Machine Learning

Elton Massahiro Saito Loures

2020-12-02

Contents

Prefácio

0.1 Por que ler esse livro?

0.2 Estrutura

0.3 Informações a respeito do conteúdo

0.4 Agradecimentos

```
install.packages("bookdown")
# or the development version
# devtools::install_github("rstudio/bookdown")
```

Remember each Rmd file contains one and only one chapter, and a chapter is defined by the first-level heading `#`.

To compile this example to PDF, you need XeLaTeX. You are recommended to install TinyTeX (which includes XeLaTeX): https://yihui.name/tinytex/.

Chapter 1

Introdução

Caro leitor, se você veio até esse livro é bem provável que passou e/ou ainda passa pelas mesmas dificuldades que todo estudante interessado nessa área.

Ao elevado número de pesquisas que fiz para aprender o que era a Inteligência Artificial, o que era o *Machine Learning* (Aprendizado de Máquina) e todos os outros temas similares, é nítido que ainda não está totalmente definido o conceito de cada um. É um ramo novo na área acadêmica, na indústria e em todo o mercado, com diversos temas, diversos modelos matemáticos, diversos modelos computacionais, diversos *softwares*, diversas aplicações e em diversas áreas. Diversos "diversos"... E o mais assustador é que esse campo une todos esses "diversos", tornando o universo **caótico** ainda maior. Quando destaco o termo "caótico", refiro exatamente pela ironia deste mote, todo esse universo confuso é aplicado em nosso cotidiano para organizar, analisar, diagnosticar e facilitar as coisas.

Poucos instruem como devemos enxergar todo esse cosmos que ao longo da história está passando por diversas construções para estruturar seu conceito. Com uma tentativa de trazer isso com base em artigos, livros, vídeos, podcasts e cursos, disponho este simples livro com o propósito de organizar a imagem que você, leitor, tem de Aprendizado de Máquina e entender os principais modelos utilizados tanto no meio acadêmico, quanto no mercado de trabalho.

1.1 Dicas de estudo

Não cabe a mim dizer como estudar, mas o que posso lhe aconselhar como principal ponto é a **paciência**. Temas como esse podem abranger qualquer campo, desde a filosofia até a área da saúde e portanto, do mesmo modo que se aplica a qualquer conteúdo, o mais importante é a base. Leia, releia, pesquise, veja vídeos, ouça um podcast, converse e discuta com colegas e professores a

respeito. Não se cobre de que precisa aprender o mais rápido possível, mas preze a qualidade do estudo.

Com intuito de explicar sobre Aprendizado de Máquina. Na seção **AQUI VOU COLOCAR A REFERENCIA DA SESSAO**, para faciliar o leitor dependendo de sua demanda de conteúdo, busquei separar em subseções a lógica computacional e a matemática. Tornando mais prático para o público que não tem interesse no modelo matemático e que busca o conhecimento de determinado assunto quanto ao público que demanda esse conteúdo.

1.2 Dicionário

- **Escalares e Vetores:**
- **Espaço Vetorial e Transformação Linear:**
- **Assimetria e Curtose:**
- **Variância e Desvio padrão (Erro padrão):**
- **Covariância:** A covariância mede a relação linear entre duas variáveis. É possível utilizar a covariância para compreender a direção da relação entre as variáveis. Valores de covariância positivos indicam que valores acima da média de uma variável estão associados a valores médios acima da outra variável e abaixo dos valores médios são igualmente associado. Valores de covariância negativos indicam que valores acima da média de uma variável estão associados com valores médios abaixo da outra variável.
- **Distribuição normal:**
- **Distribuição binomial:**
- **Disitrbuição de Poisson:** (Banzatto and Kronka, 1992) Quando número de plantas daninhas por parcela, número de insetos capturados em armadilhas luminosas, número de pulgões ou ácaros por folhas, etc.
- **Teorema de Bayes:** quando tratamos de probabilidades, $P(A|B)$ e $P(B|A)$ podem ser parecidos, mas possuem grande diferença entre as probabilidades que representam. Por exemplo $P(A|B)$ pode se referir sobre a probabilidade de uma pessoa que cometeu um furto (B) ser condenada (A) e $P(B|A)$ seria a probabilidade de uma pessoa que foi condenada por furto ter efetivamente cometido um crime. A causa se torna o efeito e o efeito se torna a causa (Freund, 2009).

Pela regra geral de multiplicação que afirma que a probabilidade da ocorrência de dois eventos é o produto da probibilidade da ocorrência de um deles pela probabilidade condicional da ocorrência do outro evento, temos:

$$P(A \bigcap B) = P(A).P(B|A) \text{ ou } P(A \bigcap B) = P(B).P(A|B) \tag{1.1}$$

Igualando ambas expressões, temos: \$ P(A). P(B|A) = P(B). P(A|B)\$ e portanto, divindo por $P(B)$, obtém-se o Teorema de Bayes que descreve a probabilidade de um evento, baseado em um conhecimento a *priori* que pode estar relacionado ao evento:

$$P(A|B) = \frac{P(A).P(B|A)}{P(B)} \tag{1.2}$$

Para B_n e A_k atributos, podemos reescrever:

$$P(A_k|B_1, ..., B_n) = \frac{P(A_k).P(B_1, ..., B_n|A_k)}{P(B_1, ..., B_n)} \tag{1.3}$$

Exemplo: este exemplo pode ser encontrado em Freund (2009). Numa certa empresa, 4% dos homens e 1% das mulheres têm mais de 1,75m de altura, respectivamente, sendo que 60% dos trabalhadores são mulheres. Um trabalhador é escolhido ao acaso.

A) Qual a probabilidade de que tenha mais de 1,75m?

Solução: Temos de informação de que 60% dos trabalhadores são mulheres e que 1% delas possuem mais de 1,75m. Portanto 40% dos trabalhadores são homens, sendo 4% deles com mais de 1,75m. Logo temos que:

$P(> 1,75m) = (0,04.0.4)+(0,01.0.6) = 0,022 \rightarrow 2,2\%$ de probabilidade de que tenha mais de 1,75m.

B) E que seja homem dado que o trabalhador escolhido tenha mais de 1,75m?

Solução: pelo enunciado "que seja homem dado que o trabalhador escolhido tenha mais de 1,75m", podemos perceber que já possuímos uma afirmação que já foi escolhido uma pessoa que tenha mais que 1,75m e queremos saber se é homem. Por meio da questão anterior sabemos a probabilidade P(> 1,75m). Portanto:

$$P(H| > 1,75m) = \frac{P(> 1,75m|H).P(H)}{P(> 1,75m)} = \frac{0,04.0,4}{0,022}$$

$\rightarrow 72,73\%$ de probabilidade de ser homem dado que seja maior que 1,75m.

- **Função de verossimilhança:** a verossimilhança L de um conjunto de parâmetros θ, com dada informação x. É igual a probabilidade da mesma observação x ter ocorrido dados os valores dos mesmos parâmetros θ. Conhecendo um parâmetro θ, a probabilidade condicional de x é $P(x|\theta)$, mas se o valor de x é conhecido, pode-se realizar inferências sobre o valor de θ (Bolfarine and Sandoval, 2001).

$$L(\theta|x) = P(x|\theta) \tag{1.4}$$

Para "n" valores:

$$L(\theta|x_1, ..., x_n) = \prod_{i=1}^{n} P(x_i|\theta) \tag{1.5}$$

Geralmente utiliza-se o logaritmo natural em verossimilhança $L(\theta|x) = lnL(\theta|x)$ como função suporte e facilitar em seu estudo.

Para facilitar a compreensão, considere a observação de que você esteja ouvindo barulho em sua sala de estar num dia de natal (observação x), você parte da hipótese inicial que poderia ser o "Papai Noel" lhe entregando presentes (hipótese θ). A probabilidade de ser Noel lhe entregando presente apenas porque ouviu o barulho, isto é, $P(\theta|x)$ é baixa. No entanto o contrário, você com a afirmação de que é o Noel lhe entregando presentes, a probabilidade de haver barulho em sua sala de estar é bem alta, logo a verossimilhança $L(\theta|x) = P(x|\theta)$.

- **Parâmetros:** podem ser vistos como características númericas de um modelo ou população. Os valores não podem ser mensurados diretamente mas que podem ser estimados através dos dados de uma amostra.
- **Paramétrico:**
- **Correlação:**
- **Supervisionada x Não supervisionada:**

Chapter 2

Inteligência Artificial (IA)

2.1 O que é IA? De onde veio esse conceito?

Humano (taxonomicamente Homo sapiens), termo que derivado do latim "homem sábio". Pensamos, analisamos, aprendemos , prevemos e manipulamos. Somos seres **inteligentes**. Já pesquisou o significado de "inteligência" no dicionário?

É importante entender o conceito de inteligência, pois nem tudo que o ser humano faz pode ser classificado como inteligente. Aprender somar para calcular a soma de $2+2$ é uma ação inteligente, mas copiar o resultado e colocar em sua folha de resultados que é 4 pode não ser tanto assim. Da mesma forma uma calculadora que executa um código passado por um humano, contendo dentro todos os passos a serem executados (algoritmos) para resolver esse cálculo, não é considerada.

Quando tratamos da inteligência artificial não é fácil definir o que ela é. O seu próprio conceito vem sendo discutido e moldado ao longo do tempo. A idéia de construir uma máquina pensante ou um ser artificial que se assemelhasse aos humanos é muito antigo. O mito do Golem, por exemplo, um dos primeiros seres artificiais criados pelo homem. Dizia a lenda que o mito do Golem surgiu no século XIII quando uma matéria informe tornou-se num homúnculo a partir da invocação mágica de Elijah de Chelm que escreveu em sua fronta *"Shemhamforash"* - nome secreto de Deus (MOSER, 2006). Na literatura foi publicado o famoso romance *Frankensteins* (Shelley, 1818) que relata a história de um estudante que constrói um monstro em seu laboratório. Mas como ela realmente surgiu?

O primeiro trabalho a ser reconhecido como IA foi elaborado por McCulloch and Pitts (1943) que tinha como propósito estudar como os neurônios podiam funcionar, modelando uma rede neural simples com circuitos elétricos. Os mesmos

autores sugeriram que as redes neurais definidas em conformidade poderiam ser capazes de aprender. Por seguinte, Hebb (1949) escreveu *The Organization of Behavior* que fortalecia as teorias de que o condicionamento psicológico estava presente em qualquer parte dos animais. Teve como a premissa de que dois neurônios participantes de uma sinapse, têm ativação simultânea, então a força da conexão entre eles deve ser seletivamente aumentada, ou seja, os caminhos neurais são fortalecidos cada vez que são utilizados.

Em 1950, o matemático Claude E. Shannon publicou um artigo sobre como "ensinar" seu computador a jogar xadrez (Shannon, 1950); no mesmo ano Alan Turing, em "Computing Machinery and Intelligence" (TURING, 1950), sugeriu que, ao invés de perguntarmos se as máquinas podem pensar, devemos perguntar se as máquinas podem passar por um teste de inteligência comportamental, o teste de Turing. Uma forma de avaliar se uma máquina consegue se passar por um humano em uma conversa por escrito com um avaliador passando no teste caso o avaliador não conseguisse identificar se estava conversado com um computador ou com outro ser humano. No ano seguinte, os estudantes Marvin Minsky e Dean Edmonds construíram o SNARC, o primeiro computador de rede neural que simulava uma rede de 40 neurônios.

Em 1956 houve a conferência de verão em Dartmouth College (Hanover, New Hamphire), foi oficializada o nascimento da IA. John McCarthy, Minsky, Claude Shannon e Nathaniel Rochester elaboram uma proposta a fim de reunir pesquisadores dos Estados Unidos interessados em teoria de redes neurais, autômatos e estudo da inteligência:

> Propusemos que um estudo de dois meses e dez homens sobre inteligência artificial fosse realizado durante o verão de 1956 no Dartmouth College, em Hanover, New Hampshire. O estudo foi para prosseguir com a conjectura básica de que cada aspecto de aprendizado ou qualquer outra característica da inteligência pode, em princípio, ser descrita com tanta precisão a ponto de que uma máquina pode ser feita para simulá-la. Será realizada uma tentativa para descobrir como fazer com que as máquinas usem a linguagem, a partir de abstrações e conceitos, resolvam os tipos de problemas hoje reservados aos seres humanos e se aperfeiçoarem. Achamos que poderá haver avanço significativo em um ou mais desses problemas se um grupo cuidadosamente selecionado de cientistas trabalhar em conjunto durante o verão.
>
> — "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence" , McCarthy et al. (2006) , Agosto de 1955.

Entre diversas ideias e apresentações, Allen Newell e Herbert Simon apresentaram o programa logic theorist, capaz de provar diversos teoremas e segundo Simon, capaz de pensar não numericamente. Apesar de muitos editores não se agradarem, esta importante proposta trouxe nos próximos anos, uma dominação nesse campo (Russel and NORVIG, 2004):

- General Problem Solver (GPS), projetado por Newell and Shaw (1959), é um sistema que buscava imitar o homem na forma de resolver problemas. Concluíram de que a forma em como dividia um objetivo em sub objetivos e possíveis ações era similar à forma em como o homem fazia. Esta pesquisa ajudou a estabelecer os fundamentos teóricos dos sistemas de símbolos e forneceram à área da IA uma série de técnicas de programação voltadas à manipulação simbólica, por exemplo, as técnicas de busca heurística;

- a IBM produz alguns dos primeiros programas de IA, entre os quais, em 1959 o Geometry Theorem Prover;

- Arthur Samuel desenvolveu um programa capaz de jogar damas ao nível de um jogador de torneio. O programa jogava melhor do que o seu autor;

- John McCarthy no MIT, em 1958, define a linguagem de programação Lisp (List Processing) que se transformou na linguagem dominante da IA e publicou um artigo intitulado "Programs with common sense" (McCarthy, 1968), onde descrevia um programa hipotético designado por "Advice taker", o qual pode ser visto como o primeiro sistema completo da IA;

- Slagle (1963), com o programa SAINT, foi capaz de resolver problemas de cálcuo integral;

- Evans (1964) e Bobrow (1967), com os respectivos programas ANALOGY e STUDENT, resolviam problemas de análises geométricas semelhantes aos testes de QI e problemas clássicos de álgebra.

