4	143		
10	5	164		
11	6	160		
12	7	152		

Normal View   Ready

**Figura 8-12:** Começando com os parâmetros suavizadores configurados em 0,5

Ao adicioná-los em D5 e C5 respectivamente, você obtém a planilha exibida na Figura 8-13.

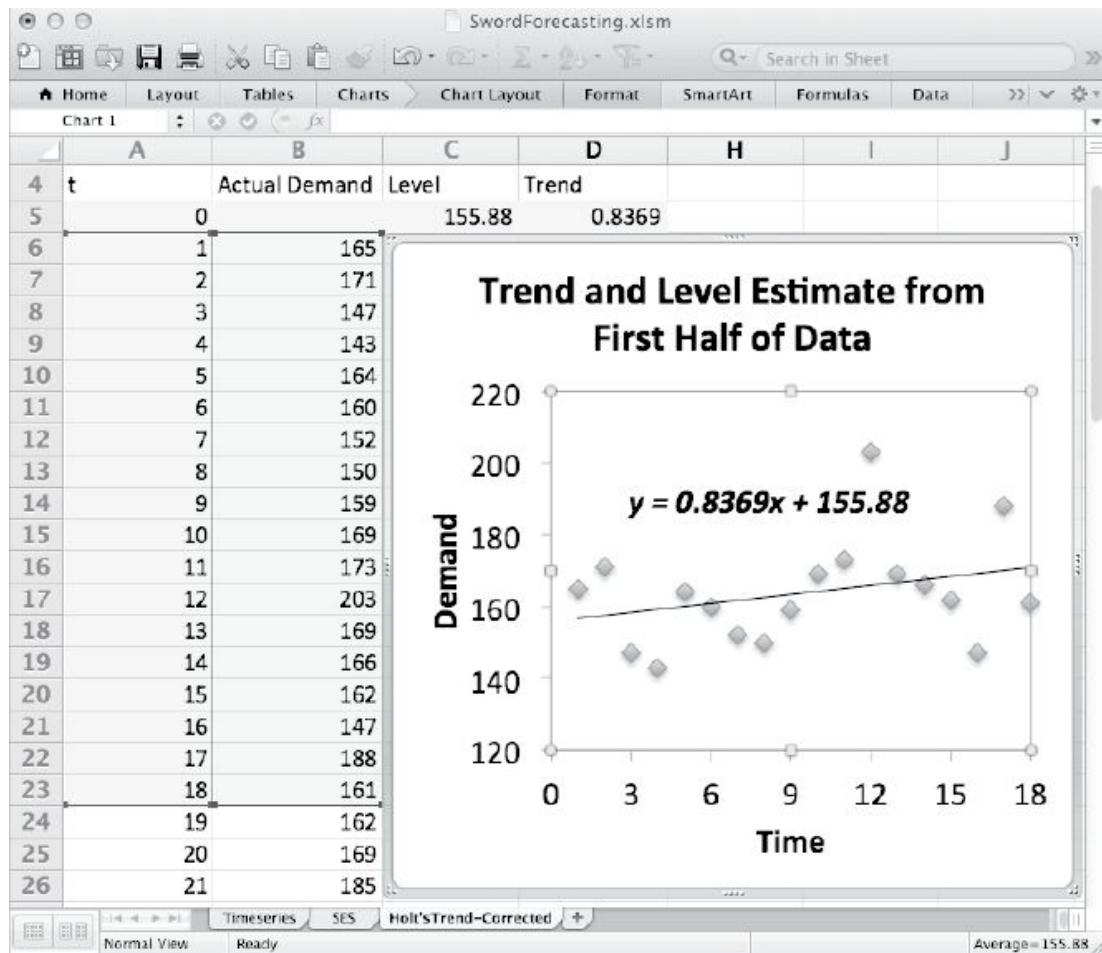


Figura 8-13: Os valores iniciais de nível e tendência

Agora, nas colunas E e F, adicione uma previsão de um passo e a previsão das colunas de erro. Se observar a linha 6, a previsão de um passo é apenas o nível anterior mais a tendência de um mês usando a estimativa anterior — ou seja, C5+D5. E o erro de previsão é o mesmo da suavização exponencial simples; F6 é somente a demanda atual menos a previsão de um passo — B6-E6.

Pode-se atualizar o nível na célula C6 como o nível anterior mais a tendência anterior mais **alfa** vezes o erro:

$$=C5+D5+C\$2*F6$$

A tendência em D6 é atualizada como a tendência anterior mais **gama** vezes **alfa** vezes o erro:

$$=D5+D\$2*C\$2*F6$$

Repare que você precisa usar as referências absolutas em ambos **alfa** e **gama** a fim de arrastar as fórmulas. Você fará isso agora — arraste C6:F6 por todo o mês 36. Vemos isso na Figura 8-14.

		Level smoothing parameter (alpha)	Trend smoothing parameter (gamma)			
1	Total months	0.5	0.5			
2	36					
3		Actual Demand	Level	Trend	One-step Forecast	Forecast Error
4	t					
35	30	193	197.585842	-2.65641058	202.171684	-9.1716838
36	31	211	202.964716	1.361231594	194.929431	16.0705687
37	32	208	206.162974	2.279744781	204.325947	3.67405275
38	33	216	212.221359	4.169065179	208.442718	7.55728159
39	34	218	217.195212	4.571459083	216.390424	1.60957562
40	35	264	242.883336	15.12979126	221.766671	42.2333287
41	36	304	281.006563	26.62650954	258.013127	45.9868731

**Figura 8-14:** Arrastando os cálculos de erros de nível, tendência, previsão e erro

## Prevendo Períodos Futuros

Para fazer previsões a partir do mês 36, é possível adicionar ao nível final (em que um **alfa** e um **gama** de 0,5 é 281) a quantidade de meses que você está prevendo vezes a estimativa da tendência final. Pode-se calcular a quantidade de meses entre o mês 36 e o mês que você quer subtraindo um mês da coluna A do outro.

Por exemplo, para prever o mês 37 na coluna B42, você usaria:

=C\$41 + (A42 - A\$41) \* D\$41

Ao usar as referências absolutas para o mês 36, a tendência final, e o nível final, você pode arrastar a previsão por todo o mês 48, gerando a planilha exibida na Figura 8-15:

Screenshot of Microsoft Excel showing a spreadsheet titled "SwordForecasting.xlsm". The formula bar displays: MIN -C\$41+(A53-A\$41)\*D\$41. The table below shows historical data and forecasts for 12 months.

	A	B	C	D	E	F
1	Total months		Level smoothing parameter (alpha)	Trend smoothing parameter (gamma)		
2	36		0.5	0.5		
4	t	Actual Demand	Level	Trend	One-step Forecast	Forecast Error
38	33	216	212.221359	4.169065179	208.442718	7.55728159
39	34	218	217.195212	4.571459083	216.390424	1.60957562
40	35	264	242.883336	15.12979126	221.766671	42.2333287
41	36	304	281.006563	26.62650954	258.013127	45.9868731
42	37	307.633				
43	38	334.26				
44	39	360.886				
45	40	387.513				
46	41	414.139				
47	42	440.766				
48	43	467.392				
49	44	494.019				
50	45	520.645				
51	46	547.272				
52	47	573.898				
53	48	D\$41				

**Figura 8-15:** Prevendo meses futuros com a Suavização Exponencial com Tendência de Holt

Tal como a aba de suavização exponencial simples, pode-se representar em gráfico a demanda histórica e a previsão como duas séries em um gráfico de dispersão com linha reta, como mostra a Figura 8-16.

Com um **alfa** e um **gama** de 0,5, essa previsão parece um pouco doida, não é? Ela está aumentando onde o mês inicial termina e cresce rapidamente a partir dali. Talvez você tenha que otimizar os parâmetros suavizadores.



**Figura 8-16:** Gráfico da previsão com os valores padrões de *alfa* e *gama*

### Otimizando o Erro de um Passo

Como você fez na suavização exponencial simples, adicione o erro de previsão quadrada na coluna G. Em F2 e G2, pode calcular a soma dos erros quadrados e o erro padrão para previsão de um passo exatamente como antes. Salvo que, desta vez, o modelo tem dois parâmetros suavizadores. Então você dividirá o SSE por  $36 - 2$  antes de calcular a raiz quadrada:

$$=\text{SQRT}(\text{F2} / (36 - 2))$$

Isso tem como resultado a planilha da Figura 8-17.

A configuração de otimização é idêntica à suavização exponencial simples exceto que você otimizará *alfa* e *gama* juntos, como mostra a Figura 8-18.

Ao resolver, obtém-se um valor *alfa* ideal de 0,66 e um valor *gama* ideal de 0,05. A previsão ideal é mostrada no gráfico de dispersão com a linha reta na Figura 8-19.

SwordForecasting.xlsm

	A	B	C	D	E	F	G
1	Total months		Level smoothing parameter (alpha)	Trend smoothing parameter (gamma)			Standard Error
2	36		0.5	0.5	15315.3154	21.2238181	
4	t	Actual Demand	Level	Trend	One-step Forecast	Forecast Error	Squared Error
5	0	155.88	0.8369				
6	1	165	160.85845	2.907675	156.7169	8.2831	68.6097456
7	2	171	167.383063	4.71614375	163.766125	7.233875	52.3289475
8	3	147	159.549603	-1.55865781	172.099206	-25.099206	629.970154
9	4	143	150.495473	-5.30639414	157.990945	-14.990945	224.728441
10	5	154	154.594539	-0.60366377	145.189079	18.8109215	353.850767
11	6	160	156.995438	0.898617358	153.990875	6.00912451	36.1095774
12	7	152	154.947028	-0.57489642	157.894055	-5.8940551	34.7398856
13	8	150	152.186066	-1.6679292	154.372131	-4.3721311	19.1155307
14	9	150	154.750062	0.457526702	150.518136	9.48186263	71.0420107

Figura 8-17: Calculando o SSE e o erro padrão

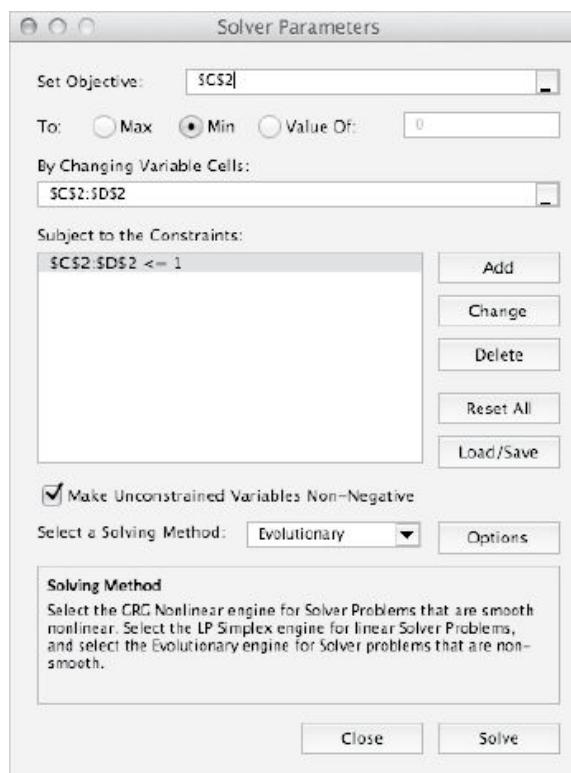


Figura 8-18: Configuração da otimização para Suavização Exponencial com Tendência

Holt



**Figura 8-19:** Gráfico da previsão ideal de Holt

A tendência que você está usando a partir da previsão é um adicional de cinco espadas vendidas por mês. A razão pela qual essa tendência é o dobro da que você encontrou usando a linha de tendência na aba anterior é porque a suavização com tendência favorece mais os pontos recentes, e, neste caso, os pontos de demanda mais recentes têm sido muito “tendenciosos”.

Observe que essa previsão começa muito próxima da previsão SES para o mês 37 – 290 versus 292. Mas, rapidamente, a previsão com tendência começa a crescer tal como você esperaria que acontecesse.

## Então Você Acabou? Considerando a Autocorrelação

Tudo bem. Isso é o melhor que pode fazer? Você já explicou tudo?

Bem, uma forma de verificar se você possui um bom modelo para fazer previsões é verificar os erros de um passo. Se esses erros forem aleatórios, você concluiu o seu trabalho. Mas, se tiver um padrão escondido no erro — algum tipo de comportamento repetitivo em um intervalo regular — talvez haja algo sazonal nos dados da demanda que está sem explicação.

E por “padrão no erro”, quero dizer que se você pegasse o erro e o alinhasse com ele mesmo modificado por um ou dois ou doze meses, eles se moveriam em sincronização?

Esse conceito de erro correlacionado à versão de tempo modificado dele mesmo é chamado de **autocorrelação** (auto significa “si mesmo” em grego. É também um bom prefixo para dispensar as vogais ao jogar palavras-cruzadas).

Então, para começar, crie uma nova aba chamada **Holt's Autocorrelation**. E, nessa aba, cole os meses 1 ao 36 junto com os erros de um passo da previsão de Holt nas colunas A e B.

Debaixo desses erros em B38, calcule a média do erro. Isso tem como resultado a planilha da Figura 8-20.

	A	B	C
1	t	One-step error	
26	25	-32.23760721	
27	26	16.13259318	
28	27	-29.94279772	
29	28	2.398381625	
30	29	10.33952442	
31	30	-15.315321	
32	31	11.47512271	
33	32	-0.793752233	
34	33	6.051576831	
35	34	2.173287229	
36	35	44.77509259	
37	36	51.73049718	
38		3.576406042	
39			

**Figura 8-20:** Meses e erros associados da previsão de um passo

Na coluna C, calcule os desvios de cada erro na coluna B a partir da média em B38. Esses desvios no erro de um passo a partir da média são onde os padrões colocarão a sua cara feia a mostra. Por exemplo, pode ser que todo mês de dezembro o erro da previsão esteja substancialmente

acima da média — esse tipo de padrão sazonal apareceria nesses números.

Na célula C2, então, o desvio do erro em B2 a partir da média seria:

=B2 - B\$38

Você pode arrastar essa fórmula para gerar todos os desvios da média. Na célula C38, calcule a soma dos desvios quadrados assim:

=SUMPRODUCT (\$C2 : \$C37 , C2 : C37)

Isso tem como resultado a planilha da Figura 8-21.

Agora, na coluna D, “atrasse” os desvios de erro em um mês. Nomeie a coluna D com um 1. Deixe a célula D2 em branco e configure a célula D3 para:

=C2

E, então, arraste a fórmula até que D37 se iguale a C36. Isso gera a Figura 8-22.

	A	B	C
1	t	One-step error	Deviations from mean
27	26	16.13259318	12.55618714
28	27	-29.94279772	-33.51920376
29	28	2.398381625	-1.178024417
30	29	10.33952442	6.763118375
31	30	-15.315321	-18.89172704
32	31	11.47512271	7.898716673
33	32	-0.793752233	-4.370158274
34	33	6.051576831	2.47517079
35	34	2.173287229	-1.403118813
36	35	44.77509259	41.19868655
37	36	51.73049718	48.15409114
38		3.576406042	13636.81634

**Figura 8-21:** A soma dos desvios das médias quadradas dos erros de previsão de Holt

Screenshot of Microsoft Excel showing a data table titled "SwordForecasting.xlsm". The table has columns A, B, C, and D. Column A contains values from 1 to 38. Column B contains the header "One-step error" and values from -32.23760721 to 3.576406042. Column C contains the header "Deviations from mean" and values from -35.81401325 to 13636.81634. Column D contains the value 1. The formula bar shows "=C36". The ribbon tabs include Home, Layout, Tables, Charts, SmartArt, Formulas, and others. The status bar shows "Holt's Autocorrelation" and "Normal View".

	A	B	C	D
1	t	One-step error	Deviations from mean	1
26	25	-32.23760721	-35.81401325	28.0123854
27	26	16.13259318	12.55618714	-35.814013
28	27	-29.94279772	-33.51920376	12.5561871
29	28	2.398381625	-1.178024417	-33.519204
30	29	10.33952442	6.763118375	-1.1780244
31	30	-15.315321	-18.89172704	6.76311838
32	31	11.47512271	7.898716673	-18.891727
33	32	-0.793752233	-4.370158274	7.89871667
34	33	6.051576831	2.47517079	-4.3701583
35	34	2.173287229	-1.403118813	2.47517079
36	35	44.77509259	41.19868655	-1.4031188
37	36	51.73049718	48.15409114	=C36
38		3.576406042	13636.81634	

**Figura 8-22:** Os desvios de erro em um mês atraso

Para atrasar em dois meses, selecione D1:D37 e arraste-o para a coluna E. Da mesma forma, para atrasar 12 meses, arraste a seleção pela coluna O. Fácil! Isso apresenta uma matriz em cascata de desvios de erros atrasados, como mostra a Figura 8-23:

Screenshot of Microsoft Excel showing a matrix of deviations from the mean for 36 observations across 12 time periods. The matrix is a 36x12 grid where each row represents an observation and each column represents a period. The first column is labeled "Deviations from mean". The last cell in the matrix contains the value 13636.81634.

	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Deviations from mean	1	2	3	4	5	6	7	8	9	10	11	12
2		4.71											
3		4.12	4.71	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4		-26.35	4.12	4.71	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5		-15.94	-26.35	4.12	4.71	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
6		13.04	-15.94	-26.35	4.12	4.71	0.00	0.00	0.00	0.00	0.00	0.00	0.00
7		-2.66	13.04	-15.94	-26.35	4.12	4.71	0.00	0.00	0.00	0.00	0.00	0.00
8		-12.04	-2.66	13.04	-15.94	-26.35	4.12	4.71	0.00	0.00	0.00	0.00	0.00
9		-8.95	-12.04	-2.66	13.04	-15.94	-26.35	4.12	4.71	0.00	0.00	0.00	0.00
10		3.30	8.95	-12.04	-2.66	13.04	-15.94	-26.35	4.12	4.71	0.00	0.00	0.00
11		8.23	3.30	-8.95	-12.04	-2.66	13.04	-15.94	-26.35	4.12	4.71	0.00	0.00
12		3.50	8.23	3.30	-8.95	-12.04	-2.66	13.04	-15.94	-26.35	4.12	4.71	0.00
13		27.64	3.50	8.23	3.30	-8.95	-12.04	-2.66	13.04	-15.94	-26.35	4.12	4.71
14		-29.23	27.64	3.50	8.23	3.30	-8.95	-12.04	-2.66	13.04	-15.94	-26.35	4.12
15		-16.71	-29.23	27.64	3.50	8.23	3.30	-8.95	-12.04	-2.66	13.04	-15.94	-26.35
16		-12.99	-16.71	-29.23	27.64	3.50	8.23	3.30	-8.95	-12.04	-2.66	13.04	-15.94
17		-22.39	-12.99	-16.71	-29.23	27.64	3.50	8.23	3.30	-8.95	-12.04	-2.66	13.04
18		31.07	-22.39	-12.99	-16.71	-29.23	27.64	3.50	8.23	3.30	-8.95	-12.04	-2.66
19		-19.92	31.07	-22.39	-12.99	-16.71	-29.23	27.64	3.50	8.23	3.30	-8.95	-12.04
20		-8.73	-19.92	31.07	-22.39	-12.99	-16.71	-29.23	27.64	3.50	8.23	3.30	-8.95
21		1.26	-8.73	-19.92	31.07	-22.39	-12.99	-16.71	-29.23	27.64	3.50	8.23	3.30
22		13.50	1.26	-8.73	-19.92	31.07	-22.39	-12.99	-16.71	-29.23	27.64	3.50	8.23
23		4.07	13.50	1.26	-8.73	-19.92	31.07	-22.39	-12.99	-16.71	-29.23	27.64	3.50
24		9.59	4.07	13.50	1.26	-8.73	-19.92	31.07	-22.39	-12.99	-16.71	-29.23	27.64
25		28.01	9.59	4.07	13.50	1.26	-8.73	-19.92	31.07	-22.39	-12.99	-16.71	-29.23
26		-35.81	28.01	9.59	4.07	13.50	1.26	-8.73	-19.92	31.07	-22.39	-12.99	-16.71
27		12.56	-35.81	28.01	9.59	4.07	13.50	1.26	-8.73	-19.92	31.07	-22.39	-12.99
28		-33.52	12.56	-35.81	28.01	9.59	4.07	13.50	1.26	-8.73	-19.92	31.07	-22.39
29		-1.18	-33.52	12.56	-35.81	28.01	9.59	4.07	13.50	1.26	-8.73	-19.92	31.07
30		6.76	-1.18	-33.52	12.56	-35.81	28.01	9.59	4.07	13.50	1.26	-8.73	-19.92
31		-18.89	6.76	-1.18	-33.52	12.56	-35.81	28.01	9.59	4.07	13.50	1.26	-8.73
32		7.90	-18.89	6.76	-1.18	-33.52	12.56	-35.81	28.01	9.59	4.07	13.50	1.26
33		-4.37	7.90	-18.89	6.76	-1.18	-33.52	12.56	-35.81	28.01	9.59	4.07	13.50
34		2.48	-4.37	7.90	-18.89	6.76	-1.18	-33.52	12.56	-35.81	28.01	9.59	4.07
35		-1.40	2.48	-4.37	7.90	-18.89	6.76	-1.18	-33.52	12.56	-35.81	28.01	9.59
36		41.20	-1.40	2.48	-4.37	7.90	-18.89	6.76	-1.18	-33.52	12.56	-35.81	28.01
37		48.15	41.20	-1.40	2.48	-4.37	7.90	-18.89	6.76	-1.18	-33.52	12.56	-35.81
38		13636.81634											28.01

Figura 8-23: Uma linda matriz em cascata dos desvios de erros atrasados digna de um rei

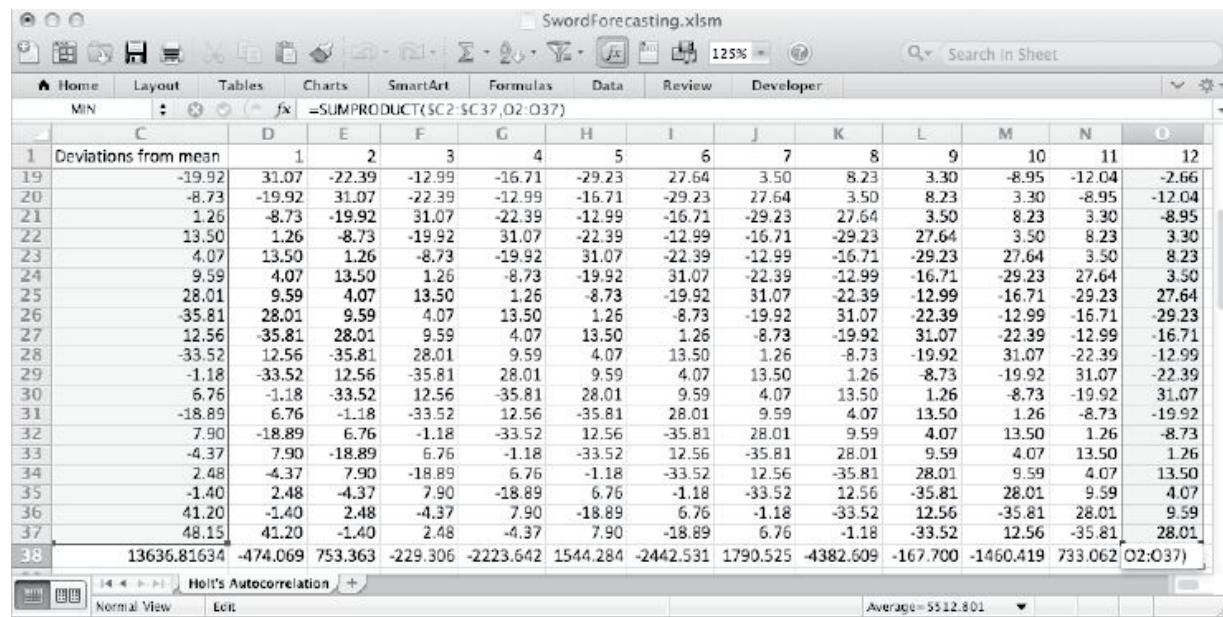
Agora que você possui esses atrasos, imagine o que eles significam para aquelas colunas se “moverem em sincronização” com a coluna C. Por exemplo, pegue um atraso de um mês na coluna D. Se essas duas colunas estivessem em sincronização quando uma fica negativa, a outra também deveria ficar. E quando uma fica positiva, a outra também deveria. Isso significa que o produto de duas colunas resultaria em muitos números positivos (resultados de um negativo vezes um negativo ou um positivo vezes um positivo em um número positivo).

Você pode somar esses produtos e, quanto mais próximo essa SUMPRODUCT das colunas em atraso com os desvios originais chegar da

soma dos desvios quadrados em C38, mais sincronização e mais correlação dos erros em atraso com os originais haverá.

Você também pode obter autocorrelação negativa onde os desvios em atraso ficam negativos quando os originais são positivos e vice-versa. A SUMPRODUCT nesse caso será um número negativo maior.

Para começar, arraste SUMPRODUCT (\$C2:\$C37, C2:C37) na célula C38 até a coluna O. Observe que a referência absoluta para a coluna C manterá a coluna em seu lugar, então você obtém o SUMPRODUCT de cada coluna em atraso com o original, como mostra a Figura 8-24:



**Figura 8-24:** O SUMPRODUCT dos desvios em atraso com os originais

Calcula-se a autocorrelação para um dado mês em atraso como o SUMPRODUCT dos desvios em atraso vezes os desvios originais dividido pela soma dos desvios quadrados em C38.

Por exemplo, pode-se calcular a autocorrelação do atraso de um mês na célula D40 como:

$$=D38 / \$C38$$

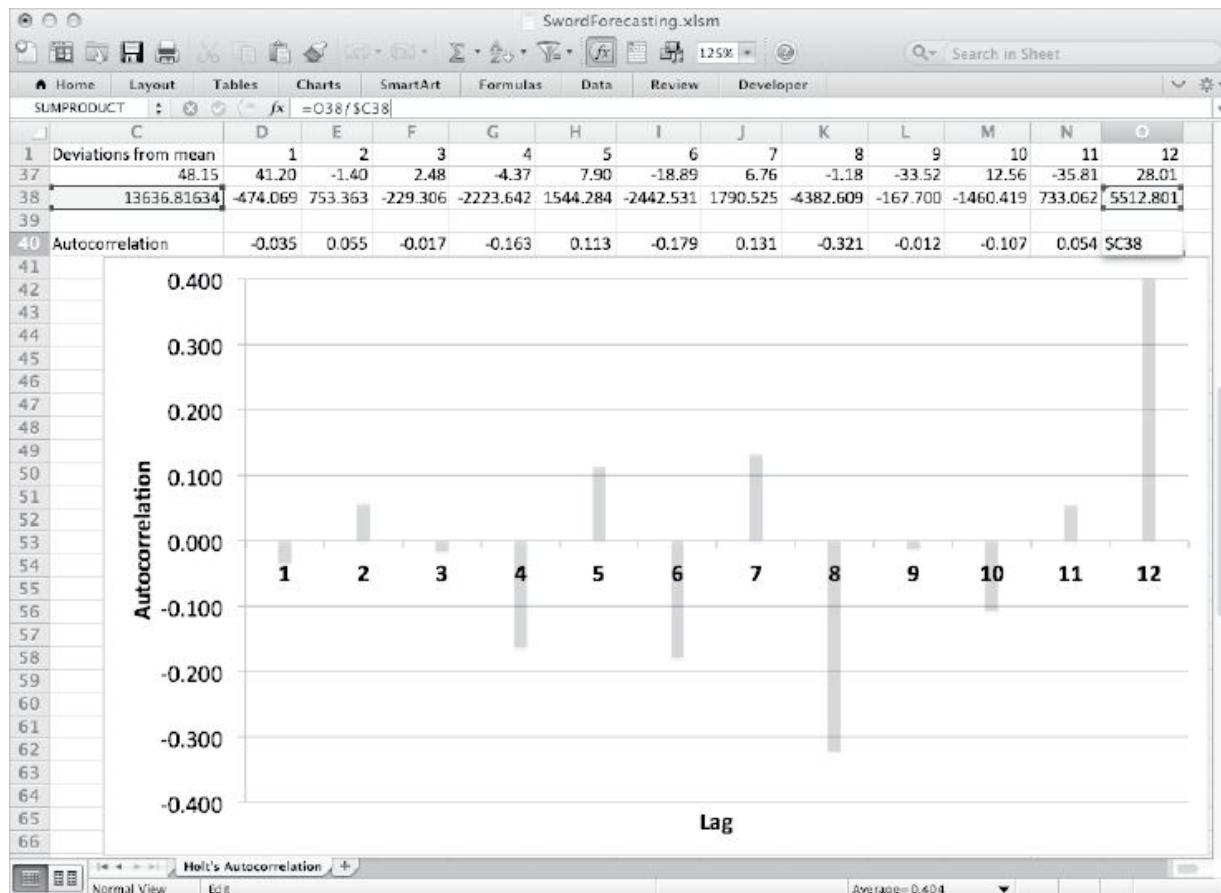
E, ao arrastá-lo, pode-se obter as autocorrelações para cada atraso.

Ao destacar D40:O40, você pode inserir um gráfico de barra na planilha como mostra a Figura 8-25 (clique com o botão direito e formate o

preenchimento da série para ter uma leve transparência se quiser ler os nomes do mês abaixo dos valores negativos). Esse gráfico de barra é chamado de **correlograma**, e ele mostra a autocorrelação para cada mês em atraso até um ano. (Na minha opinião, acho a palavra **correlograma** muito legal.)

Tudo bem, então qual autocorrelação importa? Bem, a convenção é que você se preocupe somente com as autocorrelações maiores do que 2/raiz quadrada (quantidade de pontos de dados), que nesse caso é 2/raiz quadrada(36) = 0,333. Você também deveria se importar com os que possuem a autocorrelação negativa menores do que -0,333.

Você pode somente passar os olhos no gráfico para as autocorrelações que estão acima ou abaixo desses **valores críticos**. Mas é comum em previsão representar em gráficos as linhas tracejadas desses valores críticos no correlograma. Por falta de uma figura mais bonita, mostrarei como fazer isso aqui.



**Figura 8-25:** Esse é o meu correograma; há muitos iguais a esse, mas esse é meu

Em D42, adicione  $=2/\text{SQRT}(36)$  e arraste-o até a coluna O. Faça o mesmo em D43 somente com o valor negativo  $=-2/\text{SQRT}(36)$  e arraste-o até O. Isso tem como resultado os pontos críticos para as autocorrelações, como mostra a Figura 8-26.

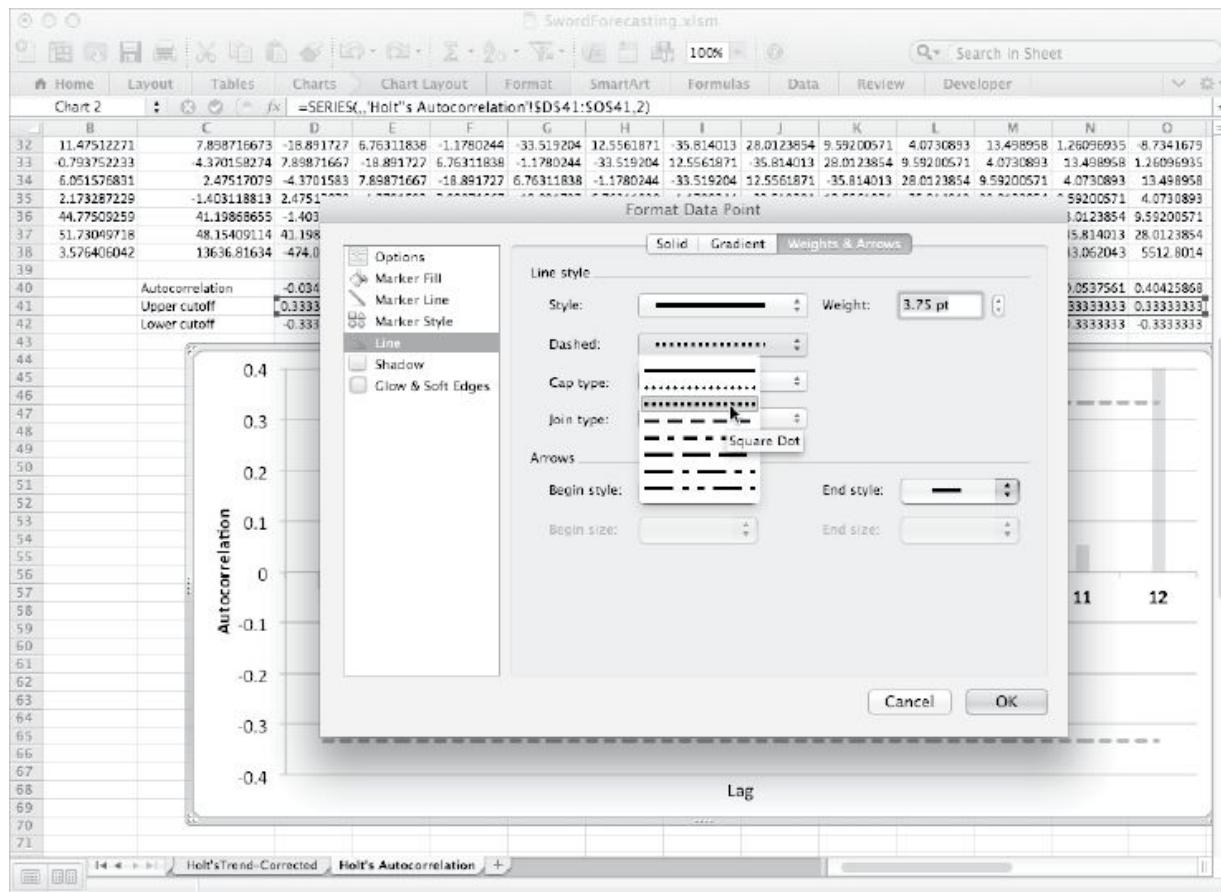
C	D	E	F	G	H	I	J	K	L	M	N	O	
1 Deviations from mean	48.15	41.20	-1.40	2.48	-4.37	7.90	-18.89	6.76	-1.18	-33.52	12.56	-35.81	28.01
37	13636.81634	-474.069	753.363	-229.306	-2223.642	1544.284	-2442.531	1790.525	-4382.609	-167.700	-1460.419	733.062	5512.801
38													
39													
40 Autocorrelation	-0.035	0.055	-0.017	-0.163	0.113	-0.179	0.131	-0.321	-0.012	-0.107	0.054	0.404	
41	0.333	0.333	0.333	0.333	0.333	0.333	0.333	0.333	0.333	0.333	0.333	0.333	0.333
42	-0.333	-0.333	-0.333	-0.333	-0.333	-0.333	-0.333	-0.333	-0.333	-0.333	-0.333	-0.333	-0.333
43													

**Figura 8-26:** Pontos críticos para as autocorrelações

Clique com o botão direito no gráfico de barra da autocorrelação e escolha Select Data. A partir da janela que aparece, pressione Add button para criar uma nova série.

