ESTATÍSTICA APLICADA AO DATA SCIENCE

PREDIÇÕES COM ÁRVORES DE DECISÃO

Autor: Dr. Antonio Gomes de Mattos

Revisor: Rafael Maltempe

INICIAR

Introdução

Sabemos que, a partir de um conjunto de dados, o qual chamamos de amostra, podemos desenvolver algoritmos preditivos e aplicá-los a situações da vida.

Especificamente, aplicamos modelos de regressão linear na predição do valor de imóveis (apartamentos) em função da sua área, do seu andar e da sua localização. Nesse caso, a variável resposta era quantitativa. Também aplicamos modelos preditivos para o caso em que a variável resposta é qualitativa, o que chamamos de problemas de classificação. Fizemos isso com modelos de regressão logística e aplicamos este tipo de modelo para a predição da inadimplência com cartões de crédito em função da renda da pessoa, seus gastos mensais com o cartão e se ela tinha ou não um emprego estável.

Aqui, veremos um outro tipo de modelo preditivo, as árvores de decisão. Aplicaremos árvores de decisão a um outro exemplo de problema de classificação. Faremos a predição do volume de vendas de um produto de varejo.

Visão Geral Sobre Aprendizagem Supervisionada

Já tivemos uma discussão inicial sobre o que são métodos de aprendizagem supervisionada e não supervisionada. Também vimos quais são os dois tipos mais importantes dos métodos de aprendizagem supervisionada. Vamos, aqui, caminhar um pouco mais sobre esse assunto, procurando ampliar a sua visão sobre esses métodos de aprendizagem.

Problemas de Regressão e de Classificação

Quando falamos sobre aprendizagem supervisionada, dois tipos de problemas são considerados principais: regressão e classificação. A diferença entre eles está centrada no conceito de variáveis quantitativas e variáveis qualitativas. Variáveis quantitativas referem-se a coisas que são medidas, tais como comprimento (metros), massa (quilos), temperatura (Kelvin), valor monetário de coisas que compramos (R$) e assim por diante. Variáveis qualitativas, às vezes chamadas de variáveis categóricas, na estatística, referem-se à qualidade das coisas, e, assim, não são coisas que se possa medir, mas são coisas que se consegue contar. Por exemplo, em uma sala de aula podemos ter um certo número de alunas e de alunos. A variável

qualitativa aqui é o sexo do aluno, que pode assumir dois valores (também chamados de níveis ou classes): feminino ou masculino. Ao observarmos a sala, podemos contar quantas alunas há, e fazer o mesmo para os alunos. Outro exemplo seria um hospital, onde observamos os pacientes e os classificamos por gravidade da doença. Por exemplo, a gravidade da doença pode assumir cinco valores (que também chamamos de níveis ou classes): sem gravidade, baixa gravidade, média gravidade, alta gravidade e altíssima gravidade. Essa classificação é feita pelos profissionais de saúde do hospital, que analisam as condições clínicas de cada paciente e os classificam em termos da gravidade da sua doença. Depois que cada paciente é classificado dessa forma, podemos contar quantos há para cada um dos níveis de gravidade da doença.

Problemas ditos de regressão são aqueles em que a variável resposta é quantitativa, e problemas de classificação são aqueles em que a variável resposta é qualitativa. Vamos ver alguns exemplos de cada um para fixarmos melhor esses conceitos.

Problemas de regressão

a. Predição do valor de um imóvel em função da sua área, do seu andar e da sua localização. Este foi o exemplo que vimos na Unidade 1. Importante destacar que, havendo dados, podemos aumentar o número de variáveis de entrada, incluindo outras, como o número de vagas de garagem, o número de banheiros, tipo da área de lazer (vários tipos, com ou sem churrasqueira, com ou sem piscina, com ou sem quadra poliesportiva etc.). Algumas empresas já usam algoritmos preditivos para avaliar preço de imóveis. Podemos citar, como exemplo provável, a empresa Quinto Andar, bastante avançada tecnologicamente nos seus processos de negócio.
b. Predição da redução de massa corporal em função da dieta alimentar e prática de atividades físicas. Este foi o exemplo que vimos no Pratique e Compartilhe da Unidade 1. Aqui também podemos incluir outras variáveis de entrada, tais como o sexo, a idade da pessoa dividida em faixas etárias, a etnia, a região em que mora etc.

c. Predição da performance atlética de um jogador profissional de futebol (ou outro esporte qualquer) em função da sua dieta alimentar e da sua rotina de treinos aeróbicos e anaeróbicos. Aqui também podemos incluir outras variáveis de entrada, tais como frequência média de jogos por semana (que, se muito alta, causa desgaste físico), número de lesões nos últimos seis meses (que afeta a performance atlética dos jogadores), variáveis biométricas (altura, peso, idade etc.). Este campo é chamado de estatística esportiva, e não se limita ao estudo da performance dos atletas e também pode ser aplicado ao estudo de táticas de jogo, por exemplo. É um campo em crescimento no mundo inteiro, havendo muitas oportunidades para a aplicação da estatística e da ciência dos dados.
d. Predição da resistência à tração de uma liga metálica, em um ensaio de tração em corpos de prova, em função da sua composição, granulometria e condições gerais do processo de fabricação da liga. Este é um problema típico da engenharia, em que problemas de regressão são muito comuns. Na engenharia e na física, é comum lidarmos com problemas em que a variável resposta é quantitativa, mas também há outros em que a variável resposta é qualitativa (categórica).
e. Predição da taxa de mortalidade anual por melanoma maligno em função da latitude. No livro *Biostatistics — A Methodology for the Health Science* (BELLE *et al.*, 2004, p. 291–300), você encontra um exemplo de estudo em que foi possível ajustar um modelo de regressão linear simples à relação entre a latitude — quanto menor a latitude de uma determinada região, maior a incidência de raios solares naquela região — e a taxa de mortalidade anual por melanoma maligno.

São muitos outros exemplos de problemas de regressão, e nunca terminaríamos essa nossa lista se tentássemos incluir todos os exemplos possíveis — estão presentes em praticamente todas as áreas da atividade humana: medicina, engenharia, economia, contabilidade, botânica, sociologia, geologia, direito etc.