- Em base de Huffman (1971), Waltz (1975), Winograd (1972), Winston (1970) e Fahlman (1974), foi elaborado o mundo de blocos, que consiste em um conjunto de blocos sólidos colocados sobre uma mesa de modo que a mão de um robô reorganize-os.

Claro que os primeiros sistema houveram dificuldades com problemas mais difíceis. Desde traduções que exigiam conhecimento profundo para solucionar ambiguidades, por exemplo, como situações de necessidade de hardwares melhores e limitações fundamentais nas estruturas simples. Com ressalva, em *Perceptrons* (Minsky and Papert, 1969) demonstra que embora suas redes neurais simples (*perceptrons*) pudessem aprender, eram capazes de representar muito pouco. Mas com exigência da formalização acadêmica na década de 70, permitiu o desenvolvimento de sistemas com grande desempenho intelectual com perspectivas industriais e comerciais, surgindo novos sistemas dispostos a resolver problemas mais complexos do que antes:

- DENDRAL (Buchanan et al., 1969), analisa compostos orgânicos a fim de determinar sua estrutura molecular;

- MYCIN (Buchanan and Shortliffe, 1984), Sistema pericial (expert system) foi capaz de diagnosticar infecções no sangue.

E sucessivamente foi crescendo este enorme e maravilhoso campo. O japão lança o projeto "*Fifth Generation*" para construir em dez anos computadores inteligentes com capacidade de fazer milhões de inferências por segundo em 1981; uso de IA na guerra do Golfo em 1991; sistemas de perícia para casos médicos no mesmo ano; sistemas para condução de veículos automotores e detectores de colisões nas ruas (1993); reserva de viagens (1994); brinquedos inteligentes (2000); computador que se comunica ao nível de uma criança com 15 meses (2001). Ao longo dos anos da história da ciência da computação, a ênfase em *algoritmos* e tratamento de dados vem aumentando.

2.2 A arte de uma IA

Atualmente, existem muitas atividades, pesquisas e aplicações em diversos temas que muitas vezes nem perbemos:

- Recomendações de mídia: com base em seu perfil de uso, o algoritmo compara filmes, músicas, clips, etc com base em vários usuários que possuem os gostos similares ao nosso. Recomedando aquilo que provavelmente irá nos agradar. Por exemplo Spotify, YouTube e Netflix.
- Reconhecimento de fala e assistentes virtuais: já refletiu sobre como funciona sua Google Assistente? Com ondas sonoras emitidas pela voz, o algoritmo reconhece palavras, frases e até mesmo o timbre, fornecendo respostas de acordo com o que recebe.
- Jogos: a inteligência artificial desenvolvida pela OpenAi conseguiu derrotar uma das melhores equipes do Dota 2 do mundo.
- Logística: a crise de 1991, por exemplo, no Golfo Pérsico. Foi utilizada a DART (Cross and Walker, 1994), uma ferramenta que envolveu até 50.000 veículos, transporte de carga aérea e pessoa simultâneamente com o objetivo de realizar um planejamento logístico automatizado levando em conta rotas, pontos de partida e resolução de conflitos.
- Reconhecimento de imagens: identificação de objetos, pessoas, animais e qualquer figura com base em exemplos prévios, como por exemplo identificador de pessoas em uma foto do Facebook.
- Verificação de compras: detecção de comportamentos suspeitos a partir do histórico e perfil do usuário, como a e-commerce.
- Automóveis autônomos: por meio do algoritmo, visualiza a estrada, as placas, condição climática, outros veículos e diversos outros obstáculos para tomar decisões de seu trajeto sem a necessidade de uma pessoa.

Poderíamos falar desde exemplos de inteligência artificial aplicados casos jurídicos, diagnósticos na área da saúde, identificadores de *fake news* (notíficas falsas) até a robótica. É uma extensa lista de exemplos na área que até hoje estão em

desenvolvimento em busca de cada vez mais melhorar. A AGI (Artificial General Intelligence), ou Inteligência Artificial Geral, trabalha na criação de uma inteligência artificial generalista, similar a humana, capaz de ser especialista em uma área, mas também aprender com facilidade outras. Uma área que se tornou uma das principais linhas de pesquisa e nos dias de hoje gera discussões sobre até onde a IA pode alcançar.

Chapter 3

Vertentes de uma IA e fundamentação filosófica

> Os filósofos têm estado por aí há muito mais tempo que os computadores e vêm tentando resolver algumas questões que se relacionam à IA: como a mente funciona? É possível que as máquinas ajam com inteligência, de modo semelhante às pessoas, e, se isso acontecer, elas realmente terão mentes conscientes? Quais são as implicações éticas de máquinas inteligentes?
>
> "Inteligência Artificial", RUSSEL and Norvig (2013).

Com todo o desenvolvimento da IA, os algoritmos podem funcionar em níveis humanos em tarefas que aparentemente envolvem julgamento humano ou, como Turing acrescentou, "aprender a partir da experiência" e a capacidade de "distinguir o certo do errado"(RUSSEL and Norvig, 2013). Paul Meehl (Meehl, 1954) analisou os processos de tomada de decisão de especialistas treinados em tarefas subjetivas como prever o sucesso de um aluno em um programa de treinamento ou a reincidência de um criminoso e descobriu que algoritmos simples de aprendizado estatístico fizeram previsões melhores que os especialistas.

A reflexão sobre "máquinas inteligentes e pensantes" é recente em nossa história e passa por longas discussões sobre o alcance dessa inteligência. Desde a classificação elaborada pelo filósofo John Searle em 1980, tomou-se na doutrina em geral a divisão do uso da inteligência artificial em **"fraca"** e **"forte"** (Searle, 1980).

A inteligência artificial **fraca** "nos permite formular e testar hipóteses de forma mais rigorosa e precisa", no entanto, ela é dependente da inserção do conhecimento fornecido pelo ser humano que a programa. A máquina não é capaz de produzir raciocínios próprios, autônomos (Searle, 1980; Guimarães, 2019). Seartle também explica que a máquina adequadamente preparada é realmente uma

mente, no sentido de que os computadores que recebem os programas certos poderiam estar, literalmente, preparados para compreender eter outros estados cognitivos (Searle, 1980).

Searle (1980) em seu *naturalismo biológico*, critica a inteligência artificial forte pois, segundo ele, as máquinas não possuem a complexidade de sistema nervoso, neurônios com axonios e dendritos e tudo mais. Para corroborar sua crítica, Searle descreve uma situação hipótetica simulando um programa que passa pelo teste de turing e que "não entende nada de suas entradas e saídas", não havendo os requisitos para ser considerada uma mente.

O sistema foi nomeado como **"quarto chinês"**. Ele se usa como exemplo com a situação de que não tem conhecimento da língua chinesa, estaria trancada e isolado num quarto recebendo uma folha de papel com ideogramas em chinês escritos. Por não conhecer a língua, não possui ideia alguma do que se trata. Em seguida, ele recebe uma seguda folha com ideiogramas chineses acompanhados de um conjunto de regras em inglês (língua nativa) que permitem a correlação da segunda folha com a primeira. Por fim, recebe uma terceira folha com ideogramas chineses, com regras em inglês que orientam a dar em respostas específicos ideogramas chineses associados a outros ideogramas da terceira folha, correlacionando os elementos da atual com as duas anteriores. As pessoas externas do quarto denominam a terceira folha como o "script", a segunda folha de "história" e a primeira folha de "questões". Essas pessoas consideram que os símbolos que Searle entregou em resposta à terceira folha são as "respostas às questões" e todo o conjunto de regras que lhe foi entrega são o "programa" (Guimarães, 2019).

Com o tempo Searle se torna melhor em dar respostas de acordo com as regras que permitem manipular os ideogramas chieses e de maneira similar ocorre com os programadores externos do quarto, que ficam bos em escrever os programas do ponto de vista externo. Qualquer pessoa que observa as resposta de Searle não contestaria de que Searle não fala chinês. Da mesma forma se o mesmo experimento fosse feito com textos em inglês, sua língua nativa, ele daria respostas em patamares semelhantes (Guimarães, 2019).

Searle conclui que no caso em chinês ele opera como um computador, respondendo corretamente mas sem a menor ideia do que está respondendo. Ao caso em inglês, ele irá responder como um ser humano e com consciência de suas respostas. O quarto se refere ao computador, o ser humano ao *software* de IA. Com isso ele assume que só seria possível produzir artificialmente uma máquina com sistema suficientemente semelhante a nós se poder duplicar exatamente as causas e seus efeitos, assim de fato seria possível produzir consciência, intencionalidade (fenômeno biológico dependente da bioquímica específica de suas origens) e tudo o mais usando princípios químicos diferentes dos usados por seres humanos (Searle, 1980) .

Em contestamento a Searle, Daniel Dennett defende o projeto de Turing porque agir inteligente consiste na capacidade de processamento de informação (Den-

nett, 2009). Segundo Dennett, o problema da mente deve ser abordado com base na teoria evolutiva darwiniana pois o que entendemos por mental está relacionado ao tipo de resposta que nosso organismo dá para as demandas que estão para além daquelas que dizem respeito à manutenção da vida (da Silveira, 2013). Para ele, como ele denomina de *intencionalidade intrínseca*, Seartle errou em atribuir aos humanos a intencionalidade produzida exclusivamente pela interação das partes que constituem uma totalidade complexa, não necessitando de influências ou interferências externas. Para Dennet nossa intencionalidade não é original (Dennett, 2009).

Para Dennet o principal argumento criticando o argumento do quarto chinês, é a forma como investigamos os fenômenos mentais. É uma região que possibilita infinitas especulações, sendo o método das ciências empíricas o mais apropriado ao estudo da mente (da Silveira, 2013).

A diferença entre ambos é de natureza filosófica com ontologias e epistemologias divergentes. É notável a importância das discussões filosóficas. O antagonismo dicotômico dos dois filósofos possuem fundamentações que auxiliam na compreensão da mente. Quando teremos estas respostas? As máquinas serão capazes de raciocinar algum dia? Até onde uma IA pode chegar?

Chapter 4

O Aprendizado de Máquina

Agora que entendemos o conceito e a origem de uma IA, podemos entrar no tão esperado **Machine Learning (ML)**. Alguns pensam erroneamente ser algo distinto de uma IA, mas é importante entender que ela é um campo específico da inteligência artificial que tem como base a ideia de que sistemas podem aprender com dados e iterações, identificar padrões para que aprimorem seu desempenho diante de problemas específicos e possam tomar decisões com a menor intervenção humana possível. Como modelos estatísticos, busca entender a estrutura dos dados modelos que atendam a certos pressupostos - muitas vezes não temos conhecimento de como essa estrutura se parece.

Samuel (1959), engenheiro do MIT popularizou o termo *"Machine Learning"* (Aprendizado de Máquina), descrevendo o conceito com "um campo de estudo que dá aos computadores a habilidade de aprender sem terem sidos programados para tal" (Simon, 2013). Com a expansão da internet e seu abundante armazenamento de dados na web, o *Big data*, foi necessário - ainda é - aprimorar sistemas de organização, classificação, análise de dados e identificação de padrões para tratá-los. Isso fez com que o Aprendizado de Máquina entrasse em destaque e passasse a ser uma das áreas mais importantes. Na seção 2 foi apresentado alguns exemplos de aplicações de IA, o mesmo se aplicam para o ML.

Um aprendizado de máquina **não** é o mesmo que uma lista de instruções. Imagine uma criança aprendendo a andar de bicicleta, ela pode até receber algumas instruções para melhorar seu aprendizado, mas provável que ela irá aprender melhor com a tentativa e erro. Pedala, cai, levanta, pedala novamente e assim sucessivamente até ela realmente saber andar. Da forma similar ocorre com o Aprendizado de Máquina.

4.1 Como a máquina aprende?

Para facilitar a compressão, imagine que você é um vendedor e está interessado em clientes "bons pagadores" e "maus pagadores". Para cada cliente, possui um conjunto de dados como: idade, quantidade de faturas pagas antes do vencimento nos últimos 12 meses, quantidade de faturas atrasadas nos últimos 12 meses, região que reside, tempo de cadastro, etc. Você já se encontra com um banco de dados muito grande de clientes com seus respectivos dados e classificações como bons pagadores e maus pagadores e pretende utilizar um algoritmo de ML para aprender com esses dados de modo que, quando você receber o banco de dados de um novo cliente, esse algoritmo pode prever se a tendência desse cliente seria de bom pagador ou mau pagador.

Primeiramente, você iria alimentar seu algoritmo de ML com os dado históricos que passaram por toda uma análise se havia dados faltantes, redundantes, etc e já classificados entre cliente bom pagador e mau pagador e suas respectivas características para treiná-lo. Com estes dados o algoritmo irá aprender por meio de com quais condições são necessárias para o cliente ser classificado como bom pagador ou mau pagador. Importante ressaltar que existem diferentes algoritmos de Aprendizado de Máquina que poderiam resolver esse problema, de acordo com modelos estatísticos e comandos computacionais que atendam a certos pressupostos.