Para uma série selecione a série D42:O42 como os valores y. Adicione uma terceira série usando D43:O43. Isso adicionará mais dois conjuntos de barras ao gráfico.

Ao clicar com o botão direito em cada uma das novas séries de barra, você pode selecionar Change Series Chart Type e selecionar Line chart para transformá-lo em uma linha sólida em vez de barras. Clique com o botão direito nessas linhas e selecione Format Data Series. Depois navegue para a opção Line (Line Style em algumas versões do Excel) na janela. Nesta seção, você pode configurar a linha para tracejado, como mostra a Figura 8-27.



**Figura 8-27:** Trocando os valores críticos das barras para linhas tracejadas

Isso resulta em um correograma com os valores críticos traçados, como mostra a Figura 8-28.

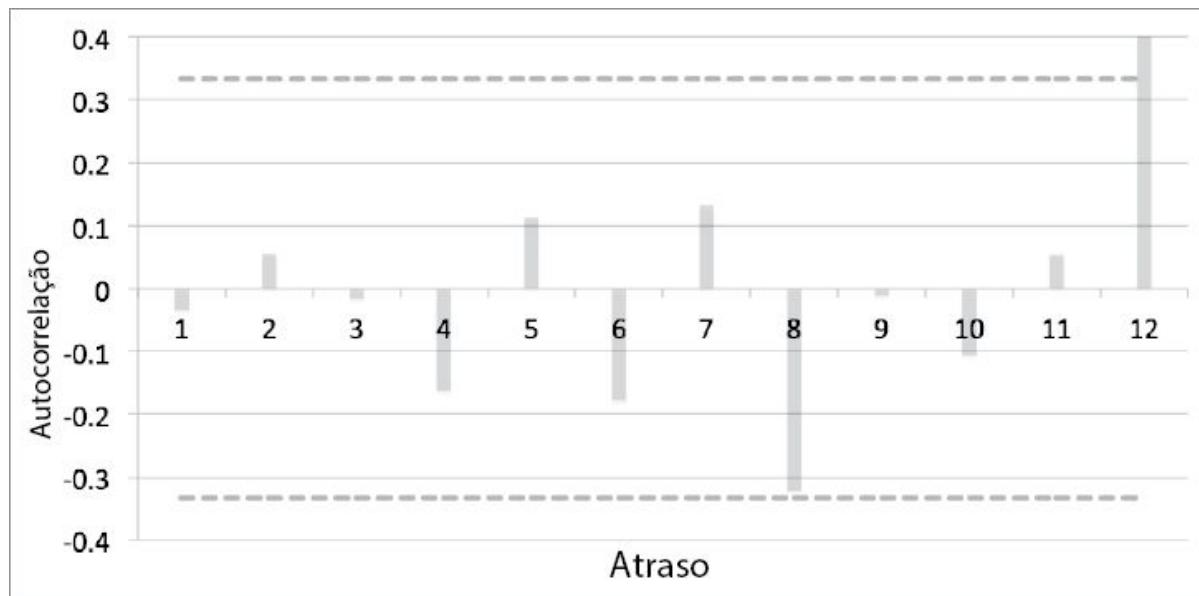
E o que você vê?

Há exatamente uma autocorrelação que está acima do valor crítico, e está nos 12 meses.

*O erro deslocado por um ano está correlacionado com ele mesmo.*

Isso indica um **ciclo sazonal de 12 meses**. Isso não deveria ser uma surpresa. Se você observar o deslocamento da demanda na aba Timeseries, está claro que há picos em cada Natal e variam até Abril/Maio.

Você precisa de uma técnica de previsão que possa explicar a sazonalidade. E você não saberia — existe uma técnica de suavização exponencial para isso.



**Figura 8-28:** Correlograma para os valores críticos

## Suavização Exponencial Multiplicativa Holt-Winters

A *Suavização Exponencial Multiplicativa Holt-Winters* é a extensão lógica da Suavização com Tendência Corrigida de Holt. Ela é responsável por um nível, uma tendência e a necessidade de adaptar a demanda para mais ou para menos regularmente devido às flutuações sazonais. Repare que a flutuação sazonal não precisa ser a cada 12 meses como nesse exemplo. No caso da MailChimp, temos flutuações de demanda periódicas todas as quintas (as pessoas devem achar que quinta-feira é um bom dia para enviar e-mails de marketing). Usando Holt-Winters, podemos explicar tal ciclo de 7 dias.

Agora, na maioria das situações, você não pode somente adicionar ou subtrair uma quantia fixa de demanda sazonal para adaptar a previsão. Se as vendas do seu negócio crescerem de 200 para 2.000 espadas por mês, você não adaptaria a demanda de Natal em ambos os contextos ao adicionar 20 espadas. Não, adaptações sazonais geralmente precisam ser multiplicadoras. Em vez de adicionar 20 espadas, talvez fosse melhor

*multiplicar* a previsão por 120%. É por isso que é chamado Holt-Winters **Multiplicativo**. Veja como essa previsão imagina a demanda:

*Demandas no tempo t = (nível + t\*tendência) \* adaptação sazonal para o tempo t \* quaisquer adaptações irregulares que não podemos explicar*

Portanto, você ainda possui o nível idêntico e a estrutura de tendência que possuía na Suavização Exponencial com Tendência Corrigida de Holt, mas a demanda é adaptada para a sazonalidade. E já que não pode explicar as variações irregulares na demanda, tal como os atos de Deus, você não vai.

Holt-Winters também é chamado de **suavização exponencial tripla** porque, você adivinhou, existem três parâmetros suavizadores desta vez. Ainda há parâmetros **alfa** e **gama**, mas agora você tem um fator de adaptação sazonal com uma equação de atualização e um fator chamado **delta**.

Agora, as equações de adaptação de três erros são ligeiramente mais complexas do que foi visto até agora, mas você reconhecerá algumas partes.

Antes de começar, quero deixar claro uma coisa — até agora você usou níveis e tendências dos períodos anteriores para prever o próximo e adaptar. Mas com as adaptações sazonais, você não considera os períodos anteriores. Em vez disso, você considera as estimativas anteriores do fator de adaptação para aquele ponto no ciclo. Nesse caso, são 12 períodos anteriores em vez de um.

Isso significa que se você estiver no mês 36 e está prevendo três meses à frente de 39, tal previsão ficará assim:

*Previsão para o mês 39 = (nível36 + 3\*tendência36)\*sazonalidade27*

Sim, você está vendo a **sazonalidade27** corretamente. É a estimativa mais recente para a adaptação sazonal de março. Você não pode usar **sazonalidade36** porque essa é para dezembro.

Tudo bem, então é assim que a previsão futura funciona. Vamos mergulhar nas equações de atualização começando pelo nível. Você precisa somente dos *nível0* e *tendência0* iniciais, mas também precisa de **doze fatores** de sazonalidade iniciais, sazonalidade-11 até *sazonalidade0*.

Por exemplo, a equação de atualização para o *nível1* depende da estimativa inicial da adaptação sazonal de janeiro:

$$\text{nível1} = \text{nível0} + \text{tendência0} + \text{alfa} * (\text{demanda1} - (\text{nível0} + \text{tendência0}) * \text{sazonalidade-11}) / \text{sazonalidade-11}$$

Há muitos componentes familiares nesse cálculo de nível. O nível atual é o nível anterior mais a tendência anterior (assim como na suavização exponencial dupla) mais *alfa* vezes o erro de previsão de um passo (*demanda1* – (*nível0* + *tendência0*) \* *sazonalidade-11*), em que o erro obtém uma adaptação sazonal ao ser dividido pela *sazonalidade-11*.

E, enquanto você caminha para a frente no tempo, o próximo mês seria:

$$\text{nível2} = \text{nível1} + \text{tendência1} + \text{alfa} * (\text{demanda2} - (\text{nível1} + \text{tendência1}) * \text{sazonalidade-10}) / \text{sazonalidade-10}$$

No geral, o nível é calculado desta forma:

$$\begin{aligned} \text{nível período atual} &= \text{nível período anterior} + \text{tendência período} \\ &\quad \text{anterior} + \text{alfa} * (\text{demanda período atual} - (\text{nível período anterior} + \text{tendência período anterior}) * \text{sazonalidade último período} \\ &\quad \text{relevante}) / \text{sazonalidade último período relevante} \end{aligned}$$

A tendência é atualizada em relação ao nível exatamente da mesma forma como na suavização exponencial dupla:

$$\begin{aligned} \text{tendência período atual} &= \text{tendência período anterior} + \text{gama} * \text{alfa} * \\ &\quad (\text{demanda período atual} - (\text{nível período anterior} + \text{tendência} \\ &\quad \text{período anterior}) * \text{sazonalidade último período} \\ &\quad \text{relevante}) / \text{sazonalidade último período relevante} \end{aligned}$$

Assim como na suavização exponencial dupla, a tendência atual é a tendência anterior mais *gama* vezes a quantia de erro incorporada na

equação de atualização do nível.

E agora vamos para a equação de atualização do fator sazonal. Não é igual à equação de atualização de tendência, *salvo que* ela ajusta o último fator sazonal relevante usando *delta* vezes o erro que as atualizações do nível e da tendência *ignoraram*:

$$\begin{aligned} \text{sazonalidade período atual} &= \text{sazonalidade último período relevante} \\ &+ \text{delta} * (1 - \text{alfa}) * (\text{demanda período atual} - (\text{nível período anterior} + \text{tendência período anterior}) * \text{sazonalidade último período relevante}) / (\text{nível período anterior} + \text{tendência período anterior}) \end{aligned}$$

Nesse caso, você está atualizando a adaptação sazonal com o fator correspondente dos 12 meses anteriores, mas está duplicando em *delta* vezes qualquer erro que foi deixado no chão da sala de corte da atualização do nível. *Salvo que*, repare que em vez de adaptar o erro sazonalmente, você está dividindo pelos valores anteriores do nível e da tendência. Ao “adaptar o nível e a tendência” do erro de um passo, você está colocando o erro na mesma escala multiplicadora que os fatores sazonais.

## Configurando os Valores Iniciais para o Nível, a Tendência e a Sazonalidade

Configurar os valores iniciais para SES e a suavização exponencial dupla foi mamão com açúcar. Mas, agora, você tem que perceber qual a tendência e qual a sazonalidade a partir das séries temporais. E isso significa que configurar os valores iniciais para essa previsão (um nível, uma tendência e 12 fatores de adaptação sazonais) é um pouco complicado. Há maneiras simples (e erradas!) de fazer isso. Vou lhe mostrar uma boa maneira de inicializar Holt-Winters, presumindo que você tem ao menos dois ciclos sazonais dignos de dados históricos. Neste caso, você tem três ciclos.

Eis o que você fará:

1. Suavize os dados históricos usando o que chamamos de medida móvel  $2 \times 12$ .
2. Compare uma versão suavizada das séries temporais com o original para estimar a sazonalidade.
3. Usando as estimativas sazonais iniciais, dessazonalise os dados históricos.
4. Estime o nível e a tendência usando uma linha de tendência nos dados dessazonalizados.

Para começar, crie uma nova aba chamada *HoltWintersInitial* e cole os dados em série temporal nas duas primeiras colunas. Agora você precisa suavizar alguns dados em série temporal usando uma média móvel. Devido ao fato da sazonalidade estar em um ciclo de 12 meses, faz sentido usar uma média móvel de 12 meses nos dados.

O que quero dizer com uma média móvel de 12 meses?

Para uma medida móvel, você pega a demanda de um mês em particular e também a demanda próxima desse mês nas duas direções e calcula a média. Isso amortece quaisquer picos diferentes na série.

Mas há um problema na medida móvel de 12 meses. Doze é um número par. Se você está suavizando a demanda para o mês 7, deveria calcular a média como a demanda de meses de 1 **até** 12 ou a demanda de meses de 2 **até** 13? De qualquer forma, o mês 7 não está muito no meio. Não existe meio!

Para ajustar isso, você terá que suavizar a demanda com uma “média móvel  $2 \times 12$ ”, ou seja, a média de ambas as possibilidades — meses 1 até 12 e 2 **até** 13. (O mesmo vale para qualquer outro número par de períodos de tempo em um ciclo. Se o seu ciclo tem um número ímpar de períodos, a parte “ $2 \times$ ” da medida móvel é descartável e você pode usar uma média móvel simples.)

Agora observe que para os primeiros seis meses de dados e os últimos seis, isso não é nem possível. Eles não têm seis meses de dados de

nenhum lado. Você pode suavizar somente os meses do meio do conjunto de dados (nesse caso são os meses 7–30). É por isso que você precisa de, no mínimo, dois anos de dados dignos, a fim de obter um ano de dados suavizados.

Para começar com o mês 7, use a seguinte fórmula:

$$= (\text{AVERAGE}(\text{B3 : B14}) + \text{AVERAGE}(\text{B2 : B13})) / 2$$

Essa é a média do mês 7 com os 12 meses próximos a ela, exceto que os meses 1 e 13 contam pela metade válida do que os outros meses contariam, o que faz sentido, uma vez que os meses 1 e 13 estariam no mesmo mês se eles fossem validados duas vezes, logo, você teria janeiro representado mais de uma vez na média móvel.

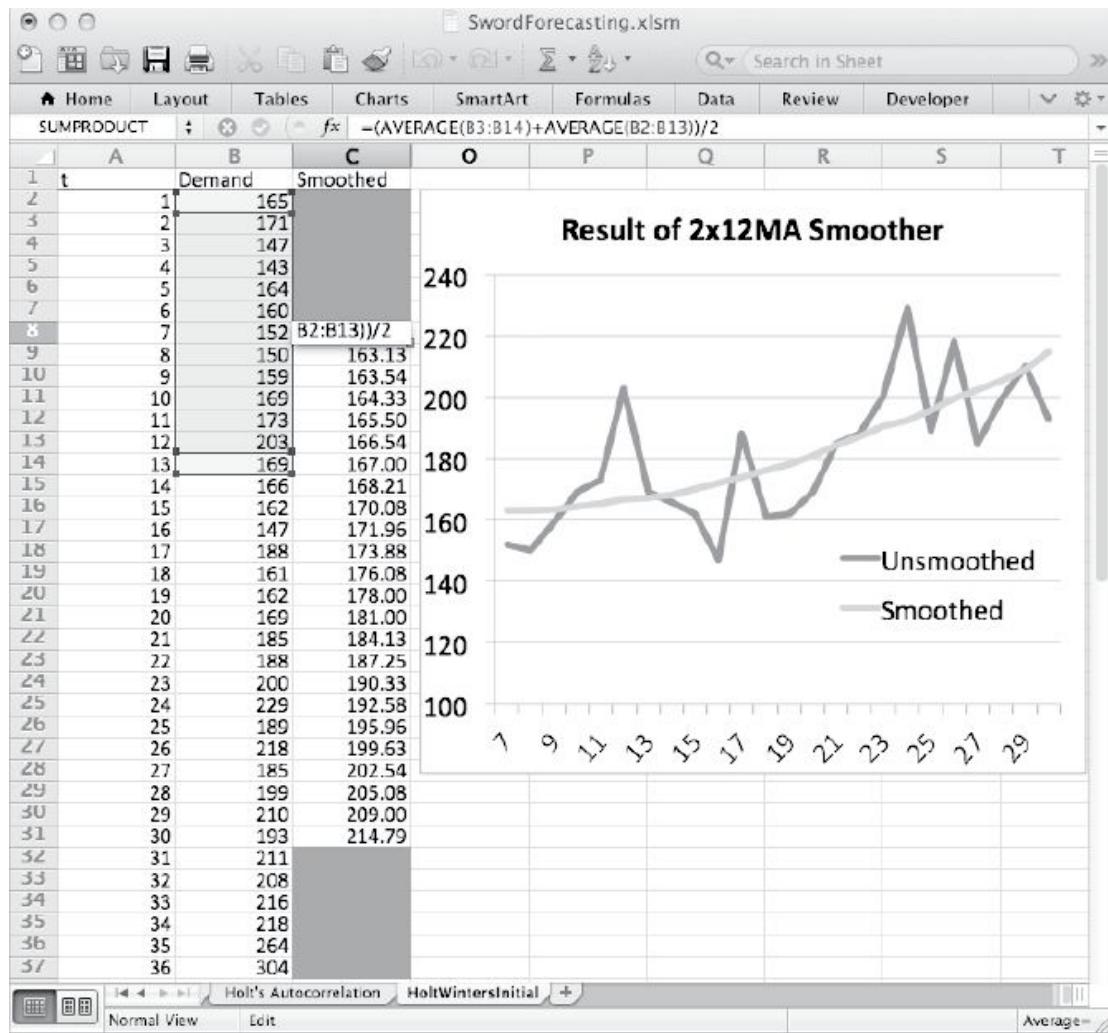
Arrastando essa fórmula por todo o mês 30 e representando em gráfico o original e os dados suavizados em um gráfico de dispersão com linha reta, temos a planilha exibida na Figura 8-29. No meu gráfico, eu nomeei as duas séries como suavizada e não suavizada. Fica claro que, ao observar a linha suavizada, qualquer variação sazonal presente nos dados foi, mais ou menos, suavizada.

Agora, na coluna D, você pode dividir o valor original pelo valor suavizado para obter uma estimativa do fator de adaptação sazonal.

Começando no mês 7, você tem para a célula D8:

B8 / C8

E você pode arrastar até o mês 30. Repare que nos meses 12 e 24 (dezembro) você tem picos em torno de 20% do normal, enquanto que tem alguns na primavera.



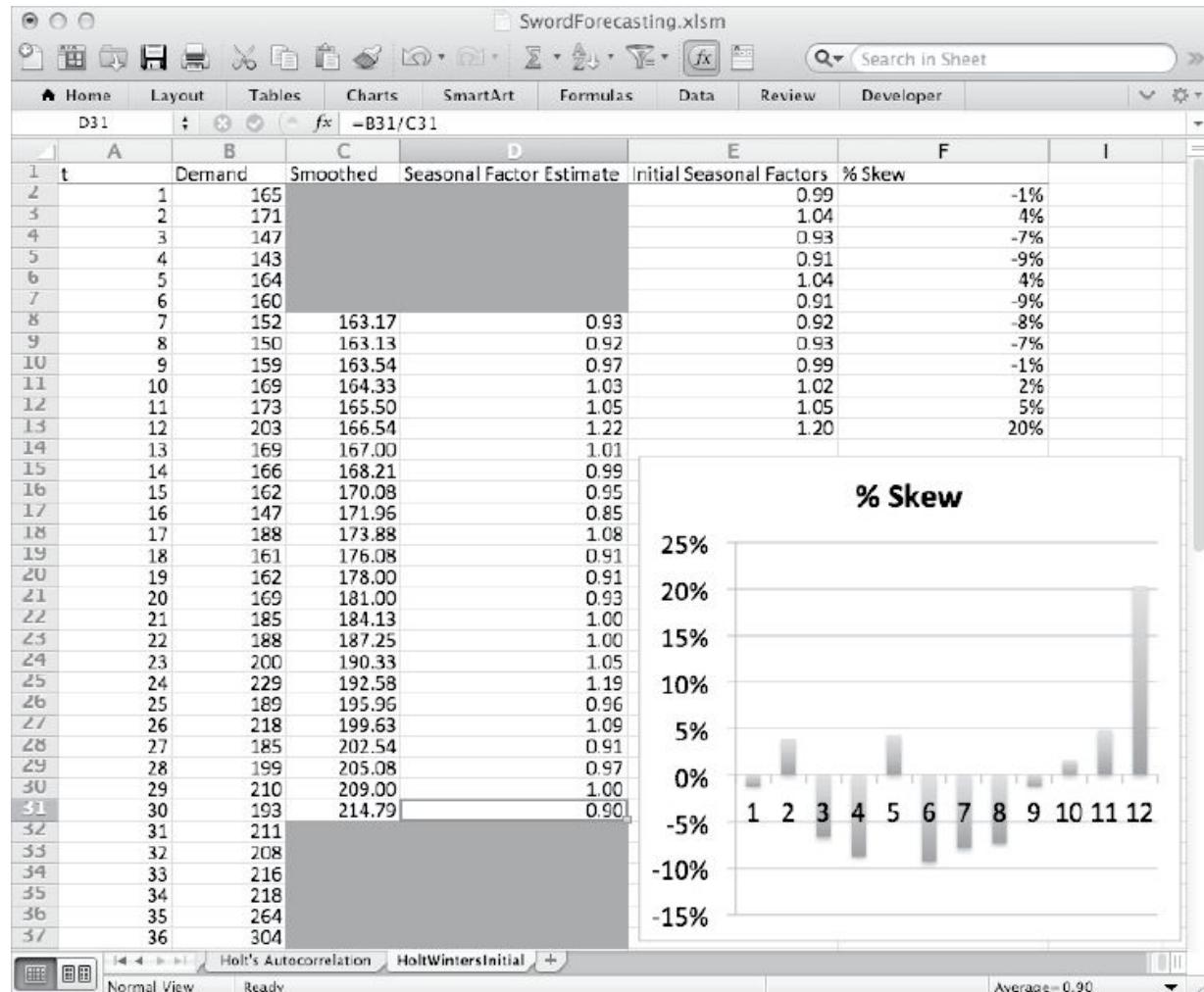
**Figura 8-29:** Os dados de demanda suavizados

Essa técnica de suavização gerou duas estimativas de ponto para cada fator de sazonalidade. Na coluna E, vamos calcular a média desses dois pontos juntos em um único valor que será o fator sazonal inicial usado em Holt-Winters.

Por exemplo, em E2 (janeiro), calcula-se a média dos dois pontos de janeiro na coluna D, que são D14 e D26. Visto que os dados suavizados começam no meio do ano na coluna D, não é possível arrastar essa média. Em E8 (julho), deve-se pegar a média de D8 e D20 por exemplo.

Uma vez que tenha os 12 fatores de adaptação na coluna E, pode-se subtrair 1 de cada um deles na coluna F e formatar as células como porcentagens (selecione as séries e clique com o botão direito em Format

Cells) para ver como esses fatores movem a demanda para cima ou para baixo em cada mês. Pode-se até inserir um gráfico de barra para esses movimentos na planilha, como mostra a Figura 8-30.



**Figura 8-30:** Um gráfico de barra das variações sazonais estimadas

Agora que possui essas adaptações sazonais iniciais, você pode usá-las para **desazonalizar** os dados em séries temporais. Uma vez que toda a série esteja desazonalizada, pode-se jogar uma linha de tendência em cima e usar a inclinação e o intercepto como tendência e nível iniciais.

Para começar, cole os valores de adaptação sazonais adequados para cada mês de G2 até G37. Essencialmente, você está apenas colando E2:E13 três vezes seguidas até a coluna G (certifique-se de colar somente os valores). Na coluna H, você pode dividir a série original na coluna B

pelos fatores sazonais em G para remover a sazonalidade estimada presente nos dados. Essa planilha é exibida na Figura 8-31.

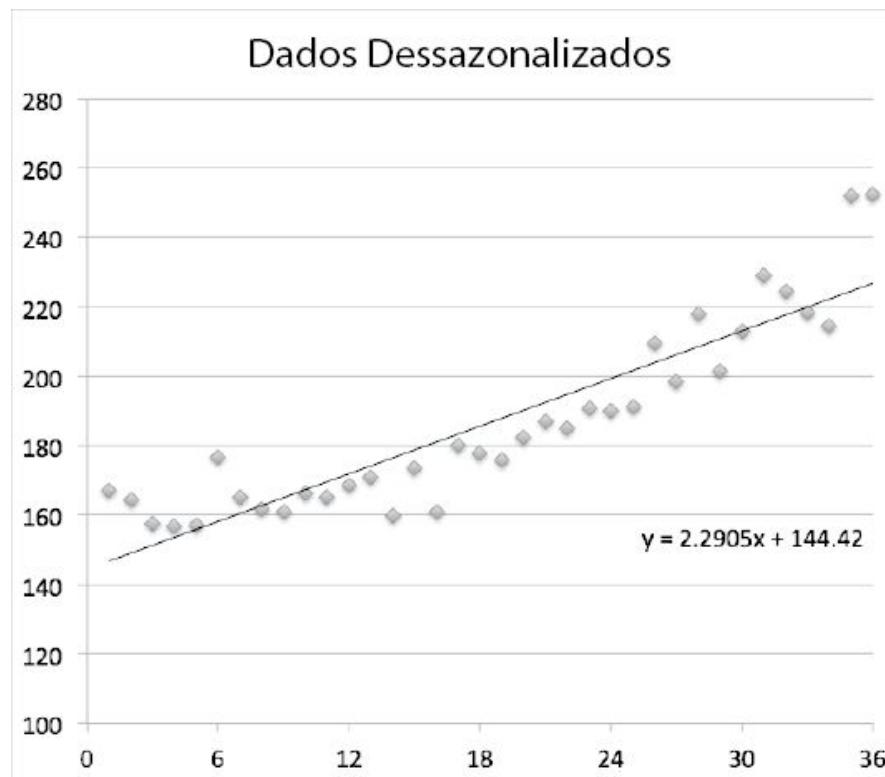
	A	B	C	D	E	F	G	H
1	t	Demand	Smoothed	Seasonal Factor Estimate	Initial Seasonal	% Skew	Initial Seasonal Factors x3	Deseasonalized Data
2	1	165			0.99	-1%	0.99	=B2/G2
3	2	171			1.04	4%	1.04	164.51
4	3	147			0.93	-7%	0.93	157.57
5	4	143			0.91	-9%	0.91	156.70
6	5	164			1.04	4%	1.04	157.24
7	6	160			0.91	-9%	0.91	176.51
8	7	152	163.17	0.93	0.92	-8%	0.92	165.07
9	8	150	163.13	0.92	0.93	-7%	0.93	161.88
10	9	159	163.54	0.97	0.99	-1%	0.99	160.85
11	10	169	164.33	1.03	1.02	2%	1.02	166.31
12	11	173	165.50	1.05	1.05	5%	1.05	165.07
13	12	203	166.54	1.22	1.20	20%	1.20	168.60
14	13	169	167.00	1.01			0.99	171.01
15	14	166	168.21	0.99			1.04	159.70
16	15	162	170.08	0.95			0.93	173.65
17	16	147	171.96	0.85			0.91	161.08
18	17	188	173.88	1.08			1.04	180.25
19	18	161	176.08	0.91			0.91	177.62
20	19	162	178.00	0.91			0.92	175.93
21	20	169	181.00	0.93			0.93	182.38
22	21	185	184.13	1.00			0.99	187.15
23	22	188	187.25	1.00			1.02	185.00
24	23	200	190.33	1.05			1.05	190.83
25	24	229	192.58	1.19			1.20	190.20
26	25	189	195.96	0.96			0.99	191.25
27	26	218	199.63	1.09			1.04	209.72
28	27	185	202.54	0.91			0.93	198.30
29	28	199	205.08	0.97			0.91	218.06
30	29	210	209.00	1.00			1.04	201.34
31	30	193	214.79	0.90			0.91	212.92
32	31	211					0.92	229.14
33	32	208					0.93	224.47
34	33	216					0.99	218.51
35	34	218					1.02	214.52
36	35	264					1.05	251.90
37	36	304					1.20	252.49

**Figura 8-31:** As séries temporais dessazonalizadas

Depois, como feito nas abas anteriores, insira um gráfico de dispersão na coluna H e desenhe uma linha de tendência nele. Ao exibir a equação da linha de tendência no gráfico, você obtém a estimativa de tendência inicial de 2,29 da venda de espadas adicionais por mês e uma estimativa de nível inicial de 144,42 (veja a Figura 8-32).

## Continuando com a Previsão

Agora que você possui os valores iniciais para todos os parâmetros, crie uma nova aba chamada **HoltWintersSeasonal**, onde você começará colando os dados em série temporal na linha 4 assim como nas duas técnicas de previsão anteriores.



**Figura 8-32:** Estimativas iniciais de nível e tendência por meio de uma linha de tendência nas séries dessazonalizadas

Nas colunas C, D e E coloque o nível, a tendência e valores sazonais, respectivamente, próximas às séries temporais. E, para começar, ao contrário das abas anteriores onde você precisava inserir uma nova linha 5 em branco, dessa vez é preciso inserir linhas em branco da 5 até a 16, nomeando-as como slots de tempo de **-11** até 0 na coluna A. Depois você pode colar os valores iniciais da aba anterior em seus respectivos lugares, como mostra a Figura 8-33.

	A	B	C	D	E
1	Total months				
2	36				
4	t	Actual Demand	Level	Trend	Seasonal Adjustment
5	-11			0.9882334	
6	-10			1.03945951	
7	-9			0.93293329	
8	-8			0.91259776	
9	-7			1.0430106	
10	-6			0.90644245	
11	-5			0.92083759	
12	-4			0.92662094	
13	-3			0.98849075	
14	-2			1.01620145	
15	-1			1.04805266	
16	0	144.42	2.2095	1.20400491	
17	1	165			
18	2	171			
19	3	147			
20	4	143			

**Figura 8-33:** Todos os valores iniciais de Holt-Winters tem um lugar

Na coluna F, você fará uma previsão de um passo. Então, para o período de tempo, é o nível anterior em C16 mais a tendência anterior em D16. Mas os dois são adaptados pela estimativa sazonalizada de janeiro adequada 12 linhas acima em E5. Portanto, F17 fica desta forma:

$$= (C16 + D16) * E5$$

O erro de previsão em G17 pode então ser calculado como:

$$=B17 - F17$$

Agora você está pronto para começar a calcular o nível, a tendência e a sazonalidade contínua. Então, nas células C2:E2, coloque os valores de **alfa**, **gama** e **delta** (como sempre, começarei com 0,5). A Figura 8-34 exibe a planilha.

Screenshot of a Microsoft Excel spreadsheet titled "SwordForecasting.xlsm". The spreadsheet contains data for Holt-Winters forecasting. The top row shows formulas: F17 =-(C16+D16)\*E5, C16 = (C15 + D15) \* E5, D16 = D15 + C\$2\*D15, and E16 = E15 + C\$2\*E15. The table below has columns for Total months, Actual Demand, Level, Trend, Seasonal Adjustment, One-step Forecast, and Forecast Error.

			Level smoothing parameter (alpha)	Trend smoothing parameter (gamma)	Seasonal smoothing parameter (delta)		
1	Total months	36		0.5	0.5	0.5	
4	t	Actual Demand	Level	Trend	Seasonal Adjustment	One-step Forecast	Forecast Error
5	-11				0.9882334		
6	-10				1.03945951		
7	-9				0.93293329		
8	-8				0.91259776		
9	-7				1.0430106		
10	-6				0.90644245		
11	-5				0.92083759		
12	-4				0.92662094		
13	-3				0.98849075		
14	-2				1.01620145		
15	-1				1.04805266		
16	0		144.42	2.2095	1.20400491		
17	1	165				144.9042	20.0958
18	2	171					
19	3	147					
20	4	143					

**Figura 8-34:** A planilha com parâmetros suavizadores e os primeiros erros e previsão de um passo

O primeiro item que você calculará enquanto continua pelos períodos de tempo é uma nova estimativa de nível para o período 1 na célula C17:

$$=C16+D16+C\$2*G17/E5$$

Tal como vimos na seção anterior, o nível novo é igual ao nível anterior mais a tendência anterior mais **alfa** vezes o erro da previsão dessazonalizada. A tendência atualizada em C17 é similar:

$$=D16+D\$2*C\$2*G17/E5$$

Você tem a tendência anterior mais **gama** vezes a quantia de erro dessazonalizado incorporado à atualização do nível.

E para a atualização do fator sazonal de janeiro você tem:

$$=E5+E\$2*(1-C\$2)*G17/(C16+D16)$$

Esse é o fator de janeiro anterior ajustado por ***delta*** vezes o erro ignorado pela correção do nível dimensionado assim como os fatores sazonais pela divisão pelo nível e a tendência anteriores.

Observe que em todos os três parâmetros dessas fórmulas ***alfa***, ***gama*** e ***delta*** são mencionados pelas referências absolutas, para que eles não se movam enquanto você arrasta os cálculos. Ao arrastar C17:G17 por todo o mês 36, você obtém a planilha exibida na Figura 8-35.