Problemas de classificação

a. Predição de inadimplência com cartões de crédito em função da renda da pessoa, seus gastos mensais com o cartão de crédito e se ela tinha ou não um emprego estável. Fizemos isso com modelos de regressão logística na Unidade 2. Este é um problema de classificação, em que a predição é da probabilidade de a pessoa ficar inadimplente. Também vimos na Unidade 2 que há outros tipos de classificadores que, em vez de fazerem a predição de uma probabilidade para uma das classes da variável resposta, fazem a predição direta da classe. Por exemplo, neste exemplo do estudo de inadimplência com o cartão, esse tipo de classificador simplesmente diria se uma determinada pessoa ficaria ou não inadimplente. Ele classificaria cada pessoa em uma das duas classes possíveis, adimplente ou inadimplente, e não a sua probabilidade de ficar inadimplente, como fazem os modelos de regressão logística.
b. Predição de falha de um componente estrutural em função da temperatura. Este foi o exemplo dado no Pratique e Compartilhe da Unidade 2. Discutimos lá o famoso caso do acidente com o ônibus espacial Challenger, para o qual aplicamos um modelo de regressão logística, que consegue predizer a probabilidade de falha dos anéis de vedação dos foguetes de propelente sólido em função da temperatura na hora do lançamento do Space Shuttle. Esse caso foi intensamente estudado, e todos dados e discussões sobre ele estão amplamente difundidos na literatura técnica mundial. Ou seja, é um caso de domínio público.
c. Predição de se a pessoa é diabética, em função de dados gerais sobre seu quadro clínico. Um classificador pode ser alimentado com dados históricos armazenados em um banco de dados sobre pessoas sem e com diabetes. Com esses dados, ele aprenderá a fazer uma predição de se a pessoa é ou não diabética. Há muitas aplicações desse tipo na área da saúde, e são particularmente importantes em estudos de saúde pública, e valiosos também como ferramenta auxiliar no diagnóstico aplicado a pessoas com suspeita de estarem doentes; podem ser usados para quaisquer tipos de doença. Para isso precisamos, basicamente, contar com dados suficientes para treinar o algoritmo.

d. Predição de se uma determinada ação negociada na bolsa de valores subirá ou descerá no pregão de um determinado dia, em função dos resultados de subida ou queda de todas as ações negociadas no pregão do dia anterior e de alguma notícia relevante nos âmbitos político ou econômico, publicada nas primeiras horas do dia em questão, em jornais de grande circulação. Este é um algoritmo dos sonhos de muitos investidores, pois ficariam ricos com ele. E, na verdade, se considerarmos todas grandes corretoras operando no Brasil ou no mundo, elas já contam com algoritmos de alta performance que fazem esse tipo de predição. São chamados de robôs (são robôs virtuais), e, como competem arduamente entre si, não fica tão fácil acertarem sempre, pois os movimentos de um são monitorados pelos outros. Mas essas corretoras têm uma vantagem competitiva clara em comparação com investidores pequenos, que não têm tais algoritmos preditivos à sua disposição.
e. Predição de se um e-mail que chega à sua caixa de entrada é ou não um spam. Ser ou não ser um spam é uma variável qualitativa dicotômica. Graças a algoritmos de classificação, que ficam vigiando os e-mails que chegam à sua caixa de entrada, vários deles são classificados como spam e armazenados em um arquivo específico. Dessa forma previne-se que a sua caixa fique cheia com e-mails que são meros *spams*. Porém, como já dissemos, todos algoritmos têm uma certa taxa de falha. É por isso que os *spams* não são simplesmente descartados; eles ficam armazenados em uma pasta, para a eventualidade de você querer se certificar de que houve ou não um engano na classificação de alguns deles e poder transferi-los, no caso de equívoco de classificação, para a sua caixa de entrada.

Aqui também nossa lista não terminaria nunca, pois são infinitas as possibilidades de aplicação de algoritmos de classificação. Para encerrar esta seção, vamos lembrar que, assim como os seres humanos, esses algoritmos erram — tanto os de regressão quanto os de classificação. A sua performance preditiva depende de fatores como: os dados disponíveis para o seu treino, em quantidade e qualidade; do próprio jeito de funcionar do algoritmo, pois cada um tem suas particularidades; e dos critérios usados para medir sua

performance preditiva nas fases de treino e de teste. Além disso, se o fenômeno ou processo estudado mudar com o tempo, naturalmente a performance preditiva dos algoritmos treinados e testados com dados anteriores tende a piorar. Eles precisam ser, nessas situações, treinados periodicamente.

Aprendizagem de Máquina e Aprendizagem Profunda

Como já discutimos, a estatística influenciou e foi influenciada pela ciência da computação e pela ciência dos dados, com vantagens mútuas para todas essas áreas do conhecimento humano. Especificamente quanto à influência exercida pela ciência da computação, é dela que herdamos a denominação "aprendizado de máquina", em inglês, *machine learning* (ML). Essa área, pouco a pouco, foi mostrando-se muito poderosa, e mesmo técnicas já desenvolvidas anteriormente pela estatística passaram a ser classificadas conforme os jargões criados pela área de aprendizado de máquina. *Machine learning* é considerado um subcampo da grande área de estudos de IA — inteligência artificial.

saiba mais

Saiba mais

A Accenture é uma famosa empresa de consultoria internacional. No link indicado abaixo você poderá ler sobre o que ela pensa acerca do potencial de impacto de aplicações de inteligência artificial em países da América do Sul. Acesso em: 24 jan. 2020.

Fonte: Accenture (2020).