- **Como verificarmos se os dados já estão bons para aplicar o algoritmo? Quais modelos podemos aplicar? Como sabemos que essas previsões são confiáveis? Como evitar problemas de um modelo ruim?**

Chapter 5

Pré-processamento

Para o profissional que trabalha com Aprendizado de Máquina ou outras áreas, embora exigindo boa parte do tempo nesta etapa, é uma das mais importantes. O pré-processamento é um conjunto de atividades que buscar preparar, organizar e estruturar o banco de dados (*dataset*) para que possa trabalhar com os dados. Ela torna a informação de seus dados mais consistentes, com organização rígida e geralmente classificados de acordo com o seu formato (caracteres, binários, númericos, etc). Podemos dizer que ele é um conjunto de técnicas do campo de **Mineração de dados (*Data mining*)**, uma outra área além de Inteligência Artificial (que engloba Aprendiza de Máquina) – que já é grande por si só - , que trat-se de uma outra dimensão de estudos e metodologias, isso sem falarmos de outros campos além destes dois. Neste tópico, vamos abordar algumas delas que são muito utilizadas nesta área. Note que em todos os procedimentos de Aprendizado de Máquina existe inúmeras metodologias para serem aplicadas em cada etapa e, de acordo com o interesse do pesquisador, pode ser utilizado diferentes estratégias com diferentes combinações. Não há uma receita de bolo, sabemos que precisamos extrair dados, pré-processalos (aplicar uma(s) estratégia para analisar, classificar os atributos, eliminar os redundantes, preencher ou eliminar os faltantes), desenvolver seus modelos de Aprendizado de Máquina, treiná-los e por fim, avaliar todo o seu modelo. É... Não é fácil, mas todo esse procedimento é fundamental para que se obtenha um modelo adequado. Portanto nesta seção busquei separar em alguns tópicos para facilitar a compreensão, porém entenda que **TODAS** as metodologias e estratégias podem ser combinadas e estão entrelaçadas. É como vários conjuntos em um *Diagrama de Venn* que estão dentro do Pré-processamento que está dentro de Mineração de dados e que está interseccionada com Aprendizado de Máquina (dentro de IA).

Não se assuste: No último capítulo deste livro estará um diagrama e uma explicação mais "cronológica" de todo esse cosmos, com suas "gálaxias" e sistemas "solares" de conteúdo.

5.1 Dados faltantes e a Limpeza de dados

Durante o desenvolvimento destes modelos é comum se deparar com dados faltantes em seu banco de dados e que podem ser ocasionadas por razões diversas como não preenchimento cadastral, problemas de armazenamento de dados ou até mesmo situações aleatórias não identificadas. A escolha da forma de tratar esses dados faltantes é fundamental para o modelo. Os valores faltantes total quando todas as informações são perdidas ou parcial quando somente uma parte delas são perdidas

(Little and Rubin, 2019), descrevem que os motivos de aparecimento de dados faltantes são comumente classificados em:

1. ***Missing Completely at Random (MCAR)***: neste caso, as observações faltante surgiram de maneira aleatória, portanto as razões para as perdas não são relacionadas às respostas do sujeito. O uníco problema gerado pelos dados faltantes é a perda de poder da análise a ser realizada. Por exemplo, um jovem que deixou de responder uma questão de sua prova sem querer, sem motivo algum.
2. ***Missing at Random (MAR)***: os dados faltantes dependem das variáveis preenchidas e, portanto, podem ser totalmente explicadas pelas variáveis presentes no conjunto de dados. É possível não viesar a análise, considerando as informações que causam estes dados faltantes. Como por exemplo uma pesquisa elaborada por uma universidade com a finalidade de analisar a renda das mulheres em sua cidade porém não possui recursos financeiros suficiente para entrevistar todas as mulheres. A pesquisa é respondida por uma parcela de mulheres na cidade e todas as envolvidas estão com os dados completamente observados, seria analisado uma amostra aleatória de mulheres.
3. ***Missing Not at Random (MNAR)***: nesta situação os dados faltante são gerados de forma não mensurável, isto é, de eventos que o pesquisador não consegue observar e não tem controle. É o pior caso e algumas vezes, é necessário técnica mais robustas. Em geral,dados situados nos extremos da distribuição são mais propensos a serem faltantes (muito baixos ou altos em relação ao padrão da amostra).

5.1.1 Tratamento de dados faltantes

Existem diversas metodologias de tratamentos em dados faltantes. Quando os dados são faltantes em um conjunto de dados, existem cinco grandes categorias de tratamento de análise que um pesquisador deve escolher. Como mencionado anteriormente e ainda reforço, a escolha do tratamento de análise de dados faltantes tem implicações importantes para a acurácia e o viés das estimativas. A Tabela 1 resume as definições e os maiores problemas nos cinco tipos de análises (de Andrade et al., 2019).

Table 5.1: **Tabela 1** : Metodologia de dados faltantes. Determinados termos estão na seção 1.2 e alguns outros serão apresentados ao longo do livro.

Técnicas de Análise para dados faltantes	**Definições**	**Maiores Problemas**
Listwise Deletion	Exclui todos os casos para os quais alguns dados estão faltando	Descarta dados de respondentes com respostas parciais. Menor amostra, menor potência. Viés em MAR e MNAR.
Pairwise Delection	Calcula as estimativas (médias, EP, correlações) usando todos os casos disponíveis com dados relevantes para cada estimativa.	Diferentes correlações representam misturas de subpopulação. Às vezes, a matriz de covariância não é definida positiva. Viés em MAR e MNAR. Nenhuma amostra faz sentido para a matriz de correlação (EP impreciso).
Imputação Simples	Preenche cada valor faltante, por exemplo média, por regressão, etc.	A imputação média (entre casos) e a imputação por regressão são ambas tendenciosas sob MCAR! Nenhuma amostra faz sentido para a matriz de correlação (EP impreciso). EP's subestimados se você tratar o conjunto de dados como completo.

Técnicas de Análise para dados faltantes	Definições	Maiores Problemas
Máxima Verossimilhança (MV)	Estima diretamente os parâmetros de interesse a partir de uma matriz de dados incompleta; ou calcula estimativas como média, desvio padrão, ou correlação usando algum algoritmo.	Não-viesada sob MCAR e MAR. Melhora à medida que adiciona mais variáveis ao modelo de imputação. Número de variáveis deve ser menor que 100. EP'S preciso para FIML. para o algoritmo EM, nenhuma amostra faz sentido para a matriz de correlação (EP impreciso).
Imputação Múltipla (IM)	Imputa valores faltantes várias vezes, cria-se m conjuntos de dados completamente imputados. Executa a análise em cada conjunto de dados imputado. Combina os m resultados para obter estimativas de parâmetros e erros padrão.	Imparcial sob MCAR e MAR. Melhora à medida que adiciona mais variáveis ao modelo de imputação. O número de variáveis deve ser menor que 100. EP's precisos. Fornece estimativas ligeiramente diferentes a cada vez que analisa os dados. Em Equações Estruturais, piora a convergência.

- ***Listwise* deletion**: exclui todos os casos para os quais alguns dados estão faltando . A eliminação dos casos frequentemente reduz muito o tamanho da amostra e o poder estatístico do teste de hipóteses. Importante o pesquisador se atentar que mesmo quando o poder do teste parece adequado, este método pode produzir estimativas de parâmetros tendenciosas sob dados faltantes sistemático (MAR e MNAR). O *listwise deletion* restringe a população-alvo do estudo, assim em geral quase nunca se utiliza esse procedimento. Uma vez que ele descarta dados que custaram tempo, disponibilidade dos participantes e até mesmo recursos financeiros, a eliminação desses participantes da pesquisa pode violar o princípio ético da pesquisa (Rosenthal, 1994).

Resumo geral: elimina todos os casos que possuem dados faltantes em sua pesquisa.

- ***Pairwise* deletion**: este método tenta minimizar a perda que ocorre em *Listwise deletion*. Como exemplo a matriz de correlação. Uma correlação como explicada em 1.2, mede a força da relação entre duas variáveis. Para cada par de variáveis para os quais os dados estão disponíveis, o coeficiente de correlação indicará a força. Em *Listwise* será o mesmo tamanho para todas as correlações excluindo toda observação faltante, em *Pairwise deletion* irá variar. Ela exclui apenas os casos que não tem respostas completas dentro da observação, aproveitando o maior número de casos possíveis.

Resumo geral: ao invés de eliminar as observações (coluna ou linha inteira da matriz) com dados faltantes, como *listwise deletion*, este metodo elimina apenas os casos que não tem respostas completas nas combinações das observações, aproveitando o maior número possível.

- **Imputação simples***: envolve o preenchimento de cada dado faltante com uma suposição de qual deve ser o valor que está faltando no conjunto de dados. Os exemplos mais comuns de imputação simples são: imputação pela média - substituição de cada valor faltante pela média do grupo para a variável correspondente; imputação* hot deck* - substituição de cada dado faltante por um valor "doador" que possui um escore similar em outras variáveis; e imputação por regressão – substituindo cada valor faltante por um valor predito com base em um modelo de regressão múltipla (será explicado conceito de regressão posteriormente), obtido a partir dos valores observados (de Andrade et al., 2019). A maioria das técnicas de imputação simples é tendenciosa. Por exemplo, a imputação pela média insere uma média constante para cada valor faltante, as estimativas da variância e da correlação serão tendenciosas – mesmo que o mecanismo de dados faltantes seja completamente aleatório (MCAR). A imputação por regressão leva à subestimação da variância e superestimação da correlação (pois os valores imputados estarão exatamente na linha de regressão). Pode-se melhorar ao caso de regressão adicionando um termo de erro aleatório aos valores imputados (regressão estocástica), no entanto, ainda são imprecisas. Ao caso dos testes de hipóteses, não estima com precisão o erro padrão (de Andrade et al., 2019).

Resumo geral: envolve o preenchimento de cada dado faltante com uma "boa adivinhação" de qual deve ser o valor que está faltando no conjunto de dados, sendo essa estimação de acordo com o pesquisador e sua pesquisa (média, regressão, etc).

- **Imputação múltipla (IM)**: cada valor faltante é substituído por dois ou mais valores imputados e ordenados a fim de representar a incerteza sobre qual valor imputar, permitindo que as estimativas das variâncias estimadas sejam calculadas com dados completos (Rubin, 2004). Assim, m imputações atribuídas a cada valor faltante gera n conjuntos de dados completados que são analisados inerente aos valores observados da amostra.

Muitos utilizam este método, visto que aumenta a eficiência de estimação, facilita o estudo direto da sensibilidade de inferências, abrange uma variedade de análises e geralmente válidas por incorporar incertezas devido à falta de dados. Tornando-os mais eficientes que a imputação simples, porém mais trabalhosa e ocupa mais espaço de armazenamento. Em desvantagem desse método,pode surgir discrepância na variância quando se admite pressupostos equivocados (modelo escolhido não consistente com os dados), com isso um m pequeno se torna mais adequado com menor gravidez. Uma das característica mais importantes desse método é que os valores faltantes para cada envolvido é predito a partir de seus próprios valores observados, com o ruído aleatório adicionado para preservar uma correta quantidade de variabilidade nos dados imputados (Schafer and Graham, 2002).

Schafer (1999) recomenda que a quantidade necessária de imputações para que a estimativa de conjunto de dados tenha relativa eficiência, com a seguinte equação:

$$RE = \sqrt{1 + \frac{\lambda}{m}} \tag{5.1}$$

onde, é m é a quantidade do conjunto de dados completados e λ é a taxa de informação - caso fosse 50% dos dados faltantes, $\lambda = 0,5$.

Claro que o método para mensurar a quantidade necessária **varia de acordo com o tema da pesquisa e a escolha do pesquisador**. Dependendo área que o pesquisador está interessado, pode-se haver outras recomendações para mensurar a quantidade.

A IM é composto basicamente por três passos (Assunção, 2012):

1. **Imputação dos dados:** são gerados **m** bancos de dados completos através de técnicas adequadas que devem levar em conta ao máximo a relação entre os dados faltantes e os observados. Existe diversos métodos que podem ser utilizadas para este primeiro passo, um dos mais utilizados atualmente é o método de **regressão linear bayesiana** - ao caso de não entender o que são as técnicas de Regressão linear nem de Bayes, as seções XXXXXXXXXXX instruem.

Este método tem como resposta a variável que possui dados faltantes (Y) e como variáveis preditoras são utilizadas as demais variáveis presentes ($X_1, X_2, ..., X_k$), com k número de preditoras. Na abordagem Bayesiana, a regressão linear é formulada através de distribições de probabilidade ao invés da abordagem clássica. Seu modelo será:

$$Y_i \sim N(\beta^T X_k, \sigma^2 I)$$

A variável dependente Y_i é gerada a partir de uma Distribuição Normal (Gaussiana) 1.2 caracterizada pela média e variância (σ^2. A média é o produto entre

os parâmetros β e variáveis independentes X_k. O objetivo deste método é determinar a distribuição posterior para os parâmetros do modelo ao invés de encontrar um único valor. A resposta e seus parâmetros são gerados por meio de uma distribuição de probabilidade.

Para encontrar as distribuições dos parâmetros do modelo, a inferência bayesiana utiliza o Teorema de Bayes para combinar informações prévias ao experimento e dados de amostra com o objetivo de deduzir as propriedades sobre um parâmetro de interesse a partir dos dados de entrada X_k e de saída Y. A aplicação de Bayes neste contexto seria:

$$P(\beta|y, X) = \frac{P(y|\beta, X)P(\beta|X)}{P(y|X)} \tag{5.2}$$

onde $P(\beta|X)$ reflete a incerteza de β. Qualquer informação que se tenha inicialmente sobre o parâmetro é tratado como ela (pode ser utilizada como não informativa). Em $P(y|\beta, X)$ é a verossimilhança que diz respeito a distribuição característica dos dados (interpretada como no caso clássico). O denominado $P(y|X)$ é tratada como uma constante de normalização para a equação e reflete a probabilidade que pode-se obter qualquer dado.