SwordForecasting.xlsx						
Home Layout Tables Charts SmartArt Formulas Data Review						
	E52		-E40+E\$2*(1-C\$2)*G52/(C51+D51)			
A	B	C	D	E	F	G
1	Total months		Level smoothing parameter (alpha)	Trend smoothing parameter (gamma)	Seasonal smoothing parameter (delta)	
2	36		0.5	0.5	0.5	
3						
t	Actual Demand	Level	Trend	Seasonal Adjustment	One-step Forecast	Forecast Error
40	24	229	190.684064	1.51579481	1.20638287	233.1156 -4.1156
41	25	189	188.831061	-0.1686042	1.01014428	195.8661 -6.8661
42	26	218	200.281424	5.64087967	1.06046465	194.0932 23.9068
43	27	185	201.840563	3.60000903	0.92621147	192.6368 -7.6368
44	28	199	215.021228	8.3903371	0.90667161	182.0228 16.9772
45	29	210	211.354069	2.36158946	1.02527183	235.4101 -25.41
46	30	193	211.469762	1.23864123	0.91760994	197.1435 -4.1435
47	31	211	222.161688	5.96537953	0.93123693	193.7759 17.2241
48	32	208	226.994721	5.3991103	0.91862962	210.0859 -2.0859
49	33	216	225.858533	2.13146073	0.97100001	228.8725 -12.873
50	34	218	221.570015	-1.0785282	0.99898038	231.01 -13.01
51	35	264	235.798618	6.57503737	1.0878437	231.8137 32.1863
52	36	304	247.183312	8.97986548	1.21835258	292.3954 11.6046

**Figura 8-35:** Levando a equação de atualização por todo o mês 36

Agora que possui as estimativas sazonais, o nível e a tendência finais, você pode prever a demanda do próximo ano. Começando no mês 37 na célula B53 você tem:

$$=(C\$52+(A53-A\$52)*D\$52)*E41$$

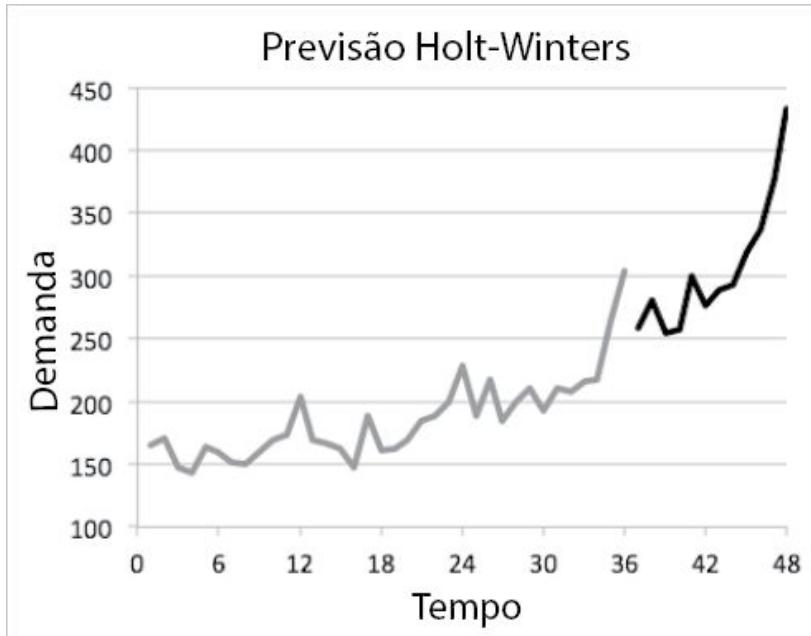
Tal como na Suavização com Tendência Corrigida de Holt, você levará a última estimativa de nível e a adicionará à tendência **vezes** a quantidade de meses decorridos desde a estimativa de tendência mais recente. A única diferença é que você está redimensionando a previsão inteira pelo multiplicador sazonal mais atualizado para janeiro, que está na célula E41. E, enquanto o nível em C\$52 e a tendência em D\$52 usarem referências absolutas para que eles não se desloquem enquanto você arrasta a previsão, a referência sazonal em E41 precisa se mover para baixo enquanto arrasta a previsão pelos próximos 11 meses. E então, ao arrastar o cálculo para baixo, você obtém a previsão exibida na Figura 8-36.

The screenshot shows a Microsoft Excel spreadsheet titled "SwordForecasting.xlsm". The formula bar displays the formula  $=(C52+(A64-A52)*D$52)*E52$ . The spreadsheet contains data for 12 months (t=50 to t=63) and a forecast for month t=64. The columns represent Actual Demand, Level, Trend, Seasonal Adjustment, One-step Forecast, and Forecast Error. The Level column uses the formula  $=C52+(A64-A52)*D$52$ , the Trend column uses  $=D$52$ , and the Seasonal Adjustment column uses  $=E52$ . The One-step Forecast column is calculated by multiplying the Level and Trend values. The Forecast Error column shows the difference between the actual demand and the forecast.

	A	B	C	D	E	F	G
1	Total months		Level smoothing parameter (alpha)	Trend smoothing parameter (gamma)	Seasonal smoothing parameter (delta)		
2	36		0.5	0.5	0.5		
3							
4	t	Actual Demand	Level	Trend	Seasonal Adjustment	One-step Forecast	Forecast Error
50	34	218	221.570015	-1.0785282	0.99898038	231.01	-13.01
51	35	264	235.798618	6.57503737	1.0878437	231.8137	32.1863
52	36	304	247.183312	8.97986548	1.21835258	292.3954	11.6046
53	37	258.76					
54	38	281.17					
55	39	253.90					
56	40	256.68					
57	41	299.46					
58	42	276.26					
59	43	288.72					
60	44	293.06					
61	45	318.49					
62	46	336.64					
63	47	376.35					
64	48	D\$52)*E52					

**Figura 8-36:** Obtendo a previsão Holt-Winters para os próximos meses

Pode-se representar essa previsão usando o gráfico de dispersão de linha reta do Excel tal como nas duas técnicas anteriores (veja a Figura 8-37).



**Figura 8-37:** O gráfico da previsão Holt-Winters

## E...↑Optimize!

Você pensou que tinha terminado, mas não. Hora de configurar aqueles parâmetros suavizadores. Então, assim como nas duas técnicas anteriores, jogue o SSE na célula G2, e coloque o erro padrão em H2.

Dessa vez, a única diferença é que você tem três parâmetros suavizadores, logo, o erro padrão é calculado desta forma:

$$=\text{SQRT}(\text{G2} / (36 - 3))$$

Isso tem como resultado a planilha exibida na Figura 8-38.

Quanto à configuração do Solver (exibida na Figura 8-39), dessa vez você otimizará H2 variando os três parâmetros suavizadores. Pode-se chegar a um erro padrão de quase metade das técnicas anteriores. O gráfico da previsão (veja a Figura 8-40) cai bem aos olhos, não é? Você está localizando com a tendência e as flutuações sazonais. Muito bem.

SwordForecasting.xlsx

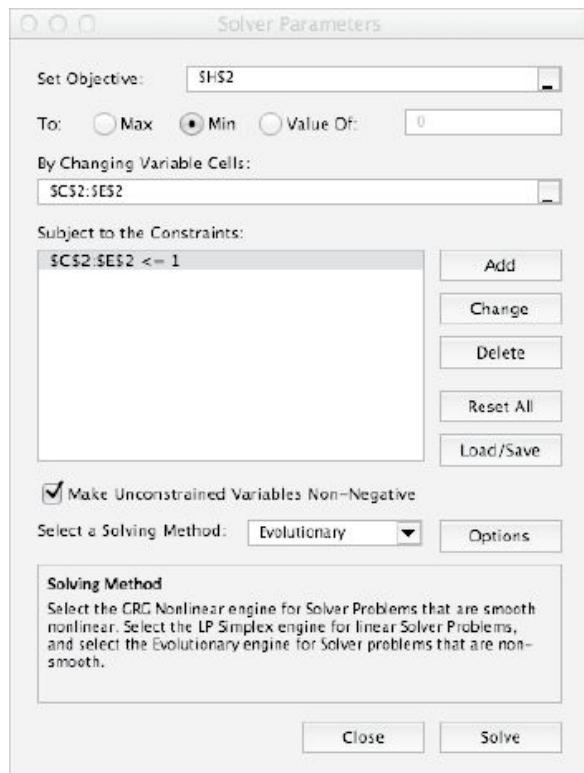
The screenshot shows a Microsoft Excel spreadsheet titled "SwordForecasting.xlsx". The data is organized into several sections:

- Section 1 (Rows 1-3):** Parameters and Summary Statistics.
 

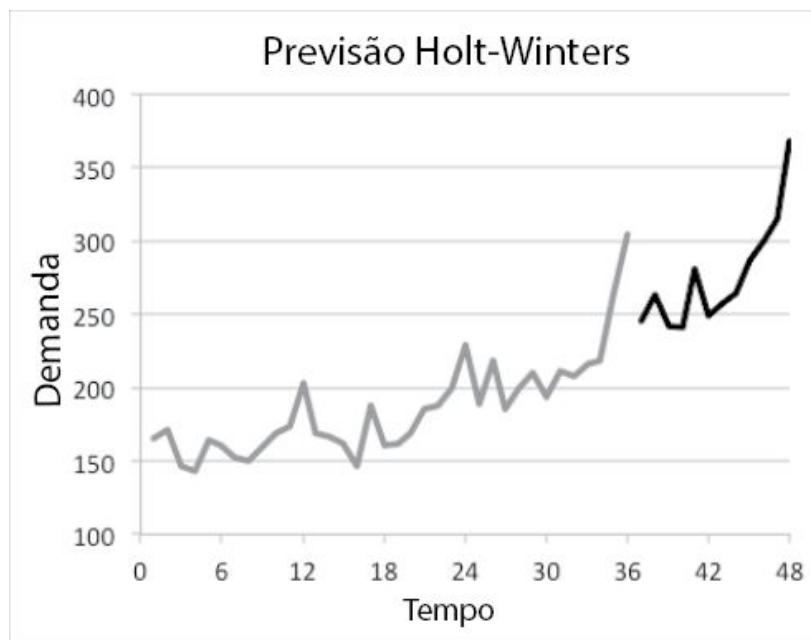
	A	B	C	D	E	F	G	H
1	Total months		Level smoothing parameter (alpha)	Trend smoothing parameter (gamma)	Seasonal smoothing parameter (delta)			Standard Error
2	36			0.5	0.5	0.5	5212.59778	12.5681147
- Section 2 (Rows 4-18):** Actual Data and Forecasts.
 

t	Actual Demand	Level	Trend	Seasonal Adjustment	One-step Forecast	Forecast Error	Squared Error	
5	-11			0.9882334				
6	-10			1.03945951				
7	-9			0.93293329				
8	-8			0.91259776				
9	-7			1.0430106				
10	-6			0.90644245				
11	-5			0.92083759				
12	-4			0.92662094				
13	-3			0.98849075				
14	-2			1.01620145				
15	-1			1.04805266				
16	0	144.42	2.2095	1.20400491				
17	1	165	156.797053	7.29327646	1.02249634	144.904169	20.0958308	403.842415
18	2	171	164.299451	7.39783704	1.04012187	170.565254	0.43474592	0.18900402
- Bottom Row (Row 19):** Summary statistics and average error.

Figura 8-38: Adicionando SSE e o erro padrão



**Figura 8-39:** A configuração do Solver para o Holt-Winters

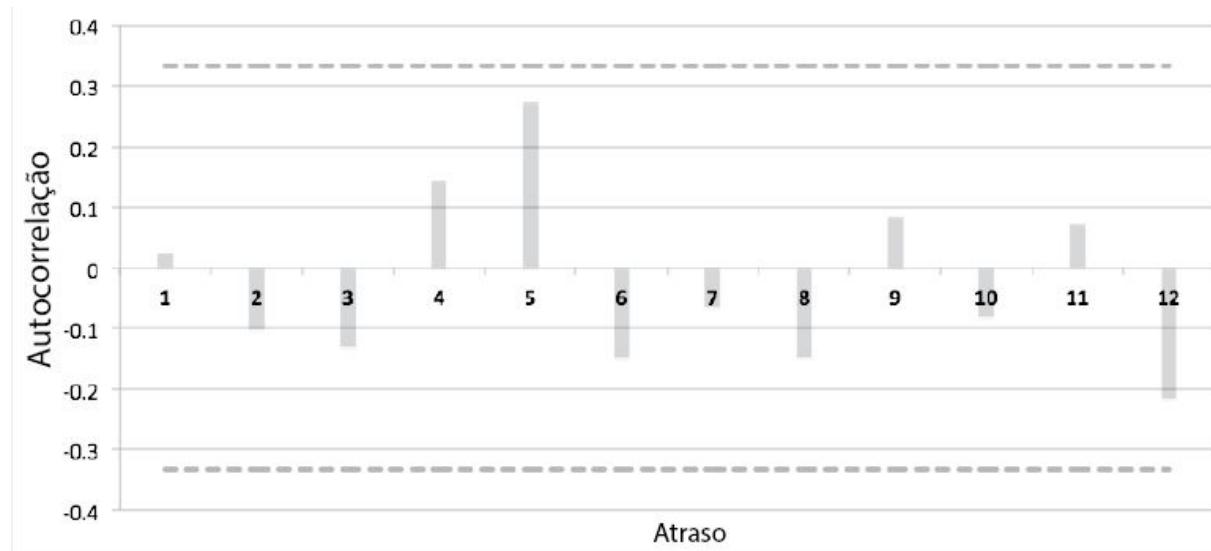


**Figura 8-40:** A previsão Holt-Winters otimizada

Por Favor, Diga que Terminamos Agora!!!

Agora você precisa verificar as autocorrelações para essa previsão. Embora já tenha configurado a planilha de autocorrelação, dessa vez você precisa fazer uma cópia dela e colar os novos valores de erro.

Faça uma cópia da aba Holt's Autocorrelation e a chame de **HW Autocorrelation**. Depois, cole especial os valores da coluna de erro G na planilha de autocorrelação na coluna B. Isso resulta no correlograma exibido na Figura 8-41.



**Figura 8-41:** Correlograma para o modelo Holt-Winters

Pronto! Como não há autocorrelações acima do valor crítico de 0,33, você sabe que o modelo está fazendo um bom trabalho ao capturar a estrutura nos valores da demanda.

## Colocando um Intervalo de Previsão em volta da Previsão

Tudo bem, então você tem uma previsão que ajusta bem. Como você coloca limites altos e baixos em volta dela para que possa usar para configurar expectativas realísticas com o chefe?

Você fará isso por meio da simulação Monte Carlo, que você já viu no Capítulo 4. Essencialmente, você gerará cenários futuros de como a demanda se pareceria e determinará a banda em que 95% desses cenários

caem. A pergunta é como você começa a simular a demanda futura? Na verdade, é bem fácil.

Comece fazendo uma cópia da aba HoltWintersSeasonal e a chame de *PredictionIntervals*. Apague todos os gráficos na aba, eles não são utilizáveis. Além do mais, limpe a previsão nas células B53:B64. Você colocará a demanda “atual” (mas simulada) nesses lugares.

Agora, como disse no início deste capítulo, a previsão está sempre errada. Sempre haverá um erro. Mas você sabe como esse erro será distribuído. Você possui uma previsão bem ajustada e pode presumir que o erro de um passo tenha a média 0 (neutra) com um desvio padrão de 10,37, como calculado na aba anterior.

Tal como no Capítulo 4, pode-se gerar um erro simulado usando a função NORMINV. Nos meses futuros, pode-se acrescentar a média (0), o desvio padrão (10,37 na célula H\$2) e um número aleatório entre 0 e 1 à função NORMINV, e ele retirará um erro da curva de sino. (Veja a discussão sobre as funções de distribuição cumulativa no Capítulo 4 para saber como isso funciona.)

Tudo bem, então jogue um erro de um passo simulado na célula G53:

=NORMINV(RAND(), 0, H\$2)

Arraste-o por todo o G64 para obter 12 meses de erros simulados da previsão de um passo. Isso tem como resultado a planilha exibida na Figura 8-42 (a sua terá valores simulados diferentes desses).

Aqui é onde as coisas ficam analiticamente sinistras. Agora você tem um erro de previsão simulado e uma previsão de um passo. Portanto, se adicionar um erro em G para a previsão em F, você pode, de fato, suspender uma demanda simulada para aquele período de tempo.

Logo, B53 seria:

=F53+G53

	A	B	C	D	E	F	G	H
1	Total months		Level smoothing parameter (alpha)	Trend smoothing parameter (gamma)	Seasonal smoothing parameter (delta)			Standard Error
2	36		0.30719534	0.22854493	0		3550.66489	10.3728446
3								
4	t	Actual Demand	Level	Growth Rate	Seasonal Adjustment	One-step Forecast	Forecast Error	Squared Error
49	33	216	225.443778	3.53306794	0.98849075	225.886034	-9.8860337	97.7336631
50	34	218	224.53712	2.51839098	1.01620145	232.686604	-14.686604	215.696336
51	35	264	234.686314	4.26237232	1.04805266	237.966131	26.0338687	677.762319
52	36	304	243.108719	5.21312676	1.20400491	287.695391	16.3046092	265.840281
53	37						-14.046532	
54	38						-4.7917394	
55	39						1.77645454	
56	40						8.90077396	
57	41						10.6184145	
58	42						-2.5858345	
59	43						17.8924253	
60	44						-8.9456186	
61	45						-22.866944	
62	46						-7.4538358	
63	47						10.9658722	
64	48						13.6237852	

**Figura 8-42:** Erros de um passo simulados

E você pode arrastá-lo por todo B64 para obter todos os valores de demanda dos 12 meses (veja a Figura 8-43).

Uma vez que surja esse cenário, os valores de demanda mudam ao simplesmente atualizar a planilha. Você pode gerar múltiplos cenários de demanda futuros apenas copiando e colando um dos cenários em qualquer lugar e vendo a planilha atualizar sozinha.

Para começar, nomeie a célula A69 como **Simulated Demand** e a célula A70:L70 como os meses **37 ao 48**. Você pode fazer isso copiando A53:A64 e colando especial com valores transpostos em A70:L70.

Da mesma forma, cole especial os valores transpostos no primeiro cenário de demanda em A71:L71. Para inserir um segundo cenário,

apenas clique com o botão direito na linha 71 e selecione Insert para inserir uma nova linha 71 em branco. Depois cole especial mais alguns valores de demanda simulados (eles deveriam atualizar quando você colou o último conjunto).

Continue fazendo essa operação para gerar quantos cenários de demanda futuros quiser. É entediante o suficiente. Em vez disso, você pode gravar um macro.

			C	D	E	F	G	H
1	Total months		Level smoothing parameter (alpha)	Trend smoothing parameter (gamma)	Seasonal smoothing parameter (delta)			Standard Error
2	36		0.30719534	0.22854493	0		3550.66489	10.3728446
3	t	Actual Demand	Level	Growth Rate	Seasonal Adjustment	One-step Forecast	Forecast Error	Squared Error
49	33	216	225.443778	3.53306794	0.98849075	225.886034	-9.8860337	97.7336631
50	34	218	224.53712	2.51839098	1.01620145	232.686604	-14.686604	215.696336
51	35	264	234.686314	4.26237232	1.04805266	237.966131	26.0338687	677.762319
52	36	304	243.108719	5.21312676	1.20400491	287.695391	16.3046092	265.840281
53	37	256	251.689034	5.98268058	0.9882334	245.399942	10.8320911	
54	38	266	257.186607	5.87181167	1.03945951	267.839315	-1.6414631	
55	39	250	264.407893	6.18022715	0.93293329	245.415956	4.09826962	
56	40	238	267.671012	5.51353701	0.91259776	246.938111	-8.6659704	
57	41	266	267.631851	4.2444959	1.0430106	284.934382	-18.852901	
58	42	258	275.63589	5.10372039	0.90644245	246.440262	11.0932975	
59	43	256	279.800648	4.88912539	0.92083759	258.515586	-2.8145982	
60	44	262	284.257492	4.79032959	0.92662094	263.799507	-1.3039303	
61	45	292	291.09096	5.25727845	0.98849075	285.721099	6.57439373	
62	46	300	295.972126	5.17131991	1.01620145	301.14951	-1.2441783	
63	47	312	299.96847	4.90278522	1.04805266	315.614188	-4.0086421	
64	48	366	304.722067	4.86868907	1.20400491	367.066488	-0.5847192	

**Figura 8-43:** Demanda futura simulada

Assim como no Capítulo 7, transforme os passos seguintes em um macro:

1. Insira uma linha 71 em branco.

- 2.** Copie B53:B64.
- 3.** Cole especial os valores transpostos na linha 71.
- 4.** Pressione o botão de gravação Stop.

Uma vez que você tenha gravado tais teclas, pode apertar qualquer atalho de macro que tenha selecionado (veja o Capítulo 7) várias vezes até você obter toneladas de cenários. Você ainda pode segurar a tecla de atalho — 1.000 cenários devem ser suficientes. (Se a ideia de segurar um botão é horripilante, você pode aprender a colocar um loop em volta do código do seu macro usando o Visual Basic for Applications. Procure ele no Google.)

Quando estiver tudo pronto, sua planilha deve se parecer com a Figura 8-44.

The screenshot shows a Microsoft Excel spreadsheet titled "SwordForecasting.xlsxm". The data is organized into several sections:

- Top Row:** Total months (36), Level smoothing parameter (alpha) (0.30719534), Trend smoothing parameter (gamma) (0.22854493), Seasonal smoothing parameter (delta) (0), SSE (3550.66), Standard Error (10.3728).
- Header Row 4:** t, Actual Demand, Growth Rate, Seasonal Adjustment, One-step Forecast, Forecast Error, Squared Error.
- Data Rows 61-64:** Actual Demand values: 292, 300, 312, 366; Growth Rate values: 291.09096, 295.972126, 299.96847, 304.722067; Seasonal Adjustment values: 5.25727845, 5.17131991, 4.90278522, 4.86868907; One-step Forecast values: 285.7211, 301.1495, 315.6142, 367.0665; Forecast Error values: 6.57439, -1.24418, -4.00864, -0.58472.
- Section 69:** Simulated Demand, spanning columns A through L for rows 69 to 76.
- Bottom Row:** Holt's Autocorrelation, HoltWintersInitial, HoltWintersSeasonal, HW Autocorrelation, PredictionIntervals, Average.

**Figura 8-44:** Eu tentei 1.000 cenários de demanda

Uma vez com seus cenários para cada mês, você pode usar a função `PERCENTILE` para obter os limites inferiores e superiores no meio dos 95% dos cenários para criar um intervalo de previsão.

Por exemplo, acima do mês 37 em A66 você pode colocar a fórmula:

```
=PERCENTILE(A71:A1070, 0.975)
```

Isso gera o 97,5º percentil da demanda para esse mês. Na minha planilha, aparece aproximadamente 264. Em A67 você pode obter o 2,5º percentil assim:

```
=PERCENTILE(A71:A1070, 0.025)
```

Repare que estou usando A71:A1070 porque tenho 1.000 cenários de demanda simulados. Você talvez tenha mais ou menos, dependendo da agilidade do seu dedo indicador. Para mim, o limite inferior resulta em aproximadamente 224.

Isso significa que apesar da previsão para o mês 37 ser 245, o intervalo de previsão de 95% é de 224 a 264.

Você pode arrastar essas equações de percentil até o mês 48 na coluna L para obter o intervalo inteiro (veja a Figura 8-45). Agora você pode fornecer a seus superiores uma série tradicional mais uma previsão se você desejar! Sinta-se à vontade para substituir 0,025 e 0,975 por 0,05 e 0,95 para um intervalo de 90% ou 0,1 e 0,9 para um intervalo de 80%, e assim por diante.

L67 : -PERCENTILE(L71:L1070,0.025)

	A	B	C	D	E	F	G	H	I	J	K	L
1	Total months		Level smoothing parameter (alpha)	Trend smoothing parameter (gamma)	Seasonal smoothing parameter (delta)							
2	36		0.30719534	0.22854493		0		3550.66	10.3728			
3												
4	t	Actual Demand	Level	Growth Rate	Seasonal Adjustment	One-step Forecast	Forecast Error	Squared Error				
61	45	292	291.09096	5.25727845	0.98849075	285.7211	6.57439					
62	46	300	295.972126	5.17131991	1.01620145	301.1495	-1.24418					
63	47	312	299.96847	4.90278522	1.04805266	315.6142	-4.00864					
64	48	366	304.722067	4.86868907	1.20400491	367.0665	-0.58472					
65												
66	263.592	284.285	264.700112	266.197426	309.378978	279.5458	291.232	300.17	326.65	348.21	370.74	436.82
67	223.96	240.793	218.676482	216.819849	251.661462	220.213	224.33	230.729	243.53	254.13	263.29	302.77
68												
69	Simulated Demand											
70	37	38	39	40	41	42	43	44	45	46	47	48
71	253.895	257.511	236.488212	235.119858	287.358643	242.2021	274.59	274.032	297.02	299.91	341.44	373.5
72	231.758	255.289	225.791095	235.913185	250.313833	236.7738	237.426	218.736	245.4	244.94	261.9	294.14
73	240.58	264.497	247.950805	225.703194	285.538599	237.3215	240.553	267.418	262.65	281.69	291.79	344.46
74	222.041	248.807	255.754605	246.224517	282.653295	258.222	240.712	255.152	285.54	282.17	221.72	262.81

Figura 8-45: O intervalo de previsão para Holt-Winters

## Criando um Fan Chart para Efeito

Agora, esse último passo não é necessário, mas as previsões com intervalos de previsão geralmente são exibidas em algo chamado *fan chart* (*leque*). Você pode criar esse gráfico no Excel.

Para começar, crie uma nova aba chamada **Fan Chart** e, nessa aba, cole os meses 37 a 48 na linha 1 e então cole os valores do limite superior do intervalo de previsão da linha 66 da aba **PredictionInterval** na linha 2. Na linha 3, cole especial os valores transpostos para a previsão atual da aba **HoltWintersSeasonal**. E, na linha 4, cole os valores do limite inferior do intervalo de previsão da linha 67 da planilha de intervalos.

Então você tem os meses, o limite inferior do intervalo, a previsão e o limite superior na mesma linha (veja a Figura 8-46).

The screenshot shows a Microsoft Excel spreadsheet titled "SwordForecasting.xlsxm". The ribbon at the top has tabs for Home, Layout, Tables, Charts, SmartArt, Formulas, Data, Review, Developer, and a search bar. Below the ribbon is a data table with columns labeled A through L and rows labeled 1 through 4. The data values are as follows:

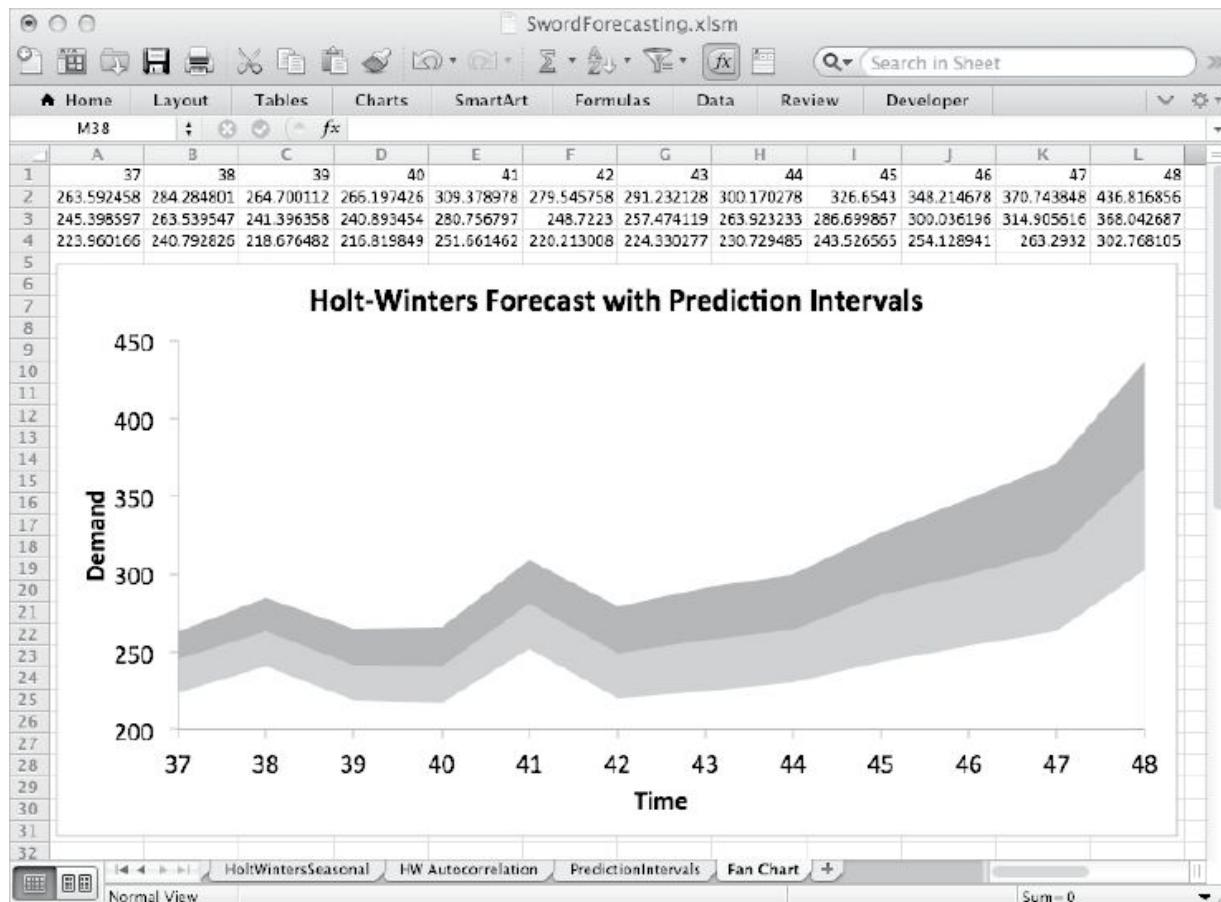
	A	B	C	D	E	F	G	H	I	J	K	L
1	37	38	39	40	41	42	43	44	45	46	47	48
2	263.59	284.28	264.7	266.2	309.38	279.55	291.23	300.17	326.65	348.21	370.74	436.82
3	245.4	263.54	241.4	240.89	280.76	248.72	257.47	263.92	285.7	300.04	314.91	368.04
4	223.96	240.79	218.68	216.82	251.66	220.21	224.33	230.73	243.53	254.13	263.29	302.77

The ribbon also shows tabs for Holt's Autocorrelation, Holt/WintersInitial, HoltWintersSeasonal, HW Autocorrelation, Prediction/Intervals, and Fan Chart. The "Fan Chart" tab is selected.

**Figura 8-46:** A previsão colada pelo intervalo de previsão

Ao destacar A2:L4 e selecionar Area Chart do menu de gráficos do Excel, você obtém três áreas sólidas de gráficos uma sobre a outra. Clique com o botão direito em uma das séries e escolha Select Data. Você pode mudar o nome do eixo Category (X) para uma das séries para ser A1:L1 a fim de adicionar os nomes mensais corretos ao gráfico.

Agora, clique com o botão direito na série de limite inferior e formate-a para ter um preenchimento branco. Remova também as linhas de grade do gráfico em favor da consistência. Sinta-se à vontade para adicionar nomes aos eixos e um título. Isso resulta no fan chart exibido na Figura 8-47.



**Figura 8-47:** O fan chart possui uma beleza incrível

O mais legal sobre esse fan chart é a conveniência de ambos os intervalos e a previsão estarem em uma única imagem. Bem, você poderia colocar em camadas um intervalo de 80% se quisesse mais sombreado em cinza. Há dois itens interessantes que se destacam nesse gráfico:

- Os erros ficam maiores conforme o tempo passa. Faz sentido. A incerteza de mês para mês é atenuada.
- Da mesma forma, há mais erros em termos absolutos durante os períodos de demanda sazonal alta. Quando a demanda cai um pouco, os erros são mais restritos.

## Resumindo

Este capítulo abordou diversos conteúdos:

- Suavização exponencial simples (SES)
- Realizar um teste t em uma regressão linear para verificar a tendência linear em séries temporais
- Suavização Exponencial com Tendência Corrigida de Holt
- Calcular autocorrelações e representar em gráfico um correlograma usando uma medida móvel  $2 \times 12$
- Iniciar a Suavização Exponencial Multiplicativa Holt-Winters usando uma medida móvel  $2 \times 12$
- Fazer previsões com Holt-Winters
- Criar intervalos de predição próximo a previsão usando a simulação Monte Carlo
- Representando em gráfico um intervalo de predição em um fan chart

Se você passou pelo capítulo inteiro, parabéns. De verdade, é muita previsão para um capítulo.

Agora, se quiser ir além com previsões, há alguns livros didáticos excelentes por aí. Eu gosto do *Forecasting, Time Series, and Regression by Bowerman* et al. (CengageLearning, 2004). Hyndman tem um livro didático online gratuito em <http://otexts.com/fpp/> em inglês, e seu blog (incrivelmente chamado de “HyndSight”) é uma fonte excelente. Para perguntas, <http://stats.stackexchange.com/> em inglês, é o lugar de destino.

Quando se trata de fazer previsões em uma configuração de produção, existem inúmeros produtos por aí. Para trabalhos leves, sinta-se à vontade para permanecer no Excel. Se você tiver toneladas de produtos ou SKUs, usar alguns códigos seria conveniente.