ACESSAR

Algoritmos de aprendizagem de máquina nos permitem tratar de problemas que seriam difíceis de tratar por regras criadas por seres humanos, mas que, curiosamente, ficam relativamente fáceis de serem tratados por algoritmos criados pelos seres humanos. Esses problemas também são chamados de tarefas de aprendizado de máquina e não se limitam àqueles já abordados aqui. Com o intuito de reforçar seu entendimento sobre essas tarefas, apresentamos adiante uma lista de algumas dessas tarefas (GOODFELLOW; BENGIO; COURVILLE, 2017), com mais tipos de tarefas do que já havíamos discutido antes:

- **Regressão:** Já vimos como modelos de regressão funcionam — capazes de predizer valores para variáveis resposta quantitativas.
- **Classificação:** Também já vimos como modelos de classificação funcionam — capazes de predizer uma classe da variável resposta ou a probabilidade de aquela classe acontecer.
- **Classificação com dados faltantes:** Imagine que você esteja analisando dados relativos a 100 pacientes de um hospital, e para cada um deles você coleta informações sobre 20 variáveis (idade, peso, pressão, índice glicêmico etc.). Pode acontecer que, para alguns deles, você não encontre todas as 20 informações, e, neste caso, algumas delas estarão faltando na tabela final com os dados desses 100 pacientes. Este tipo de tarefa de classificação é mais difícil de ser resolvido, quando faltam alguns dados para algumas das observações, mas isso é comum.
- **Transcrição**: Neste tipo de tarefa, pede-se que o algoritmo de aprendizado de máquina transcreva um conjunto de dados não tão bem estruturados em uma forma discreta bem estruturada.
- **Tradução:** Em tradução, chamada de tradução de máquina, a entrada é uma sequência de símbolos escritos em alguma linguagem natural. Pede-se que o algoritmo converta essa sequência em uma sequência de símbolos de outra linguagem natural. Linguagens naturais são as linguagens faladas por seres humanos.

Um outro tipo de aprendizado de máquina é a chamada aprendizagem profunda, em inglês *deep learning* (DL). Segue muitos dos paradigmas dos

algoritmos de aprendizado de máquina já descritos, porém com uma grande distinção: usa como modelos as redes neurais. Esses modelos de redes neurais foram inicialmente desenvolvidos para emular a inteligência humana, constituindo-se de um dos campos da chamada IA — inteligência artificial; em inglês AI — *artificial intelligence.* Pouco a pouco deixaram de tentar reproduzir a arquitetura biológica do cérebro humano e ganharam independência, sendo usados como algoritmos de aprendizagem em situações em que os algoritmos tradicionais de aprendizado de máquina não têm boa performance, particularmente nos campos de visão computacional (CV — *computer vision*) e processamento de linguagem natural. De alguns anos para cá, as redes neurais vêm rompendo barreiras e conquistando o mundo com aplicações dessa natureza. Os algoritmos que têm como base redes neurais são chamados de aprendizagem profunda por serem construídos em várias camadas. Quanto mais camadas, mais profunda é a aprendizagem.

Algoritmos de *deep learning* não serão estudados aqui, mas é importante você saber que existem e que estão ao seu lado o tempo todo, no seu dia a dia, tais como quando usamos tradutores de textos ou em situações de reconhecimento facial, ao sermos identificados visualmente. Por trás dessas câmeras, há um desses algoritmos nos vigiando.

praticar

Vamos Praticar

Uma *rede neural artificial* (ou rede neural) é um modelo preditivo motivado pela forma como o cérebro funciona. Pense no cérebro como uma coleção de neurônios conectados. Cada neurônio olha para a saída de outros neurônios que o alimentam, faz um cálculo e, então, dispara (se o cálculo exceder algum limite) ou não (se não exceder) [...] Redes neurais podem resolver uma variedade de problemas como

reconhecimento de caligrafia e detecção facial, e elas são muito usadas em *deep learning* (aprendizado profundo), uma das subáreas mais populares de *data science*. Entretanto, a maioria das redes neurais são “caixas-pretas” — inspecionar seus detalhes não lhe fornece muito entendimento de *como* elas estão resolvendo um problema. E grandes redes neurais podem ser difíceis de treinar. Para a maioria de problemas que você encontrará como um cientista de dados, elas provavelmente não são a melhor opção [...]

GRUS, J. *Data science do zero:* primeiras regras com o Python. Rio de Janeiro: Alta Books, 2016, p. 213.

Está correto o que se afirma em:

- ○ **a)** Redes neurais possuem alta interpretabilidade.
- ○ **b)** Grandes redes neurais são fáceis de ser treinadas.
- ○ **c)** Redes neurais não consistem em neurônios artificiais
- ○ **d)** Redes neurais servem para reconhecimento de caligrafia e detecção facial.
- ○ **e)** Redes neurais não são modelos de aprendizagem profunda (*deep learning*).

Estudo de Caso: Volume de Vendas de um Produto de Varejo

Nesta seção descreveremos um caso em que árvores de decisão são usadas para predizer o volume de vendas de um produto de varejo — uma boneca falante. Também aqui, duas personagens nos ajudarão com isso. Uma delas é a gerente comercial do fabricante dessa boneca, e a outra é uma economista especializada no mercado varejista, que domina ferramentas estatísticas. Vamos ver o que elas têm a nos contar.

Árvores de Decisão para Regressão e Classificação

Árvores de decisão são muito usadas nos campos da economia, administração, pesquisa operacional, engenharia e ciência dos dados. A forma de usar árvores de decisão nesses campos varia ligeiramente. Vamos ver aqui como elas são usadas na ciência dos dados e na estatística (JAMES *et al*., 2013).

Na ciência de dados, árvores de decisão são usadas como algoritmos preditivos, tanto para variáveis quantitativas — problemas de regressão — quanto para variáveis qualitativas — problemas de regressão. Começaremos vendo como elas funcionam quando aplicadas a problemas de regressão. Para isso, vamos usar os dados relativos aos valores dos imóveis da Unidade 1. Selecionamos aleatoriamente metade daqueles dados e, com a ajuda do software estatístico R, obtivemos a árvore exibida na Figura 3.1.

O ponto forte das árvores de decisão é sua fácil interpretabilidade. Normalmente elas são exibidas de cabeça para baixo, como essa da Figura 3.1. De cima para baixo, há um primeiro nó, que se bifurca em dois ramos, e nós intermediários, sempre com dois ramos, até chegarmos aos nós terminais, as folhas da árvore. Nos nós terminais, há a predição da variável resposta. Neste exemplo, são as predições para os valores dos apartamentos. Você lembra bem quais eram as variáveis de entrada: a área do apartamento (em metros quadrados), seu andar (1, 2, 3, ...) e sua localização (Bairro = 0 ou Centro = 1).