Ressalto que existe diversos métodos nesta primeira etapa e recomendo o leitor interessado, buscar outras literaturas.

2. **Análise dos bancos de dados gerados pelo passo 1:** ao criar o conjunto de dados imputados, é importante fazer uma análise separadamente para cada um dos m banco de dados da mesma forma como tradicionalmente se faz, o modelo pode variar de acordo com o pesquisador - são apresentadas na seção SEIS AQUI COLOCAR A SEÇÃO DEPOIS.

3. **Combinar os resultados:** com as análises realizadas, precisa-se combinar os resultados apropriados para obter a inferência da imputação repetida. Por meio do passo 2, obtém-se estimativas para o parâmetro de interesse D. Estas estimativas podem ser qualquer medida escalar como médias, variâncias, correlações, coeficientes de regressão por exemplo. A estimativa D será a combinação será a média das estimativas individuais.

$$\overline{D} = \frac{1}{m}\sum_{s=}^{m} \hat{D}_s \tag{5.3}$$

Em seguida, a variância combinada é calculada:

$$T = \overline{E} + (1 + \frac{1}{m})F \tag{5.4}$$

em que $\overline{E} = \frac{1}{m}\sum_{s=}^{m} E_s$ é a média das variâncias que preserva a variabilidade natural (E) do parâmetro de interesse nos m banco de dados e $F = \frac{1}{(m+1)}\sum_{s=}^{m}(\hat{D}_s - \overline{D})^2$ o componentes que estima a incerteza causada pelos dados faltantes. Se F for muito pequeno as estimativas dos parâmetros são muito semelhantes, com menos incerteza. Do contrário as incertezas variam muito.

Resumo geral: a imputação múltipla executa uma rotina de imputação simples repetidamente (múltiplas associações sobre os valores plausíveis) e consegue estimar sem víes o erro padrão. Ocorre as imputações muitas vezes contabilizando a imprecisão de cada imputação.

- **Método de máxima verossimilhança (EM - Expecativa-maximização)**: proposto por Fisher (1912) , é um método paramétrico (ver 1.2) que parte do princípio de especificar como a função de verossimilhança (ver 1.2) deveria ser utilizada como um instrumento de redução de dados Casella and Berger (2010). Este método consiste na escolha do conjunto de valores para os parâmetros que torne um máximo a função de verossimilhança. A inferência de verossimilhança pode ser considerada como um processo de obtenção de informação sobre um vetor de parâmetros θ, a partir do ponto x do conjunto amostral, por meio da função de verossimilhança. Vários vetores podem produzir a mesma verossimilhança, reduzindo a informação de θ (Cordeiro, 1999).

O objetivo é encontrar uma estimativa do parâmetro θ, $\hat{\theta}$, que maximize a verossimilhança. Portanto, utiliza-se o conceito de derivada (diferenciação) e igualamos a zero (Bolfarine and Sandoval, 2001).

$$L'(\theta; x) = \frac{\delta L(\theta; x)}{\delta \theta} = 0 \tag{5.5}$$

Para inferir se é um ponto máximo, aplica-se a segunda derivada e verificar se o resultado é menor que zero (Bolfarine and Sandoval, 2001).

$$L''(\hat{\theta}; x) = \frac{\delta^2 log L(\theta; x)}{\delta \theta^2} < 0 \tag{5.6}$$

Com algoritmo EM (Expectativa-maximização), por Dempster et al. (1977) é um procedimento que realiza a estimativa dos parâmetros (vetor de médias e a matriz de covariância) por meio da máxima verossimilhança em conjuntos amostrais incompletos (dados faltantes) e pode ser utilizado como uma ferramenta para inserção de dados. Por um processo iterativo, na etapa E(Estimação/Esperança) se estima os dados faltantes para completar a matriz dos dados, no caso calcula-se a esperança condicional (média condicional) da função de log-verossimilhança; no passo M (Maximização), com os dados

completados, encontra-se um $\hat{\theta}$ que maximiza a esperança condicional da log-verossimilança e então seu resultado é usado para fazer a inferência no passo E e assim sucessivamente até que o algoritmo processado tenha convergido, ou seja, a diferença entre o valores da verossimilhança dos dados incompletos na k-ésima e na $(k+1)$-ésima iteração seja tão pequena (Enders, 2010, ; Pereira, 2019).

Resumo geral: o algoritmo EM, faz a etapa E com a função de verossimilhança para encontrar um valor médio e preencher os dados faltantes, faz a etapa M utilizando a máximização de verossimilhança para encontrar um valor médio com o menor erro possível e continua, a partir do resultado do segundo passo, sucessivamente até convergir no melhor valor e menor erro possível (global) para preencher os dados faltantes.

Além de dados faltantes, é possível lidarmos com grande volume de dados. Por isso, o processamento computacional se torna cada vez mais complexo e para aumentarmos a eficiência e reduzir os custos usamos o processo de redução de dados ou a hierarquização para separarmos os conjuntos a serem estudados. Pode-se por meio de **Agregação de cubo de dados** (atividade de construção de um cubo de dados) que apesar de gerar maior necessidade de armazenamento, permite um processamento mais rápido por não necessitar varrer toda a base em busca de determinado valor. A **Seleção de subconjuntos de atributos** para utilizar os atributos altamente relevantes em detrimento dos menos relevantes (como por exemplo verificar pela significância). Ou também **reduzir a numerosidade** ou **dimensionalidade** que permitem que os dados seja estimados por alternativas de representação de dados menores e compactados e alguns métodos para hierarquizar as variáveis. Na seção de XXXXXXXXXXXX serão apresentados as principais estratégias.

5.1.2 *Outlier*

Um *outlier* é um valor que se encontra distante da normalidade e que provavelmente causará anomalias nos resultados obtidos, pois pode viesar negativamente todo o resultado de uma análise e que seu comportamento pode ser justamente o que está sendo procurado. São basicamente dados que se diferenciam drasticamente dos outros, conhecidos como anomalias, pontos fora da curva, dados discrepantes, ruídos, e que estão fora da distribuição normal.

Pode-se verificar dados incomuns apenas verificando a taebla, mas dependendo do tamanho de seu banco de dados não é uma boa recomendação. Uma das melhores maneiras de identificarmos dados *outliers* é utilizando gráficos. Ao plotar um gráfico o analista consegue verificar que existe algo diferente. Como exemplo, um estudo no sistema de saúde brasileiro pela AQUARELA (2017) utilizando dados da prefeitura de Vitória no Espírito Santo, analisando fatores que levam as pessoas a não comparecerem em consultas agendadas no sistema público de saúde da cidade. Padrões encontrados de que mulheres comparecerem muito mais que os homens e crianças faltam poucos às consultas, porém,

uma senhora *outlier*, com 79 anos agendou uma consulta e com 365 dias de antecedência apareceu à consulta. Neste caso, convém ser estudado o *outlier* pelo comportamento trazer informações relevantes que podem ser adotadas para aumentar a taxa de assiduidade nos agendamentos. *Outlier* do caso indicado pela seta vermelha 5.1.

feitos/Livro ML/bookdown-demo-master/Figuras/outlier.bb

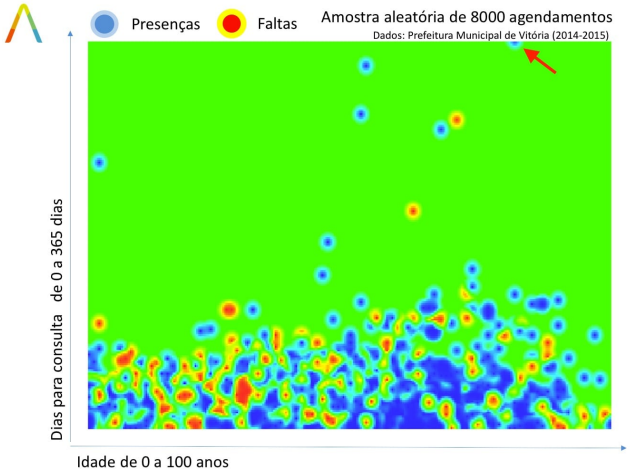


Figure 5.1: "Gráfico de estudo no sistema de saúde apresentando *outlier* (AQUARELA, 2017)."

Por diversos motivos pode ocorrer de ter presença de *outlier* nos dados e podem viesar negativamente todo resultado de uma análise e seu comportamente pode muitas vezes ser o que justamente o pesquisador está procurando. Há possibilidade do *outlier* ser importante para o pesquisador entender o por que da anomalia estar acontecendo, ou para identificar algum dado extraído erroneamente, por exemplo.

Uma maneira mais complexa e muito precisa, é de identificá-los através de análise dos dados. Encontrando a distribuição estatísticas que mais e aproxima à distribuição dos dados e utilizar métodos estatísticos para detectar as anomalias. Como por exemplo o uso de histograma e a distribuição normal para verificar os dados que estão dentro e fora do intervalo de confiança (ver 1.2 Distribuição normal).

5.2 Transformação de dados

5.2.1 Tipos de *datasets*

A escolha das medidas estatísticas para sua análise ou modelo de Aprendizado de Máquina dependem muito dos tipos de dados das variáveis em observação.

Estes tipos de dados podem ser numéricos (como uma sala de aula, com alunos que variam sua altura de 1,51 metros a 1,98 metros) e categórico (como uma classificação num hospital de pacientes doentes ou não doentes), embora esses dois tipos podem ser subdivididos como números inteiros e ponto flutuante para variáveis numéricas e booleano, ordinal ou nominal para variáveis categóricas.

As subdivisões mais comuns são: - Variáveis Numéricas: 1. Variáveis inteiras (exemplo: $1, 2, 3, ..., n$); 2. Variáveis de ponto flutuante (parte fracionária, por exemplo: 1,17; 0,10; 47,2).

- Variáveis categóricas:
  1. Variáveis booleanas (dicotômicas, binárias: Verdadeiro e Falso).
  2. Variáveis ordinais (1^{o}, 2^{o}, 3^{o}, etc).
  3. Variáveis nominais (não possuem ordenação como por exemplo, cor dos olhos: azuis, castanhos, pretos e verdes).

Importante ressaltar que quando trabalhamos dentro da programação, possuem mais tipos além de *int* (númericos inteiros) *char* (caracteres) e *float* (pontos flutuantes), como o *double* que armazena números com ponto flutuantes com precisão dubla com o dobro da capacidade de *float*, *string* como cadeia de caracteres.

Muitos algoritmos possuem a limitação de trabalhar somente com atributos qualitativos (variáveis categóricas), com isso muitas vezes é necessário aplicar algum método capaz de transformar um atributo quantitativo em um atributo qualitativo (faixas de valores). Uma estratégia que cresce ao longo do tempo é o processo de **discretização** que transforma atributos contínuos em atributos discretos como por exemplo, dividir alturas entre menor que 1,70 metros e maior igual que 1,70 metros. Dependendo do estudo pode ser adequado, embora o pesquisador precisa tomar muito cuidado pois é provável que possar perder algumas informações. De mesmo modo, é possível transformar variáveis categóricas em númericas, como por exemplo classificar tamanhos como pequeno = 1, médio = 2 e grande = 3 possibilitando por meio do mapeamento manter a ordem dos valores (Batista et al. (2003)).

É bem comum estes tipos de tratamento de dados ao caso de datas, como trabalhos que aplicam-se **séries temporais** em que o pesquisador precisa estudar a sazonalidade de algum objeto de estudo. A soja por exemplo pode-se analisar sua tendência ao longo dos anos, mas quando tratamos os dados e analisamos em outro período podemos verificar que possui sazonalidades em sua produção. Em análises para investimentos também, atentar o comportamento mensal e diário das ações de uma empresa, muitas vezes está com tendência de alta num âmbito mensal, porém ao analisar diariamente é possível que esteja em baixa.

Para facilitar a compreensão, considere a série temporal *AirPassengers* que representa o número de passageiros mensalmente em uma empresa de transporte aéreo ao período de 1949 a 1960 (Box and Jenkins, 1976).

Para o campo de transformação de dados e séries temporais, ao leitor que pre-

feitos/Livro ML/bookdown-demo-master/Figuras/airpassengers.bb

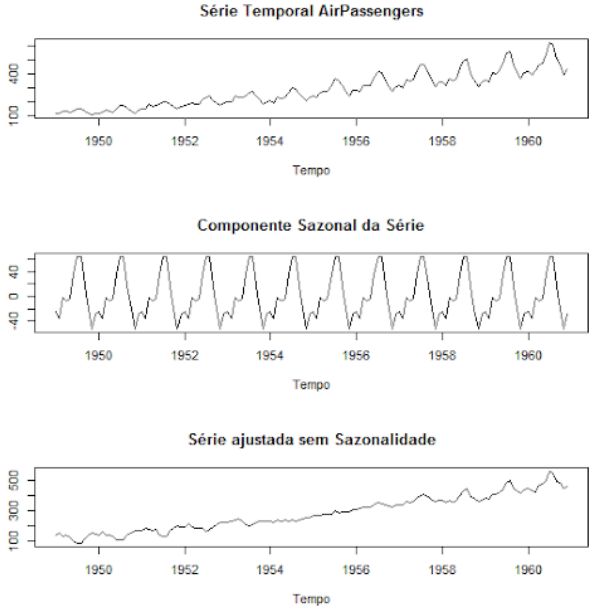


Figure 5.2: "Número de passageiros tratados mensalmente (Box and Jenkins, 1976)."

tende ir mais a fundo nestes outros "galhos" de estudos. Recomendo buscar outras literaturas que tem como foco este temas. Em discretizações por exemplo, Dougherty et al. (1995) e Garcia et al. (2012) abordam diversos métodos que podem agradá-lo.