SAS e R possuem pacotes excelentes para fazer previsões. Os que estão em R foram escritos pelo Hyndman (veja o Capítulo 10), que desenvolveu o apoio estatístico para como fazer intervalos de previsão nas técnicas de suavização exponencial.

E é isso! Espero que você esteja encorajado a seguir em frente e a “organizar sua ignorância”!

## 9

# Detecção↑de↑Valor↑Atípico:↑Só Porque↑Eles↑São↑Estranhos↑Não Significa↑que↑Não↑São Importantes

**O**s valores atípicos são os pontos estranhos em um conjunto de dados — aqueles que não se encaixam de alguma forma.

Historicamente, eles são os valores **extremos**, quantidades tão grandes ou tão pequenas para terem sido criados naturalmente a partir dos mesmos processos das outras observações no conjunto de dados.

A única razão pela qual as pessoas costumavam se importar com os valores atípicos era porque elas queriam livrar-se deles. Há cem anos, os estatísticos tinham muito em comum com o Borg: um ponto de dados precisava assimilar ou ele morreria. Entretanto, isso era feito com boa vontade (no caso dos estatísticos) — os valores atípicos podem mover médias e mexer com as medidas de dispersão nos dados. Um ótimo exemplo de remoção de valor atípico é na ginástica artística, em que os mais altos e mais baixos escores dos juízes são sempre retirados dos dados antes de calcular a média do escore.

Os valores atípicos possuem o dom de bagunçar os modelos de aprendizado de máquina. Por exemplo, nos Capítulos 6 e 7 você observou a previsão das clientes grávidas baseada em seus dados de compras. Mas, e se uma loja processou erroneamente alguns itens nas prateleiras das farmácias e estavam registrando compras de um multivitamínico como se fossem ácido fólico? As clientes com esses vetores de compras defeituosas são os valores atípicos que alteram a

relação de compras-de-gravidez-com-ácido-fólico de forma que prejudica o entendimento do modelo IA.

Certa vez quando eu prestava consultoria para o governo, minha empresa encontrou um local de armazenamento de água que os Estados Unidos tinham em Dubai e foi avaliado em bilhões e bilhões de dólares. O valor da propriedade era um valor atípico que estava prejudicando os resultados da nossa análise — acontece que alguém o digitou no banco de dados com muitos zeros a mais.

Então esse é um motivo para se importar com os valores atípicos: *facilitar a limpeza da modelagem e da análise de dados.*

No entanto, há outro motivo para dar importância aos valores atípicos. Eles mesmos são bem interessantes!

## Os↑Valores↑Atípicos↑São↑(Más)↑Pessoas Também

Observe quando a empresa do seu cartão de crédito liga após você ter feito uma transação que seja potencialmente fraudulenta. O que a empresa do seu cartão de crédito está fazendo? Eles estão detectando a transação como se fosse um valor atípico baseado em seu comportamento passado. Em vez de ignorar a transação por ela ser um valor atípico, eles estão sinalizando uma fraude em potencial e tomando atitudes.

No MailChimp, quando previmos os spams antes de eles serem enviados, nós estávamos prevendo os valores atípicos. Esses spams são enviados por grupos de pessoas cujo comportamento difere do que nós como uma empresa consideramos como normal. Usamos modelos supervisionados similares àqueles dos Capítulos 6 e 7 para prever com base em ocorrências passadas quando um novo usuário vai enviar um spam.

Portanto, no caso do MailChimp, um valor atípico nada mais é do que uma classe pequena mas compreendida de dados na população que pode ser prevista usando os dados em treinamento. Mas e os casos nos quais

você não sabe o que está procurando? Como esses compradores de ácido fólico com rótulos errados? Os fraudadores geralmente mudam seu comportamento para que a única coisa que você possa esperar deles seja algo inesperado. Se tal erro nunca aconteceu antes, como encontrar os pontos estranhos pela primeira vez?

Esse tipo de detecção de valor atípico é um exemplo de *aprendizagem não supervisionada* e mineração de dados. É o lado negativo da intuição da análise realizada nos Capítulos 2 e 5 deste livro nos quais você detectou agrupamentos de pontos. Na análise de agrupamentos, procura-se por um grupo de amigos dos pontos de dados e para a análise. Na detecção do valor atípico, você se importa com os pontos de dados que diferem dos grupos. Eles são estranhos e especiais ao mesmo tempo.

Este capítulo começa com uma forma simples e padrão de calcular valores atípicos em dados unidimensionais normais. Logo depois ele segue para gráficos dos k vizinhos mais próximos (kNN) para detectar os valores atípicos em dados multidimensionais, similar a como você usou os gráficos r-vizinhança para criar os agrupamentos no Capítulo 5.

## O↑Caso↑Fascinante↑de↑Hadlum↑vs Hadlum

### NOTA

A pasta de trabalho do Excel utilizado nesta seção, “Pregnancy Duration.xlsx”, está disponível para download no site da editora, em [www.altabooks.com.br](http://www.altabooks.com.br), procurando pelo título do livro. Mais adiante neste capítulo, você mergulhará em uma planilha ainda maior, “SupportCenter.xlsx”, também disponível no mesmo site.

Na década de 1940, um rapaz britânico chamado Sr. Hadlum foi para a guerra. Alguns dias depois, 349 pra falar a verdade, sua esposa Sra. Hadlum deu à luz. Agora, a média de uma gravidez dura

aproximadamente 266 dias. Isso coloca Sra. Hadlum 12 semanas **atrasada** da data prevista. Não consigo imaginar uma mulher solteira que aturaria isso somado ao desconforto hoje em dia, mas naquele tempo, induzir o parto não era tão comum.

Agora, o Sr. Hadlum garante que ela não teve nada além de uma gravidez excepcionalmente longa. Nada mais justo.

Mas o Sr. Hadlum concluiu que a gravidez deve ter sido o resultado de outro homem enquanto ele estava fora — uma gravidez de 349 dias era uma anomalia que não poderia ser justificada devido à distribuição da duração de nascimentos regulares.

Então, se você tinha uma data para a gravidez, qual a forma mais rápida de decidir se a gravidez da Sra. Hadlum deveria ser considerada um valor atípico?

Bem, estudos comprovam que a duração de uma gestação é mais ou menos uma variável aleatória distribuída normalmente com uma média de 266 dias após a concepção, com um desvio padrão de aproximadamente 9. Você pode calcular a função de distribuição acumulada normal (CDF) introduzida no Capítulo 4 para obter a probabilidade de um valor menor do que 349 de ocorrência. No Excel, isso é avaliado usando a função NORMDIST:

```
=NORMDIST(349, 266, 9, TRUE)
```

A função NORMDIST é preenchida com o valor cuja probabilidade cumulativa você quer, a média, o desvio padrão e uma sinalização configurada para TRUE, o que configura a função para fornecer o valor cumulativo.

Essa fórmula retorna um valor de 1,000 afora até quanto as trilhas decimais do Excel permitirem. Isso significa que aproximadamente todos os bebês nascidos daqui até a eternidade nascerão com ou abaixo de 349 dias. Subtraindo esse valor de 1:

```
=1 - NORMDIST(349, 266, 9, TRUE)
```

Você obtém 0,0000000 até onde a vista alcança. Em outras palavras, é quase impossível para um bebê humano permanecer em gestação todo esse tempo.

Nunca saberemos com certeza, mas eu aposto um bom dinheiro que a Sra. Hadlum escondeu alguma coisa. O engraçado é que a corte a favoreceu, alegando que tal longa gestação, mesmo que improvável, ainda era possível.

## O↑Teste↑de↑Tukey

Esse conceito dos valores atípicos serem pontos improváveis quando amostrados a partir da curva de sino, liderou uma regra geral para a detecção de valores atípicos chamado de Teste de Tukey (*Tukey Fence*). Os Testes de Tukey são fáceis de serem verificados e codificados. Eles são usados pelos pacotes estatísticos mundo afora para identificar e remover os pontos de dados falsos de qualquer conjunto de dados que se ajuste em uma curva de sino regular.

Estas são as técnicas do Teste de Tukey em sua integridade:

- Calcule as porcentagens de 25% e 75% em qualquer conjunto de dados que você quiser encontrar valores atípicos. Esses valores também são chamados de *primeiro quartil* e *terceiro quartil*. O Excel calcula esses valores usando a função `PERCENTILE`.
- Subtraia o primeiro quartil do terceiro para obter a medida de dispersão dos dados, que é chamada de *Intervalo do Interquartil* (IQR, do inglês *Interquartile Range*). O IQR é legal porque ele é relativamente robusto em comparação com os valores extremos como uma medida de dispersão, diferente do cálculo do desvio padrão típico que você usou para medir a dispersão nos Capítulos anteriores deste livro.
- Subtraia  $1,5 * \text{IQR}$  do primeiro quartil para obter o delimitador (cerca) interno inferior. Adicione  $1,5 * \text{IQR}$  ao terceiro quartil para obter o delimitador interno superior.

- Da mesma forma, subtraia 3\*IQR do primeiro quartil para obter o delimitador externo inferior. Adicione 3\*IQR ao terceiro quartil para obter o delimitador externo superior.
- Quaisquer valores abaixo do delimitador interno ou acima do delimitador externo são extremos. Em dados distribuídos normalmente, você veria ao menos 1 a cada 100 pontos fora do delimitador interno, mas somente 1 a cada 500.000 pontos fora do delimitador externo.

## Aplicando o Teste de Tukey em uma Planilha

Eu incluí uma planilha chamada `PregnancyDuration.xlsx` para download no site da editora para que você possa aplicar essa técnica aos seus dados atuais. Se abri-la, verá uma aba chamada `Pregnancies`, com uma amostra de 1.000 durações na coluna A.

O período de gestação da Sra. Hadlum de 349 dias está na célula A2. Na coluna D, coloque todos os resumos estatísticos e os delimitadores. Comece com a mediana (o valor do meio), pois é uma estatística mais robusta da centralidade do que a média (as médias podem ser distorcidas pelos valores atípicos).

Nomeie C1 como Median e, em D1, calcule a mediana desta forma:

```
=PERCENTILE(A2:A1001, 0.5)
```

Essa seria a 50<sup>a</sup> porcentagem. Abaixo da média, você pode calcular o primeiro e o terceiro quartil assim:

```
=PERCENTILE(A2:A1001, 0.25)
```

```
=PERCENTILE(A2:A1001, 0.75)
```

E o intervalo do interquartil é a diferença entre eles:

```
=D3 - D2
```

Adicionando 1,5 e 3 vezes o IQR ao primeiro e terceiro quartil respectivamente, pode-se calcular então todos os delimitadores:

```
=D2 - 1.5*D4
```

```
=D3 + 1.5*D4
```

=D2 - 3 \* D4

=D3 + 3 \* D4

Se nomear todos esses valores, você terá a planilha exibida na Figura 9-1.

	A	B	C	D
1	Birth Duration		Median	267
2		349	1st Quartile	260
3		278	3rd Quartile	272
4		266	Interquartile Range	12
5		265	Lower Inner Fence	242
6		269	Upper Inner Fence	290
7		263	Lower Outer Fence	224
8		278	Upper Outer Fence	308
	257			

Figura 9-1: O Teste de Tukey para a duração de algumas gestações

Agora você pode aplicar uma formatação condicional à planilha e ver quem sai dos limites. Comece com o delimitador interno. Para destacar os valores extremos, selecione Conditional Formatting a partir da aba Home, escolha Highlight Cells Rules, e selecione Less Than, como mostra a Figura 9-2.

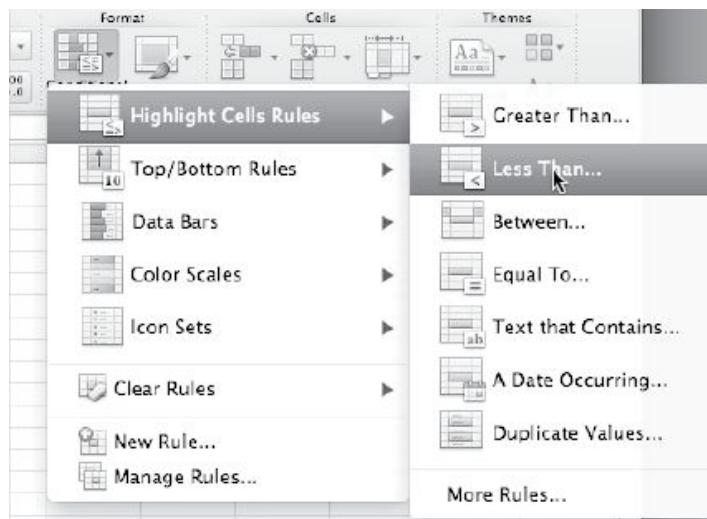


Figura 9-2: Adicionando formatação condicional para os valores atípicos

Ao especificar o delimitador interno inferior, fique à vontade para escolher uma cor de destaque que agrade seu olhar (eu vou escolher preenchimento amarelo para os delimitadores internos e vermelho para os externos, porque eu gosto de semáforos). Similarmente, adicione formatação para os outros três delimitadores (se estiver usando o Excel 2011 para Mac, pode usar a regra Not Between para adicionar formatação com duas regras em vez de quatro).

Como mostra a Figura 9-3, a Sra. Hadlum torna-se vermelha, significando que sua gravidez foi radicalmente extrema. Ao rolar para baixo, você não encontrará nenhuma outra gravidez em vermelho, mas há nove amarelas. Isso combina perfeitamente com 1 de 100 pontos que você esperaria que fossem sinalizados em dados normais pela regra.

	A	B	C	D
1	Birth Duration		Median	267
2	349		1st Quartile	260
3	278		3rd Quartile	272
4	266		Interquartile Range	12
5	265		Lower Inner Fence	242
6	269		Upper Inner Fence	290
7	263		Lower Outer Fence	224
8	278		Upper Outer Fence	308
9	257			

**Figura 9-3:** Uh-oh, Sra. Hadlum. O que você diz sobre essa formatação condicional?

## As Limitações dessa Simples Abordagem

O Teste de Tukey funciona somente quando estas três coisas forem verdadeiras:

- Os dados são vagamente distribuídos normalmente. Eles não têm que ser perfeitos, mas deveriam ser em formato de curva de sino e, esperançosamente, simétricos sem nenhuma parte sobressaindo de um lado.

- A definição de um valor atípico é um valor extremo no perímetro de uma distribuição.
- Você está observando dados unidimensionais.

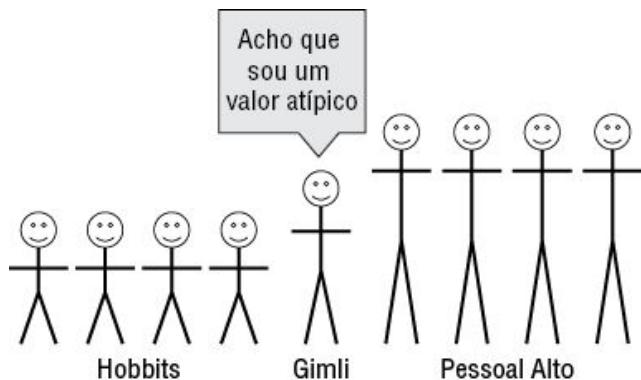
Vamos ver um exemplo de um valor atípico que viola as duas primeiras pressuposições.

Em *O Senhor dos Anéis: A Sociedade do Anel*, quando os aventureiros finalmente formam uma única sociedade (pela qual o livro recebeu o nome), todos eles ficam em um grupo enquanto o líder dos elfos, Elrond, declara quem eles são e qual a sua missão.

Esse grupo contém quatro pessoas altas: Gandalf, Aragorn, Legolas e Boromir. Também há quatro pessoas baixas, os hobbits Frodo, Merry, Pippin e Sam.

E, entre eles, há somente um anão: Gimli. Ele é menor do que os homens e maior do que os hobbits por alguns centímetros (veja a Figura 9-4).

No filme, quando nos apresentam esse grupo pela primeira vez, Gimli é o valor atípico pela altura. Ele não pertence a nenhum grupo.



**Figura 9-4:** Gimli, filho de Gloin, não é um valor atípico

Mas como ele é um valor atípico? Sua altura não é nem a menor e nem a maior. Na verdade, ela é a mais próxima da média do grupo.

Entenda, essa distribuição de altura não chega nem perto da normal. Se for o caso, você poderia chamar de “multi-modal” (uma distribuição com

picos múltiplos). Gimli é um valor atípico não porque sua altura é extrema, e sim porque ele está entre picos múltiplos. Esses tipos de pontos de dados podem ser ainda mais difíceis de representar quando se está explorando diversas dimensões.

Esse tipo de valor atípico ocorre em fraudes com frequência. Alguém que seja *muito normal* para na verdade *ser normal*. Bernie Madoff é um bom exemplo disso. Apesar da maioria dos esquemas de Ponzi oferecerem taxas de valor atípico com um retorno de 20% a mais, Madoff ofereceu retornos modestos com confiança que se misturaram ao ruído a cada ano — ele não estava pulando nenhum Teste de Tukey. Mas, ao longo dos anos, seu retorno plurianual e sua confiabilidade se tornaram um valor atípico multidimensional.

Como encontrar valores atípicos no caso de dados multimodelos e multidimensionais (você poderia simplesmente chamar de dados do “mundo real”)?

Uma maneira fantástica de abordagem é tratar os dados como um gráfico, assim como você fez no Capítulo 5 para encontrar agrupamentos nos dados. Pense sobre isso. O que define Gimli como um valor atípico é a sua relação com os outros pontos de dados; a distância deles em relação à distância de cada um.

Todas essas distâncias, cada ponto alternado, define os vértices em um gráfico. Usando esse gráfico, você pode provocar os pontos isolados. Para fazer isso, você pode começar criando um gráfico de k vizinho mais próximo (kNN) e partir dali.

## Terrível↑em↑Nada,↑Ruim↑em↑Tudo

Para esta próxima seção, imagine que você gerencia um enorme call center de suporte ao cliente. Cada chamada, e-mail ou bate-papo de um cliente gera uma venda, e cada membro do time de suporte é designado para lidar com ao menos 140 vendas por dia. Ao final de cada interação, uma oportunidade é dada ao cliente para classificar o funcionário de

suporte em uma escala de cinco estrelas. A equipe de suporte deve manter a média acima de 2, ou são demitidos.

Altos padrões, eu sei.

A empresa mantém registros de muitas outras métricas sobre cada funcionário também. Quantas vezes eles chegaram atrasados durante o ano. Em quantos turnos da noite e de fim de semana eles ficaram em nome do grupo. Quantos dias eles tiraram de licença médica, e, dentro desses, quantos foram em uma sexta-feira. A empresa até registra quantas horas o funcionário usa para cursos de treinamento interno (eles vão até 40 horas pagas) e quantas vezes eles pediram para trocar o turno ou foram bons samaritanos e atenderam ao pedido de outro funcionário.

Você tem esses dados para todos os 400 funcionários do call center em uma planilha. As perguntas são: quantos funcionários são valores atípicos e o que eles lhe ensinam sobre ser um funcionário de call center? Existem algumas escorregadas que não foram selecionadas pelas exigências da venda e as taxas mínimas do cliente? Talvez os valores atípicos lhe ensinem como fazer regras melhores.

Se você abrir a planilha para esta seção do capítulo (SupportCenter.xlsx está disponível para download na página da editora em [www.altabooks.com.br](http://www.altabooks.com.br), procurando pelo título do livro), encontrará todos esses dados de desempenho registrados na planilha SupportPersonnel (veja a Figura 9-5).

---

### PADRONIZANDO USANDO MEDIDAS SÓLIDAS DE CENTRALIDADE E ESCALA

Nem todos os dados que você quer dimensionar são distribuídos normalmente. Subtrair a média e dividir pelo desvio padrão tende a funcionar bem de qualquer forma. Mas os valores atípicos podem bagunçar o cálculo da média e do desvio padrão, portanto, às vezes, as pessoas gostam de padronizar ao subtrair estatísticas sólidas de centralidade (o “meio” do dado) e dividir por mais medidas sólidas de dimensão/dispersão estatística (a dispersão dos dados).

Estes são alguns cálculos de centralidade que funcionam melhor em valores atípicos unidimensionais do que a média:

- Mediana — Sim, apenas o percentil de 50%
- Média dos Quartis (*Midhinge*) — A média dos 25% e 75%
- Média Interna (*Trimean*) — A média da mediana e da média dos quartis. Gosto dessa porque parece algo inteligente.
- Média Aparada (*trimmed/truncated mean*) — A média, mas você descarta os pontos N superiores e inferiores ou as porcentagens dos pontos superiores e inferiores. Vemos isso bastante nos esportes (lembre-se da ginástica artística na qual eles descartam os escores mais altos e mais baixos). Se você descartar a parte superior e inferior em 25%, e fizer a média do meio com 50% dos dados, isso tem o próprio nome: a média interquartil (IQM, do inglês interquartile mean).
- Média Winsorizada (de Charles P. Winsor) — Como a média aparada, mas em vez de descartar pontos muito grandes ou muito pequenos, os substitui por um limite.

Para as medidas sólidas da dimensão, estas são algumas que valem a pena serem usadas em vez do desvio padrão:

- Intervalo interquartil — Você viu isso anteriormente neste capítulo. É só o 75º percentil menos o 25º percentil nos dados. Você pode usar outros percentis também. Por exemplo, se você usa o 90º e o 10º percentis, você obtém o intervalo interdecil.
- Desvio absoluto da mediana (MAD, do inglês *median absolute deviation*) — Tire a mediana dos dados. E então pegue o valor absoluto da diferença de cada ponto a partir da mediana. A mediana desses padrões é o MAD. E como se fosse a resposta da mediana para o desvio padrão.

	A	B	C	D	E	F	G	H	I	J	K
1	Employee ID	Avg Tix / Day	Customer rating	Tardies	Graveyard Shifts Taken	Weekend Shifts Taken	Sick Days Taken	% Sick Days Taken on Friday	Employee Dev. Hours	Shift Swaps Requested	Shift Swaps Offered
2	144624	151.8	3.32	1	0	2	3	0%	0	2	1
3	142619	155.2	3.16	1	3	1	1	0%	12	1	2
4	142285	164.2	4.00	3	3	1	0	0%	23	2	0
5	142158	159	2.77	0	3	1	2	50%	13	1	0
6	141008	155.5	3.52	4	1	0	3	67%	16	1	0
7	145082	153.8	3.90	3	2	1	3	100%	5	1	0
8	139410	162.1	3.45	3	3	1	3	0%	13	2	1
9	135014	154	3.67	0	3	1	1	0%	18	1	2
10	139356	157.5	3.40	0	1	1	4	25%	14	0	3

Figura 9-5: Dados multidimensionais de desempenho dos funcionários

## Preparando os Dados para o Gráfico

Há um problema com esses dados de desempenho. Você não pode medir a distância entre os funcionários a fim de descobrir quem está “do lado de fora” quando cada coluna é dimensionada com tanta diferença. O que quer dizer ter uma diferença de 5 entre dois funcionários em suas médias de vendas versus uma diferença de 0,2 na classificação dos clientes? Você precisa *padronizar* cada coluna para que os valores estejam próximos do mesmo centro e dispersão.

As colunas dos dados geralmente são padronizadas da seguinte forma:

1. Subtrair a média da coluna de cada observação.
2. Dividir cada observação pelo desvio padrão da coluna.

Para dados distribuídos normalmente, isso centraliza os dados em 0 (gera uma média de 0) e gera um desvio padrão de 1. De fato, a distribuição normal com média 0 e desvio padrão 1 é chamada de *distribuição normal padrão*.

Para começar, calcule a média e o desvio padrão de cada coluna na parte inferior da planilha SupportPersonnel. O primeiro valor que você vai querer em B402 é a média das vendas tiradas por dia, que pode ser escrita desta forma:

=AVERAGE(B2:B401)

E abaixo dela você tira um desvio padrão da coluna assim:

=STDEV(B2:B401)

Copiando essas fórmulas por toda a coluna K, tem-se a planilha exibida na Figura 9-6.

	A	B	C	D	E	F	G	H	I	J	K
1	Employee ID	Avg Tix / Day	Customer rating	Tardies	Graveyard Shifts Taken	Shifts Taken	Sick Days Taken	% Sick Days Taken on Friday	Employee Dev. Hours	Shift Swaps Requested	Shift Swaps Offered
395	141343	159.1	3.60	4	1	0	0	0%	17	0	2
396	143981	160.3	3.70	1	2	1	0	0%	1	2	4
397	139820	162.6	3.37	2	3	1	1	100%	6	1	3
398	144780	159.6	3.50	1	2	1	2	0%	15	1	1
399	138420	155.4	4.29	3	3	1	2	50%	18	1	3
400	131547	150.7	3.99	1	2	1	4	25%	30	2	3
401	137942	160.6	3.87	1	1	1	2	100%	16	1	0
402	Mean	156.086	3.50	1.465	1.985	0.9525	1.875	35%	11.97	1.4475	1.76
403	Std. Dev.	4.41664	0.46	0.9727	0.79457749	0.548631	1.673732	39%	7.4708523	0.99987155	1.8126263

Figura 9-6: A média e o desvio padrão para cada coluna

Crie uma nova aba chamada Standardized e copie os nomes das colunas da linha 1 e os IDs dos funcionários da coluna A. Você pode começar a padronizar os valores na célula B2 usando a fórmula STANDARDIZE do Excel. Essa fórmula pega o valor original, um centro, e uma medida de dispersão e retorna o valor com o centro subtraído e dividido pela dispersão. Logo, em B2 temos:

```
=STANDARDIZE(SupportPersonnel!B2,  
SupportPersonnel!B$402, SupportPersonnel!B$403)
```

Observe que você está usando referências absolutas nas linhas para a média e o desvio padrão, para que eles estejam preparados para quando você copiar a fórmula. Entretanto, quando você copia a fórmula, a coluna mudará.

Copie e cole B2 por todo K2, realce a série, e dê um clique duplo para enviar os cálculos para K401. Isso produz o conjunto de dados

padronizados exibido na Figura 9-7.

The screenshot shows a Microsoft Excel spreadsheet titled "SupportCenter.xlsx". The active sheet is labeled "B2" and contains a formula in cell B2: =STANDARDIZE(SupportPersonnel!B2,SupportPersonnel!B\$402,SupportPersonnel!B\$403). The data is organized in columns A through K. Columns A, B, C, D, E, F, G, I, J, and K have headers. Column A is "Employee ID", B is "Avg Tix / Day", C is "Customer rating", D is "Tardies", E is "Graveyard Shifts Taken", F is "Weekend Shifts Taken", G is "% Sick Days Taken on Friday", H is "Employee Dev. Hours Requested", I is "Shift Swaps Offered", and K is "Shift Swaps Offered". Rows 1 through 8 contain numerical data for each column. Row 9 is a summary row with the formula =AVERAGE(B2:B9) in B9. The status bar at the bottom shows "Average = -0.970364853".

	A	B	C	D	E	F	G	H	I	J	K
1	Employee ID	Avg Tix / Day	Customer rating	Tardies	Graveyard Shifts Taken	Weekend Shifts Taken	Sick Days Taken on Friday	Employee Dev. Hours Requested	Shift Swaps Offered	Shift Swaps Offered	
2	144624	-0.9704	-0.37206	-0.478	-2.498183	1.909297	0.672151	-0.8965586	-1.602227	0.55257098	-0.4192811
3	142619	-0.2005	-0.73098	-0.478	1.2774085	0.086579	-0.52278	-0.8965586	0.0040156	-0.4475575	0.13240457
4	142285	1.8372	1.083058	1.5781	1.2774085	0.086579	-1.12025	-0.8965586	1.4764045	0.55257098	-0.9709668
5	142158	0.65983	-1.57789	-1.506	1.2774085	0.086579	0.074683	0.37633793	0.1378691	-0.4475575	-0.9709668
6	141008	-0.1326	0.062075	2.6062	-1.239653	-1.736139	0.672151	0.80063678	0.5394298	-0.4475575	-0.9709668
7	145082	-0.5175	0.867104	1.5781	0.018878	0.086579	0.672151	1.64923448	-0.932959	-0.4475575	-0.9709668
8	139410	1.36173	-0.10399	1.5781	1.2774085	0.086579	0.672151	0.8965586	0.1378691	0.55257098	-0.4192811
Normal View Ready											

Figura 9-7: Os dados de desempenhos do conjunto de funcionários padronizado

## Criando um Gráfico

Um gráfico nada mais é do que alguns vértices (nós) e linhas (arestas). Nesse caso, cada funcionário é um nó, e para começar, você pode desenhar linhas entre todos eles. O tamanho das linhas é a distância euclidiana entre os dois funcionários usando seus dados de desempenho padronizados.

Como vimos no Capítulo 2, a distância euclidiana (em linha reta) entre dois pontos é a raiz quadrada da soma das diferenças quadradas do valor de cada coluna para os dois.

Em uma nova planilha chamada Distances, crie uma matriz de distância funcionário-por-funcionário exatamente da mesma forma do Capítulo 2, usando a fórmula OFFSET.

Para começar, numere os funcionários de 0 a 399 começando em A3 e descendo e atravessando C1. (Dica: digite 0, 1 e 2 nas primeiras três células e então realce-as e arraste-as para baixo e atravessando. O Excel preenche o restante para você, porque ele é muito esperto.) Próximo a esses valores de desvio, cole as IDs dos funcionários (cole especial os valores transpostos para as colinas). Isso cria a matriz vazia exibida na Figura 9-8.

Para preencher essa matriz, vamos começar na primeira célula C3 de distância. Essa é a distância entre o funcionário 144624 e ele mesmo.

Agora, para todos esses cálculos, você usará a fórmula `OFFSET` ancorada na primeira linha dos dados dos funcionários padronizados:

```
OFFSET(Standardized!$B$2:$K$2, algum número de linha, 0 colunas)
```

No caso da célula C3, `Standardized!$B$2:$K$2` é a linha que você quer para o funcionário 144624. Você pode tirar a diferença entre esse funcionário e ele mesmo usando a fórmula de desvio como:

```
OFFSET(Standardized!$B$2:$K$2, Distances!$A3, 0) -
```

```
OFFSET(Standardized!$B$2:$K$2, Distances!C$1, 0)
```

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1			0	1	2	3	4									
2			144624	142619	142285	142158	141008									
3	0	144624														
4	1	142619														
5	2	142285														
6	3	142158														
7	4	141008														
8	5	145082														
9	6	139410														
10	7	135014														
11	8	139356														
12	9	137368														
13	10	141982														
14	11	144753														
15	12	132229														
16	13	137744														

**Figura 9-8:** Matriz de distância dos funcionários vazia

Na primeira fórmula de desvio, você move as linhas usando o valor em \$A3, enquanto que na segunda fórmula de desvio você usa o valor em C\$1 para mover a fórmula `OFFSET` para outro funcionário. As referências absolutas são usadas nesses valores nas posições apropriadas para que quando você copie a fórmula pela planilha, ainda esteja lendo desvios de linhas a partir da coluna A e linha 1.

É preciso elevar esse cálculo da diferença ao quadrado, somá-lo e tirar sua raiz quadrada para obter a distância euclidiana completa:

```
{=SQRT(SUM((OFFSET(Standardized!$B$2:$K$2,Distances!$A3,0)
-OFFSET(Standardized!$B$2:$K$2,Distances!C$1,0))^2))}
```

Repare que esse cálculo é uma fórmula array devido à diferença de linhas inteiras umas das outras. Você deve pressionar Ctrl+Shift+Enter (Command+Return em um Mac) para funcionar.

A distância euclidiana do funcionário 144624 dele mesmo é, naturalmente, 0. Essa fórmula pode ser copiada e colada por todo OL2. Realce essa série e dê um clique duplo no canto inferior para enviar o cálculo por toda a célula OL402. Isso tem como resultado a Figura 9-9.

É isso aí! Agora você possui um gráfico funcionários-por-funcionários. Você pode exportá-lo para Gephi, como fez no Capítulo 5, e dar uma olhadinha nele, mas já que ele tem 16.000 arestas e somente 400 nós, seria uma bagunça.

Da mesma forma como no Capítulo 5, você construiu um gráfico r-vizinhança a partir da matriz de distância, neste capítulo você focalizará somente nos k vizinhos mais próximos de cada funcionário a fim de encontrar os valores atípicos.

O primeiro passo é classificar a distância de cada funcionário em relação uns aos outros. Essa classificação produzirá o primeiro e o mais básico método para realçar os valores atípicos no gráfico.

	A	B	C	D	E	F	G	H	I	J	K	L	M	
1			0	1	2	3	4	5	6	7	8	9	10	
2			144624	142619	142285	142158	141008	141982	139410	135014	139356	137368	141982	144624
3	0	144624	0.0	4.9	6.7	5.4	5.8	5.0	5.3	5.1	5.5	5.0	5.7	5.0
4	1	142619	4.9	0.0	4.1	2.4	5.1	4.3	4.9	4.6	3.1	4.2	6.2	4.9
5	2	142285	6.7	4.1	0.0	4.9	5.5	3.1	2.6	3.9	5.1	4.2	5.7	6.7
6	3	142158	5.4	2.4	4.9	0.0	5.5	4.9	3.7	4.3	5.1	4.0	5.5	5.0
7	4	141008	5.8	5.1	4.9	5.5	0.0	4.6	3.9	3.0	5.7	3.1	5.0	4.5
8	5	145082	4.9	4.3	4.9	4.6	5.3	0.0	4.6	3.0	5.7	3.1	4.2	4.0
9	6	139410	5.5	3.1	2.6	3.9	4.2	5.3	0.0	5.7	5.1	4.0	6.2	5.0
10	7	135014	5.3	1.7	4.3	3.0	5.7	4.3	5.0	0.0	5.5	4.0	5.7	4.5
11	8	139356	4.1	3.6	5.8	3.7	4.3	4.3	4.0	5.7	0.0	5.1	6.2	4.5
12	9	137368	6.9	3.5	3.7	4.3	4.3	4.3	4.0	5.7	5.1	0.0	6.2	5.0
13	10	141982	4.9	2.0	3.4	3.6	4.5	4.5	4.0	5.7	5.1	5.0	0.0	4.5

Figura 9-9: A matriz de distância do funcionário

## Obtendo os K Vizinhos Mais Próximos

Crie uma nova aba chamada de **Rank**. Cole as IDs do funcionário em A2 até B1 para formar uma grade, como na aba anterior.