A leitura da árvore se faz da seguinte forma:

- Começamos no nó superior. Lá encontramos a área do imóvel como sendo a variável mais importante, neste estágio, que a árvore considerou para a predição do valor do apartamento. Se o valor for igual ou superior a 73,6 metros quadrados, caminhamos pelo ramo à direita, e o valor estimado para o apartamento é de 448,8 mil reais.

- Caminhando pelo ramo à esquerda da partição do primeiro nó, que é quando a área for menor que 73,6 metros quadrados, vemos que agora a árvore não prediz um valor, porém indica um nó intermediário, no qual está a segunda variável que ela, neste estágio, considera mais importante: a localização do imóvel.
- Se a localização for no bairro (Bairro = 0), caminhamos pelo ramo à esquerda do nó, e o valor estimado para o imóvel é de 390,4 mil reais.
- Se a localização for no centro (Centro = 1), também aqui a árvore não faz uma estimativa do valor do imóvel, mas indica outro nó intermediário, no qual está a terceira variável que ela considera, neste estágio, a mais importante: o andar do imóvel.
- Se o andar for menor que 6,5, ou seja, igual ou menor que 6, caminhamos pelo ramo à esquerda, e o valor estimado para o apartamento é de 319,2 mil reais.
- Se o andar for maior ou igual a 6,5, ou seja igual ou maior que 7, caminhamos pelo ramo à direita do nó, e o valor estimado para o apartamento é de 366,5 mil reais.

O número de nós terminais é definido por quem está construindo a árvore — o especialista, o estatístico ou o cientista de dados. Se for muito grande, a árvore fica um pouco confusa e perde sua maior qualidade, que é a sua fácil interpretabilidade. Se for muito pequena, a predição pode ficar muito grosseira. Há vários métodos e critérios de otimização da performance preditiva da árvore, mas, aqui, o nosso interesse é apenas o de apresentar essas árvores de decisão a você, mostrar como você deve “ler uma árvore de decisão”, e como elas são aplicadas à ciência de dados. Lembre que, na estatística e na ciência dos dados, essas árvores são um pouco diferentes daquelas usadas nos campos da economia, administração, na engenharia etc. (HILLIER et al., 2017). Aquelas têm nós de decisão — as nossas, aqui, não têm esses nós de decisão; e, naquelas, os nós referentes às alternativas probabilísticas podem ter mais que dois ramos — os nós inicial e intermediários das nossas árvores só têm dois ramos.

Outro ponto a deixar mais claro é como caminhar pelos ramos da árvore. Todos eles nascem de nós, ou do primeiro nó ou dos nós intermediário. Em cada um deles, há uma variável de entrada, com seu domínio bipartido. Se a variável desse nó for quantitativa, o valor da partição é um valor quantitativo. No ramo esquerdo está a predição da variável resposta para valores menores que o valor da partição, ou então outro nó intermediário; e no ramo direito está a predição da variável resposta para valores maiores ou iguais que o da partição, ou outro nó intermediário. Nos nós terminais sempre haverá a predição da variável resposta — uma quantidade, se ela for quantitativa, ou uma classe (ou nível), se ela for qualitativa.

Quando a partição no nó é feita para uma variável qualitativa, então os níveis (ou classes) indicados à direita do nome da variável são aqueles do ramo que se estende à esquerda do nó intermediário, e os níveis omissos são aqueles do ramo que se estende à direita do nó intermediário. Com um pouco de prática isso fica automático.

O exemplo de árvore que vimos aqui foi o de uma árvore de decisão aplicada a um problema de regressão, em que a variável resposta é quantitativa. As predições da árvore, que são os valores exibidos nos nós terminais, são os valores estimados da variável resposta para a qual o modelo foi desenvolvido. Uma árvore de decisão para classificação funciona da mesma forma, com a única diferença sendo a variável resposta, que deve ser agora qualitativa. Para este caso, os valores estimados da variável resposta serão seus níveis (ou classes), também exibidos nos nós terminais da árvore. Isso será visto, na prática, no estudo de caso desta unidade, que é o de predição de volume de vendas de um produto de varejo, logo em seguida nesta unidade.

Estudo de Caso: Produto de Varejo

Acima, você foi apresentado à árvore de decisão aplicada a um problema de regressão. Agora, veremos a sua aplicação a um problema de classificação, na predição de venda de um produto de varejo — uma boneca falante. A ideia da gerente comercial da fábrica que a produzia era otimizar as vendas da boneca. Para isso, ela tinha dados relativos ao volume de vendas em 200

pontos comerciais, assim como dados relativos a diversos fatores que poderiam influenciar nesse volume de vendas. Mas ela não sabia como destrinchar esses dados. Felizmente tinha uma colega economista que era especialista em varejo. Essa economista dominava técnicas estatísticas e já as havia usado, por diversas vezes, na análise de dados e no desenvolvimento de modelos preditivos. Então, combinaram que a gerente passaria todos os dados para ela, e assim foi feito. As variáveis dessa base de dados eram, para cada ponto de venda:

Volume de vendas da boneca falante (em unidades por mês).

Preço da boneca falante (em reais).

Preço da boneca do concorrente mais forte (em reais).

Gasto com publicidade (em mil reais por mês).

Idade média da população local (em anos).

Local de exposição da boneca na loja (ruim, médio, bom).

A variável resposta de interesse é o volume de vendas, que, na base de dados da gerente, era uma variável qualitativa. A economista era bastante experiente e sabia que fazer algo mais simples inicialmente poderia ser vantajoso. Então combinou com a gerente que discretizaria essa variável qualitativa, o volume de vendas em unidades por mês, transformando-a em uma variável qualitativa dicotômica, vendas altas ou baixas (Altas=1 e Baixas=0). Assim, ela não usou como variável resposta a variável quantitativa volume de vendas, em unidades vendidas por mês, mas, sim, no seu lugar, a variável qualitativa vendas altas ou baixas.

A economista organizou os dados cedidos pela gerente comercial em uma tabela, conforme o Quadro 3.1. Neste quadro, temos alguns exemplos dos dados coletados nos 200 pontos de venda durante um determinado mês.