5.2.2 Normalização e padronização

Muitos conjuntos de dados apresentam atributos contínuos que espalham-se em diferentes faixas de valores ou possuem distintas variações, devido às suas naturezas ou escalas em que foram medidas. Estas diferenças podem ser fundamentais e precisam ser levadas em conta (CARVALHO et al., 2011). Em situações também para validarmos a análise variância precisa-se dos requisitos de atiditividade, independência, normalidade e homogeneidade de variâncias - será apresentada em ANOVA seção XXXXXXXX. Quando alguma das características mencionadas acontece ou não verifica seus requisitos o pesquisador, antes de fazer uma análise não-paramétrica (1.2), pode-se transformar seus dados (Banzatto and Kronka, 1992).

1. **Normalização por reescala:** através de um valor mínimio e um máximo, gera um novo intervalo onde os valores de um atributo estão contidos. Um intervalo entre 0 e 1.

$$x_{ij} = \frac{x_{ij} - min_j}{max_j - min_j} \tag{5.7}$$

sendo x_i a observação de ordem i, min_j e max os valores mínimos e máximos do atributo j respectivamente.

2. **Transformação de raiz quadrada:** frequentemente utilizada para dados de contagens que geralmente segue uma distribuição de Poisson (1.2), onde a média é igual à variância (Banzatto and Kronka, 1992).

$$\sqrt{x_i} \tag{5.8}$$

sendo x_i representando as observações do banco de dados. Quando ocorrem zeros ou valores baixos (menores que 10 ou 15), recomenda-se $\sqrt{x + 0,5}$ ou$\sqrt{x + 1,0}$ (Banzatto and Kronka, 1992).

3. **Transformação angular:** recomenda-se para dados expressos em porcentagens, que geralmente seguem a distribuição binomial (1.2). Atualmente existe tabelas apropriadas para essa transformação (Banzatto and Kronka, 1992). Segundo Banzatto and Kronka (1992) porcentagens entre 30% e 70% ou as porcentagens são resultantes da divisão dos valores observados nas parcelas por um valor constante tornam-se desnecessárias e pode-se analisar diretamente os dados originais, mas atente-se pois algumas vezes variar essas exceções de acordo com sua área e pesquisador que a propõe.

$$arc\ sen\sqrt{\frac{x}{100}} \tag{5.9}$$

4. **Transformação logaritmica:** quando verificada determinada proporcionalidade entre as médias e desvios padrões dos diversos tratamentos. É geralmente utilizada para problemas de assimetria (1.2). Em casos, por exemplo, tratamentos com amplitude alta como uma população numerosa que varia de 1.000 a 10.000 indivíduos ou tratamentos de baixa amplitude de 10 a 100 indivíduos. Esta trasformação pode ser útil.

$$log(x)\ \text{ou} ln(x) \tag{5.10}$$

Uma vez transformados os dados em logaritmos, a soma de dados logarítmicos não tem o mesmo valor que a soma de seus antilogaritmos, mas representa o produto destes.

5. **Padronização:** é um método muito utilizado por diversas áreas de pesquisa. Neste caso diferentes atributos podem abranger diferentes intervalos, porém possuir os mesmos valores para alguma medida de posição e de variação (CARVALHO et al., 2011). Imagine você como economista interessado em avaliar o desempenho da produção de soja com as variáveis econômicas e monetárias o Brasil e possui as seguintes variáveis: produção de soja anual medida em milhares de toneladas, taxa básica de juros SELIC medida em porcentagem, receita média anual em milhares de reais, área plantada de soja medida em hectares. Já podemos perceber que todos possuem medidas e grandezas bem diferente uma das outras. Este o propósito da padronização, deixar com que todas as variáveis tenham uma medida em comum.

$$x_{ij} = \frac{x_{ij} - \overline{X}}{S_j} \tag{5.11}$$

em que $\overline{X_j}$ e S_j representam a média e o desvio padrão do atributo j respectivamente. Após a transformação todos os atributos terão a média zero e desvio-padrão unitário.

Caso transformado seu banco de dados e seu banco de dados apresentarem uma distribuição contínua não-normal, ou não-homogênea ou não-aditiva, não há outra alternativa senão utilizar a estatística não-paramétrica.

Resumo geral: Muitos conjuntos de dados apresentam atributos contínuos que espalham-se em diferentes faixas de valores ou possuem variações diferentes, por motivo de suas naturezas ou escalas medidas. Estas diferenças podem ser muito importantes e precisam ser levadas em conta para não causar erros em sua pesquisa. Para isso usam-se alguns métodos para transformar seus dados para que possam ser trabalhados, apresentados os principais neste livro. Em

situações para fazermos análise variância precisa-se também ser transformado seus dados caso não cumpra seus requisitos. Caso o problema ainda persistir, precisa-se utilizar estatística não-paramétrica.

5.3 Features Selection - Seleção de atributos (SA)

Uma literatura que achei bastante interessante foi Parmezan et al. (2012). Seguindo sua estrutura a respeito de Seleção de atributos. Podemos definir SA como a determinação de um subconjunto ótimo de atributos, partindo de algum critério ou medida de importância, que representa a informação importante dos dados (Parmezan et al., 2012). Extraímos um subconjunto de P atributos a partir de um conjunto original de N atributos, sendo $P \leq M$ (Parmezan et al., 2012; Liu and Motoda, 1998; Lee, 2005). A cada conjunto de dados com M atributos, existem 2^M subconjuntos de atributos candidatos (Langley et al., 1994).

Existem diversas metodologias para selecionarmos os atributos que podem variar em sentido de buscas e estratégias para a seleção. Repare que os tópicos mencionados anteriormente também são utilizados para remoção e seleção, foi fragmentado apenas para facilitar a compreensão.

O "sentido de busca" influencia na determinação do(S) ponto(s) de partida no espaço de busca, ou seja, na direção em que a busca será realizada e os operadores que serão utilizados. Elas são categorizadas, seguindo Parmezan et al. (2012) e Liu and Motoda (2008), em:

• **Forward Selection - Seleção para Frente:** o estado inicial é estabelecido como vazio (subconjunto vazio de atributos), e os atributos são incluídos um por vez;

• **Backward Elimination - Eliminação por Trás:** o ponto de partida é iniciado com o conjunto de todos os atributos (completo), tais quais são removidos sucessivamente;

• **Bidirectional Search - Pesquisa Bidirecional:** como o próprio nome diz, duas buscas são processadas simultâneamente. Ambas terminam quando atingem o centro do espaço de busca, ou quando uma das buscas encontra os melhores atributos antes de alcançar o centro do espaço de busca;

• **Random Search - Pesquisa Aleatória:** com o propósito de evitar que a busca fique restrita a ótimos locais. Não tem uma direção específica para buscar, pois o ponto de partida da busca e o modo de adicionar ou remover atributos são decididos aleatoriamente.

Além dos sentidos de busca, existem diversas abordagens que avaliam subconjuntos de atributos e que podem remover tanto atributos irrelevantes quanto

redundantes (Parmezan et al., 2012; Liu and Motoda, 2008). A seguir, as principais abordagens:

- ***Filter* - Filtro:**

Com a finalidade de filtrar atributos não importantes, essa abordagem é feita antes da construção dos modelos. A ideia é simplesmente receber como entrada o conjunto de exemplos descrito utilizando somente o subconjunto de atributos importantes identificados. Ela ocorre antes do aprendizado de máquina (John et al., 1994) e utiliza-se métodos estatísticos diversos para esta seleção, como por exemplo árvores de decisão ou as "medidas de importância" que são apresentadas na próxima seção.

- ***Wrapper*- Empacotar:** ocorre também externamente ao algoritmo de aprendizado. Este método gera um subconjuto candidato de atributos, executa o algoritmo de aprendizado considerado somente esse subcojunto selecionado de treinamento e avalia a precisão desse classificador. Repete-se esse processo para cada subconjunto de atributos até buscar um bom modelo. Como exemplo temos a análise por arvores de decisão e florestas aleatórias (serão apresentadas mais a frente). Tem como desvantagem o custo operacional desta abordagem. Exemplo de aplicações: *Naive Bayes* e Máquina de vetores de suporte para classificação.

- ***Embedded* - Embutida:** é realizada internamente pelo próprio algoritmo de extração de padrões. Esta estratégia seleciona o subconjunto de atributos no processo de construção do modelo de classificação, durante a fase de treinamento, e geralmente são específicos para um dado algoritmo de aprendizado. A principal diferença dos métodos do tipo *embedded* e *wrapper*, é que em *embedded* depende em relação a um modelo preditivo específico, assim não permite a sua implementação em combinação com outros modelos (Souza, 2014).

Observação e resumo geral: Note que o que muitas vezes confunde o leitor é o excesso de categorias - que ironicamente tem o propósito de organizar e facilitar. Basicamente são estratégias diferentes com sentidos diferentes de se iniciar a busca de atributos que podem ser irrelevantes ou relevantes: antes de criar um modelo de Aprendizado de maquina; usa-se um modelo de aprendizado para selecionar os atributos antes de iniciar uma etapa de análise [pode-se até mesmo realizar outro algoritmo de aprendizado após este algoritmo de seleção] ou a própria seleção com a análise [mesmo algoritmo para selecionar e concluir]). Quando misturamos esta estratégia, denominamos de **híbridos**.

feitos/Livro ML/bookdown-demo-master/Figuras/diferenssa.bb

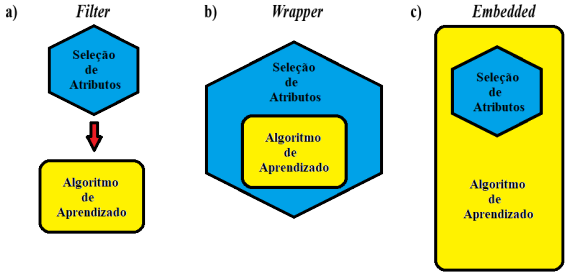


Figure 5.3: "Diferença de *Filter*, *Wrapper* e *Embedded* respectivamente (modificado de Covões (2010))."

Chapter 6

Algoritmos de Aprendizagem - Parte I

Existe uma infinidade de algoritmos utilizados em machine learning, cada um com uma finalidade específica. Há também características que podem inviabilizar a escolha do modelo mais preciso para determinado problema, como a utilização alto poder computacional.

Aqui vai a Parte I de Algoritmos de Aprendizagem, neste capítulo serão apresentados:

1. Medidas de Importância:

- Medidas de Informação
- Medidas de Distância
- Medidas de Dependência
- Medidas de Precisão
- Medidas de Consistência

2. Teste de Hipóteses

-

3. Naive Bayes

4. Regressão

- Regressão Linear Simples
- Regressão Múltipla
- Modelo de Probabilidade Linear
- Gradiente Descendente

6.1 Medidas de Importância

> Um atributo é dito importante se quando removido a medida de importância considerada em relação aos atributos restantes é deteriorada , seja a precisão da medida, consistência, informação, distância ou dependência
>
> Tradução de Liu and Motoda (2012).

É fundamental estimarmos a importância de um atributo, tanto uma avaliação individual quanto à avaliação de subconjuntos de atributos. É uma questão complexa e multidimensional (Liu and Motoda, 2012). Podemos avaliar se os atributos selecionados pela etapa do pré-processamento auxiliam a melhorar a precisão do classificador ou a simplifcar algum modelo construído. A seguir, apresenta-se algumas medidas utilizadas (Lee, 2005).

6.1.1 Medidas de Informação

As medidas de informação determinam o ganho de informação a partir de um atributo. O ganho de informação é definido como a diferença entre a incerteza a *priori* e a incerteza a *posteriori* considerando-se o atributo X_i. X_i é preferido ao atributo X_j se seu ganho de informação for maior que de X_j. Uma das mais utilizadas é a entropia que normalmente é usada na teoria da informação para medir a pureza ou impureza de um determinado conjunto.

Shannon (1948), tomou como "ponto de partida" encontrar uma forma matemática de medir o quanto de informação existe na transmissão de uma mensagem de um ponto a outro, denominando-a entropia. Sua proposta baseava-se na ideia de que o aumento da probabilidade do próximo símbolo diminuiria o tamanho da informação. Com isso, a entropia pode ser definida como a quantidade de incerteza que há em uma mensagem e que diminui à medida que os símbolos são transmitidos (vai se conhecendo a mensagem), tendo-se então a informação, que pode ser vista como redução da incerteza (Shannon, 1948; Paviotti and Magossi, 2019). Por exemplo: ao utilizarmos como idioma a nossa língua portuguesa e ao transmitir como símbolo a letra "q", a probabilidade do próximo símbolo ser a letra "u" é maior que a de ser qualquer outro símbolo, enquanto que a probabilidade de ser novamente a letra "q" é praticamente nula (Paviotti and Magossi, 2019).

Shannon define que a entropia pode ser calculada por meio da soma das probabilidades de ocorrência de cada símbolo pela expressão $\sum p_i = 1 = 100\%$, em que p_i representa a probabilidade do i-ésimo símbolo que compõe a mensagem. Segundo ele, estes símbolos devem ser representados através de sequências binárias, utilizando das propostas de Nyquist (1924) e Hartley (1928). Sua proposta consistia em representar símbolos de um alfabeto através de um logaritmo de acordo com suas respectivas unidades de informação. A entropia proposta por ele é obtida pela média das medidas de Hartley (Moser and Chen, 2012).