Agora você precisa classificar cada funcionário pela maior distância dele para cada funcionário na coluna A. Começando as classificações em 0, para que a classificação 1 irá para **outro** funcionário atual, e todos os 0s ficarão na diagonal do gráfico (devido às distâncias deles mesmos sempre sendo as menores).

Começando em B2, a classificação do funcionário 144624 em relação a ele/a mesmo/a é escrita usando a fórmula RANK:

```
=RANK(Distances!C3, Distances!$C3:$OL3, 1) - 1
```

Esse -1 no final da fórmula dá a essa distância dela mesma uma classificação de 0 em vez de 1. Repare que você trava as colunas C até OL na aba Distances com referências absolutas, permitindo que copie a fórmula para a direita.

Ao copiar a fórmula um para a direita, C2, está classificando o funcionário 142619 em relação à sua distância de 144624:

```
=RANK(Distances!D3, Distances!$C3:$OL3, 1) - 1
```

Isso retorna uma classificação de 194 de um total de 400, então esse pessoal não é exatamente amigo (veja a Figura 9-10).

A	B	C	D	E	F
	144624	142619	142285	142158	141008
2	144624	0	194		
3	142619				
4	142285				
5	142158				
6	141008				
7	145082				
8	139410				
9	135014				
10	139356				
11	137368				
12	141982				
13	144753				
14	132229				

Figura 9-10: O funcionário 142619 classificado pela distância em relação a 144624

Copie essa fórmula por toda a planilha. Você obterá uma classificação completa da matriz exibida na Figura 9-11.

A	B	C	D	E	F	G
	144624	142619	142285	142158	141008	145082
2	144624	0	194	382	279	328
3	142619	367	0	286	29	381
4	142285	389	86	0	206	199
5	142158	360	6	316	0	357
6	141008	350	252	214	313	0
7	145082	317	207	314	275	40
8	139410	387	49	9	154	195
9	135014	374	5	262	56	385
10	139356	203	86	387	107	345
11	137368	393	20	31	103	357
12	141982	382	6	152	195	342
13	144753	367	32	255	51	268
14	132229	232	199	378	283	238

Figura 9-11: Cada funcionário da coluna classificatório em relação a cada linha

Método 1 de Detecção do Valor Atípico em Gráfico: Apenas Use o Indegree

Se você quiser reunir o gráfico dos k vizinhos mais próximos (kNN) usando as planilhas Distances e Rank, tudo o que precisa fazer é apagar todas as arestas na planilha Distances (deixe a célula em branco) cuja classificação tenha sido maior do que k. Para k = 5, você desistiria das distâncias com uma classificação na planilha Rank que fossem 6 ou mais.

O que significaria ser um valor atípico nesse contexto? Bem, um valor atípico não seria escolhido com a frequência de um “vizinho mais próximo”, seria?

Digamos que você tenha criado um gráfico 5NN, e manteve as arestas com uma classificação de 5 ou menos. Se você rolar por uma coluna, tal como a coluna B para o funcionário 144624, quantas vezes esse funcionário acabaria nos top 5 da classificação de todos os outros funcionários? Ou seja, quantos funcionários escolhem 144624 como um dos top cinco vizinhos? Não muitos. Não estou vendo nenhum, na verdade, salvo pela sua própria distância na diagonal com uma classificação de 0, que você pode ignorar.

Que tal se você fizesse um 10NN? Bem, no caso do funcionário 139071 na linha 23, considere 144624 seu nono vizinho mais próximo. Isso significa que no gráfico 5NN o funcionário 144624 possui um indegree (grau de entrada) de 0, enquanto que no gráfico 10NN o funcionário 144624 possui um indegree de 1.

*O indegree é a contagem de números de arestas indo para um nó em um gráfico.* Quanto menor o indegree, maior é o valor atípico, porque nenhum outro quer ser seu vizinho.

No final da coluna B na planilha Rank, conte os indegree para o funcionário 144624 para os casos 5, 10 e 20 dos gráficos dos vizinhos mais próximos. Você pode fazer isso usando uma simples fórmula COUNTIF (subtraindo 1 da própria distância na diagonal que você está ignorando). Então, por exemplo, para contar o indegree para o funcionário 144624 em um gráfico 5NN, você usaria a seguinte fórmula na célula B402:

=COUNTIF(B2:B401, "<=5") -1

Da mesma forma, abaixo dela, você poderia calcular o indegree do funcionário se fizesse um gráfico 10NN:

```
=COUNTIF(B2:B401, "<=10") - 1
```

E abaixo dela um gráfico 20NN:

```
=COUNTIF(B2:B401, "<=20") - 1
```

Na verdade, você poderia escolher qualquer k que quisesse entre 1 e a quantidade de funcionários que tem. Mas pode continuar com 5, 10 e 20 por enquanto. Usando o menu de formatação condicional, você pode realçar as células cujas contagens são 0 (isso significa que não há pontas de entrada para o nó para um gráfico daquele tamanho). Esse cálculo do funcionário 144624 gera a guia exibida na Figura 9-12.

Realçando B402:B404, você pode arrastar os cálculos para a direita por meio da coluna OK. Rolando pelos resultados, você pode ver que alguns funcionários podem ser considerados valores atípicos na marca 5NN mas não necessariamente na marca 10NN (se você definir um valor atípico como um funcionário com um indegree 0 — poderia usar outro número se quisesse).

The screenshot shows a Microsoft Excel spreadsheet titled "SupportCenter.xlsx". The table has columns A through E. Row 1 contains the formula `=COUNTIF(B2:B401, "<=20") - 1`. Rows 2 through 401 contain employee IDs. Row 402 contains the question "How many top 5s?", row 403 contains "How many top 10s?", and row 404 contains "How many top 20s?". Row 405 is blank. The cells B402, C402, and D402 are highlighted in yellow, indicating they are part of a copied formula. The ribbon tabs visible are Home, Layout, Tables, Charts, SmartArt, and three others partially visible. The status bar shows "Normal View" and "Ready".

	A	B	C	D	E
1		144624	142619	142285	142158
393	141467	337	255	281	356
394	132149	363	69	303	144
395	141343	362	159	96	331
396	143981	261	52	268	228
397	139820	395	56	218	71
398	144780	351	17	212	76
399	138420	396	92	55	324
400	131547	356	116	285	263
401	137942	321	243	287	154
402	How many top 5s?	0			
403	How many top 10s?	1			
404	How many top 20s?	1			
405					

**Figura 9-12:** A contagem indegree para três gráficos diferentes de vizinhos mais próximos

Há somente dois funcionários que mesmo no nível do gráfico 20NN ainda não possuem entrada nos vértices. Ninguém os considera nos 20 primeiros vizinhos mais próximos. É bem distante!

Aquelas duas IDs de funcionários são 137155 e 143406. Você pode investigar voltando à aba SupportPersonnel. O funcionário 137155 está na linha 300 (veja a Figura 9-13). Eles têm uma média alta de vendas, classificação alta dos clientes, e parecem ser bons samaritanos. Eles trabalharam milhares de turnos nos fins de semanas e se ofereceram em sete ocasiões para trocar de turno com um funcionário que precisava. Legal! Eles são pessoas que em múltiplas dimensões são excepcionais o suficiente e que não estão nem entre as 20 primeiras distâncias de algum outro funcionário. Isso é sensacional. Talvez esse funcionário mereça um rodízio de pizza ou algo parecido.

The screenshot shows a Microsoft Excel spreadsheet titled "SupportCenter.xlsx". The active sheet is labeled "137155". The table has the following columns and data:

	Employee ID	Avg Tix / Day	Customer rating	Tardies	Graveyard Shifts Taken	Weekend Shifts Taken	Sick Days Taken	% Sick Days Taken on Friday	Employee Dev. Hours	Shift Swaps Requested	Shift Swaps Offered
300	137155	165.3	4.49	1	3	2	1	0%	30	1	7
301	142940	155.7	3.06	1	1	1	6	0%	12	2	1
302	141231	158.2	3.46	2	3	1	1	100%	23	0	0
303	134409	154.1	3.94	0	2	1	2	50%	0	0	0

Figura 9-13: Os dados de desempenho do funcionário 137155

E o outro funcionário — 143406? Ele está na linha 375 e é um contraste interessante com o funcionário anterior (veja a Figura 9-14). Nenhuma métrica por si só é suficiente para demiti-lo, mas, dito isso, seu número de vendas é dois desvios padrões abaixo da média, sua classificação de cliente é provavelmente também uns dois desvios padrões abaixo da distribuição. Seu atraso é acima da média e tirou cinco dos seis dias de licença médica às sextas-feiras. Hmmm.

Esse funcionário participou muito do desenvolvimento de funcionários, e isso é algo positivo. Mas talvez seja porque ele apenas gosta de se destacar nas vendas. Talvez o funcionário deva ser avaliado. Ele pediu

para trocar de turno quatro vezes sem oferecer-se para trocar com ninguém.

Esse funcionário acha que está trabalhando dentro do sistema. Ele possui o mínimo necessário para o emprego (repare que ele não está pulando nenhum Teste de Tukey aqui), ele parece patinar pelo limite fraco de cada distribuição.

	A	B	C	D	E	F	G	H	I	J	K
1	Employee ID	Avg Tix / Day	Customer rating	Tardies	Graveyard Shifts Taken	Weekend Shifts Taken	Sick Days Taken	% Sick Days Taken on Friday	Employee Dev. Hours	Shift Swaps Requested	Shift Swaps Offered
375	143405	145	2.33	3	1	0	6	83%	30	4	0
376	145176	151.7	3.23	2	2	1	2	100%	15	1	1
377	143091	159.3	2.92	1	3	2	0	0%	21	2	4
378	138759	153.4	3.96	1	2	0	0	0%	6	3	3

Figura 9-14: Os dados de desempenho do funcionário 143406

## Método 2 de Detecção do Valor Atípico em Gráfico: Sendo Flexível com a k-Distance

Uma das desvantagens do método anterior é que para um gráfico kNN ou você recebe uma aresta de entrada de alguém ou você não recebe. Isso significa que você obtém grande deslocamento em quem é um valor atípico e quem não é, dependendo do valor de k que você escolher. Esse exemplo acabou tentando em 5, 10 e 20 antes de ser deixado somente com dois funcionários. E, desses dois funcionários, qual deles era o maior valor atípico? Não sei! Ambos tinham indegree de 0 em 20NN, então eles estavam em uma disputa bem acirrada, não é?

O que seria bom era ter o cálculo que atribuía a um funcionário um grau contínuo de atipicidade. Nos dois próximos métodos você perceberá a tentativa de fazer apenas isso. Primeiro, observará a classificação dos valores atípicos usando uma quantidade chamada *k-distance*.

A *k-distance* é a distância de um funcionário para o seu *k*-ésimo vizinho.

Simples e eficaz, mas já que ele está retornando uma distância em vez de uma contagem, pode-se obter uma boa classificação do valor. Crie uma nova aba na pasta de trabalho chamada **K-Distance** para dar uma olhada.

Para k, use 5, o que significa que você pegará a distância de todos até o quinto vizinho mais próximo. Uma forma de pensar sobre isso é se a vizinhança na qual eu vivo tem cinco vizinhos e eu, quanto solo aquela vizinhança ocupa? Se eu tiver que andar 30 minutos para chegar à quinta casa, então eu talvez more na roça.

Nomeie A1 como How many employees are in my neighborhood? e coloque um 5 em B1. Esse é o nosso valor k.

Começando em A3, nomeie a coluna Employee ID e cole as Ids do funcionário. Você começará a calcular a k-distance com aquele funcionário 144624 na célula B4.

Agora, como você calcula a distância entre 144624 e seu quinto vizinho mais próximo? O quinto vizinho mais próximo será classificado como 5 na linha 2 (a linha do 144624) na aba Rank. Portanto, pode usar a instrução IF para configurar aquele valor para 1 em um vetor de 0s, e multiplicar esse vetor pela linha de distância para 144624 na aba Distances. Por fim, some tudo.

Assim, em B4 teremos:

```
{=SUM(IF(Rank!B2:OK2=$B$1, 1, 0)*Distances!C3:OL3)}
```

Observe que o valor k na célula B1 é fechado com referências absolutas, então você pode copiar a fórmula. Além disso, isso é uma fórmula array uma vez que a instrução IF está verificando um array inteiro de valores.

Dê um duplo clique na fórmula, arraste-a pela planilha e aplique uma formatação condicional para realçar as distâncias maiores. Mais uma vez, os dois valores atípicos da seção anterior vão para o topo (veja a Figura 9-15).

	A	B	C	D	E	F	G
1	How many employees are in my neighborhood?	5					
2							
3	Employee ID	k-distance					
373	134999	2.1347712					
374	137910	2.9087565					
375	136944	2.4926029					
376	136145	2.8666611					
377	143406	4.8453539					
378	145176	1.8906392					
379	143091	2.6930362					
380	136750	2.5420207					

Figura 9-15: O funcionário 143406 tem uma distância alta de 5

Desta vez, você obtém um pouco mais de flexibilidade. Pode-se ver nesta lista única que o funcionário fraco, 143406, está substancialmente mais distante do que 137155, e ambos os valores são maiores do que o próximo valor maior de 3,53.

A desvantagem desse método está exibida na Figura 9-16. Meramente usando uma k-distance dá uma sensação global de atipicidade, ou seja, você pode realçar os pontos que estão mais longe dos seus vizinhos do que quaisquer outros pontos. Mas, quando olhar a Figura 9-16, o ponto triangular é claramente o valor atípico, e, ainda, sua k-distance será menor do que os pontos em forma de diamante.

Aqueles diamantes são mais estranhos do que os triângulos? Não para os meus olhos!

A questão aqui é que o triângulo não é um *valor atípico global*, enquanto é um *valor atípico local*. O motivo de os seus olhos pescarem logo o ponto diferente é que ele está próximo ao agrupamento de círculos. Se o triângulo estivesse entre os diamantes espaçados, estaria tudo bem. Mas não. Em vez disso, eles se parecem com vizinhos circulares.

Isso nos leva a uma técnica de alta tecnologia chamada *fatores do valor atípico local* (LOF, do inglês *local outlier factors*).



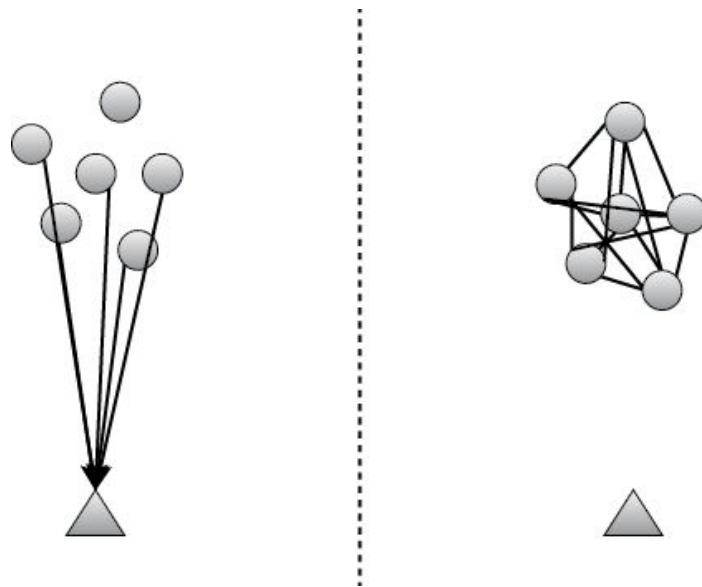
**Figura 9-16:** k-distance falha com os valores atípicos locais

### Método 3 de Detecção do Valor Atípico em Gráfico: Fatores do Valor Atípico Local Estão Onde Estão

Assim como o uso da k-distance, os fatores do valor atípico fornecem um escore único para cada ponto. Quanto maior o escore, mais atípicos os valores são. Mas LOF proporciona algo mais legal do que isso: quanto mais próximo o escore é de 1, mais comum é o ponto localmente. Conforme o escore aumenta, o ponto deve ser considerado menos comum e mais como um valor atípico. E, diferente da k-distance, esse fato “1 é comum” não muda não importa o tamanho ou dimensão do seu gráfico, o que é bem legal.

Em alto nível é assim que funciona: *você é um valor atípico se seus k vizinhos mais próximos considerarem você mais distante do que seus vizinhos os consideram*. O algoritmo se importa com o ponto do amigo e o ponto do amigo do amigo. É assim que ele define “local”.

Voltando à Figura 9-16, é exatamente aqui que um triângulo se torna um valor atípico, não é? Ele pode não ter a k-distance mais alta, mas a proporção da distância do triângulo para seus vizinhos mais próximos sobre a distância um do outro é bastante alta (veja a Figura 9-17).



**Figura 9-17:** O triângulo não é tão alcançável pelos seus vizinhos quanto eles são um do outro

### *Começando com a Distância de Alcance*

Antes de juntar seus fatores de valor atípico local para cada funcionário, você precisa calcular mais alguns conjuntos de números, chamados de *distâncias de alcance (reachability distances)*.

*A distância de alcance do funcionário A com respeito ao funcionário B é a distância normal, a menos que A esteja dentro da vizinhança da k-distance de B, caso em que a distância de alcance é simplesmente a k-distance de B.*

Em outras palavras, se A estiver dentro da vizinhança de B, você arredonda a distância de A para B para o tamanho da vizinhança de B; caso contrário, deixe como está.

Usar a distância de alcance em vez da distância normal para LOF ajuda a estabilizar o cálculo um pouco.

Crie uma nova aba chamada **Reach-dist** e substitua as distâncias da aba Distances pelas novas distâncias de alcance.

A primeira coisa a fazer é colar especial os valores transpostos da aba K-Distance pelo topo da aba, e então colar a grade funcionários por funcionário, como na aba Distances começando na linha 3. Isso gera a planilha vazia exibida na Figura 9-18.

K Distance	2.9826459	1.7983162	2.5526728	2.3543933	3.10512	2.22456
144624						
142619						
142285						
142158						
141008						
145082						
139410						
135014						

**Figura 9-18:** O esqueleto da aba da distância de alcance

Começando pela célula B4, você utilizará a distância de 144624 para ela mesma na aba Distances (`Distances!C3`) a não ser que seja menor do que a k-distance acima de B1. É uma fórmula simples:

```
=MAX(B$1,Distances!C3)
```

A referência absoluta na k-distance permite que você copie a fórmula por toda a planilha. Copiando a fórmula por OK4, pode-se realçar os cálculos na linha 4 e dar um clique duplo para enviá-los por toda a linha 403. Isso preenche todas as distâncias de alcance, como mostra a Figura 9-19.

### *Juntando os Fatores de Valor Atípico Local*

Agora você está pronto para calcular cada fator de valor atípico local de cada funcionário. Para começar, crie uma nova aba chamada `LOF` e cole as IDs dos funcionários na coluna A.

Como vimos antes, os fatores de valor atípico local avaliam como um ponto é visto pelos seus vizinhos versus como seus vizinhos são vistos por seus vizinhos. Se eu estou 30 milhas fora da cidade, meus vizinhos mais próximos podem me ver como um caipira, mas eles podem ser vistos pelos seus vizinhos como membros da comunidade. Isso significa que, localmente, eu sou mais visto como um valor atípico do que meus vizinhos. Você quer capturar esse fenômeno.

Esses valores dependem da média de alcance de cada funcionário em relação aos seus  $k$  vizinhos mais próximos.

	A	B	C	D	E	F	G
1	K Distance	2.9826459	1.7983162	2.5526728	2.3543933	3.10512	2.2245
2							
3		144624	142619	142285	142158	141008	145
4	144624	2.9826459	4.8591492	6.6932738	5.3903329	5.835023	4.884
5	142619	4.8591492	1.7983162	4.0513505	2.3931206	5.0612464	4.2977
6	142285	6.6932738	4.0513505	2.5526728	4.8804977	4.8588879	4.8656
7	142158	5.3903329	2.3931206	4.8804977	2.3543933	5.5303203	4.6458
8	141008	5.835023	5.0612464	4.8588879	5.5303203	3.10512	3.1060
9	145082	4.884647	4.2977011	4.8656301	4.6458954	3.1050262	2.2245
10	139410	5.5084968	3.1323987	2.6351269	3.9318224	4.1619487	3.8718
11	135014	5.2891944	1.7983162	4.2862748	2.954915	5.6902001	4.8473
12	139356	4.1406441	3.5947127	5.8435713	3.727379	5.0802883	4.7081
13	137368	6.8699305	3.4895474	3.7090906	4.316935	5.1635246	6.2932
14	141982	4.9278756	2.0400034	3.3580127	3.6257353	4.488423	3.3905
15	144753	6.072065	3.4595083	5.2336009	3.7026766	5.3056604	5.5693
16	132229	3.9423729	3.8121432	4.9951808	4.1701209	3.9795078	3.518

**Figura 9-19:** Todas as distâncias de alcance

Preste atenção no funcionário 144624 na linha 2. Você já configurou  $k$  para 5, então a pergunta é: qual é a média da distância de alcance de 144624 *em relação aos cinco* vizinhos mais próximos daquele funcionário?

Para calcular isso, retire um vetor de 1s da aba Rank para os cinco funcionários mais próximos de 144624 e de 0s para todos os demais (igual ao que fez na aba K-Distance). Tal vetor pode ser criado usando a fórmula IF para pegar os vizinhos melhores classificados enquanto exclui o funcionário atual:

```
IF(Rank!B2:OK2<='K-Distance'!B$1,1,0)*IF(Rank!B2:OK2>0,1,0)
```

Multiplique esse vetor indicador pelas distâncias de alcance do 144624, some o produto e divida-o por k=5. Na célula B2, você tem:

```
=SUM(IF(Rank!B2:OK2<='K-Distance'!B$1,1,0)*  
IF(Rank!B2:OK2>0,1,0)*  
'Reach-dist'!B4:OK4) / 'K-Distance'!B$1}
```

Assim como calculou a k-distance, essa é uma fórmula de array. É possível enviar essa fórmula pela planilha ao dar um clique duplo nela (veja a Figura 9-20).

Portanto, essa coluna indica como os cinco vizinhos mais próximos de cada funcionário os visualiza.

*O fator do valor atípico local para um funcionário é a média das proporções da média da distância de alcance do funcionário dividido pela média das distâncias de alcance de cada um dos seus k vizinhos.*

	Average Reachability Distance w.r.t.					
1	Employee ID	Neighbors	D	E	F	G
2	144624	2.8063634				
3	142619	1.7937304				
4	142285	2.4495722				
5	142158	2.3786975				
6	141008	3.0674455				
7	145082	2.1982516				
8	139410	2.2267825				
9	135014	1.772618				
10	139356	2.411504				

**Figura 9-20:** A média de alcance para cada funcionário em relação aos seus vizinhos

Você pode lidar com o cálculo LOF para o funcionário 144624 na célula C2 primeiro. Assim como nos cálculos anteriores, a instrução IF a seguir

produz um vetor de 1s para os cinco primeiros vizinhos mais próximos de 144624:

```
IF(Rank!B2:OK2<='K-Distance'!B$1,1,0)*IF(Rank!B2:OK2>0,1,0)
```

Agora multiplique a proporção da média de alcance de 144624 dividido pela média de alcance de cada vizinho assim:

```
IF(Rank!B2:OK2<='K-Distance'!B$1,1,0)
*IF(Rank!B2:OK2>0,1,0)*B2/TRANSPOSE(B$2:B$401)
```

Observe que as distâncias de alcance dos vizinhos referenciados na série B2:B401 na parte de cima da proporção são transpostas para que a coluna seja transformada em uma linha, assim como os vetores gerados da instrução IF na equação.

Você pode fazer a média dessas proporções somando e dividindo por k:

```
{=SUM(IF(Rank!B2:OK2<=
'K-Distance'!B$1,1,0)
*IF(Rank!B2:OK2>0,1,0)
*B2/TRANSPOSE(B$2:B$401))/'K-Distance'!B$1}
```

Repare nas chaves já que essa é uma fórmula de array. Pressione Control+Shift+Enter

(Command+Return no Mac) para ter de volta o fator LOF para 144624.

É 1,34, que de alguma forma está acima do valor de 1, isso significa que esse funcionário tem um pouco de valor atípico local.

É possível enviar essa fórmula pela planilha ao dar um clique duplo e então verificar os outros funcionários. A formatação condicional é útil para realçar os valores atípicos mais significantes.

A surpresa é que, ao rolar para baixo, você descobrirá que o funcionário 143406, o companheiro preguiçoso, é o ponto mais atípico com um LOF de 1,97 (veja a Figura 9-21). Seus vizinhos o veem como duas vezes tão distantes quanto eles são vistos pelos seus vizinhos. É bem fora da comunidade.

The screenshot shows a Microsoft Excel spreadsheet titled "SupportCenter.xlsx". The formula bar contains a complex formula for cell C375:

$$=SUM(IF(Rank!B375:OK375 <= K-Distance!B$1, 1, 0)*IF(Rank!B375:OK375 > 0, 1, 0)*B375 / TRANSPOSE(B$2:B$401)) / K-Distance!B$1$$

The table below has columns A, B, and C. Row 375 is highlighted in grey, showing the value 1.9731194. The data includes:

	A	B	C
372	137910	2.4900125	1.1226723
373	136944	2.2989401	1.0407792
374	136145	2.6712797	1.1712167
375	143406	4.6095274	1.9731194
376	145176	2.1687322	0.9971866
377	143091	2.6925553	1.2049894
378	138759	2.5927569	1.1853399
379	144013	2.4982843	1.0018662
380	138843	2.6052016	1.1437859
381	133915	2.8014458	1.1308304
382	131731	2.4314684	1.076688
383	144570	1.9851454	0.9984509
384	139219	1.8808297	1.0435078
385	140601	2.0646032	1.0344514

**Figura 9-21:** LOFs para os funcionários. Alguém está batendo na porta do 2

E por hoje é só! Agora você tem um único valor atribuído para cada funcionário que os classifica como um valor atípico local e é dimensionado da mesma forma, não importando o tamanho do gráfico. Incrível para caramba.

## Resumindo

Entre o Capítulo da modularidade do gráfico e este de detecção de valor atípico, você foi exposto ao poder de analisar um conjunto de dados ao “representar em gráfico” os seus dados, ou seja, atribuindo distâncias e arestas entre suas observações.

Apesar de ter explorado grupos de pontos relacionados para insights nos capítulos de agrupamento, aqui você minerou os dados para pontos fora das comunidades. Viu o poder de algo tão simples quanto um indegree para demonstrar quem é influente e quem é isolado.

Para mais sobre detecção de valor atípico, verifique pesquisas de 2010 feitas por Kriegel, Kroger e Zimek em

<http://www.siam.org/meetings/sdm10/tutorial3.pdf> — em inglês, para a

conferência SIAM 2010. Todas as técnicas deste capítulo estão lá junto com várias outras.

Observe que essas técnicas não requerem nenhum tipo de processo de longa execução arbitrário da forma que a otimização de modelos requer. Há uma certa quantidade de passos para obter LOFs, então esse é o tipo da coisa que pode ser codificada em produção no topo de um banco de dados facilmente.

Se você está procurando por uma boa linguagem de programação para fazer isso, R é o caminho a seguir. A função bplot em R fornece gráficos de caixa de dados com Teste de Tukey embutidas. A habilidade de representar Teste de Tukey graficamente é tão dolorosa no Excel que eu nem me preocupei em mencionar isso no livro, portanto a função bplot é um ponto a mais para R.

Além do mais, em R, o pacote DMwR (que acompanha o excelente livro *Data Mining with R* by Torgo [Chapman and Hall, 2010]) inclui uma implementação de LOF em uma função chamada *lofactor*. Para construir e analisar o grau de nós em um gráfico, o pacote igraph em Python e R é o melhor caminho.

# 10

# Trocando das Planilhas para R

**D**epois de passar os nove capítulos anteriores injetando Excel diretamente nas suas veias, tenho que dizer pra você deixar tudo para lá. Bem, nem tudo, mas vamos ser honestos, o Excel não é o ideal para as tarefas analíticas.

O Excel é ótimo para aprender análise, porque você pode tocar nos dados e vê-los em cada estado conforme o algoritmo muda da entrada para a saída. Mas você veio, você viu, você aprendeu. Você realmente precisa seguir todos esses passos manualmente todas as vezes? Por exemplo, você realmente precisa cozinhar sua própria formulação de otimização para ajustar suas próprias regressões logísticas? Você precisa inserir as definições das similaridades do cosseno sozinho?

Agora que já aprendeu isso tudo, você está liberado para trapacear e ter alguém para fazer tudo aquilo para você! Imagine se você fosse o Buddy Valastro. Ele cozinha em todos os seus restaurantes? Eu espero que não; caso contrário, suas habilidades variam muito do aeroporto para o mundo real. Agora que já aprendeu bastante, você também deveria sentir-se confortável ao usar a implementação dos algoritmos de outras pessoas.

E é por isso, dentre outras coisas (por exemplo, referenciar uma tabela inteira de dados usando uma palavra) que vale a pena mudar do Excel para a linguagem de programação focada em análise *R*.

Este capítulo percorre as análises dos capítulos anteriores em R em vez do Excel — os mesmos dados, algoritmos, mas um ambiente diferente. Você verá como isso tudo é fácil!

Mas, só um aviso, este capítulo *não é* um tutorial de introdução ao R. Moverei a 1.000 km/hora para acertar alguns poucos algoritmos em um

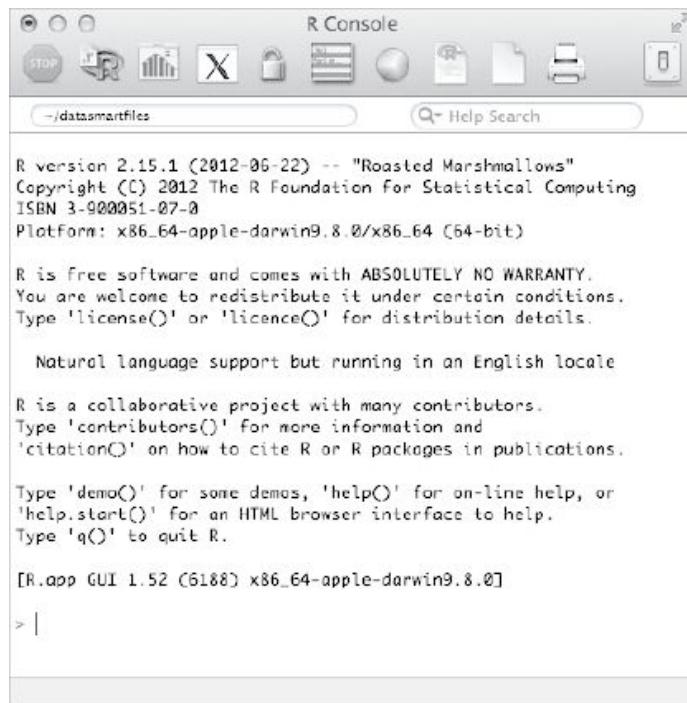
único capítulo. Se quiser uma introdução mais completa, verifique os livros que recomendo no final do capítulo.

E se você não leu os capítulos anteriores, este capítulo não fará o menor sentido, pois estou presumindo que você já está familiarizado com os dados, problemas e técnicas dos capítulos anteriores. Este livro não é um romance “escolha sua aventura”. Leia todo o conteúdo e volte aqui!

## Botando para Funcionar com R

Você pode fazer o download do R em seu website [www.r-project.org](http://www.r-project.org) — em inglês. Clique no link de download, escolha um mirror mais próximo e baixe o instalador para o seu sistema operacional.

Abra o instalador (no Windows, instale o software como administrador) e abra a aplicação. No Windows e no Mac, o console do R vai carregar. Ele é similar à Figura 10-1.



**Figura 10-1:** O console do R no Mac OS

Dentro do console do R, digite comandos no prompt > e pressione Return para que o sistema faça tudo para você. Estes são alguns

exemplos:

```
> print("No regrets. Texas forever.")  
[1] "No regrets. Texas forever."  
> 355/113  
[1] 3.141593
```

Pode-se usar a função `print` para fazer com que o sistema imprima o texto. Pode-se também digitar direto em aritmética para efetuar cálculos. Agora, meu fluxo de trabalho padrão para usar R é:

- 1.** Trazer os dados para R.
- 2.** Fazer coisas data-cientescas com os dados.
- 3.** Jogar os resultados do R onde outra pessoa ou processo possa usá-los.

Quando se trata do primeiro passo, trazer os dados para R, existem várias opções mas, a fim de entender as variáveis e os tipos de dados, você começará simplesmente inserindo os dados manualmente.

## Apenas↑um↑Pouco↑de↑Trabalho↑Manual

A maneira mais fácil de obter dados em R é a mesma do Excel. Digitando com seus dedos e armazenando tais teclas em algum lugar. Pode-se começar armazenando um único valor em uma variável:

```
> almostpi <- 355/113  
> almostpi  
[1] 3.141593  
> sqrt(almostpi)  
[1] 1.772454
```

Nesse pequeno trecho de código, você está armazenando 355/113 em uma variável chamada `almostpi`. É possível digitar a variável de volta ao console e pressionar Return, e então imprimir seu conteúdo. Você pode agir nessa variável com diversas funções (esse exemplo mostrou a raiz quadrada).