PrecoCon (reais)	GastosPu (mil reais)	Preço (reais)	LocalExp (B, M e R)	Idade Média (anos)	Vendas Altas (SIM ou NAO)
315,00	32	287,50	Médio	37	SIM
350,00	23	347,50	Bom	50	SIM
235,00	30	192,50	Medio	41	SIM
302,50	0	332,50	Bom	63	NAO
290,00	34	360,00	Ruim	61	NAO
192,50	0	60,00	Medio	40	SIM
397,50	0	415,00	Médio	36	NÃO
...	....	...	...	...	...
277,50	0	302,50	Ruim	33	NÃO
332,50	0	210,00	Medio	40	SIM

Quadro 3.1 - Dados relativos a vendas da boneca falante
Fonte: Elaborado pelo autor.

Com isso, ela já estava em condições de prosseguir para o próximo passo, que era o de fazer uma análise descritiva dos dados antes de tentar desenvolver um modelo preditivo.

praticar

Vamos Praticar

O vice-presidente de talentos da DataSciencester entrevistou um número de candidatos para emprego do site, com níveis de sucesso variados. Ele coletou um conjunto de dados com vários atributos (qualitativos) de cada candidato, bem como se o candidato se saiu bem ou mal na entrevista. Você poderia usar esses dados para construir um modelo identificando quais candidatos farão boas entrevistas, para que ele não precise perder tempo fazendo entrevistas? Isso parece perfeito para uma árvore de decisão, outra ferramenta de modelagem de previsão no kit de um cientista de dados.

GRUS, J. **Data science do zero:** primeiras regras com o Python. Rio de Janeiro: Alta Books, 2016, p. 201.

Está correto o que se afirma em:

- ○ **a)** Não é possível usar árvore de decisão como modelo preditivo para este caso, pois não é possível usar atributos como dados de entrada de uma árvore.
- ○ **b)** Árvores de decisão não podem ser usadas como modelos preditivos para esse caso, pois o texto fala em previsão e não predição.
- ○ **c)** Para esse caso não é possível usar árvores de decisão como modelo preditivo; só modelos preditivos com base em regressão logística.
- ○ **d)** Árvore de decisão não poderia ser usada como modelo preditivo para essa situação, pois se sair bem ou mal é uma variável resposta qualitativa.
- ○ **e)** Árvores de decisão são, sim, uma boa alternativa como modelo preditivo para saber qual candidato se sairia bem ou mal na entrevista de emprego.

Análise Descritiva dos Dados

Nesta seção, veremos como a economista, especializada em comércio de varejo, fez a análise descritiva dos dados relativos às vendas da boneca falante, em 200 pontos de vendas de um determinado mês.

Análise Descritiva de Cada Variável Individualmente

Ela começou examinando em quantos pontos de venda as vendas foram altas. Para isso usou da função table() do software estatístico R aplicada à variável indicativa de vendas altas ou baixas (variável VendasAltas no Quadro 3.1) e obteve o seguinte resultado:

NAO SIM

110 90

Logo percebeu que havia um grande potencial para aumento das vendas, pois, na maior parte dos pontos de venda, as vendas foram baixas. A outra

variável qualitativa da sua amostra referia-se ao local de exposição da boneca nas lojas (variável LocalExp no Quadro 3.1). Ela aplicou a função table() do R a essa variável e obteve:

Bom Medio Ruim

45 114 41

Em seguida, decidiu examinar o preço praticado nesses 200 pontos de vendas. Para isso, ela usou a função summary() do software estatístico do R aplicada à variável preço da boneca falante (variável Preco no Quadro 3.1) e obteve:

Min. 1st Qu. Median Mean 3rd Qu. Max.

60.0 249.4 297.5 291.7 330.6 477.5

Logo percebeu também que havia uma grande variação de preços praticados por este fabricante, onde a sua colega trabalhava como gerente comercial.

Aqui estamos vendo pela primeira vez o uso da função summary() do R, uma de suas funções mais úteis. Quando aplicada a uma variável quantitativa contínua, como é o caso aqui para o preço da boneca, seu *output* sempre aparece nesta ordem:

1. o valor mínimo observado da variável em questão;
2. o valor que delimita a fronteira do primeiro quartil;
3. a mediana, que é a fronteira que delimita o segundo quartil;
4. a média (valor médio dos dados observado para a variável em questão);
5. o valor que delimita a fronteira do terceiro quartil;
6. o valor máximo observado da variável em questão.

Vamos explicar os quartis. Lembre-se de que foram coletados dados de preços de venda praticados em 200 pontos de venda em um determinado mês. O primeiro quartil delimita a fronteira até qual estão contidos 25% (um quarto) dos 200 preços amostrados. Este valor é 249,20 reais. A mediana é

igual à fronteira que delimita o segundo quartil, ou seja, o valor até qual estão contidos 50% (dois quartos) dos 200 preços amostrados. Finalmente, o terceiro quartil delimita a fronteira até qual estão contidos 75% (três quartos) dos 200 preços amostrados.

A nossa economista resolveu então analisar o preço do concorrente e aplicou a função summary() do R à variável relativa ao preço da boneca do maior concorrente (variável PrecoCon no Quadro 3.1) e obteve:

Min. 1st Qu. Median Mean 3rd Qu. Max.

192.5 290.0 313.8 314.3 338.1 437.5

Percebeu que os preços da boneca do maior concorrente variavam menos que os da sua boneca, porém o preço médio da sua boneca (igual a R$291,77) era menor que o da boneca do concorrente (igual a R$314,30). Isso certamente contribui para vendas mais altas.

Finalmente, ela realizou análise similar relativamente às variáveis de idade média da população local e de gastos com publicidade, mas resolveu que seria melhor concentrar sua atenção na relação dessas variáveis com vendas altas. Acreditava que esta análise traria informações mais interessantes. É isto que veremos na seção a seguir.

Análise Descritiva da Relação entre as Variáveis

A primeira relação que a economista decidiu examinar foi a de vendas altas (sim ou não) versus o preço da boneca. Como se trata da relação entre uma variável qualitativa e uma quantitativa, já vimos que boxplot é uma ótima opção para retratar essa relação. Foi isso que a economista fez, e obteve o gráfico exibido na Figura 3.3.