Se A é discreto com distribuição de probabilidade $p(A)$, a entropia será:

$$H(A) = -\sum p(A) log_2(p(A)) \tag{6.1}$$

Para facilitar a compreensão, vamos supor um exemplo de um questionário com resposta binária entre "sim" e "não": quanto mais distribuído as probabilidades das respostas, mais desorganizada é, logo maior suaa entropia, do contrário caso for uma probabilidade de ser zero "sim"/"não" ou de ser 1 (100%), ou seja, ter apenas uma opção de resposta, será menos distribuído e portanto menor usa entropia.

feitos/Livro ML/bookdown-demo-master/Figuras/entropia.bb

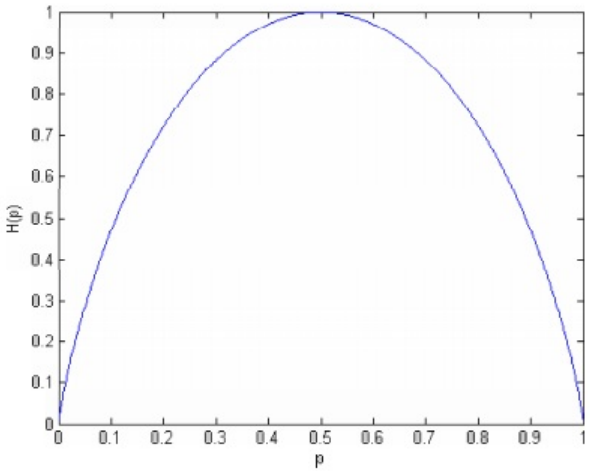


Figure 6.1: Gráfico de Probabilidade x Entropia.

O ganho de informação portanto mede a redução da entropia (nesse caso) causada pela partição dos exemplos de acordo com os valores do atributo.

$$\text{Ganho de Informação}(D, T) = \text{entropia}(D) - \sum_{i=1}^{k} \frac{|D_i|}{|D|}.\text{entropia}(D_i) \tag{6.2}$$

É muito utilizado em algoritmo de **Árvore de decisão** que será apresentado na seção 7 mesma seção com um exemplo de seu uso.

6.1.2 Medidas de Distância

Também conhecidas com medidas de separabilidade, discriminação e divergência. Em caso de duas classes, um atributo X_i é preferido ao atributo X_j se

fornece uma diferença maior que X_j entre as probabilidades condicionais das duas classes. Uma das mais utilizadas é a distância Euclidiana.

6.1.3 Medidas de Dependência

feitos/Livro ML/bookdown-demo-master/Figuras/correlacao.bb

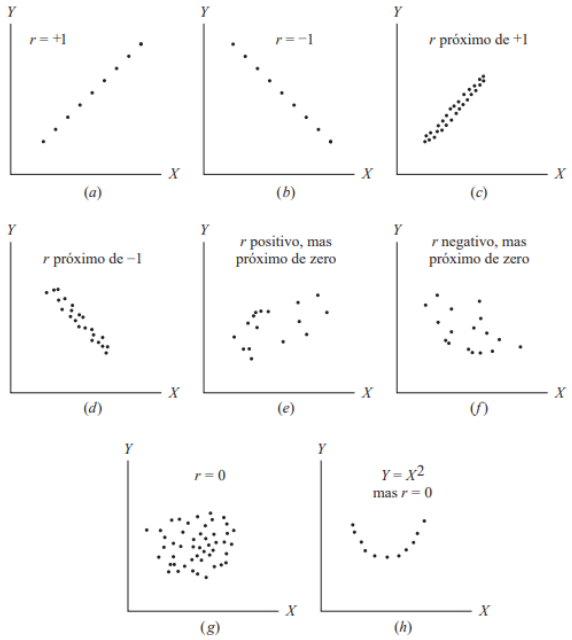


Figure 6.2: Padrões de correlação. Elaborado por Gujarati and Porter (2011) e adaptado Henri (1978).

6.1.4 Medidas de Precisão

6.1.5 Medidas de consistência

6.2 Teste de hipóteses e Análise de Variância

6.3 Naive Bayes

Antes de falarmos sobre este algoritmo, vamos para o conceito matemático. Em (1.2) tratamos do Teorema de Bayes para n atributos. Colocando-o como probabilidade condicional:

$$p(A|B_1, ..., B_n) = p(A)p(B_1|A)p(B_2|A, B_1), p(B_3|A, B_1, B_2)...p(B_n|A, B_1, B_2, ..., B_{n-1}) \tag{6.3}$$

Assumindo que cada atributo B_i é condicionalmente independente de todos os outros B_j para $j \neq i$ e $p(B_i|A, B_j) = p(B_i|A)$ o modelo poderá ser expresso como:

$$p(A_k|B_1, ..., B_n) = p(A_k)p(B_1|A_k)p(B_2|A_k), ... = p(A_k)\prod_i^n p(B_i|A_k) \ k \in 1, ..., k \tag{6.4}$$

Por fim para podermos classificar, aplicamos argumento de máxima para otimizarmos a função, assim obtém-se o classificador de Naive Bayes:

$$\text{classificador } \hat{y} = argmax \ p(A_k)\prod_{i=1}^n p(B_i|A_k) \quad k \in 1, ..., k \tag{6.5}$$

Lembrando que para cada atributo, a sua distribuição de probabilidades é assumida como normal.

O Naive Bayes é uma técnica de classificação baseado no teorema de Bayes com uma suposição de independência entre os preditores, ou seja, este classificador assume que a presença de uma característica particular em uma classe não está relacionada com a presença de qualquer outro fator. Por exemplo, uma fruta verde, redonda e com um tamanho de diâmetro X pode ser uma melancia, porém mesmo que estas variáveis dependam uns dos outros e de outras características, todas estas propriedades contribuem de forma independente para a probabilidade de que seja uma melancia. Este modelo é muito utilizado devido que é fácil de construir e particularmente útil para grandes volumes de dados. Porém a própria independência entre os preditores a torna desvantajosa na prática e caso haja variáveis categóricas num conjunto de dados de teste que não forem treinadas, o modelo não irá estimar estas novas variáveis.

Exemplo: para facilitar, podemos supor que estamos trabalhando no diagnóstico de uma nova doença e que foi feito testes em 100 pessoas aleatórias (exemplo de Orgânica Digital (2019)).

Após coletarmos a análise, descobrimos que das 100 pessoas, 20 possuíam a doença (20%) e 80 pessoas estavam saudáveis (80%), sendo que das pessoas que possuíam a doença, 90% receberam o resultado positivo no teste da doença, e 30% das pessoas que não possuíam a doença também receberam o teste positivo. Caso uma nova pessoa realizar o teste e receber um resultado positivo, qual a probabilidade de ela realmente possuir a doença?

Com o algoritmo de Naive Bayes, buscamos encontrar uma probabilidade da pessoa possuir a doença dado que ela recebeu um resultado positivo, multiplicando a probabilidade de possuir a doença pela probabilidade de "receber um resultado positivo, dado que tem a doença". De mesmo modo verificar a probabilidade de não possuir a doença dado que recebeu um resultado positivo.

feitos/Livro ML/bookdown-demo-master/Figuras/bayes.bb

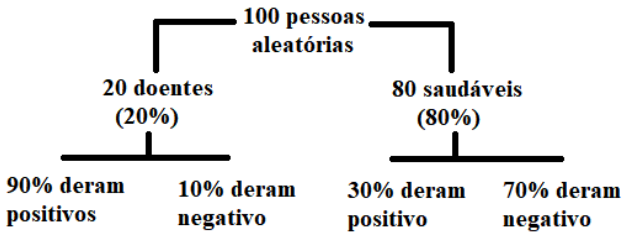


Figure 6.3: Dados coletados de uma amostra de 100 pessoas aleatórias.

Ou seja, ao caso de ter a doença dado que o resultado deu positivo:

$$P(doena|positivo) = 20\%.90\%$$

$$P(doena|positivo) = 0,2 * 0,9$$

$$P(doena|positivo) = 0,18$$

Para o caso de não ter a doença, dado que deu positivo:

$$P(no\ doena|positivo) = 80\%.30\%$$

$$P(no\ doena|positivo) = 0,8 * 0,3$$

$$P(no\ doena|positivo) = 0,24$$

Após isso precisamos normalizar os dados, para que a soma das duas probabilidades resulte 1 (100%). Como vimos em pré-processamento 5, a **Normalização por reescala** por meio de um valor mínimio e um máximo, gera um novo intervalo onde os valores de um atributo estão contidos. Um intervalo entre 0 e 1. Portanto, dividimos o resultado pela soma das duas probabilidades.

$$P(doena|positivo) = 0,18/(0,18 + 0,24) = 0,4285$$

$$P(nodoena|positivo) = 0,24/(0,18 + 0,24) = 0,5714$$

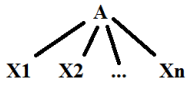
Logo, podemos concluir que se o resultado do teste da nova pessoa for positivo, ela possui aproximadamente 43% (0,4285) de chance de estar doente.

Observação e resumo geral: Naive Bayes é uma técnica de classificação baseado no teorema de Bayes com uma **suposição de independência entre os preditores** diferentemente do caso em 1.2 (Teorema de Bayes), ou seja, O Naive Bayes assume que a presença de uma característica particular em uma classe não está relacionada com a presença de qualquer outro fator. Ao caso da melancia, uma fruta verde, redonda e com um tamanho de diâmetro X é possível ser ela, porém mesmo que estas variáveis dependam uma das outras e de outras

características, elas contribuem de forma independente para a probabilidade de que seja uma melancia. É um modelo simples de construir e útil para grandes volumes de dados. Porém a própria independência entre os preditores a torna desvantajosa para apliação prática e que variáveis categóricas num conjunto de dados de teste que não foram treinadas, não irá estimar essa nova variável.

Por isso *Naive* vem do significado "ingênuo", pois como a Figura 6.4 demonstra, os atributos contribuem de forma independente para a probabilidade de A.

feitos/Livro ML/bookdown-demo-master/Figuras/naive.bb



$$p(A_k|B_1,\ldots,B_n) = p(A_k)p(B_1|A_k)p(B_2|A_k),\ldots = p(A_k)\prod_i^n p(B_i|A_k)\ k \in 1,\ldots,k$$

Figure 6.4: Gráfico de Probabilidade x Entropia.

6.4 Regressão

6.4.1 Análise de Regressão Linear Simples

A análise de variância, pressupõe a independência dos efeitos dos diversos tratamentos utilizados no experimento. Quando a hipótese não é verificada, necessitamos refletir a dependência entre os efeitos dos tratamentos. No caso de experimentos quantitativos, frequentemente justifica a existência da equação de regressão, que une os valores dos tratamentos aos analisados. Em grande parte, trata de estimação e/ou previsão do valor médio (para população) da variável dependente com base nos valores conhecidos da variável explanatória, ela é supervisionada.

Como na prática não conseguimos análisar uma população, trabalhamos em cima de amostras e estimamos para o todo, para que possamos fazer uma aproximação. Partimos da ideia de estimarmos uma função com dados amostrais com o menor erro possível. Portanto, o Y_i (população) observado pode ser expresso como:

$$Y_i = \hat{Y}_i + \hat{\mu}_i \tag{6.6}$$

E o modelo para função de regressão amostral:

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\mu}_i \tag{6.7}$$

em que:

$\hat{Y}_i$ é o valor observado com i níveis de X (estimador da esperança $E(Y|Xi)$), $\hat{\beta}_0$ a constante de regressão estimado e intercepto de $\hat{Y}$, $\hat{\beta}_1$ o coeficiente de regressão estimado que seria a variação de $\hat{Y}$ em função da variação de cada unidade de X, X_i com i níveis da variável independente e $\hat{\mu_i}$ é o erro associado à distância entre o valor observado e o correspondente ponto na curva. Note que os "chapéis" em cima das variáveis é utilizado quando referimos a estimações, ou seja, são variáveis de dados amostrais e não a população.

Mas como estimâmetros os parâmetros da função de forma que fique mais próxima possível e com o menor erro? Com o **Método dos Mínimos Quadrados (MMQ)** atribuído ao Carl Friedrich Gauss - matemático alemão - torna-se possível estimar os melhores β_0 e β_1 que minimizam os erros.

Como não podemos observar a função de regressão populacional (FRP), precisamos estimálo por meio da função de regressão amostral:

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\mu_i} Y_i = \hat{Y}_i + \hat{\mu_i} \text{Logo temos que} \rightarrow \ \hat{\mu_i} = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$$

Podemos ver que os erros $\hat{\mu_i}$ (resíduos) são basicamente as diferenças entre os valores observados e estimados de Y. Ao caso de dados com n pares de observações de Y e X, queremos encontrar a FRA que se encontra o mais próximo possível do Y observado, ou seja, escolher a FRA de modo que a soma dos resíduos $\sum \hat{\mu_i} = \sum (Y_i - \hat{Y}_i)$ seja a menor possível. Porém, como se pode ver pelo diagrama de dispersão na Figura 6.5, os erros possuem a mesma importância com variações entre sinais positivos e negativos e sua somatória será zero. Isso dificultari a possibilidade de minimizarmos.

feitos/Livro ML/bookdown-demo-master/Figuras/mmq.bb

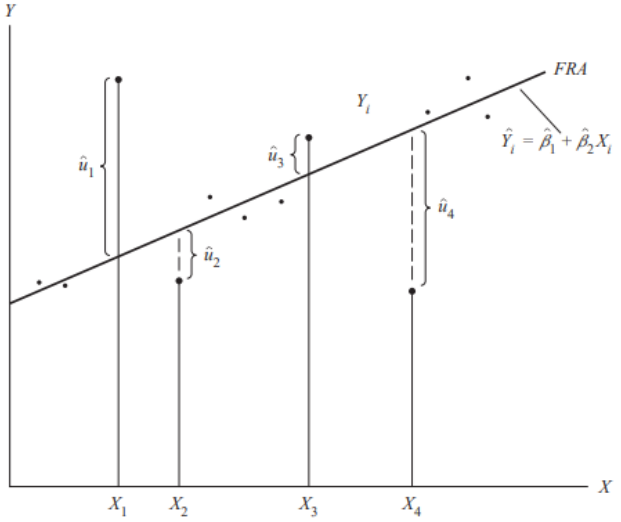


Figure 6.5: Critério do minímos quadrados Gujarati and Porter (2011).