Para uma rápida referência das muitas das funções embutidas de R (funções disponíveis sem o pacote de carregamento... algo com o qual você constrói para um objetivo), verifique o cartão de referência de R em <http://cran.r-project.org/doc/contrib/Short-refcard.pdf> — em inglês.

Para entender o que uma função faz, digite um ponto de interrogação antes de colocar no console:

```
> ?sqrt
```

Uma janela Help aparecerá sobre a função (veja a Figura 10-2 para a janela Help sobre sqrt).

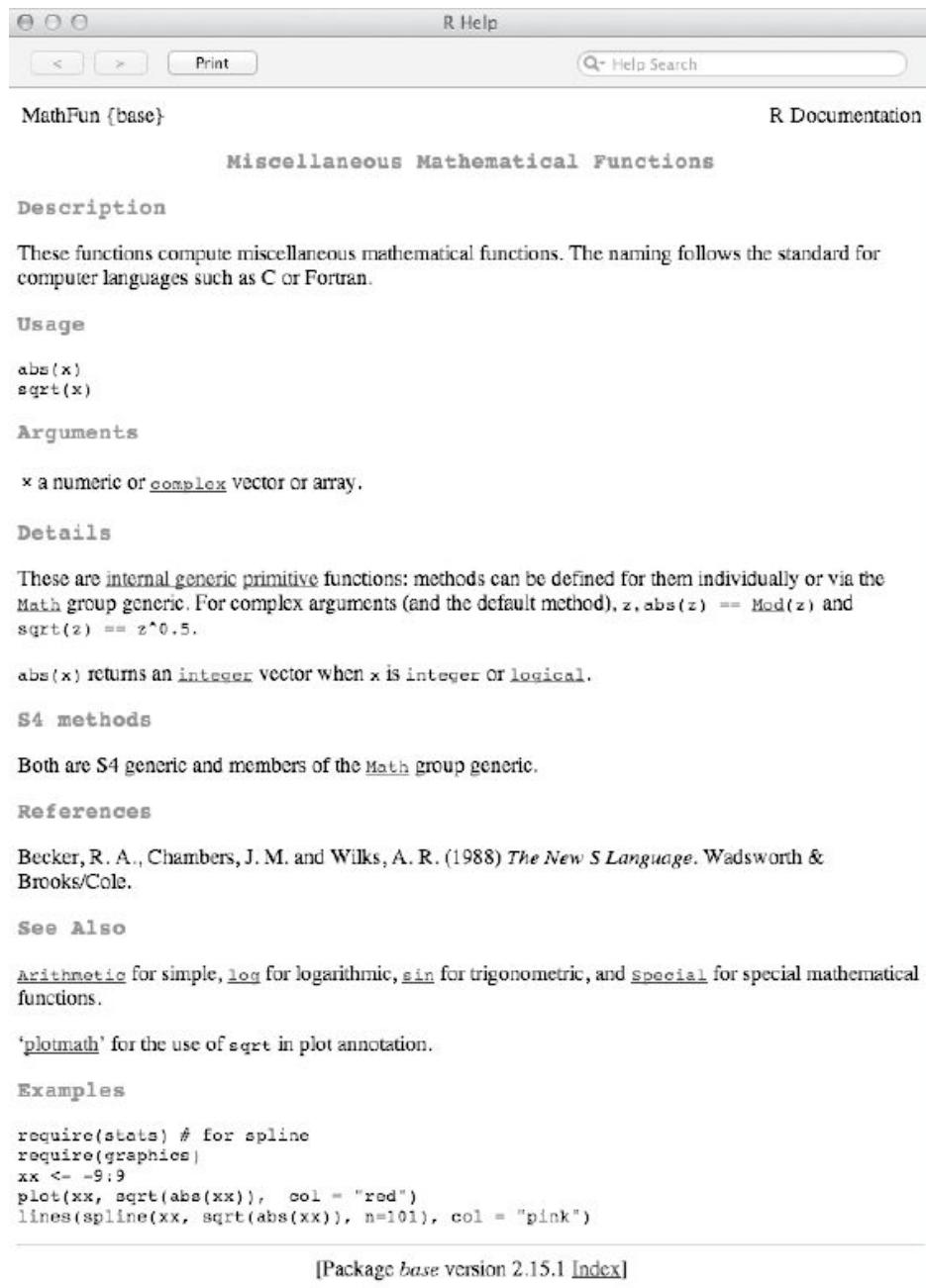
Você também pode digitar dois pontos de interrogação na frente das funções para fazer uma busca por informações, desta forma:

```
> ??log
```

A pesquisa por log tem como resultado a Figura 10-3.

### NOTA

Há todos os tipos de recursos para descobrir quais funções e pacotes estão disponíveis para você em R além do blá blá blá do ?. Por exemplo, [rseek.org](http://rseek.org) é um ótimo mecanismo de pesquisa para conteúdos relacionados a R. Você pode postar perguntas específicas em [stackoverflow.com](http://stackoverflow.com) (veja <http://stackoverflow.com/questions/tagged/r> — em inglês) e a lista



**Figura 10-2:** A janela Help para a função raiz quadrada

## Vetor Matemática e Fatoração

Você pode inserir um vetor de números usando a função `c()` — `c` significa **combine** (combinar). Insira alguns números primos em uma variável:

```
> someprimes <- c(1, 2, 3, 5, 7, 11)
> someprimes
```

```
[1] 1 2 3 5 7 11
```

Help topics matching 'log'			
Topic	Package	Description	
subtreeplot	ape	Zoom on a Portion of a Phylogeny by Successive Clicks	
subtrees	ape	All subtrees of a Phylogenetic Tree	
summary.phylo	ape	Print Summary of a Phylogeny	
vcv	ape	Phylogenetic Variance-covariance or Correlation Matrix	
yule	ape	Fits the Yule Model to a Phylogenetic Tree	
zoom	ape	Zoom on a Portion of a Phylogeny	
log	base	Logarithms and Exponentials	
!	base	Logical Operators	
as.data.frame	base	Coerce to a Data Frame	
logical	base	Logical Vectors	
aml	boot	Remission Times for Acute Myelogenous Leukaemia	
inv.logit	boot	Inverse Logit Function	
logit	boot	Logit of Proportions	
neuro	boot	Neurophysiological Point Process Data	
BloodBrain	caret	Blood Brain Barrier Data	
predictors	caret	List Predictors used in the model	

logical {base}

R Documentation

Logical Vectors

**Description**

Create or test for objects of type "logical", and the basic logical constants.

**Usage**

```
TRUE  
FALSE  
T; F  
  
logical(length = 0)  
as.logical(x, ...)  
is.logical(x)
```

**Figura 10-3:** O resultado da busca pela palavra log

Usando a função `Length()`, você pode contar a quantidade de elementos que o seu vetor possui:

```
> length(someprimes)  
[1] 6
```

Também pode referenciar valores únicos no vetor usando os colchetes:

```
> someprimes[4]  
[1] 5
```

Isso retorna o quarto valor no vetor, que é 5. Você pode fornecer vetores de índices usando a função `c()` ou o caractere `:` para especificar um intervalo:

```
> someprimes[c(4, 5, 6)]  
[1] 5 7 11
```

```
> someprimes[4:6]
[1] 5 7 11
```

Em ambos os casos, você está pegando o quarto e o sexto valores do vetor. Também é possível usar instruções lógicas para retirar valores. Por exemplo, se quisesse todos os números primos menos o sete, você poderia usar a função `which()` para retornar seus índices:

```
> which(someprimes<7)
[1] 1 2 3 4
> someprimes[which(someprimes<7)]
[1] 1 2 3 5
```

Uma vez que tenha colocado os dados em uma variável, você pode efetuar operações no conjunto de dados inteiro e armazenar os resultados em uma nova variável. Por exemplo, você pode multiplicar todos os dados por dois:

```
> primestimes2 <- someprimes*2
> primestimes2
[1] 2 4 6 10 14 22
```

Imagine como fazer isso no Excel. Você insere a fórmula em uma coluna adjacente e a copia. R permite que você nomeie essa coluna ou linha de dados e opere naquela variável como uma entidade única, o que é sensacional.

Uma função útil para verificar seus dados para as entradas confusas é a função `summary`:

```
> summary(someprimes)
Min. 1st Qu. Median Mean 3rd Qu. Max.
1.000 2.250 4.000 4.833 6.500 11.000
```

Você também pode trabalhar com dados de texto:

```
> somecolors <- c("blue", "red", "green", "blue",
  "green", "yellow", "red", "red")
> somecolors
[1] "blue" "red" "green" "blue" "green" "yellow" "red" "red"
```

Se resumir `somecolors`, você obtém dados descritivos:

```
> summary(somecolors)
Length Class Mode
8 character character
```

Mas você pode tratar essas cores como categorias e transformar esse vetor em dados categóricos ao “fatorá-lo”:

```
> somecolors <- factor(somecolors)
> somecolors
[1] blue red green blue green yellow red red
Levels: blue green red yellow
```

Agora quando resumir os dados, você obtém de novo a contagem para cada “nível” (essencialmente, um nível é uma categoria):

```
> summary(somecolors)
blue green red yellow
2 2 3 1
```

### ***Matrizes Bidimensionais***

Os vetores com os quais você brincou até agora são unidimensionais. O mais semelhante a uma planilha em R deve ser uma matriz, um array bidimensional de números. Você pode construir uma com a função `matrix`:

```
> amatrix <- matrix(data=c(someprimes,primestimes2),nrow=2,ncol=6)
> amatrix
[,1] [,2] [,3] [,4] [,5] [,6]
[1,] 1 3 7 2 6 14
[2,] 2 5 11 4 10 22
```

Pode contar colunas e linhas:

```
> nrow(amatrix)
[1] 2
> ncol(amatrix)
[1] 6
```

Se quiser transpor os dados (assim como fez no decorrer do livro usando a funcionalidade de transpor do colar especial do Excel), pode usar a função `t()`:

```
> t(amatrix)
[,1] [,2]
[1,] 1 2
[2,] 3 5
[3,] 7 11
[4,] 2 4
[5,] 6 10
[6,] 14 22
```

Para pegar registros e intervalos individuais você usa os colchetes, a não ser que separe as referências de coluna e linha com uma vírgula:

```
> amatrix[1:2,3]
[1] 7 11
```

Isso retorna as linhas 1 e 2 para a coluna 3. Mas você não precisa referenciar as linhas 1 e 2 já que essas são todas as linhas que você tem — pode deixar a parte do colchete em branco e todas as linhas serão impressas:

```
> amatrix[,3]
[1] 7 11
```

Ao usar as funções `rbind()` e `cbind()`, você pode lançar novas linhas e colunas de dados para dentro da matriz:

```
> primestimes3 <- someprimes*3
> amatrix <- rbind(amatrix,primestimes3)
> amatrix
[,1] [,2] [,3] [,4] [,5] [,6]
1 3 7 2 6 14
2 5 11 4 10 22
primestimes3 3 6 9 15 21 33
```

Até aqui você criou uma nova linha de dados (`primetimes3`) e usou `rbind()` na variável `amatrix` para prender `primetimes3` nela e atribuir o resultado de volta para `amatrix`.

## **O Melhor Tipo de Dados de Todos os Tempos: o Dataframe**

Um dataframe é a forma ideal de trabalhar com o mundo real, um banco de dados estilo tabela em R. Um dataframe em R é uma versão específica do tipo de dados “lista”. Então o que é uma lista? Uma *lista* é uma coleção de objetos em R que pode ser de diferentes tipos. Por exemplo, esta é uma lista com algumas informações sobre uma pessoa:

```
> John <- list(gender="male", age="ancient", height = 72,
  spawn = 3, spawn_ages = c(.5,2,5))
> John
$gender
[1] "male"
$age
[1] "ancient"
$height
[1] 72
$spawn
[1] 3
$spawn_ages
[1] 0.5 2.0 5.0
```

Um dataframe é um tipo de lista que estranhamente se parece com uma planilha do Excel. Essencialmente, é uma planilha de dados orientada por colunas bidimensionais em que as colunas podem ser tratadas como vetores numéricos ou categóricos. Pode-se criar um dataframe ao chamar a função `data.frame()` para arrays dos dados importados ou inseridos manualmente. O exemplo a seguir usa os dados dos filmes do James Bond para demonstrar. Primeiro, crie alguns vetores:

```
> bondnames <-
c("connery", "lazenby", "moore", "dalton", "brosnan", "craig")
> firstyear <- c(1962,1969,1973,1987,1995,2006)
```

```

> eyecolor <- c("brown", "brown", "blue", "green", "blue", "blue")
> womenkissed <- c(17, 3, 20, 4, 12, 4)
> countofbondjamesbonds <- c(3, 2, 10, 2, 5, 1)

```

Então, neste momento, você possui cinco vetores — alguns textuais, alguns numéricos — e todos possuem o mesmo tamanho. Pode-se combiná-los em um único dataframe chamado `bonddata` como este:

```

> bonddata <- data.frame(bondnames, firstyear, eyecolor, womenkissed,
  countofbondjamesbonds)
> bonddata
bondnames firstyear eyecolor womenkissed countofbondjamesbonds
1 connery 1962 brown 17 3
2 lazenby 1969 brown 3 2
3 moore 1973 blue 20 10
4 dalton 1987 green 4 2
5 brosnan 1995 blue 12 5
6 craig 2006 blue 4 1

```

A função `data.frame` tratará de reconhecer quais dessas colunas são fatores e quais são numéricas. É possível ver essa diferença ao chamar as funções `str()` e `summary()` (`str` significa estrutura):

```

> str(bonddata)
'data.frame': 6 obs. of 5 variables:
 $ bondnames : Factor w/ 6 levels "brosnan", "connery", ...
 2 5 6 4 1 3
 $ firstyear : num 1962 1969 1973 1987 1995 ...
 $ eyecolor : Factor w/ 3 levels "blue", "brown", ...
 2 2 1 3 1 1
 $ womenkissed : num 17 3 20 4 12 4
 $ countofbondjamesbonds: num 3 2 10 2 5 1
> summary(bonddata)
bondnames firstyear eyecolor womenkissed countofbondjamesbonds
brosnan:1 Min. :1962 blue :3 Min. : 3.00 Min. : 1.000

```

```
connery:1 1st Qu.:1970 brown:2 1st Qu.: 4.00 1st Qu.: 2.000
craig :1 Median :1980 green:1 Median : 8.00 Median : 2.500
dalton :1 Mean :1982 Mean :10.00 Mean : 3.833
lazenby:1 3rd Qu.:1993 3rd Qu.:15.75 3rd Qu.: 4.500
moore :1 Max. :2006 Max. :20.00 Max. :10.000
```

Rpare que o ano está sendo tratado como um número. Você poderia fatorizar essa coluna usando a função `factor()` se quisesse tratá-la categoricamente.

Uma das coisas incríveis sobre os dataframes é que você pode referenciar cada coluna usando o caractere \$ mais o nome da coluna, assim:

```
> bonddata$firstyear <- factor(bonddata$firstyear)
> summary(bonddata)

bondnames firstyear eyecolor womenkissed countofbondjamesbonds
brosnan:1 1962:1 blue :3 Min. : 3.00 Min. : 1.000
connery:1 1969:1 brown:2 1st Qu.: 4.00 1st Qu.: 2.000
craig :1 1973:1 green:1 Median : 8.00 Median : 2.500
dalton :1 1987:1 Mean :10.00 Mean : 3.833
lazenby:1 1995:1 3rd Qu.:15.75 3rd Qu.: 4.500
moore :1 2006:1 Max. :20.00 Max. :10.000
```

Portanto, quando você executa a função `summary`, os anos se enrolam em contagens de categorias em vez de dados de distribuição. Além disso, lembre-se que sempre que transpor um dataframe, o resultado será uma matriz bidimensional em vez de outro dataframe. Isso faz sentido, já que a versão transposta dos dados Bond não teriam tipos de dados consistentes em cada coluna.

## Lendo↑dados↑em↑R

## NOTA

O arquivo CSV utilizado nesta seção, “WineKMC.csv”, está disponível para download na página da editora em [www.altabooks.com.br](http://www.altabooks.com.br), procurando pelo título do livro.

Certo, então você aprendeu a introduzir os dados em vários tipos de dados manualmente, mas como ler os dados dentro dos arquivos? A primeira coisa que precisa entender é o ***diretório de trabalho***. Ele é uma pasta na qual você pode colocar os dados para que o console de R possa encontrá-los e lê-los. A função `getwd()` exibe o diretório de trabalho a seguir:

```
> getwd()  
[1] "/Users/johnforeman/RHOME"
```

Se não gostar do diretório de trabalho atual, você pode mudá-lo com o comando `setwd()`. Lembre-se que, mesmo nas máquinas com Windows, o R espera que os caminhos dos diretórios sejam especificados com barras. Por exemplo:

```
> setwd("/Users/johnforeman/datasmartfiles")
```

Use esse comando para configurar seu diretório de trabalho para uma posição em que você fique feliz em introduzir alguns dados. Comece substituindo o arquivo WineKMC.csv baixado naquele diretório. Esse arquivo delimitado por vírgula possui dados da aba Matrix no agrupamento k-means da pasta de trabalho no Capítulo 2. Leia e observe.

Para ler os dados, use a função `read.csv()`:

```
> winedata <- read.csv("WineKMC.csv")
```

Esses dados devem se parecer com os da aba Matrix do Capítulo 2, para quando você imprimir as primeiras poucas colunas (escolhi nove para caberem nesta página), vir os dados descritivos sobre cada uma das 32 ofertas seguidas de alguns vetores de cliques dos clientes em colunas:

```
> winedata[,1:9]
```

Offer Mth Varietal MinQty Disc Origin PastPeak Adams Allen

1 1 Jan Malbec 72 56 France FALSE NA NA

2 2 Jan Pinot Noir 72 17 France FALSE NA NA

3 3 Feb Espumante 144 32 Oregon TRUE NA NA

4 4 Feb Champagne 72 48 France TRUE NA NA

5 5 Feb Cab. Sauv. 144 44 NZ TRUE NA NA

6 6 Mar Prosecco 144 86 Chile FALSE NA NA

7 7 Mar Prosecco 6 40 Australia TRUE NA NA

8 8 Mar Espumante 6 45 S. Africa FALSE NA NA

9 9 Apr Chardonnay 144 57 Chile FALSE NA 1

10 10 Apr Prosecco 72 52 CA FALSE NA NA

11 11 May Champagne 72 85 France FALSE NA NA

12 12 May Prosecco 72 83 Australia FALSE NA NA

13 13 May Merlot 6 43 Chile FALSE NA NA

14 14 Jun Merlot 72 64 Chile FALSE NA NA

15 15 Jun Cab. Sauv. 144 19 Italy FALSE NA NA

16 16 Jun Merlot 72 88 CA FALSE NA NA

17 17 Jul Pinot Noir 12 47 Germany FALSE NA NA

18 18 Jul Espumante 6 50 Oregon FALSE 1 NA

19 19 Jul Champagne 12 66 Germany FALSE NA NA

20 20 Aug Cab. Sauv. 72 82 Italy FALSE NA NA

21 21 Aug Champagne 12 50 CA FALSE NA NA

22 22 Aug Champagne 72 63 France FALSE NA NA

23 23 Sept Chardonnay 144 39 S. Africa FALSE NA NA

24 24 Sept Pinot Noir 6 34 Italy FALSE NA NA

25 25 Oct Cab. Sauv. 72 59 Oregon TRUE NA NA

26 26 Oct Pinot Noir 144 83 Australia FALSE NA NA

27 27 Oct Champagne 72 88 NZ FALSE NA 1

28 28 Nov Cab. Sauv. 12 56 France TRUE NA NA

29 29 Nov P. Grigio 6 87 France FALSE 1 NA

30 30 Dec Malbec 6 54 France FALSE 1 NA

31 31 Dec Champagne 72 89 France FALSE NA NA

```
32 32 Dec Cab. Sauv. 72 45 Germany TRUE NA NA
```

Está tudo aí! Porém, você perceberá que os espaços em branco nos vetores de compra (o Excel os trata como zeros) se tornaram valores NA. Você precisa transformar esses valores NA em 0 usando a função `is.na()` dentro dos colchetes:

```
> winedata[is.na(winedata)] <- 0  
> winedata[1:10,8:17]  
Adams Allen Anders Bailey Baker Barnes Bell Bennett Brooks Brown  
1 0 0 0 0 0 0 0 0 0 0  
2 0 0 0 0 0 0 1 0 0 0  
3 0 0 0 0 0 0 0 0 1 0  
4 0 0 0 0 0 0 0 0 0 0  
5 0 0 0 0 0 0 0 0 0 0  
6 0 0 0 0 0 0 0 0 0 0  
7 0 0 0 1 1 0 0 0 0 1  
8 0 0 0 0 0 0 0 1 1 0  
9 0 1 0 0 0 0 0 0 0 0  
10 0 0 0 0 1 1 0 0 0 0
```

Bum! NA se torna 0.

## Praticando um Pouco de Data Science Real

Até este momento você aprendeu como trabalhar com variáveis e tipos de dados, dados inseridos manualmente, e ler dados a partir de um CSV. Mas como usar de fato os algoritmos vistos anteriormente neste livro? Já que você possui os dados dos vinhos carregados, começaremos com um pouco de agrupamento k-means.

## K-Means Esférico nos Dados dos Vinhos em Poucas Linhas

Nesta seção, você fará agrupamentos baseados em similaridade do cosseno (*também chamado de k-means esférico*). E, em R, há um pacote de k-means esférico que você pode carregar, chamado `skmeans`. Mas ele não vem pronto para usar em R; ele foi escrito por terceiros como um pacote que você pode carregar em R e usar. Essencialmente, esses gênios fizeram todo o trabalho, você só tem que apoiá-los.

Como a maioria dos pacotes de R, você pode lê-los e instalá-los a partir da rede de distribuição do R, o CRAN (do inglês **Comprehensive R Archive Network**). O CRAN é um repositório de muitos dos pacotes úteis que podem ser carregados em R para ampliar a sua funcionalidade. Uma lista de todos os pacotes que podem ser baixados do CRAN está disponível aqui: <http://cran.r-project.org/web/packages/> — em inglês.

Procure por “spherical k-means” em `rseek.org` e um PDF com explicação em inglês sobre o pacote aparece logo no primeiro resultado. Há uma função chamada `skmeans()` que você precisa.

R é inicialmente configurado para baixar pacotes do CRAN, então para obter o pacote `skmeans` você só precisa instalar a função `install.packages()` (R talvez peça para configurar uma biblioteca pessoal na primeira vez que fizer isso):

```
> install.packages("skmeans", dependencies = TRUE)
trying URL 'http://mirrors.nics.utk.edu/cran/bin/macosx/leopard/
contrib/2.15/skmeans_0.2-3.tgz'
Content type 'application/x-gzip' length 224708 bytes (219 Kb)
opened URL
=====
downloaded 219 Kb

The downloaded binary packages are in
/var/.../downloaded_packages
```

Você pode ver no código que configurei `dependencies = TRUE` na chamada de instalação. Isso assegura que o pacote `skmeans` é dependente de qualquer outro pacote, o R também baixa tais pacotes. A chamada

baixa o pacote adequado para minha instalação do R (versão 2.15 no Mac) a partir de um mirror e o coloca no lugar certo.

Você pode então carregar o pacote usando a função `library()`:

```
> library(skmeans)
```

Você pode procurar como usar a função `skmeans()` usando a chamada `?skmeans`. A documentação especifica que `skmeans()` aceita uma matriz em que cada linha corresponda a um objeto do agrupamento.

Seus dados, por outro lado, são orientados por colunas com um monte de descritores de negociações que o algoritmo não vai querer ver. Então é preciso transpõe-los (repare que a função `transposta` leva a matriz para fora do `dataframe`).

Ao usar a função `ncol()`, pode-se ver que as colunas do cliente vão para a coluna 107, então você pode isolar os vetores de compra como linhas para cada cliente ao transpor os dados da coluna 8 para a 107 e introduzindo nela uma nova variável chamada `winedata.transposed`:

```
> ncol(winedata)
[1] 107
> winedata.transposed <- t(winedata[, 8:107])
> winedata.transposed[1:10, 1:10]
 [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
 Adams 0 0 0 0 0 0 0 0 0 0
 Allen 0 0 0 0 0 0 0 0 1 0
 Anders 0 0 0 0 0 0 0 0 0 0
 Bailey 0 0 0 0 0 0 1 0 0 0
 Baker 0 0 0 0 0 0 1 0 0 1
 Barnes 0 0 0 0 0 0 0 0 0 1
 Bell 0 1 0 0 0 0 0 0 0 0
 Bennett 0 0 0 0 0 0 0 1 0 0
 Brooks 0 0 1 0 0 0 0 1 0 0
 Brown 0 0 0 0 0 0 1 0 0 0
```

Depois você pode chamar `skmeans` para o conjunto de dados, especificando cinco médias e o uso do algoritmo genético (bem parecido com o algoritmo que você usou no Excel). Você atribuirá os resultados de volta ao objeto chamado `winedata.clusters`:

```
> winedata.clusters <- skmeans(winedata.transposed, 5,  
method="genetic")
```

Digitando o objeto de volta ao console, pode obter um resumo do seu conteúdo (seus resultados podem variar devido à otimização do algoritmo):

```
> winedata.clusters  
A hard spherical k-means partition of 100 objects into 5 classes.  
Class sizes: 16, 17, 15, 29, 23  
Call: skmeans(x = winedata.transposed, k = 5, method = "genetic")
```

Ao chamar `str()` para o objeto dos agrupamentos mostra que as atribuições de agrupamento atuais estão armazenadas dentro da lista de “agrupamento” do objeto:

```
> str(winedata.clusters)  
List of 7  
 $ prototypes: num [1:5, 1:32] 0.09 0.153 0 0.141 0 ...  
 ..- attr(*, "dimnames")=List of 2  
 ... .$. : chr [1:5] "1" "2" "3" "4" ...  
 ... .$. : NULL  
 $ membership: NULL  
 $ cluster : int [1:100] 5 4 1 5 2 2 1 3 3 5 ...  
 $ family :List of 7  
 ...$. description: chr "spherical k-means"  
 ...$. D :function (x, prototypes)  
 ...$. C :function (x, weights, control)  
 ...$. init :function (x, k)  
 ...$. e : num 1  
 ...$. .modify : NULL  
 ...$. .subset : NULL
```

```
... - attr(*, "class")= chr "pclust_family"
$ m : num 1
$ value : num 38
$ call : language skmeans(x = winedata.transposed,
k = 5, method = "genetic")
- attr(*, "class")= chr [1:2] "skmeans" "pclust"
```

Então, por exemplo, se você quisesse retirar a atribuição do agrupamento da linha 4, você usaria a notação da matriz no vetor de agrupamento:

```
> winedata.clusters$cluster[4]
[1] 5
```

Agora, cada vetor está nomeado com o nome do cliente (porque eles foram nomeados quando você os leu com a função `read.csv()`), então também pode retirar as atribuições por nome usando a função `row.names()` combinada com a função `which()`:

```
> winedata.clusters$cluster[
  which(row.names(winedata.transposed) == "Wright")
]
[1] 4
```

Legal! Além disso, pode escrever todas essas atribuições de agrupamentos usando a função `write.csv()` se quisesse. Use `?`  para aprender como usá-la. Spoiler: é igual a `read.csv()`.

Agora, a principal forma de entender os agrupamentos no Excel era entendendo os padrões dos descritores das negociações que os define. Você contava todas as negociações tomadas de cada agrupamentos e ordenava. Como fazer algo parecido em R?

Para efetuar as contagens, use a função `aggregate()` e no campo “by” (por) especifique as atribuições do agrupamento — significando “agregar compras por atribuição”. Você também precisa especificar que o tipo de agregação que quer é uma soma em contraposição a uma média, mínima, máxima, mediana e assim por diante:

```
aggregate(winedata.transposed, by=list(winedata.clusters$cluster), sum)
```

Você usará transpor para armazenar essas contagens de volta em cinco colunas (assim como eram no Excel) e cortará a primeira linha da agregação, o que devolve os nomes da atribuição do agrupamento. Então, armazene tudo isso de volta como uma variável chamada

winedata.clustercounts:

```
> winedata.clustercounts <- t(aggregate(winedata.transposed, by=list(winedata.clusters$cluster), sum) [,2:33])  
> winedata.clustercounts  
[,1] [,2] [,3] [,4] [,5]  
V1 2 5 0 3 0  
V2 7 3 0 0 0  
V3 0 2 3 0 1  
V4 0 5 1 6 0  
V5 0 0 0 4 0  
V6 0 8 1 3 0  
V7 0 3 1 0 15  
V8 0 1 15 0 4  
V9 0 2 0 8 0  
V10 1 4 1 0 1  
V11 0 7 1 4 1  
V12 1 3 0 0 1  
V13 0 0 2 0 4  
V14 0 3 0 6 0  
V15 0 3 0 3 0  
V16 1 1 0 3 0  
V17 7 0 0 0 0  
V18 0 1 4 0 9  
V19 0 4 1 0 0  
V20 0 2 0 4 0  
V21 0 1 1 1 1
```

```
V22 0 17 2 2 0  
V23 1 1 0 3 0  
V24 12 0 0 0 0  
V25 0 3 0 3 0  
V26 12 0 0 3 0  
V27 1 4 1 3 0  
V28 0 5 0 0 1  
V29 0 1 4 0 12  
V30 0 4 4 1 13  
V31 0 16 1 0 0  
V32 0 2 0 2 0
```

Tudo bem, então estas são suas contagens de negociações por agrupamento. Vamos arremessar essas sete colunas de dados descritivos nas negociações usando a função de ligamento de coluna `cbind()`:

```
> winedata.desc.plus.counts <-  
  cbind(winedata[,1:7],winedata.clustercounts)  
> winedata.desc.plus.counts  
Offer Mth Varietal MinQty Disc Origin PastPeak 1 2 3 4 5  
V1 1 Jan Malbec 72 56 France FALSE 2 5 0 3 0  
V2 2 Jan Pinot Noir 72 17 France FALSE 7 3 0 0 0  
V3 3 Feb Espumante 144 32 Oregon TRUE 0 2 3 0 1  
V4 4 Feb Champagne 72 48 France TRUE 0 5 1 6 0  
V5 5 Feb Cab. Sauv. 144 44 NZ TRUE 0 0 0 4 0  
V6 6 Mar Prosecco 144 86 Chile FALSE 0 8 1 3 0  
V7 7 Mar Prosecco 6 40 Australia TRUE 0 3 1 0 15  
V8 8 Mar Espumante 6 45 S. Africa FALSE 0 1 15 0 4  
V9 9 Apr Chardonnay 144 57 Chile FALSE 0 2 0 8 0  
V1 10 Apr Prosecco 72 52 CA FALSE 1 4 1 0 1  
V1 11 May Champagne 72 85 France FALSE 0 7 1 4 1  
V12 12 May Prosecco 72 83 Australia FALSE 1 3 0 0 1  
V13 13 May Merlot 6 43 Chile FALSE 0 0 2 0 4  
V14 14 Jun Merlot 72 64 Chile FALSE 0 3 0 6 0
```

```

V15 15 Jun Cab. Sauv. 144 19 Italy FALSE 0 3 0 3 0
V16 16 Jun Merlot 72 88 CA FALSE 1 1 0 3 0
V17 17 Jul Pinot Noir 12 47 Germany FALSE 7 0 0 0 0
V18 18 Jul Espumante 6 50 Oregon FALSE 0 1 4 0 9
V19 19 Jul Champagne 12 66 Germany FALSE 0 4 1 0 0
V20 20 Aug Cab. Sauv. 72 82 Italy FALSE 0 2 0 4 0
V21 21 Aug Champagne 12 50 CA FALSE 0 1 1 1 1
V22 22 Aug Champagne 72 63 France FALSE 0 17 2 2 0
V23 23 Sept Chardonnay 144 39 S. Africa FALSE 1 1 0 3 0
V24 24 Sept Pinot Noir 6 34 Italy FALSE 12 0 0 0 0
V25 25 Oct Cab. Sauv. 72 59 Oregon TRUE 0 3 0 3 0
V26 26 Oct Pinot Noir 144 83 Australia FALSE 12 0 0 3 0
V27 27 Oct Champagne 72 88 NZ FALSE 1 4 1 3 0
V28 28 Nov Cab. Sauv. 12 56 France TRUE 0 5 0 0 1
V29 29 Nov P. Grigio 6 87 France FALSE 0 1 4 0 12
V30 30 Dec Malbec 6 54 France FALSE 0 4 4 1 13
V31 31 Dec Champagne 72 89 France FALSE 0 16 1 0 0
V32 32 Dec Cab. Sauv. 72 45 Germany TRUE 0 2 0 2 0

```

É possível ordenar usando a função `order()` dentro dos colchetes do dataframe. Esta é uma ordenação para descobrir as negociações mais populares para o agrupamento 1 (repare que coloquei um sinal de menos na frente dos dados para uma ordenação decrescente. Como alternativa, você pode configurar a flag `decreasing=TRUE` na função `order()`):

```

> winedata.desc.plus.counts[order(-
+ winedata.desc.plus.counts[,8]),]
Offer Mth Varietal MinQty Disc Origin PastPeak 1 2 3 4 5
V24 24 Sept Pinot Noir 6 34 Italy FALSE 12 0 0 0 0
V26 26 Oct Pinot Noir 144 83 Australia FALSE 12 0 0 3 0
V2 2 Jan Pinot Noir 72 17 France FALSE 7 3 0 0 0
V17 17 Jul Pinot Noir 12 47 Germany FALSE 7 0 0 0 0
V1 1 Jan Malbec 72 56 France FALSE 2 5 0 3 0
V10 10 Apr Prosecco 72 52 CA FALSE 1 4 1 0 1

```

```

V12 12 May Prosecco 72 83 Australia FALSE 1 3 0 0 1
V16 16 Jun Merlot 72 88 CA FALSE 1 1 0 3 0
V23 23 Sept Chardonnay 144 39 S. Africa FALSE 1 1 0 3 0
V27 27 Oct Champagne 72 88 NZ FALSE 1 4 1 3 0
V3 3 Feb Espumante 144 32 Oregon TRUE 0 2 3 0 1
V4 4 Feb Champagne 72 48 France TRUE 0 5 1 6 0
V5 5 Feb Cab. Sauv. 144 44 NZ TRUE 0 0 0 4 0
V6 6 Mar Prosecco 144 86 Chile FALSE 0 8 1 3 0
V7 7 Mar Prosecco 6 40 Australia TRUE 0 3 1 0 15
V8 8 Mar Espumante 6 45 S. Africa FALSE 0 1 15 0 4
V9 9 Apr Chardonnay 144 57 Chile FALSE 0 2 0 8 0
V11 11 May Champagne 72 85 France FALSE 0 7 1 4 1
V13 13 May Merlot 6 43 Chile FALSE 0 0 2 0 4
V14 14 Jun Merlot 72 64 Chile FALSE 0 3 0 6 0
V15 15 Jun Cab. Sauv. 144 19 Italy FALSE 0 3 0 3 0
V18 18 Jul Espumante 6 50 Oregon FALSE 0 1 4 0 9
V19 19 Jul Champagne 12 66 Germany FALSE 0 4 1 0 0
V20 20 Aug Cab. Sauv. 72 82 Italy FALSE 0 2 0 4 0
V21 21 Aug Champagne 12 50 CA FALSE 0 1 1 1 1
V22 22 Aug Champagne 72 63 France FALSE 0 17 2 2 0
V25 25 Oct Cab. Sauv. 72 59 Oregon TRUE 0 3 0 3 0
V28 28 Nov Cab. Sauv. 12 56 France TRUE 0 5 0 0 1
V29 29 Nov P. Grigio 6 87 France FALSE 0 1 4 0 12
V30 30 Dec Malbec 6 54 France FALSE 0 4 4 1 13
V31 31 Dec Champagne 72 89 France FALSE 0 16 1 0 0
V32 32 Dec Cab. Sauv. 72 45 Germany TRUE 0 2 0 2 0

```

Observando as primeiras negociações, fica claro que o agrupamento 1 é o agrupamento de Pinot Noir. (Sua classificação pode variar. O algoritmo genético não dá a mesma resposta todas as vezes.)