Aparentemente, há um efeito do preço no volume de vendas, mas não parece ser tão forte assim para os dados observados. Em seguida ela passou a examinar a relação entre vendas altas (sim ou não) e os gastos com publicidade (em mil reais) no mês em questão. Novamente, usou a função gráfica boxplot() do R e obteve o gráfico exibido na Figura 3.4.

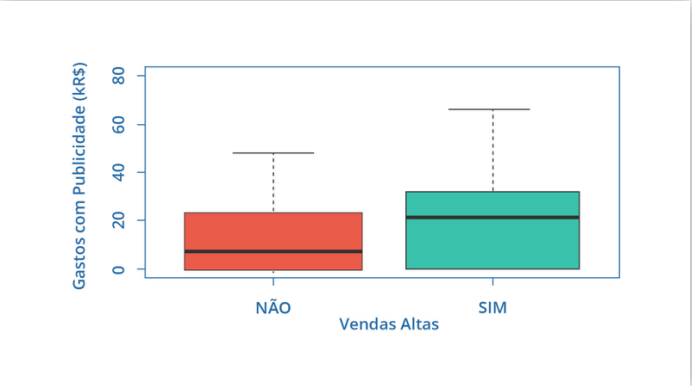


Figura 3.4 - Efeito dos gastos com publicidade nas vendas
Fonte: Elaborada pelo autor.

Parece haver um efeito dos gastos com publicidade no volume de vendas da boneca mais acentuado que o efeito do preço.

Finalmente, nesta fase de exame de relações entre variáveis, ela investigou a relação entre vendas altas (sim ou não) e o local de exposição da boneca nas lojas (bom, médio e ruim). Agora são duas variáveis qualitativas: uma

dicotômica (com dois níveis) e a outra tricotômica (com três níveis ou classes). Para investigar essa relação, a economista começou com a função table() do R e obteve:

VendasAltas

LocalExp NAO SIM

Bom 9 36

Medio 68 46

Ruim 33 8

Essa tabela que resultou da aplicação da função table() do R a essas duas variáveis indica que há um forte efeito do local de exposição no volume de vendas. Veja que 36 pontos de venda, do total de 9 + 36 = 45 pontos de vendas em que as bonecas estavam expostas em uma boa posição na loja, as vendas foram altas. Em outras palavras, nesses pontos em que a boneca estava bem posicionada na loja, 80% deles tiveram vendas altas. Por outro lado, em 8 pontos de venda, do total de 33 + 8 = 41 pontos de vendas em que as bonecas estavam expostas em uma posição ruim na loja, as vendas foram baixas. Em outras palavras, dos pontos em que a boneca estava mal posicionada na loja, só 19,5% tiveram vendas altas.

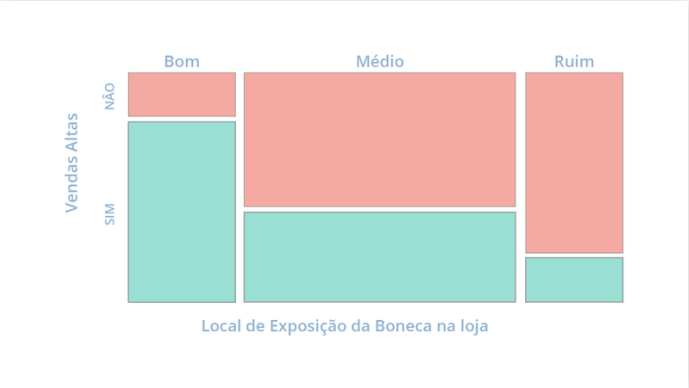


Figura 3.5 - Efeito do local de exposição nas vendas
Fonte: Elaborada pelo autor.

Isso pode ser visto graficamente por meio da função gráfica mosaicplot() do R. Foi isso que a nossa economista fez, obtendo o gráfico exibido na Figura 3.5. Esta figura é autoexplicativa e está em consonância com o resultado já discutido da função table() do R.

Ao chegar a esse ponto, a nossa economista decidiu já estar em condições de tentar desenvolver seu modelo preditivo para esta situação, para o que ela também já havia decidido usar árvores de decisão para classificação, já que a variável resposta era qualitativa (vendas altas, sim ou não), e as árvores permitem fácil interpretação.

praticar

Vamos Praticar

O quadro típico para uma análise em ciência de dados é um objeto de *dados retangulare*s, como uma planilha ou tabela de banco de dados. *Dado retangular* é basicamente uma matriz bidimensional com linhas indicando registros (caso) e colunas indicando características (variáveis). Os dados nem sempre começam dessa forma: dados não estruturados (por exemplo, texto) devem ser processados e tratados de modo a serem representados como um conjunto de características nos dados retangulares.

BRUCE, P.; BRUCE, A. **Estatística prática para cientistas de dados:** 50 conceitos iniciais. Rio de Janeiro: Alta Books, 2019, p. 5-6.

Está correto o que se afirma em:

- ○ **a)** É impossível converter dados não estruturados em representações estruturadas para que possam ser analisados pela ciência dos dados.
- ○ **b)** Dados retangulares não são uma forma típica de organização de dados para análise em ciência dos dados.

- **c)** Textos são dados estruturados, pois sempre vêm em estruturas bem padronizadas, como aquelas que estudamos em gramática.
- **d)** Dados retangulares, dados tabulares ou dados estruturados são termos usados como sinônimos na ciência dos dados.
- **e)** Dados de séries temporais, dados espaciais e dados de redes também são exemplos de dados retangulares.

Predições com Árvore de Decisão para Classificação

Nesta seção veremos a aplicação de árvores de decisão aos dados de volume de vendas da boneca falante, um produto de varejo. Será a nossa personagem economista, especializada em comércio de varejo e experiente no uso de técnicas da ciência dos dados, que desenvolverá esta aplicação.