Para evitarmos isso, utilizamos o critério dos mínimos quadrados, de modo que elevamos os resíduos ao quadrado. Fazendo isso, o método dá mais peso aos resíduos (não irão mais se anular), podendo visualizar melhor o "tamanho" do erro total e obter propriedades estatísticas mais desejáveis.

$$\sum \hat{\mu}_i^2 = \sum (Y_i - \hat{Y}_i)^2 = \sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1) X_i^2 \tag{6.8}$$

O método dos mínimos quadrados nos oferece estimativas únicas de β_0 e β_1 que proporcionam o menor valor possível (encontrando $\hat{\beta}_0$ e $\hat{\beta}_1$) de $\sum \hat{\mu}_i$. Por meio de cálculo diferenciável (recomendo o leitor interessado em se aprofundar na definição matemática buscar literaturas em foco estatísticoler, como por exemplo a seção 3A de Gujarati and Porter (2011)) encontra-se:

$$\sum Y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum X_i \tag{6.9}$$

$$\sum Y_i X_i = \hat{\beta}_0 \sum X_i + \hat{\beta}_1 \sum X_i^2 \tag{6.10}$$

AQUI VOU COLOCAR DO JEITO Q FIZ EM ECONOMETRIA COM AS DEFINCOES E AS DERIVADAS

Para que seja feito o modelo de regressão, ela depende das premissas: independência das variáveis erro, homogeneidade das variâncias, normalidade e relação linear entre as variáveis.

- **Coeficiente de determinação r^2: medir a qualidade de seu ajuste**

Estimamos os parâmetros e o erro da função, agora precisamos considerar a **qualidade do ajuste** da linha de regressão ajustada a um conjunto de dados, ou seja, vamos descobrir quão "bom" o ajuste dessa linha de regressão amostral é adequada aos dados. Se todas as observações estivessem exatamente em cima da linha de regressão, seria "perfeito", o que raramente acontece e provávelmente seria um problema de **Overfitting** (será apresentado no próximo capítulo para verificarmos a validade do modelo). O coeficiente de terminação r^2 é um medida que diz quanto a linha de regressão amostral ajusta-se aos dados.

Para entendermos melhor, vamos visualizar por Diagrama de Venn (Kennedy, 1981). O círculo Y representa a variação da variável dependente Y e o círculo X, a variação da variável explanatória X como vimos em regressão linear. A área sombreada indica o quanto em que a variação de Y é explicada pela variação de X. Quanto maior a área sobreposta, maior a parte da variação de Y é explicada por X. O coefiente de determinação r^2 é apenas a medida numérica dessa sobreposição. Na Figura 6.6, conforme move-se da esquerda para a direita, a sobreposição aumenta, ou seja, uma proporção cada vez maior da variação de Y é explicada por X (o r^2 aumenta). Sem sobreposição, $r^2 = 0$ e com total sobreposição, $r^2 = 1$, pois 100% da variação de Y é explicada por X. Portanto o coefienciente situa-se no intervalo entre 0 e 1.

feitos/Livro ML/bookdown-demo-master/Figuras/ballentine.bb

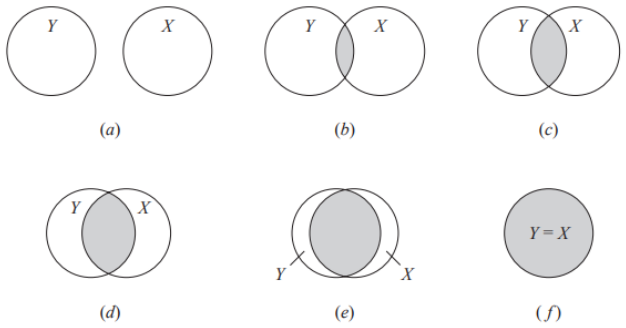


Figure 6.6: Critério do minímos quadrados Gujarati and Porter (2011).

Podemos chegar ao coeficiente de determinação apenas por manipulação algébrica:

sabemos que: $y_i = \hat{y}_i + \hat{\mu}_i$elevando ao quadrado e somando a amostra: $\sum y_i^2 = \sum \hat{y}_i^2 + \sum \hat{\mu}_i^2 + 2 \sum \hat{y}_i \hat{\mu}_i$

$$\text{podemos dizer } SQT = SQE + SQR \tag{6.11}$$

sendo SQT a soma total dos quadrados, SQE a soma do quadrados explicados e SQR soma dos quadrados dos resíduos.

dividindo a equação anterior por SQT:$1 = \frac{SQE}{SQT} + \frac{SQR}{SQT}$definindo r^2 como: $\frac{SQE}{SQT}$

$$\text{obtemos: } r^2 = 1 - \frac{SQR}{SQT} \rightarrow 1 - \frac{\sum \hat{\mu}_i}{\sum (Y_i - \overline{Y}_i)^2} \tag{6.12}$$

Por manipulação algébrica, podemos verificar também que $r^2 = \hat{\beta}_1^2 (\frac{S_x^2}{S_y^2})$, sendo S_x^2 e S_y^2 as respectivas variâncias amostrais de X e Y.

Note que ao aplicarmos a raiz quadrada no coeficiente de determinação obtemos o coeficiente de correlação visto em 6.1.3, que mede o grau de associação entre duas variáveis.

$$r = \pm\sqrt{r^2}$$

AQUI VOU COLOCAR UM EXEMPLO DE REGRESSOA PARA ENTENDER E PARTE MATEMATICA e falar de ANOVA

Não esqueça: dependendo das variáveis em estudo é possível que haja comportamento polinomial ao observarmos no gráfico, podendo ser quadrática, cúbica, etc. Os procedimentos são os mesmos de que linear, mas basicamente incluímos a variável e seu respectivo grau. Dependendo do comportamento muitas vezes é mais fácil ao invés e manter em exponencial (não linear), linearizarmos a função por meio dos logaritmos, semi-logaritmicos entre outros. Isso faz com que temos menos trabalho para tratarmos e estimarmos os parâmetros da função exponencial.

feitos/Livro ML/bookdown-demo-master/Figuras/explog.bb

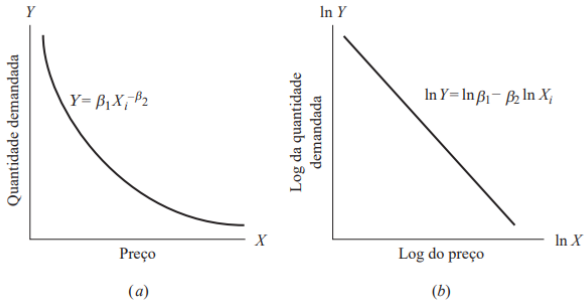


Figure 6.7: Em (a) curva de função exponencial e (b) após aplicarmos o logaritmo (Gujarati and Porter, 2011).

Atualmente é bem comum utilizarmos o modelo **log-log**, pois seu coeficiente angular β_i mede a **elasticidade** de Y em relação a X, ou seja, a variação percentual de Y correspondente a uma variação percentual em X. Por exemplo: na Figura 6.7 se Y representa a quantidade demandada de camisetas e X seu preço unitário. Em (a) temos a relação da quantidade de demanda por camisetas e o preço, mas com a transformação logaritmica teremos a estimação de $-\beta_2$ (pois é uma reta descendente) que indica a elasticidade preço (variação em $ln(Y)$ por unidade de variação em $ln(X)$). Portanto teríamos a variação percentual da quantidade demandada de camisetas dada uma variação percentual do preço. Atente-se: **porcentagem** (Gujarati and Porter, 2011).

Resumo geral: Em palavras, r2 mede a proporção ou percentual da variação total de Y explicada pelo modelo de regressão.

6.4.2 Regressão Linear Múltipla

Na prática deparamos com muitas outros fatores que podem influenciar em sua variável dependente Y. Portanto são acrescentadas dentro de seu modelo de regressão mais variáveis, o que é conhecido como **Regressão Linear Múltipla**, nada mais do que uma ampliação da regressão linear simples. Num modelo, por exemplo, com três variáveis (caso mais simples) pode ser expressa para a amostra como:

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \mu_i \tag{6.13}$$

Da mesma forma, Y_i a variável dependente, X_2 e X_3 as independentes explanatórias (explicativa), μ_i o erro estocático e i para indicar i-ésima observação. Ao caso dos parâmetros, β_0 como intercepto, β_1 e β_2 os **coeficientes parciais de regressão/angulares**. β_2 mede a variação no valor médio de Y (esperança de Y), por unidade de variação em X_2, mantendo X_3 constante, ou seja, traz o efeito "direto" de uma unidade de variação em X_2 sobre o valor médio de Y, excluindo o efeito de X_3 na média de Y. De mesmo modo, X_3 com X_2 constante.

A regressão múltipla pressupõe as mesma hipóteses de que a regressão linear simples, porém como acréscimo - e muito importante- que as variáveis independentes devem estar **ausentes de multicolinearidade**, ou seja, não devem haver relação linear entre si. Se essa relação linear existir entre X_2 e X_3 **são colineares** ou **linearmente dependentes**, do contrário **linearmente independentes**. Caso a multicolinearidade for perfeita, os coeficientes de regressão das variáveis X serão indeterminados e seus erros padrão, infinitos. Se a multicolinearidade for menos que perfeita, serão determinado mas com grandes erros padrão (em relação aos próprios coeficientes), o que trará um modelo ruim para sua estimação.

Para medirmos a multicolinearidade é comum a análise de **correlação de pearson** entre todas as variáveis, como mencionada em **Medidas de Dependência 6.1.3**, ou analisar a ocorrência de intervalo de confiança mais amplo, verificação de razões "t" insignificantes mesmo que seu R^2 esteja alto, parâmetros estimados muitos sensíveis a qualquer alteração de dados e comumente utilizado para verificar o **fator de inflação de variância (FIV)** (Montgomery et al., 2012), que pode ser expressa como:

$$VIF_j = \frac{1}{1 - r_j^2} \quad j = 1, 2, ..., p \tag{6.14}$$

sendo r^2 o coeficiente de correlação ao quadrado e j para referir as variáveis. Por exemplo, se r_{23}^2, refe-se ao coeficiente de correlação entre as variáveis X_2 e X_3. Segundo ,quando este indicador apresenta o valor acima de cinco, é possível a existência de multicolinearidade (Maroco, 2014).

De mesmo modo que em regressão linear simples, são estimados os MQO, Máxima verossimilhança e o **coeficiente de determinação múltiplo** R^2 (mesma interpratação para regressão linear simples r^2) para que se obtenha a melhor aproximação possível.

6.4.3 Modelo de Probabilidade Linear (MPL)

Considerando um modelo típico de regressão linear simples:

$$Y_i = \beta_0 + \beta_1 X_i + \mu_i$$

em que X =sua renda e $Y = 1$ de que você compre um celular e 0 não compre. Como o regressando é binário, ou dicotômico, chamamos de probabilidade linear (MPL). Pode ser interpretada como probabilidade condicional de que o evento ocorra dado X_i, isto é, Pr $(Yi = 1|Xi)$. Neste caso, é a probabilidade de você comprar um celular e cuja renda é dado por X_i.

Para entender este modelo, vamos supor $E(\hat{\mu}_i) = 0$ para evitarmos estimadores tendenciosos (erros). Portanto:

$$E(Y_i|X_i) = \beta_0 + \beta_1 X_i \tag{6.15}$$

Com P_i =probabilidade de que $Y_i = 1$(ocorrência do evento) e $(1 - P_i)$=probabilidade de $Y_i = 0$(não ocorrência do evento). Y_i possui a seguinte **distribuição de probabilidade de Bernoulli**:

Y_i	**Probabilidade**
0	$1 - P_i$
1	P_i
Total	1

Aplicando a esperança, obtemos:

$$E(Y_i) = 0(1 - P_i) + 1(P_i) = P_i \tag{6.16}$$

Igualando (6.16) com (6.15), obtemos:

$$E(Y_i|X_i) = \beta_0 + \beta_1 X_i \tag{6.17}$$

Isso verifica que a esperança condicional do modelo de regressão (6.6) pode ser interpretada como a probabilidade condicional de Yi. Note que, como explicado em 1.2 sobre **Distribuição Bernoulli** e **Distribuição Binominal**, caso haja n observações independentes, cada um com uma probabilidade p (sucesso) e probabilidade $(1-p)$ (fracasso) e X dessas observações representarem o número de sucessos, X então segue a distribuição binomial (com médi np e variância $np(1 - p)$. Lembrando que a probabilidade P_i situa-se entre 0 e 1 $\rightarrow$ $0 \leq E(Y_i|X_i) \leq 1$.