Apenas para reforçar, se você retirar toda a minha gestão, o código R a seguir replica parte do Capítulo 2 deste livro:

```
> setwd("/Users/johnforeman/datasmartfiles")
```

```

> winedata <- read.csv("WineKMC.csv")
> winedata[is.na(winedata)] <- 0
> install.packages("skmeans", dependencies = TRUE)
> library(skmeans)
> winedata.transposed <- t(winedata[,8:107])
> winedata.clusters <- skmeans(winedata.transposed, 5,
method="genetic")
> winedata.clustercounts <-
t(aggregate(winedata.transposed,
by=list(winedata.clusters$cluster),sum) [,2:33])
> winedata.desc.plus.counts <-
cbind(winedata[,1:7],winedata.clustercounts)
> winedata.desc.plus.counts[order(-
winedata.desc.plus.counts[,8]),]

```

E é isso — desde a leitura dos dados até a análise dos agrupamentos. Muito louco! E isso é porque a chamada para `skmeans()` basicamente isola toda a complexidade desse método para longe de você. Terrível de aprender mas magnífico para trabalhar.

## Construindo Modelos IA com os Dados de Gravidez

### NOTA

Os arquivos CSV usados nesta seção, “Pregnancy.csv” e “Pregnancy\_Test.csv”, estão disponíveis para download na página da editora, [www.altabooks.com.br](http://www.altabooks.com.br), procurando pelo título do livro.

Nesta seção, você replicará alguns modelos de previsão de gravidez que construiu nos Capítulos 6 e 7 deste livro. Especificamente, você construirá dois classificadores usando a função `glm()` (modelo linear geral) com uma função de ligação logística e usando a função

`randomForest()` (árvores de multiconjunto `randomForest()` podem estar em qualquer lugar desde tocos simples a árvores de decisão inteiras).

Os dados em treinamento e em teste estão separados em dois arquivos CSV, chamados `Pregnancy.csv` e `Pregnancy_Test.csv`. Vá em frente e salve-os no diretório de trabalho e então carregue-os em dois dataframes:

```
> PregnancyData <- read.csv("Pregnancy.csv")
> PregnancyData.Test <- read.csv("Pregnancy_Test.csv")
```

Você pode executar `summary()` e `str()` nos dados para ter uma ideia. É imediatamente aparente que os tipos de dados de gênero e o tipo de endereço carregaram como dados categóricos, mas, como pode ver na saída do `str()`, a variável responsável (1 para grávida, 0 para não-grávida) foi tratada como numérica em vez de duas classes distintas:

```
> str(PregnancyData)
'data.frame': 1000 obs. of 18 variables:
 $ Implied.Gender : Factor w/ 3 levels "F", "M", "U": 2 2 2 3 1 ...
 $ Home.Apt..PO.Box : Factor w/ 3 levels "A", "H", "P": 1 2 2 2 1 ...
 $ Pregnancy.Test : int 1 1 1 0 0 0 0 0 0 0 ...
 $ Birth.Control : int 0 0 0 0 0 0 1 0 0 0 ...
 $ Feminine.Hygiene : int 0 0 0 0 0 0 0 0 0 0 ...
 $ Folic.Acid : int 0 0 0 0 0 0 1 0 0 0 ...
 $ Prenatal.Vitamins : int 1 1 0 0 0 1 1 0 0 1 ...
 $ Prenatal.Yoga : int 0 0 0 0 1 0 0 0 0 0 ...
 $ Body.Pillow : int 0 0 0 0 0 0 0 0 0 0 ...
 $ Ginger.Ale : int 0 0 0 1 0 0 0 0 1 0 ...
 $ Sea.Bands : int 0 0 1 0 0 0 0 0 0 0 ...
 $ Stopped.buying.ciggies: int 0 0 0 0 0 1 0 0 0 0 ...
 $ Cigarettes : int 0 0 0 0 0 0 0 0 0 0 ...
 $ Smoking.Cessation : int 0 0 0 0 0 0 0 0 0 0 ...
 $ Stopped.buying.wine : int 0 0 0 0 1 0 0 0 0 0 ...
 $ Wine : int 0 0 0 0 0 0 0 0 0 0 ...
 $ Maternity.Clothes : int 0 0 0 0 0 0 0 0 1 0 1 ...
```

```
$ PREGNANT : int 1 1 1 1 1 1 1 1 1 1 ...
```

É melhor para `randomForest()` que você fatorize essa variável responsável em duas classes (classe 0 e classe 1) em vez de tratar os dados como um inteiro. Você pode fatorizar os dados desta forma:

```
PregnancyData$PREGNANT <- factor(PregnancyData$PREGNANT)  
PregnancyData.Test$PREGNANT <- factor(PregnancyData.Test$PREGNANT)
```

Agora, se você resumir a coluna `PREGNANT`, recebe de volta contagens de classe como se 0 e 1 fossem categorias:

```
> summary(PregnancyData$PREGNANT)  
0 1  
500 500
```

Para construir uma regressão logística, você precisa da função `glm()` que está embutida no pacote de estatística para R. Mas para a função `randomForest()`, você precisará do pacote `randomForest`. Além disso, seria bom construir as curvas ROC vistas nos Capítulos 6 e 7. Há um pacote especificamente construído para gerar tais gráficos, chamado `ROCR`. Vá em frente e instale e carregue estes dois bem rápido:

```
> install.packages("randomForest", dependencies=TRUE)  
> install.packages("ROCR", dependencies=TRUE)  
> library(randomForest)  
> library(ROCR)
```

Agora os dados e os pacotes foram baixados. É hora de começar a construir o modelo! Comece com uma regressão logística:

```
> Pregnancy.lm <- glm(PREGNANT ~ .,  
data=PregnancyData, family=binomial("logit"))
```

A função `glm()` constrói o modelo linear que foi especificada como regressão logística usando a opção `family=binomial("logit")`. Você fornece dados para a função usando o campo `data=PregnancyData`. Agora você deve estar pensando o que `PREGNANT ~ .` significa. Ela é uma **fórmula** em R. Significa “treine meu modelo para prever a coluna `PREGNANT` usando todas as outras colunas”. O `~` significa “usando” e o ponto

significa “todas as outras colunas”. Você pode especificar um conjunto de colunas também ao digitar seus nomes:

```
> Pregnancy.lm <- glm(PREGNANT ~  
  Implied.Gender +  
  Home.Apt..PO.Box +  
  Pregnancy.Test +  
  Birth.Control,  
  data=PregnancyData, family=binomial("logit"))
```

Você está usando a notação PREGNANT ~ . porque você quer usar todas as colunas para treinar o modelo.

Uma vez que o modelo linear esteja construído, você pode visualizar os coeficientes e analisar quais variáveis são significantes estatisticamente (similar aos testes t conduzidos no Capítulo 6) ao resumir o modelo:

```
> summary(Pregnancy.lm)  
  
Call:  
  
glm(formula = PREGNANT ~ ., family = binomial("logit"),  
  data = PregnancyData)  
  
Deviance Residuals:  
  
Min 1Q Median 3Q Max  
-3.2012 -0.5566 -0.0246 0.5127 2.8658  
  
Coefficients:  
  
Estimate Std. Error z value Pr(>|z|)  
(Intercept) -0.343597 0.180755 -1.901 0.057315 .  
Implied.GenderM -0.453880 0.197566 -2.297 0.021599 *  
Implied.GenderU 0.141939 0.307588 0.461 0.644469  
Home.Apt..PO.BoxH -0.172927 0.194591 -0.889 0.374180  
Home.Apt..PO.BoxP -0.002813 0.336432 -0.008 0.993329  
Pregnancy.Test 2.370554 0.521781 4.543 5.54e-06 ***  
Birth.Control -2.300272 0.365270 -6.297 3.03e-10 ***  
Feminine.Hygiene -2.028558 0.342398 -5.925 3.13e-09 ***  
Folic.Acid 4.077666 0.761888 5.352 8.70e-08 ***
```

```

Prenatal.Vitamins 2.479469 0.369063 6.718 1.84e-11 ***
Prenatal.Yoga 2.922974 1.146990 2.548 0.010822 *
Body.Pillow 1.261037 0.860617 1.465 0.142847
Ginger.Ale 1.938502 0.426733 4.543 5.55e-06 ***
Sea.Bands 1.107530 0.673435 1.645 0.100053
Stopped.buying.cig 1.302222 0.342347 3.804 0.000142 ***
Cigarettes -1.443022 0.370120 -3.899 9.67e-05 ***
Smoking.Cessation 1.790779 0.512610 3.493 0.000477 ***
Stopped.buying.win 1.383888 0.305883 4.524 6.06e-06 ***
Wine -1.565539 0.348910 -4.487 7.23e-06 ***
Maternity.Clothes 2.078202 0.329432 6.308 2.82e-10 ***
---
Signif. codes: 0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Tais coeficientes sem ao menos um \* próximos a eles são de valor duvidoso.

Da mesma forma, você pode treinar um modelo floresta aleatória usando a função `randomForest()`:

```

> Pregnancy.rf <-
randomForest(PREGNANT~., data=PregnancyData, importance=TRUE)

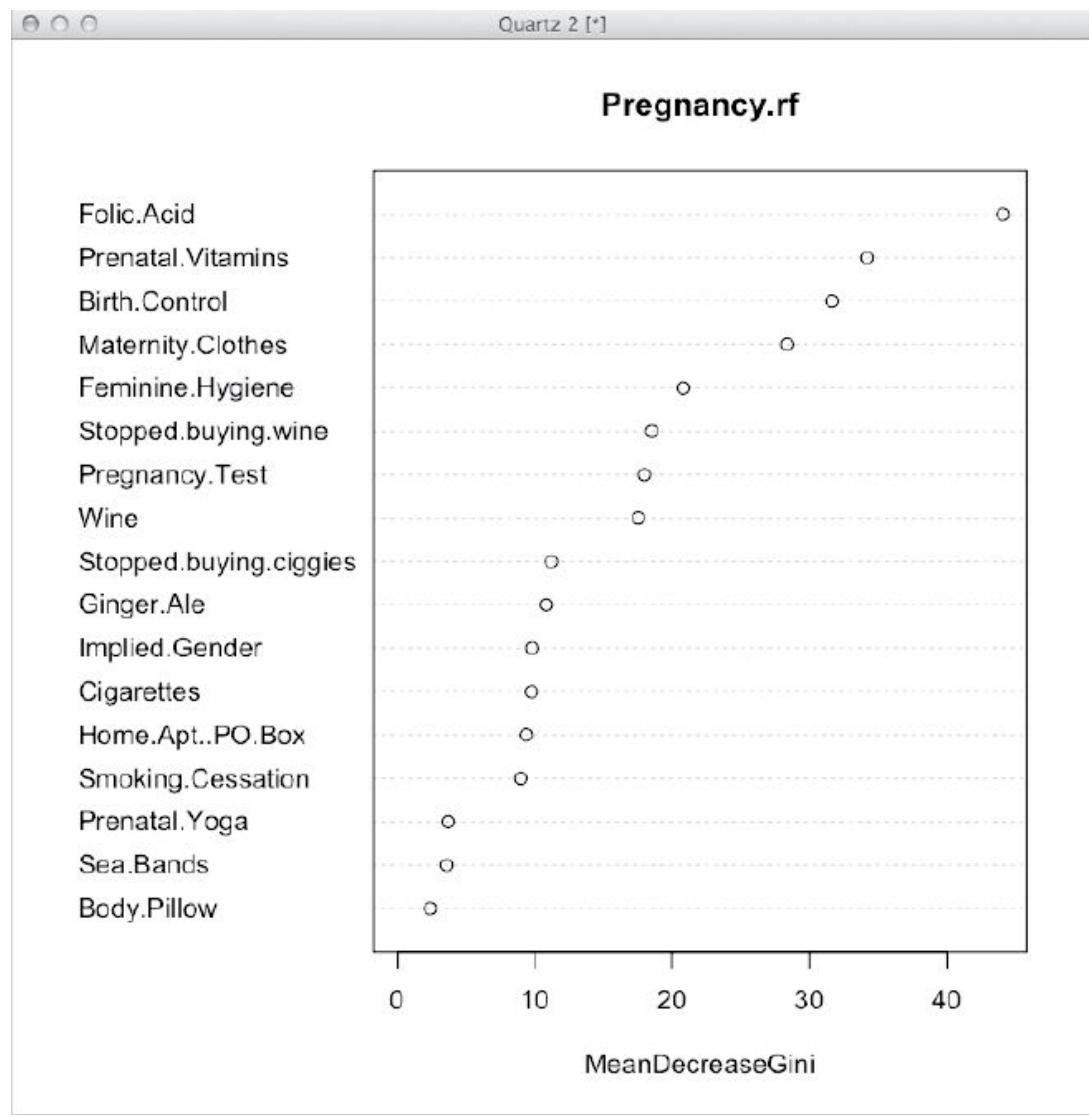
```

É a mesma sintaxe básica que a chamada `glm()` (execute `?randomForest` para aprender mais sobre contagem de árvore e profundidade). Observe `importance=TRUE` na chamada. Isso permite que você transforme em gráfico a importância da variável usando outra função, `varImpPlot()`, e, portanto, entender quais variáveis são importantes e quais são fracas.

O pacote `randomForest` permite que você observe o quanto cada variável contribui para diminuir a impureza do nó na média. Quanto mais uma variável contribui, mais útil ela é. Você pode usar isso para selecionar e parear as variáveis que talvez queira para usar em outro modelo. Para observar esses dados, use a função `varImpPlot()` com `type=2` para retirar as classificações baseadas nos cálculos de impureza do nó introduzidos no Capítulo 7 (fique à vontade para usar o comando `?varImpPlot` para ler a diferença entre `type=1` e `type=2`):

```
> varImpPlot(Pregnancy.rf, type=2)
```

Isso tem como resultado a classificação exibida na Figura 10-4. As classificações de ácido fólico vêm em primeiro e em seguida as de vitaminas pré-natais.



**Figura 10-4:** O gráfico da importância da variável em R

Agora que construiu os modelos, você pode fazer previsões com eles usando a função `predict()` em R. Chame a função e salve os resultados em duas variáveis diferentes a fim de comparar os modelos. A forma como a função `predict()` geralmente funciona aceitando um modelo, um

conjunto de dados para ser previsto, e quaisquer opções de modelo específicas:

```
> PregnancyData.Test.lm.Preds <-  
predict(Pregnancy.lm, PregnancyData.Test, type="response")  
> PregnancyData.Test.rf.Preds <-  
predict(Pregnancy.rf, PregnancyData.Test, type="prob")
```

Você pode ver nas duas chamadas de `predict` que, para cada uma, é fornecido um modelo diferente, o conjunto de dados e os parâmetros `type` que tais modelos precisam. No caso de um modelo linear, `type="response"` configura os valores retornados da previsão para serem entre 0 e 1 assim como nos valores originais de `PREGNANT`. No caso da floresta aleatória, `type="prob"` se certifica de ter de volta as probabilidades de classe — duas colunas de dados, uma probabilidade de gravidez e uma de não gravidez.

Tais saídas são ligeiramente diferentes, mas, novamente, elas são algoritmos diferentes, modelos diferentes, e assim por diante. É importante lidar com essas coisas e ler a documentação.

Este é um resumo da saída da previsão:

```
> summary(PregnancyData.Test.lm.Preds)  
Min. 1st Qu. Median Mean 3rd Qu. Max.  
0.001179 0.066190 0.239500 0.283100 0.414300 0.999200  
> summary(PregnancyData.Test.rf.Preds)  
0 1  
Min. :0.0000 Min. :0.0000  
1st Qu.:0.7500 1st Qu.:0.0080  
Median :0.9500 Median :0.0500  
Mean :0.8078 Mean :0.1922  
3rd Qu.:0.9920 3rd Qu.:0.2500  
Max. :1.0000 Max. :1.0000
```

A segunda coluna da previsão da floresta aleatória é a probabilidade associada com a gravidez (em oposição à sem gravidez), então essa é a

coluna que é semelhante às previsões da regressão logística. Usando a notação dos colchetes, você pode retirar registros individuais e conjuntos de registros e observar seus dados de entrada e previsões (eu transpus a linha para a impressão ficar mais bonita):

```
> t(PregnancyData.Test[,])  
1  
Implied.Gender "U"  
Home.Apt..PO.Box "A"  
Pregnancy.Test "0"  
Birth.Control "0"  
Feminine.Hygiene "0"  
Folic.Acid "0"  
Prenatal.Vitamins "0"  
Prenatal.Yoga "0"  
Body.Pillow "0"  
Ginger.Ale "0"  
Sea.Bands "1"  
Stopped.buying.ciggies "0"  
Cigarettes "0"  
Smoking.Cessation "0"  
Stopped.buying.wine "1"  
Wine "1"  
Maternity.Clothes "0"  
PREGNANT "1"  
> t(PregnancyData.Test.lm.Preds[1])  
1  
[1,] 0.6735358  
> PregnancyData.Test.rf.Preds[1,2]  
[1] 0.504
```

Observe que, ao imprimir a linha de entrada, eu deixo o índice da coluna em branco dentro dos colchetes [1, ] para que todos os dados das

colunas sejam impressos. Esse cliente em particular possui gênero desconhecido, mora em um apartamento, comprou alguns remédios e vinho, porém, parou de comprar vinho depois de um tempo. A regressão logística tem um escore 0,67 enquanto que a floresta aleatória fica em torno de 0,5. A verdade é que ela está grávida — um ponto pra regressão logística!

Agora que você possui dois vetores de probabilidades de classe, um para cada modo, você pode comparar os modelos em termos de taxa positiva verdadeira e taxa falsa positiva da mesma forma como fez antes no livro. Você tem sorte de, em R, o pacote `ROCR` poder computar e traçar as curvas ROC por você. Já que você carregou o pacote `ROCR`, a primeira coisa que precisa fazer é criar dois objetos de previsão `ROCR` (usando a função `ROCR prediction()`) que apenas façam a contagem das previsões de classe positiva e negativa em vários níveis de corte nas probabilidades de classe:

```
> pred.lm <-  
prediction(PregnancyData.Test.lm.Preds,  
PregnancyData.Test$PREGNANT)  
> pred.rf <-  
prediction(PregnancyData.Test.rf.Preds[,2],  
PregnancyData.Test$PREGNANT)
```

Repare na segunda chamada que você acerta a segunda coluna de probabilidades de classe a partir do objeto floresta aleatória como foi discutido antes. Você então pode transformar esses objetos de previsão em objetos de desempenho `ROCR` ao fazê-los percorrer a função `performance()`. Um objeto de desempenho pega as classificações dadas pelo modelo no conjunto de teste para vários valores de corte e as usa para montar uma curva de sua escolha (nesse caso uma curva ROC):

```
> perf.lm <- performance(pred.lm, "tpr", "fpr")  
> perf.rf <- performance(pred.rf, "tpr", "fpr")
```

## NOTA

Se você estiver curioso, `performance()` fornece outras opções além dos valores de `tpr` e `fpr`, tal como `prec` para precisão e `rec` para recall. Leia a documentação do pacote `ROCR` para mais detalhes.

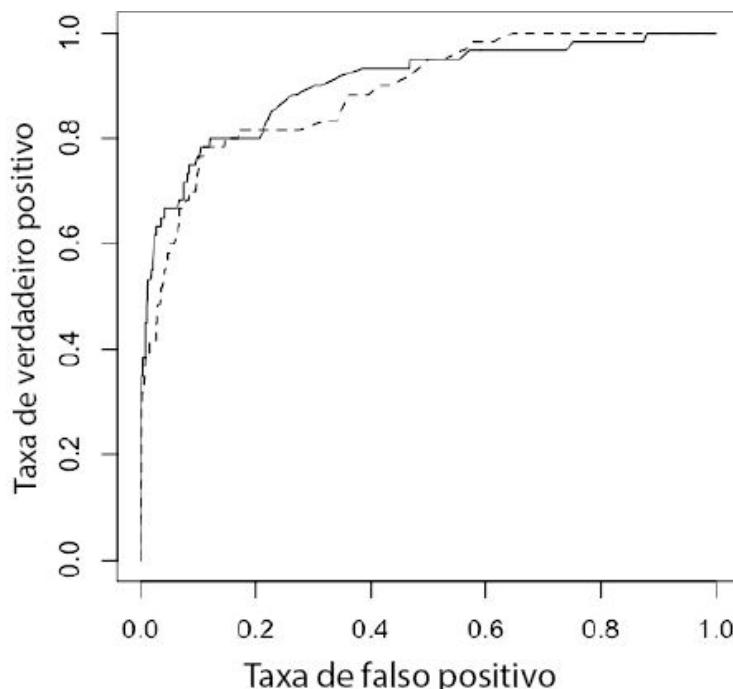
Então você pode traçar essas curvas usando a função `plot()` de R. Primeiro, a curva do modelo linear (as flags `xlim` e `ylim` são usadas para configurar os limites superior e inferior dos eixos `x` e `y` no gráfico):

```
> plot(perf.lm, xlim=c(0,1), ylim=c(0,1))
```

Você pode adicionar a curva floresta aleatória usando a flag `add=TRUE` para cobri-la e a flag `lty=2` (`lty` significa “line type”; verifique `?plot` para saber mais) para tornar essa linha tracejada:

```
> plot(perf.rf, xlim=c(0,1), ylim=c(0,1), lty=2, add=TRUE)
```

Isso sobrepõe duas curvas com o desempenho da floresta aleatória como uma linha tracejada, como mostra a Figura 10-5. Na maior parte, a regressão logística é superior com a floresta aleatória aumentando na distância direita do gráfico.



**Figura 10-5:** Recall e precisão no gráfico em R

Tudo bem, vamos recapitular um pouco, você treinou dois modelos preditivos diferentes, os usou em um conjunto de teste e comparou sua precisão versus o recall usando o código a seguir:

```
> PregnancyData <- read.csv("Pregnancy.csv")
> PregnancyData.Test <- read.csv("Pregnancy_Test.csv")
> PregnancyData$PREGNANT <- factor(PregnancyData$PREGNANT)
> PregnancyData.Test$PREGNANT <-
factor(PregnancyData.Test$PREGNANT)
> install.packages("randomForest", dependencies=TRUE)
> install.packages ("ROCR", dependencies=TRUE)
> library(randomForest)
> library(ROCR)
> Pregnancy.lm <- glm(PREGNANT ~ .,
data=PregnancyData, family=binomial("logit"))
> summary(Pregnancy.lm)
> Pregnancy.rf <-
randomForest(PREGNANT~., data=PregnancyData, importance=TRUE)
> PregnancyData.Test.rf.Preds <-
predict(Pregnancy.rf, PregnancyData.Test, type="prob")
> varImpPlot(Pregnancy.rf, type=2)
> PregnancyData.Test.lm.Preds <-
predict(Pregnancy.lm, PregnancyData.Test, type="response")
> PregnancyData.Test.rf.Preds <-
predict(Pregnancy.rf, PregnancyData.Test, type="prob")
> pred.lm <-
prediction(PregnancyData.Test.lm.Preds,
PregnancyData.Test$PREGNANT)
> pred.rf <-
prediction(PregnancyData.Test.rf.Preds[,2],
PregnancyData.Test$PREGNANT)
```

```
> perf.lm <- performance(pred.lm, "tpr", "fpr")
> perf.rf <- performance(pred.rf, "tpr", "fpr")
> plot(perf.lm, xlim=c(0,1), ylim=c(0,1))
> plot(perf.rf, xlim=c(0,1), ylim=c(0,1), lty=2, add=TRUE)
```

Bem direto, realmente. Comparado ao Excel, observe como foi fácil comparar dois modelos diferentes. Isso é bem legal.

## Fazendo Previsões em R

### NOTA

O arquivo CSV utilizado nesta seção, “SwordDemand.csv”, está disponível para download na página da editora, em [www.altabooks.com.br](http://www.altabooks.com.br), procurando pelo título do livro.

Esta próxima seção é uma loucura. Por quê? Porque você vai regenerar a previsão suavização exponencial do Capítulo 8 tão rápido que fará sua cabeça girar.

Primeiro, carregue os dados de demanda da espada de SwordDemand.csv e a imprima no console:

```
> sword <- read.csv("SwordDemand.csv")
> sword
SwordDemand
1 165
2 171
3 147
4 143
5 164
6 160
7 152
8 150
9 159
```

```
10 169  
11 173  
12 203  
13 169  
14 166  
15 162  
16 147  
17 188  
18 161  
19 162  
20 169  
21 185  
22 188  
23 200  
24 229  
25 189  
26 218  
27 185  
28 199  
29 210  
30 193  
31 211  
32 208  
33 216  
34 218  
35 264  
36 304
```

Tudo bem, você teve 36 meses de demanda carregados, simples e correto. A primeira coisa que precisa fazer é contar para R que eles são dados em série temporal. Existe uma função chamada `ts()` que é utilizada com esse propósito:

```
sword.ts <- ts(sword,frequency=12,start=c(2010,1))
```

Nesta chamada, você fornece a função `ts()` com os dados, um valor de frequência (a quantidade de observações por unidade de tempo, que nesse caso é de 12 ao ano) e um ponto inicial (esse exemplo usa Janeiro 2010).

Quando imprime `sword.ts` ao digitá-lo no terminal, R agora sabe imprimi-la em uma tabela por mês:

```
> sword.ts
```

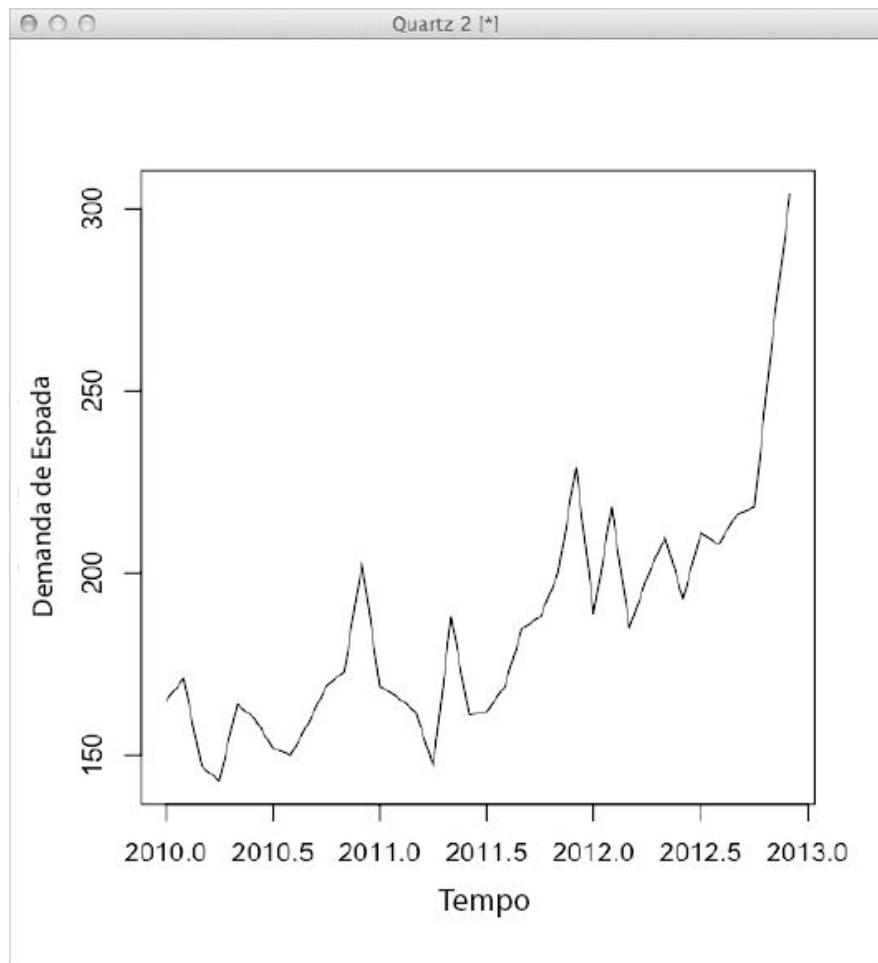
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2010	165	171	147	143	164	160	152	150	159	169	173	203
2011	169	166	162	147	188	161	162	169	185	188	200	229
2012	189	218	185	199	210	193	211	208	216	218	264	304

Legal!

Você pode traçar os dados também:

```
> plot(sword.ts)
```

Isso tem como resultado o gráfico exibido na Figura 10-6:



**Figura 10-6:** Gráfico da demanda de espada

Neste momento, você está pronto para fazer previsões usando o excelente pacote `forecast`. Sinta-se à vontade para procurar sobre CRAN (<http://cran.r-project.org/package=forecast> — em inglês), ou assistir ao autor falando sobre isso neste vídeo no YouTube

<http://www.youtube.com/watch?v=1Lh1H1BUF8k> — em inglês.

Para fazer previsões usando o pacote `forecast`, usa-se um objeto de série temporal em uma função `forecast()`. A chamada de `forecast()` foi configurada para detectar a técnica adequada a ser usada. Você se lembra de ter visto algumas técnicas no decorrer do livro? A função `forecast()` fará tudo isso por você:

```
> install.packages("forecast", dependencies=TRUE)
> library(forecast)
```

```
> sword.forecast <- forecast(sword.ts)
```

E é isso. Sua previsão está salva no objeto `sword.forecast`. Agora você pode imprimi-la:

```
> sword.forecast  
Point Forecast Lo 80 Hi 80 Lo 95 Hi 95  
Jan 2013 242.9921 230.7142 255.2699 224.2147 261.7695  
Feb 2013 259.4216 246.0032 272.8400 238.8999 279.9433  
Mar 2013 235.8763 223.0885 248.6640 216.3191 255.4334  
Apr 2013 234.3295 220.6882 247.9709 213.4669 255.1922  
May 2013 274.1674 256.6893 291.6456 247.4369 300.8980  
Jun 2013 252.5456 234.6894 270.4019 225.2368 279.8544  
Jul 2013 257.0555 236.7740 277.3370 226.0376 288.0734  
Aug 2013 262.0715 238.9718 285.1711 226.7436 297.3993  
Sep 2013 279.4771 252.0149 306.9392 237.4774 321.4768  
Oct 2013 289.7890 258.1684 321.4097 241.4294 338.1487  
Nov 2013 320.5914 281.9322 359.2506 261.4673 379.7155  
Dec 2013 370.3057 321.2097 419.4018 295.2198 445.3917  
Jan 2014 308.3243 263.6074 353.0413 239.9357 376.7130  
Feb 2014 327.6427 275.9179 379.3675 248.5364 406.7490  
Mar 2014 296.5754 245.8459 347.3049 218.9913 374.1594  
Apr 2014 293.3646 239.2280 347.5013 210.5698 376.1595  
May 2014 341.8187 274.0374 409.5999 238.1562 445.4812  
Jun 2014 313.6061 247.0271 380.1851 211.7823 415.4299  
Jul 2014 317.9789 245.9468 390.0109 207.8153 428.1424  
Aug 2014 322.9807 245.1532 400.8081 203.9538 442.0075  
Sep 2014 343.1975 255.4790 430.9160 209.0436 477.3513  
Oct 2014 354.6286 258.7390 450.5181 207.9782 501.2790  
Nov 2014 391.0099 279.4304 502.5893 220.3638 561.6559  
Dec 2014 450.1820 314.9086 585.4554 243.2992 657.0648
```

Você obtém uma previsão com intervalos de previsão embutidos! É possível imprimir a técnica de previsão atual usada ao imprimir o valor

do método no objeto `sword.object`:

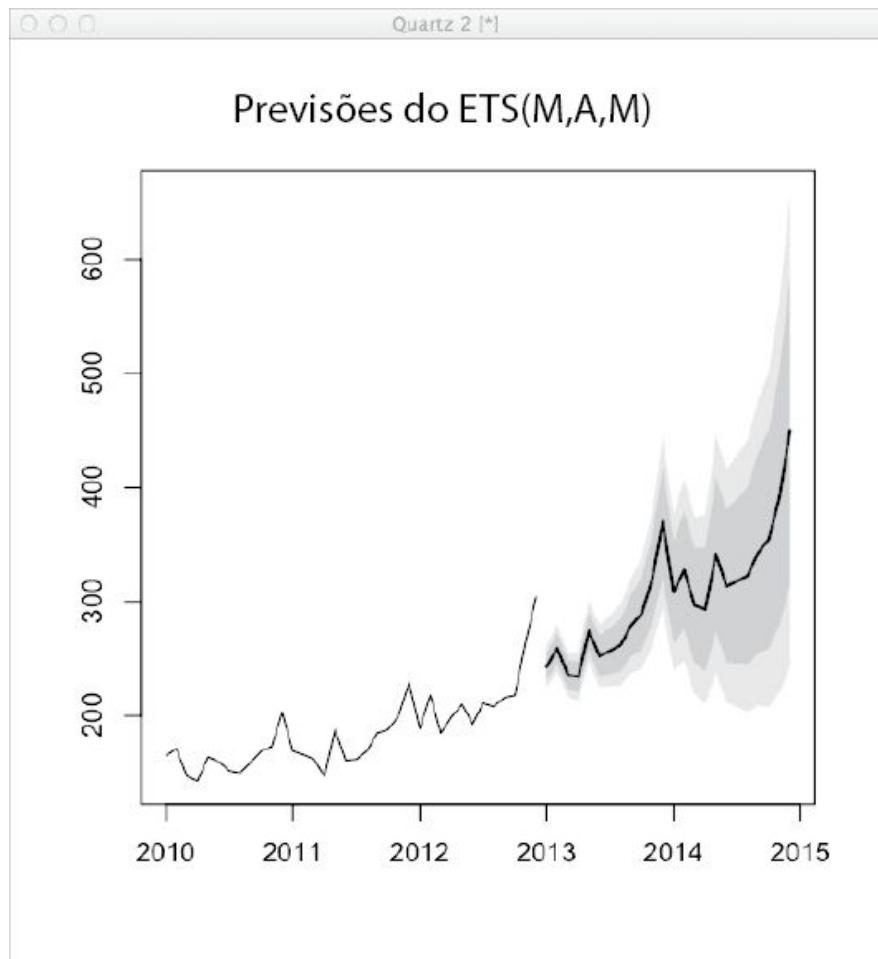
```
> sword.forecast$method  
[1] "ETS (M,A,M)"
```

MAM significa erro multiplicativo, tendência aditiva, sazonalidade multiplicativa. A função `forecast()` na verdade escolheu executar a suavização exponencial Holt-Winters! E você nem teve que fazer nada. Quando traçar, como mostra a Figura 10-7, você ganha automaticamente um fan chart:

```
> plot(sword.forecast)
```

Para recapitular, este é o código replicado do Capítulo 8:

```
> sword <- read.csv("SwordDemand.csv")  
> sword.ts <- ts(sword, frequency=12, start=c(2010,1))  
> install.packages("forecast", dependencies=TRUE)  
> library(forecast)  
> sword.forecast <- forecast(sword.ts)  
> plot(sword.forecast)
```



**Figura 10-7: Fan chart da previsão da demanda**

Louco. Mas essa é a beleza de usar pacotes que outras pessoas escreveram especialmente para fazer essas coisas.

## Observando ↑ a ↑ Detecção ↑ do ↑ Valor ↑ Atípico

### NOTA

Os arquivos CSV utilizados nesta seção, “PregnancyDuration.csv” e “CallCenter.csv”, estão disponíveis para download na página da editora em [www.altabooks.com.br](http://www.altabooks.com.br), procurando pelo título do livro.

Nesta seção, você fará mais um capítulo deste livro em R, para sentir-se em casa com esse assunto. Para começar, leia os dados de duração da

gravidez em `PregnancyDuration.csv` disponível na página da editora:

```
> PregnancyDuration <- read.csv("PregnancyDuration.csv")
```

No Capítulo 9, você calculou a mediana, o primeiro quartil, o terceiro quartil, e Teste de Tukey superior e inferior. Você pode obter os quartis ao resumir os dados:

```
> summary(PregnancyDuration)
GestationDays
Min. :240.0
1st Qu.:260.0
Median :267.0
Mean :266.6
3rd Qu.:272.0
Max. :349.0
```

Isso faz com que o intervalo do interquartil seja 272 menos 260 (como alternativa, pode chamar a função embutida `IQR()` na coluna `GestationDays`):

```
> PregnancyDuration.IQR <- 272 - 260
> PregnancyDuration.IQR <- IQR(PregnancyDuration$GestationDays)
> PregnancyDuration.IQR
[1] 12
```

Você pode calcular o Teste de Tukey superior e inferior:

```
> LowerInnerFence <- 260 - 1.5 * PregnancyDuration.IQR
> UpperInnerFence <- 272 + 1.5 * PregnancyDuration.IQR
> LowerInnerFence
[1] 242
> UpperInnerFence
[1] 290
```

Usando a função `which()` de R, é fácil determinar os pontos e seus índices que violam os delimitadores. Por exemplo:

```
> which(PregnancyDuration$GestationDays > UpperInnerFence)
[1] 1 249 252 338 345 378 478 913
```

```
> PregnancyDuration$GestationDays [  
which(PregnancyDuration$GestationDays > UpperInnerFence)  
]  
[1] 349 292 295 291 297 303 293 296
```

Claro, a melhor maneira de fazer essa análise é usando a função `boxplot()` de R. A função `boxplot()` colocará em gráficos a mediana, o primeiro e o terceiro quartis, Teste de Tukey, e quaisquer valores atípicos. Para usá-la, introduza a coluna `GestationDays` dentro da função:

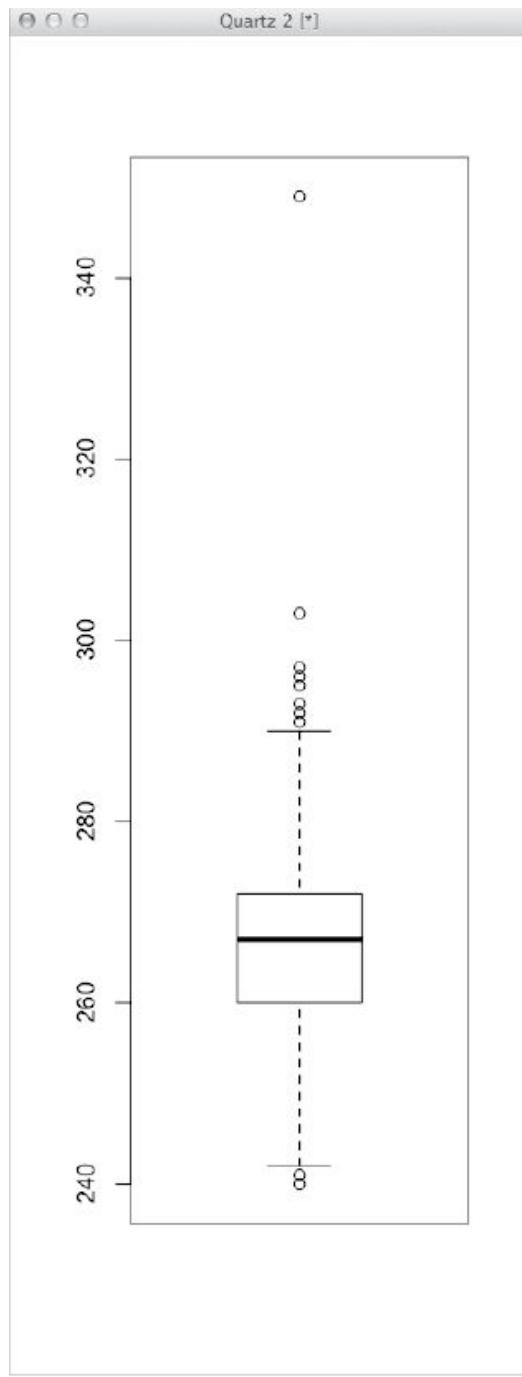
```
> boxplot(PregnancyDuration$GestationDays)
```

Isso tem como resultado a visualização exibida na Figura 10-8.

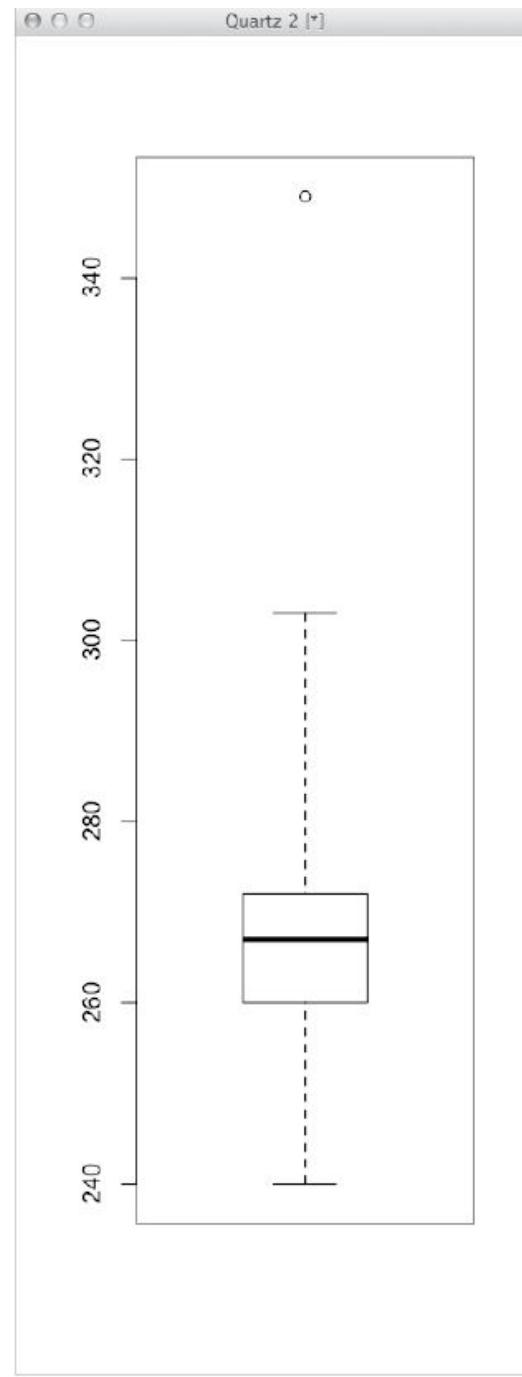
O Teste de Tukey pode ser modificado para ser delimitador “externo” ao mudar a sinalização da série na chamada do **gráfico de caixa** (o padrão é 1,5 vezes o `IQR`). Se você configurar `range=3`, o Teste de Tukey será então destinado ao último ponto dentro de três vezes o `IQR`:

```
> boxplot(PregnancyDuration$GestationDays, range=3)
```

Como na Figura 10-9, observe agora que você tem apenas um valor atípico — é a duração da gravidez de 349 dias da Sra. Hadlum.



**Figura 10-8:** Um gráfico de caixa dos dados da duração da gravidez



**Figura 10-9:** Um gráfico de caixa com Teste de Tukey usando três vezes o IQR

Você também pode retirar esses dados do boxplot no console em vez de retirar do gráfico. Ao imprimir a lista `stats`, você obtém os delimitadores e os quartis:

```
> boxplot(PregnancyDuration$GestationDays, range=3)$stats
```

```
[,1]  
[1,] 240  
[2,] 260  
[3,] 267  
[4,] 272  
[5,] 303
```

Ao imprimir a lista `out`, você obtém uma lista de valores atípicos:

```
> boxplot(PregnancyDuration$GestationDays, range=3)$out  
[1] 349
```

Tudo bem, então temos um probleminha na duração da gravidez.

Vamos prosseguir para o problema maior de encontrar os valores atípicos nos dados de desempenho dos funcionários do call center. Está na planilha `CallCenter.csv` na página da editora. Ao carregar e resumir, você obtém:

```
> CallCenter <- read.csv("CallCenter.csv")  
> summary(CallCenter)  
ID AvgTix Rating Tardies  
Min. :130564 Min. :143.1 Min. :2.070 Min. :0.000  
1st Qu.:134402 1st Qu.:153.1 1st Qu.:3.210 1st Qu.:1.000  
Median :137906 Median :156.1 Median :3.505 Median :1.000  
Mean :137946 Mean :156.1 Mean :3.495 Mean :1.465  
3rd Qu.:141771 3rd Qu.:159.1 3rd Qu.:3.810 3rd Qu.:2.000  
Max. :145176 Max. :168.7 Max. :4.810 Max. :4.000  
Graveyards Weekends SickDays PercSickOnFri  
Min. :0.000 Min. :0.0000 Min. :0.000 Min. :0.0000  
1st Qu.:1.000 1st Qu.:1.0000 1st Qu.:0.000 1st Qu.:0.0000  
Median :2.000 Median :1.0000 Median :2.000 Median :0.2500  
Mean :1.985 Mean :0.9525 Mean :1.875 Mean :0.3522  
3rd Qu.:2.000 3rd Qu.:1.0000 3rd Qu.:3.000 3rd Qu.:0.6700  
Max. :4.000 Max. :2.0000 Max. :7.000 Max. :1.0000  
EmployeeDevHrs ShiftSwapsReq ShiftSwapsOffered
```

```

Min. : 0.00 Min. :0.000 Min. :0.00
1st Qu.: 6.00 1st Qu.:1.000 1st Qu.:0.00
Median :12.00 Median :1.000 Median :1.00
Mean :11.97 Mean :1.448 Mean :1.76
3rd Qu.:17.00 3rd Qu.:2.000 3rd Qu.:3.00
Max. :34.00 Max. :5.000 Max. :9.00

```

Assim como no Capítulo 9, é preciso dimensionar e centralizar os dados. Para fazer isso, é apenas preciso usar a função `scale()`:

```

> CallCenter.scale <- scale(CallCenter[2:11])
> summary(CallCenter.scale)

AvgTix Rating Tardies Graveyards
Min. :-2.940189 Min. :-3.08810 Min. :-1.5061 Min. :-2.4981
1st Qu.:-0.681684 1st Qu.:-0.61788 1st Qu.:-0.4781 1st Qu.:-1.2396
Median :-0.008094 Median : 0.02134 Median :-0.4781 Median : 0.0188
Mean : 0.000000 Mean : 0.00000 Mean : 0.0000 Mean : 0.0000
3rd Qu.: 0.682476 3rd Qu.: 0.68224 3rd Qu.: 0.5500 3rd Qu.: 0.0188
Max. : 2.856075 Max. : 2.84909 Max. : 2.6062 Max. : 2.5359

Weekends SickDays PercSickOnFri EmployeeDevHrs
Min. :-1.73614 Min. :-1.12025 Min. :-0.8963 Min. :-1.60222
1st Qu.: 0.08658 1st Qu.:-1.12025 1st Qu.:-0.8963 1st Qu.:-0.79910
Median : 0.08658 Median : 0.07468 Median :-0.2601 Median : 0.00401
Mean : 0.00000 Mean : 0.00000 Mean : 0.0000 Mean : 0.00000
3rd Qu.: 0.08658 3rd Qu.: 0.67215 3rd Qu.: 0.8088 3rd Qu.: 0.67328
Max. : 1.90930 Max. : 3.06202 Max. : 1.6486 Max. : 2.94879

ShiftSwapsReq ShiftSwapsOffered
Min. :-1.4477 Min. :-0.9710
1st Qu.:-0.4476 1st Qu.:-0.9710
Median :-0.4476 Median :-0.4193
Mean : 0.0000 Mean : 0.0000
3rd Qu.: 0.5526 3rd Qu.: 0.6841
Max. : 3.5530 Max. : 3.9942

```

Agora que os dados estão preparados, você pode enviá-los para a função `lofactor()` que faz parte do pacote `DMwR`:

```
> install.packages ("DMwR", dependencies=TRUE)
> library(DMwR)
```

Para chamar a função `lofactor()`, forneça os dados e um valor `k` (esse exemplo usa 5, assim como o Capítulo 9), e a função cospe os LOFs:

```
> CallCenter.lof <- lofactor(CallCenter.scale,5)
```

Os dados com os fatores mais altos (LOFs geralmente ficam em torno de 1) são os pontos mais estranhos. Por exemplo, pode-se realçar os dados associados com tais funcionários cujo LOF é maior do que 1,5:

```
> which(CallCenter.lof > 1.5)
[1] 299 374
> CallCenter[which(CallCenter.lof > 1.5),]
ID AvgTix Rating Tardies Graveyards Weekends SickDays
299 137155 165.3 4.49 1 3 2 1
374 143406 145.0 2.33 3 1 0 6
PercSickOnFri EmployeeDevHrs ShiftSwapsReq ShiftSwapsOffered
299 0.00 30 1 7
374 0.83 30 4 0
```

Esses são os mesmos funcionários com valores atípicos discutidos no Capítulo 9. Mas a diferença foi grande na quantidade de linhas de código para chegar até aqui:

```
> CallCenter <- read.csv("CallCenter.csv")
> install.packages ("DMwR", dependencies=TRUE)
> library(DMwR)
> CallCenter.scale <- scale(CallCenter[2:11])
> CallCenter.lof <- lofactor(CallCenter.scale,5)
```

E isso foi tudo!

## Resumindo

Certo, essa foi uma recapitulação veloz e furiosa de algumas coisas que você pode fazer em R apenas entendendo três coisas:

- Carregar e trabalhar com dados em R
- Descobrir e instalar pacotes relevantes
- Chamar funções a partir de tais pacotes para o seu conjunto de dados

Isso é tudo o que precisa saber sobre R? Não mesmo. Não falei sobre escrever suas próprias funções, muitos e muitos gráficos, conectar bases de dados, a variação das funções disponíveis de `apply()`, e assim por diante. Mas espero que você tenha desenvolvido o gostinho por aprender mais. Se sim, existem inúmeros livros sobre R que valem a pena serem lidos como acompanhamento deste livro. Estes são alguns:

- Beginning R: The Statistical Programming Language por Mark Gardener (John Wiley & Sons, 2012)
- R in a Nutshell, 2nd Edition por Joseph Adler (O'Reilly, 2012)
- Data Mining with R: Learning with Case Studies by Luis Torgo (Chapman and Hall, 2010)
- Machine Learning for Hackers by Drew Conway and John Myles White (O'Reilly, 2012)

Vá em frente e mexa em R!

# Conclusão

## O↑que↑Aconteceu?↑Onde↑Estou?

Você deve ter começado este livro com habilidades normais em matemática e modelagem de planilhas, mas se está aqui, você sobreviveu (e não pulou os primeiros 10 capítulos). Agora imagino que você seja um conhedor de modelagem de planilhas com uma boa compreensão da variedade de técnicas de data science.

Este livro abordou tópicos desde pesquisas cruas de operações clássicas (otimização, Monte Carlo e previsão) até aprendizado não-supervisionado (detecção do valor atípico, agrupamento e gráficos) e até IA supervisionada (regressão, tocos de decisão e naïve Bayes). Você deveria sentir-se confiante ao trabalhar com dados em planilhas neste nível mais alto.

Também espero que o Capítulo 10 tenha lhe mostrado que, agora que você entende as técnicas de data science e algoritmos, é bem fácil usar tais técnicas dentro de uma linguagem de programação tal como R.

Se há algum tópico em particular que realmente pegou você neste livro, vá mais fundo! Quer mais R, mais otimização e mais aprendizado de máquina? Pegue uma das fontes que recomendo na conclusão de cada capítulo e leia. Há muito para aprender. Apenas pinciei na superfície da prática de análises neste livro.

Mas espere...

## Antes↑que↑Você↑se↑Vá

Quero usar esta conclusão para discutir alguns pensamentos sobre o que significa praticar data science no mundo real, porque apenas saber a matemática não é suficiente.

Qualquer pessoa que me conheça bem sabe que não sou uma pessoa fácil de se lidar. Minhas habilidades quantitativas são medianas, mas já vi gente bem mais esperta do que eu falhar drasticamente trabalhando com análises profissionais. O problema é que embora elas sejam brilhantes, elas não sabem o mínimo que pode levar empreendimentos técnicos a falharem dentro do ambiente de negócios. Portanto vamos cobrir esses itens mais leves que podem significar o sucesso ou a queda do seu projeto de análise ou carreira.

## Conhecer o Problema

Meu filme favorito de todos os tempos é o *Quebra de Sigilo*, de 1992. O filme gira em torno de um grupo de analistas liderados por Robert Redford, que rouba uma “caixa preta” capaz de quebrar a encriptação RSA. Uma grande brincadeira. (Eu invejo quem ainda não assistiu a esse filme, pois poderá assisti-lo pela primeira vez!)

Há uma cena na qual Robert Redford encontra um cofre eletrônico em uma porta trancada de um escritório, e ele precisa entrar.

Ele entra em contato com sua equipe usando o fone de ouvido. Eles estão esperando em uma van do lado de fora do edifício.

“Alguém já derrotou um cofre eletrônico?” ele pergunta.

“Essas coisas são impossíveis”, Sydney Poitier diz. Mas Dan Aykroyd, que também estava na van, tem uma ideia. Eles explicam os detalhes para Redford pelo fone.

Robert Redford balança sua cabeça e diz, “Certo, vou tentar”.

Ele ignora o teclado bloqueado e arromba a porta.

Veja bem, o problema não era “desvendar o bloqueio eletrônico” de forma alguma. O problema era entrar no cômodo. Dan Aykroyd entendeu isso.

Esse é o desafio fundamental da análise: entender o que realmente deve ser resolvido. Você precisa entender a situação, os processos, os dados e as circunstâncias. Você precisa categorizar tudo em volta do problema da

melhor forma que puder a fim de compreender exatamente qual é a solução ideal.

Em data science, você encontrará com frequência o “problema visto da forma errada”:

1. Alguém em uma situação encontrou um problema.
2. Eles usaram sua experiência (ou falta dela) anterior de conhecimento de análise para moldar o problema.
3. Eles entregam a concepção do problema para o analista como se fosse apenas aquilo e estivesse bem colocada.
4. O analista aceita e resolve o problema como ele está.

É possível que funcione. Mas não é o ideal, pois o problema que você pediu para ser resolvido por vezes não é o problema que precisa de solução. Se esse problema é realmente tal problema então os profissionais de análise não podem ser passivos.

Você não pode aceitar problemas como são dados a você no ambiente de negócios. Nunca permita que você seja o analista para quem problemas são “jogados”. Dedique-se às pessoas cujos desafios você está lidando para que tenha certeza de que esteja resolvendo o problema certo. Aprenda os processos de negócios e os dados que são gerados e salvos. Aprenda como as pessoas estão lidando com o problema agora e quais métricas elas usam (ou ignoram) para engrenar no sucesso.

Resolva o problema correto, e ainda com frequência desvirtuado. Isso é algo que nenhum modelo matemático jamais dirá a você. Nenhum modelo matemático poderá dizer “Olha, bom trabalho em formular esse modelo de otimização, eu acho que você deveria voltar um passo e mudar seu negócio um pouco”. E isso me leva ao próximo ponto: aprenda a se comunicar.

## Precisamos<sup>↑</sup>de<sup>↑</sup>Mais<sup>↑</sup>Tradutores

Se você terminou o livro, é seguro dizer que você sabe uma coisa ou outra sobre análise. Você está familiarizado com as ferramentas

disponíveis. Fez um protótipo delas. Isso permite que identifique oportunidades de análises melhor do que a maioria, porque você sabe o que é possível. Você não precisa esperar alguém trazer uma oportunidade. Você tem potencial para sair e correr atrás de um negócio.

Mas sem a habilidade da comunicação, fica difícil entender os desafios, articular o que é possível, e explicar o trabalho que está fazendo.

No ambiente atual de negócios, é inaceitável possuir uma única habilidade. Os cientistas de dados são ditos como poliglotas que entendem matemática, código, a linguagem popular de negócios (ou o vocabulário específico de esportes... aff). A única forma de melhorar sua comunicação com os outros é como a única forma de melhorar na matemática, é por meio da prática.

Pegue qualquer oportunidade que puder para discutir com outros sobre análise, formal e informalmente. Descubra maneiras de conversar com os outros em seu local de trabalho sobre o que eles fazem, o que você faz, e algo que você possa colaborar. Fale com os outros em reuniões sobre o que você faz. Encontre formas de articular conceitos de análise dentro do seu contexto de negócios particular.

Faça com que a gerência envolva você no planejamento e nas discussões de negócios. Com frequência o profissional de análise é abordado com um projeto depois de ele ter um escopo, mas seu conhecimento das técnicas e dados disponíveis faz com que você seja indispensável no planejamento inicial.

Invista em ser visto como uma pessoa digna de uma conversa e não apenas uma extensão de um computador cujos problemas são apenas designados à distância. Quanto mais engajado e comunicativo o analista for dentro de uma organização, mais efetivo ele é.

Durante muito tempo os analistas têm sido tratados como mulheres da época vitoriana — separados dos pontos altos dos negócios porque eles não entenderiam nada. Ah, por favor. Deixe as pessoas sentirem o peso do seu conjunto de habilidades — só porque você mastiga números igual

a um computador, não significa que você não possa discutir um slide do PowerPoint. Entre lá, faça sua parte e fale com as pessoas.

## Cuidado com o Monstro-Nerd-de-Três-Cabeças: Ferramentas, Desempenho e Perfeição Matemática

Há muitas coisas que podem sabotar o uso da análise dentro de um local de trabalho. Talvez política e conflitos internos; uma experiência ruim com projetos para “empresas, business intelligence e painel de nuvem”; ou pares que não querem suas “bruxarias” otimizadas e automatizadas por medo de seus empregos se tornarem redundantes.

Nem todos os obstáculos estão dentro do seu controle como um profissional de análise. Mas alguns estão. Há três formas primárias que vejo o pessoal de análise sabotar seu próprio trabalho: modelagem extremamente complexa, obsessão por ferramentas e fixação por desempenho.

### *Complexidade*

Muitas luas atrás, eu trabalhei em um modelo de otimização para uma cadeia de suprimentos para a empresa Fortune 500. Esse modelo foi muito maneiro se me permite dizer. Juntamos todos os tipos de regras de negócios do cliente e modelamos todo o seu processo de fretamento como um programa integral. Até modelamos a demanda futura distribuída normalmente dentro do modelo em um livro que acabou sendo publicado.

Mas o modelo foi um fracasso. Mal começou e já estava morto. Não quero dizer que ele estava morto e nem errado, mas que ele não foi usado. Francamente, uma vez que os técnicos foram embora, não sobrou ninguém na empresa que poderia manter atualizados as médias de erro de previsão cumulativas e os desvios padrões. Os soldados não entenderam apesar do treinamento pesado que demos.

Essa é uma diferença entre academia e indústria. Na academia, o sucesso não é avaliado pela utilidade. Um modelo de otimização original é valioso do seu próprio jeito, mesmo sendo muito complexo para um analista de cadeia de suprimentos manter funcionando.

Mas na indústria, a análise é uma perseguição movida a resultados e os modelos são julgados por seus valores práticos tanto quanto por sua originalidade.

Neste caso, eu passei muito tempo usando matemática complexa para otimizar a cadeia de suprimentos da empresa mas nunca de fato me atentei ao fato de que ninguém seria capaz de manter o modelo atualizado.

*A marca de um verdadeiro profissional de análise, tanto quanto a marca de uma artista, é saber quando editar.* Quando você deixa uma parte da complexidade da solução no chão da sala de cortes? Para ser bem clichê, lembre-se de que em análise o ótimo é o inimigo do bom. O melhor modelo é aquele que mantém o equilíbrio certo entre funcionalidade e manutenção. Se um modelo analítico nunca é usado, ele é inútil.

## ***Ferramentas***

Hoje em dia no mundo da análise (mesmo que você queira chamar de “data science”, “big data”, “business intelligence”, ou “blá blá blá nuvem”), as pessoas estão focadas em ferramentas e arquitetura.

Ferramentas são importantes. Elas permitem que você distribua seus produtos direcionados a dados e sua análise. Mas quando as pessoas falam sobre “a melhor ferramenta para o trabalho”, elas estão focadas na ferramenta e não no trabalho.

Softwares e companhias de serviço estão nos negócios de vender para o público soluções para problemas que eles nem possuem ainda. E, para piorar, muitos de nós têm chefes que assistem a *Pequenas Empresas Grandes Negócios*, olham para nós e dizem, “Precisamos usar esse tal de big data. Vá comprar algo e vamos Hadoopear”.

Esse é o caminho para um clima perigoso nos negócios hoje em dia em que a gerência olha para as ferramentas como provas de que as análises estão sendo feitas. Os fornecedores apenas querem vender as ferramentas que permitem as análises, mas há pouca responsabilidade que a análise realmente está sendo feita.

Portanto há uma regra simples: *identifique as oportunidades de análise com as quais você quer lidar com os maiores detalhes possíveis antes de adquirir as ferramentas.*

Você precisa de Hadoop? Bem, o seu problema requer uma agregação divida-e-conquiste de muitos dados desestruturados? Não? Então a resposta deve ser não. Não coloque a carroça na frente dos bois e compre as ferramentas (ou os consultores que são necessários para usar as ferramentas de código aberto) apenas para dizer, “Tudo bem, e agora o que fazemos com isso?”.

## *Desempenho*

Se eu ganhasse um centavo cada vez que alguém levantasse as sobrancelhas quando eu digo que no MailChimp usamos R na produção dos nossos modelos de prevenção abusivos, eu poderia comprar uma Pepsi. As pessoas acham que a linguagem não é adequada para as configurações de produção. Se eu estivesse fazendo um alto desempenho no comércio de ações, provavelmente não seria. É possível que eu codificasse tudo em C. Mas não vou.

Para o MailChimp, não passamos a maior parte do nosso tempo em R. Passamos movendo dados para serem enviados por meio de modelos IA. Não passamos a maior parte do tempo executando modelos IA, e certamente não passamos treinando o modelo IA.

Já encontrei pessoas que estavam muito preocupadas com a velocidade a qual o software pode treinar seus modelos de inteligência artificial. O modelo pode ser treinado em paralelo, em uma linguagem de baixo nível, em um ambiente vivo?

Eles nunca param de se perguntar se alguns deles são necessários e acabam gastando muito tempo lapidando a parte errada do seu projeto de

análise.

No MailChimp, nós treinamos novamente nossos modelos offline uma vez a cada quinze minutos, testamos, e então passamos para a produção. Em R, são necessárias algumas horas para treinar o modelo. E mesmo tendo terabytes de dados, o conjunto em treinamento do modelo, uma vez preparado, possui apenas 10 gigabytes, então eu posso até testar meu modelo em meu laptop. Uma loucura.

Já que este é o caso, eu não perco meu tempo na velocidade de treinamento do R. Eu foco mais em coisas importantes, como a precisão do modelo.

Não estou dizendo que você não deve ligar para o desempenho. Mas mantenha o foco e em situações em que não importa, sinta-se à vontade pra deixar para lá.

## Você↑Não↑É↑a↑Função↑Mais↑Importante↑da↑Sua Organização

Certo, então existem três coisas para se ter cuidado. Mas, geralmente, tenha em mente que a maioria das empresas não estão nos negócios fazendo análises. Eles enriquecem por outros meios, e as análises são o meio de servir a esses processos.

Você já deve ter ouvido em algum lugar que os cientistas de dados têm “o emprego mais sexy do século!”. Isso é por causa da forma como eles servem a indústria. **Servem** sendo a palavra-chave.

Considere a indústria de aviação. Eles têm feito análises de big data por décadas para tomar seu último centavo por aquela poltrona que mal cabe uma pessoa. Isso tudo é feito por meio de modelos de otimização de rendimento. É uma grande vitória para a matemática.

Mas quer saber? A parte mais importante do negócio é voar. Os produtos e serviços que uma organização vende importam mais do que os modelos que reúnem os centavos daqueles dólares. Seus objetivos devem ser coisas como usar os dados para facilitar melhores alvos,

previsões, preços, tomadas de decisão, relatórios, conformidade e assim por diante. Em outras palavras, trabalhe com o restante da sua organização para *fazer negócios melhores*, não para praticar data science por si só.

## Seja↑Criativo↑e↑Mantenha↑Contato!

Chega de bom senso. Se você trabalhou nos capítulos anteriores então você tem uma boa base para começar a sonhar, prototipar e implementar soluções para as oportunidades de análise propostas pelo seu negócio. Fale com os seus colegas de trabalho e seja criativo. Talvez haja uma solução analítica para algo que foi remendado com intuição e processos manuais. Vá para cima.

Boa briga de dados!