Predição do Volume de Vendas de Produto de Varejo

Na Seção 2.1 vimos a aplicação de uma árvore de decisão a um problema de regressão. Aqui usaremos essas árvores na predição do volume de vendas de um produto de varejo, as bonecas falantes. Como a economista decidiu transformar volume de vendas, uma variável quantitativa, em vendas altas ou baixas, uma variável qualitativa dicotômica, a nossa árvore será uma árvore de decisão de classificação.

Ao fazer o ajuste do modelo de uma árvore de decisão de classificação aos dados que lhe foram fornecidos pela colega gerente comercial da fábrica das bonecas, a nossa economista obteve como resultado a árvore exibida na Figura 3.6. Já sabemos como interpretar essa árvore, então, mãos à obra:

- O primeiro nó exibe a variável que a árvore considerou mais importante. Aqui, o local de exposição do produto. Os níveis médio e ruim se referem àqueles para o ramo da esquerda. Portanto, no ramo da direita está o nível bom para o local de exposição da boneca. Caminhando por este ramo, vemos que o próximo nó, um nó intermediário, exibe a variável preço. Se o preço ficar abaixo de 356,50 reais, as vendas serão altas. Se o preço ficar acima de 356,50 reais, as vendas serão baixas.
- Partindo do nó inicial, se o local de exposição da boneca for ruim ou médio, devemos caminhar pelo ramo esquerdo, até encontrarmos a variável preço. Se o preço for menor que 242,25 reais, o ramo esquerdo chega a um nó terminal com a predição de vendas altas para a boneca nessas condições (local de exposição médio ou ruim, porém preço abaixo de 242,25 reais).
- Voltando ao nó inicial, e ainda no cenário em que o local de exposição da boneca é ruim ou médio, devemos caminhar pelo ramo esquerdo e encontraremos novamente o preço. Vamos ver agora qual a predição se o preço, neste nó intermediário, for maior ou igual a 242,25 reais. Neste caso devemos caminhar pelo ramo à direita, até

encontrarmos a variável gastos com publicidade. Se, nessas condições, os gastos com publicidade estiveram abaixo de 78 mil reais ao mês, as vendas serão baixas. Porém, se os gastos com publicidade forem maior ou igual a 78 mil reais ao mês, encontraremos no ramo da direita a variável idade média da população local.
- Continuando deste ponto, se a idade média for abaixo de 44 anos, ou seja, uma população relativamente jovem com vários adultos com crianças que demandam brinquedos e bonecas, aí as vendas serão altas. Porém, se a idade média for maior ou igual a 44 anos, neste caso as vendas serão baixas.

Bem, é dessa forma que uma árvore de decisão faz predições. No caso aqui estudado, predições dos valores de uma variável qualitativa dicotômica — vendas altas ou baixas para um produto de varejo. A nossa economista elaborou um relatório e comunicou esses resultados à sua colega gerente comercial, que percebeu ter, agora, uma ferramenta muito boa para saber como otimizar as vendas desse seu produto. E ficou muito feliz com isso!

Teste da Performance Preditiva do Modelo

Mas o nosso caso não termina aqui. As duas combinaram testar o modelo com novos dados. A gerente tinha consigo dados relativos ao mês seguinte, em que as condições do mercado estavam idênticas às do mês para o qual a árvore foi treinada.

Para fazer esse teste, a economista simplesmente usou a função table() do R, aplicando-a aos resultados corretos de vendas altas e baixas naquele mês e aos valores altos e baixos preditos pela árvore para aquele mesmo mês, e obteve o seguinte resultado:

pred NAO SIM

NAO 91 35

SIM 19 55

Vamos ler esse resultado. São duas as linhas dessa matriz, NAO e SIM, que se referem às predições feitas pela árvore quando alimentada como os novos dados relativos aos 200 pontos de vendas. Na linha do NAO, ou seja, vendas baixas, a árvore acertou ao estimar 91 pontos comerciais com vendas baixas, mas errou ao dizer que 35 desses pontos comerciais tiveram vendas baixas, pois na verdade esses pontos tiveram vendas altas. Na linha do SIM, ou seja, predição de vendas altas, a árvore acertou ao dizer que 55 desses pontos comerciais tiveram vendas altas, mas errou ao dizer que 19 desses pontos tiveram vendas altas, pois na verdade eles tiveram vendas baixas. Do total de 200 predições que ela fez, acertou 91 + 55, levando a uma acurácia de (91 + 55) / 200 = 0,73 ou 73%. Não é uma performance excelente, mas já é boa o suficiente para ajudar a gerente comercial. A economista sabia como aumentar a capacidade preditiva dessas árvores, com algoritmos chamados de florestas randômicas, que são muitas árvores funcionando ao mesmo tempo. Mas isso ficará para uma outra oportunidade!

Só mais uma observação. Se você achou um pouco difícil interpretar essa matriz que mede a performance preditiva de um classificador determinístico, como a das árvores de decisão para classificação, não se assuste. Todos temos essa mesma dificuldade — é daí que o nome oficial dessa matriz é “matriz de confusão”.

praticar

Vamos Praticar

Os modelos de árvores, também chamados de *Árvores de Classificação e Regressão*, *árvores* de decisão ou apenas *árvores*, são um método de classificação (e regressão) efetivo e popular, inicialmente desenvolvido por Leo Breiman e outros em 1984. Os modelos de árvores e seus descendentes mais potentes, florestas aleatórias e

boosting, formam a base das ferramentas de modelagem preditiva mais potentes e amplamente usadas na ciência de dados tanto para regressão quanto para classificação.

BRUCE, P.; BRUCE, A. **Estatística prática para cientistas de dados:** 50 conceitos iniciais. Rio de Janeiro: Alta Books, 2019, p. 226.

Quanto a este assunto de árvores de decisão para classificação e regressão, analise as afirmativas a seguir:

i. Uma árvore faz partição recursiva das variáveis de entrada, selecionando uma de cada vez, de forma hierárquica, das mais importante às menos importantes, a cada estágio da sua construção, até chegar aos nós terminais, suas folhas, que exibem os valores estimados para a variável resposta.
ii. A cada estágio da construção da árvore, o nó inicial e depois os nós intermediários dividem o domínio da variável de entrada em questão, de onde bifurcam os seus ramos para a esquerda e para a direita. Os valores exibidos no nó indicam como se deve ler a árvore, ao se caminhar pelos ramos à esquerda ou à direita do nó em questão.
iii. Em cada um dos nós intermediários, assim como no nó inicial, há um valor quantitativo ou qualitativo, que representa o valor escolhido pela árvore para fazer a partição da variável tratada naquele estágio da construção da árvore.
iv. Em árvores de decisão de classificação ou regressão, folha é um termo que designa os nós terminais das árvores, nos quais são exibidos os valores estimados para a variável resposta do modelo preditivo. Cada caminho da árvore indica, dessa forma, o valor a estimar para aqueles valores das variáveis de entrada informadas no caminho do nó inicial até a folha.

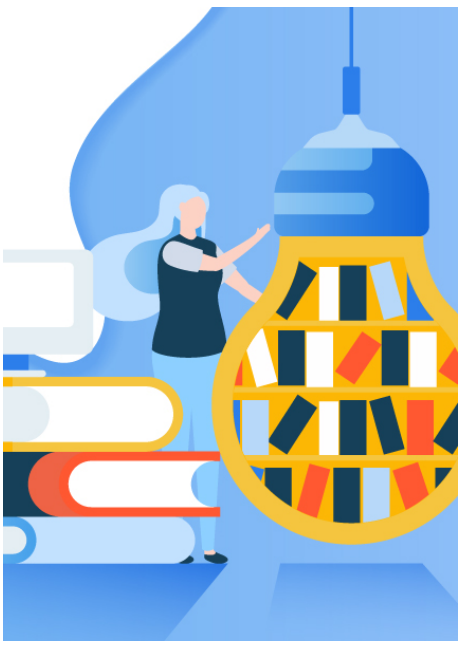
Está correto o que se afirma em:

- ○ **a)** II, III e IV, apenas.
- ○ **b)** II e III, apenas.
- ○ **c)** III e IV, apenas.
- ○ **d)** I, II, III e IV.

○ **e)** I, II e III, apenas.

indicações

Material Complementar



LIVRO

Estatística prática para cientistas de dados: 50 conceitos iniciais

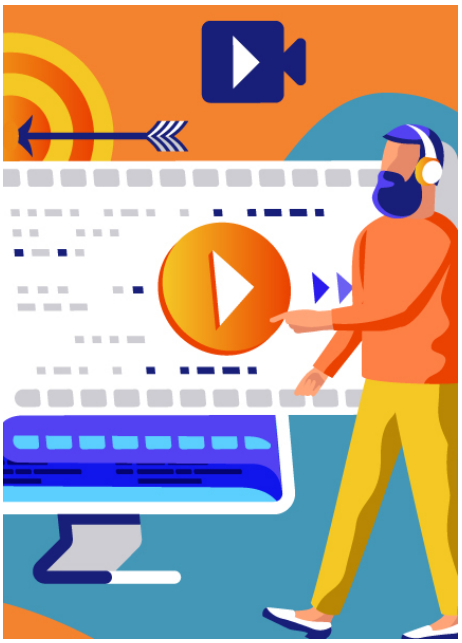
Peter Bruce & Andrew Bruce

Editora: Alta Books

ISBN: 978-85-508-0603-7

Comentário: A ciência de dados pode e é praticada por profissionais de diferentes áreas de conhecimento, tais como contadores, geólogos, engenheiros, biólogos, sociólogos, linguistas, economistas, biomédicos, advogados etc., alguns com maior familiaridade com conceitos estatísticos, outros com menos familiaridade. Esse livro se propõe a apresentar conceitos da estatística frequentes na ciência dos dados, de forma

progressiva e didática. Isso não eliminará a necessidade de esforço e dedicação, pois é um processo de aprendizado paulatino, mas o recomendamos àqueles de vocês que tiverem interesse em praticar a ciência dos dados na sua área de atuação. Você começa hoje e, dentro de alguns meses, já colherá frutos.



FILME

O jogo da Imitação

Ano: 2014

Comentário: Esse filme nos conta sobre os trabalhos e a vida de Alan Turing, um matemático inglês, durante a Segunda Guerra Mundial. Alan Turing, no filme, desenvolve uma máquina cujo propósito é decodificar as mensagens dos nazistas. Ele é considerado o pai da computação contemporânea, assim como um dos fundadores do campo da inteligência artificial.

TRAILER

conclusão

Conclusão

Nesta unidade, vimos um outro tipo de classificador: uma árvore de decisão. Usamos esse algoritmo, que é muito popular na estatística, na ciência dos dados e na mineração de dados, para uma tarefa de classificação. Mais especificamente, fizemos uma análise de predição de vendas de um produto de varejo, uma boneca falante, e vimos como a árvore pode nos ajudar a entender quais variáveis contribuem para um aumento das vendas do produto em questão. Com isso terminamos, aqui, a primeira parte do nosso passeio pelo mundo das aplicações da estatística à ciência dos dados: a modelagem preditiva. São técnicas de muito poder, que nos auxiliam em praticamente todas as áreas da atividade humana. Na próxima unidade, daremos atenção a um outro assunto: os modelos de aprendizagem não supervisionada, que têm um jeito diferente de funcionar. Vamos lá!

referências

Referências Bibliográficas

INTELIGÊNCIA ARTIFICIAL PARA acelerar o crescimento da América do Sul. Accenture. s. d. Disponível em: https://accntu.re/2tPc14t . Acesso em: 08 dez. 2019.

BELLE, G. V. *et al.* **Bioestatiscs**: a methodology for the health sciences. 2. ed. Hoboken: John Wiley & Sons, 2004.

BRUCE, P.; BRUCE, A. **Estatística prática para cientistas de dados:** 50 conceitos iniciais. Rio de Janeiro: Alta Books, 2019.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep learning.** Massachusetts: MIT Press, 2017.

GRUS, J. **Data science do zero:** primeiras regras com o Python. Rio de Janeiro: Alta Books, 2016.

HILLIER, F. S. *et al*. **Introduction to operations research.** 10. ed. Tamil Nadu: McGraw Hill Education, 2017.

JAMES, G. *et al*. **An introduction to statistical learning:** with applications in R. Nova York: Springer, 2013.