Alguns detalhes importantes:

- A hipótese de normalidade de μ_i não se verifica no caso dos modelos de probabilidade linear, pois os termos de erro assumem também apenas dois valores, seguindo a distribuição de Bernoulli. Se objetivo for a estimação pontual, a hipótese de normalidade deixa de ser necessária (Gujarati and Porter, 2011) e que conforme aumentamos o tamanho da amostra indefinidamente, os estimadores de MQO tendem geralmente a distribuir-se normalmente.
- Como sabe-se, a média e variância de uma distribuição Bernoulli possuem respectivamente p e $p(1-p)$. Logo a variância é heterocedástica $var(\mu_i) = P_i(1 - P_i)$ e portanto os estimadores de MQO não são eficientes (não possuem variância mínima). Podemos fazer a transformação para que seja homocedástico:

$$\sqrt{E(Y_i|X_i) - [1 - E(Y_i|X_i)]} = \sqrt{P_i(1 - P_i) = \sqrt{w_i}}$$

$$\frac{Y_i}{\sqrt{w)i}} = \frac{\beta_0}{\sqrt{w)i}} + \frac{\beta_1 X_i}{\sqrt{w)i}} + \frac{\mu_i}{\sqrt{w)i}} \tag{6.18}$$

Com a transformação, pode-se calcular por MQO (ponderados).

Alternativas para o MPL:

- Como mencionado, a probabilidade condicional situa-se entre 0 e 1, porém por MQO não levarem em conta esta restrição. Pode-se verificar os valores que constam entre o intervalo, considerando os valores negativos como 0 e maiores que 1 como iguais a 1 ou aplicar algum outro modelo para garanti-los dentro dos intervalos.
- O R^2 costuma-se situar muito abaixo de 1. Por ser limitado em caso de modelos binários, muitos pesquisadores buscam evitar seu uso.

Os modelos mais comuns para ser utilizado como alternativa ao MPL são o **logit** e o **probit** para evitar estes problemas.

6.4.3.1 Logit

A fim de fazer com que P_i varie entre 0 e 1 e relacione-se linearmente a X_i, a **função de distribuição logística** pode ser expressa como:

$$P_i = \frac{1}{1 + e^{-Z_i}} = \frac{e_i^Z}{1 + e_i^Z} \tag{6.19}$$

e $(1 - P_i)$ da probabilidade fracasso:

$$1 - P_i = \frac{1}{1 + e^{Z_i}} \rightarrow e^{Z_i} \tag{6.20}$$

onde $Z_i = \beta_0 + \beta_1 X_i$. Assim Z_i varia de $-\infty$ a ∞ e portanto P_i entre 0 e 1.

Para estimarmos a MQO, precisamos linearizar a função:

$$L_i = ln(\frac{P_i}{1-P_i}) = Z_i = \beta_0 + \beta_1 X_i \tag{6.21}$$

O modelo **logit** faz com que:

- A probabilidade varie entre 0 e 1, enquanto Z e L possam variar de $-\infty$ a ∞;
- Mesmo que as probabilides não sejam lineares, L é linear em X;
- Pode-se aplicar com mais regressores e com mesma interpretação angular medindo a variação de L para uma unidade variação em X e para o intercepto;
- Se L torna-se maior e positivo quando as chances do evento de interesse ocorrer aumenta, do contrário (maior e negativo) de não ocorrer;
- Como em MPL, o modelo Logit é heterocedástico precisa-se ponderar (Gujarati and Porter, 2011; Cox, 1970):

$$\sqrt{w_i}L_i = \beta_0\sqrt{w_i} + \beta_1\sqrt{w_i}X_i + \sqrt{w_i}\mu_i \tag{6.22}$$

em que, com a variância $\hat{\sigma}^2 = \frac{1}{N_i\hat{P}_i(1-\hat{P}_i)}$, W_i é o peso $N_i\hat{P}_i(1-\hat{P}_i)$. Por fim, aplicar o mínimos quadrados ponderados (da mesma forma que MQO, porém com a nova transformação de dados) e estimarmos os parâmetros normalmente.

Como o R^2 não é significativa nos modelos binários. É comum utilizar as **pseudo $R^2** [long1997regression] - existe uma variedade delas - ou o **Count** R^2 que nada mais é que o número de previsões corretas com o número total de observações. Para a hiótese nula de que todos os coeficientes angulares são simultâneamente iguais a zero, utiliza-se a **estatística da razão de verossimilhança** que segue a distribuição χ^2 que equivale ao teste F.

6.4.3.2 Probit

6.4.3.3 Tobit

6.4.4 Exemplos

6.5 Gradiente Descendente (GD)

Para a obtenção dos parâmetros de forma analítica, como regressões, muitas vezes é difícil obter os parâmetros que minimizam determinada função de interesse. Dificuldades em obter a solução do sistema na forma fechada (ou não existir) ou quando n é muito grande, o cálculo da inversa (estimando os parâmetros matricialmente) pode ser muito caro computacionalmente.

O **Gradiente Descendente (GD)** pode ser muito útil dependendo da situação, conhecido também como **máximo declive**, é um método númerico utilizado em otimização. Tem como finalidade identificar um mínimo local de uma função de modo iterativo, no qual a cada iteração toma-se a direção do gradiente. Muitas vezes serve como base para algoritmos de segunda ordem como Métodos de Newton, por exemplo.

É uma função para casos gerais, por praticidade vamos supor que temos uma função denominada custo com apenas dois parâmetros $J(\theta_0, \theta_1)$ e queremos estimar seus parâmetros que minimizam seus erros. Inicialmente atribuímos quaisquer estimativas iniciais para valores de θ_0 e θ_1, com o GD vamos alterandos os valores dos $\theta' s$ para reduzirmos $J(\theta_0, \theta_1)$ até que se chegue a um valor mínimo local.

Um exemplo que gosto muito, por NG, Andrew Y. (2019): observe a Figura 6.8 e imagine que você está em um campo, com dois montes. Mantenha sua imaginação de que está situado na cruz preta - ponto 0 - no primeiro monte vermelho. Com o GD vamos olhar 360 graus ao redor do ponto em que você está situado apenas para descobrir a resposta de que "se você fosse dar um pequeno passo em alguma direção ao seu redor com o objetivo de ir para o ponto mais baixo do campo o mais rápido possível, para qual direção você deve andar?"

Supondo que após olhar para todos os lados, com análise de GD você descobriu que seu primeiro passo será no ponto 1 da Figura 6.8. Após isso, você observa novamente para todos os lados e faz outra análise de GD para verificar aonde você vai se deslocar em seu segundo passo para chegar o mais rápido possível até concluir que será o ponto 2. Assim, sucessivamente, você vai se deslocando para os respectivos pontos 3, 4 e sucessivamente até convergir em seu objetivo Z, porém caso você iniciasse pelo ponto K, é bem possível que por meio do GD você descesse o monte por outro trajeto, encontrando outros pontos ótimos locais até chegar a outro ponto otimizado (descer por completo o monte). Esta é a ideia do Gradiente Descendente, por meio de iterações, o algoritmo vai identificando os pontos ótimos (estimadores mínimos) até convergir num ótimo local da função.

Em caso de funções simples como regressão linear, não é necessário o uso de GD. Mas em casos com muitas variáveis e ordens, pode ser bem viável.

O algoritmo pode ser expresso como:

$$\theta_j := \theta_j - \alpha \frac{d}{d\theta_j} J(\theta_0, \theta_1) \text{ com } j = \theta_0 \text{ e } j = \theta_1 \tag{6.23}$$

com j referindo-se à quantidade de observações (parâmetros que pretendemos estimar) da amostra.

O algoritmo é processado da seguinte forma: imagine na mesma Figura 6.8 que você irá dar seu primeiro passo, olhou os 360 graus e inseriu as variáveis em

feitos/Livro ML/bookdown-demo-master/Figuras/gd.bb

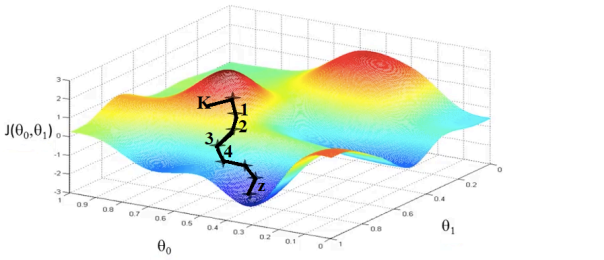


Figure 6.8: Gráfico tridimensional a exemplo de Gradiente Descendente (NG, Andrew Y., 2019).

seu algoritmo de GD e seu destino é em $Z = 10$. Seu algoritmo calcula se você passou seu destino mais do que devia ou se você está atrás de Z ainda e também verifica se precisa dar passos grandes por estar bem longe de seu destino, ou passos menores. Supondo que seu α um pouquinho alto, podemos dar um passo grande para descer o monte (1) pela diferença da observação que você inseriu com $\alpha \frac{d}{d\theta_j} J(\theta_0, \theta_1)$. Caso fosse uma taxa pequena de α, seu passo seria menor e sua derivada (taxa de variação) vai lhe dizer se você passou do ponto ótimo de Z (o quão a frente) ou está para trás (quão para trás) desse ponto ótimo.

Com o primeiro passo dado (supor passo 1 = 40), você precisa fazer o mesmo procedimento tomando agora o passo 1 como se fosse o inicial novamente, ou seja, atualizando sua função para cada θ **simultâneamente** (caso dois θ's de entrada para a função, atualiza-se para ambos) até encontrar o novo valor ótimo do próximo passo no ponto 2 = 15. Conforme vai se aproximando de Z, seus passos vão ficando cada vezes menores (de 15 para 11; de 11 para 10,50; de 10,50 para 10,10; de 10,10 para 10,05; etc) até chegar na melhor aproximação de $Z = 10$ que é o ponto ótimo da função.

Assim o algoritmo encontra os melhores parâmetros para buscar o ponto otimizado, com a estimatiza dos melhores parâmetros para a aproximação com os menores erros (sim! Podemos encontrar os parâmetros dos exemplos de regressão com este algoritmo também!)

Desta forma, atribuímos (":=") para a própria observação de entrada da função receber ela mesma subtraída α que multiplica a derivada da função em relação a observação de entrada. Para que atualize a cada passo (iteração). α**(learning rate - taxa de aprendizagem)** é um valor fixo que controla o tamanho do passo em cada iteração: quando α for pequeno, o método fica lento, quando grande ele pode falhar na convergência e até mesmo divergir. Seu valor de-

pende muito da pesquisa e de suas fundamentações teóricas, o que recomendo o leitor quando utilizar este método verificar um valor adequado, pode ser que dependendo do valor da taxa demore muito para finalizar o algoritmo pela quantidade de iterações (tamanhos de passos muito pequenos) ou divergir (tamanho de passos muito grandes). Rendle and Schmidt-Thieme (2008) divulgaram que a fatoração de matrizes para a predição de *ratings* nos dados do desafio *Netflix* precisou de 200 iterações, usando uma taxa de aprendizagem de 0,01.

Para facilitar a compreensão do efeito da taxa de variação, observe a Figura 6.9. No primeiro gráfico você inicia seu algoritmo com o valor θ e com a derivada podemos observar que inclinação da reta tangente ao ponto é positiva ($\frac{d}{d\theta}j(\theta) \geq 0$), portanto em $\theta = \theta - \alpha$.um valor positivo, faz que com que esse novo θ (segunda iteração) seja menor que o da primeira iteração, visto que terá que subtrair e deslocar-se para esquerda para tender ao ponto mínimo. Da mesma forma, ao segundo gráfico, podemos verificar que a inclinação é negativa ($\frac{d}{d\theta}j(\theta) \leq 0$), portanto $\theta = \theta - \alpha$.um valor negativo, fará com que o novo θ seja maior do que da primeira iteração, pois irá somar e deslocar-se para direita tendendo ao ponto mínimo.

feitos/Livro ML/bookdown-demo-master/Figuras/gd1.bb

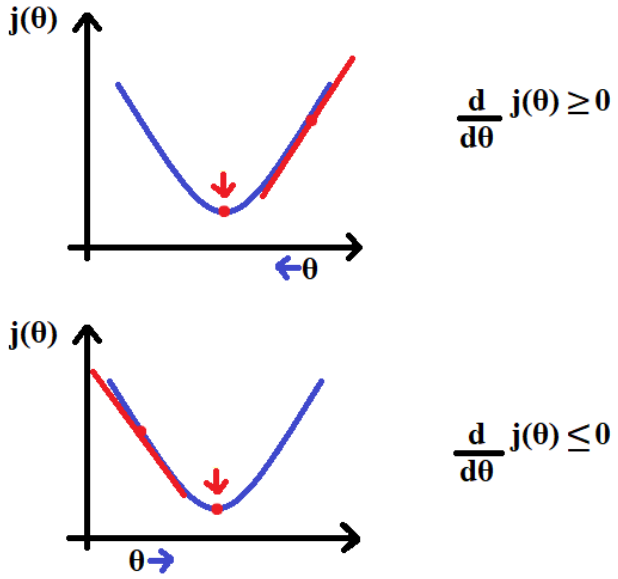


Figure 6.9: Efeito da taxa de variação no Gradiente Descendente.

Como pode-se perceber, a taxa de aprendizagem e a taxa de variação são fundamentais e complementares para o algoritmo de GD, pois elas dizem o tamanho do passo e em que posição estamos em relação ao ponto ótimo da função.

6.5.1 Exemplos

1. **Uma variável:**Vamos supor a seguinte função custo:

$$j(\theta) = \theta^2$$

Queremos minimizá-la $min\ j(\theta)$. Portanto precisamos inicialmente colocar um número aleatório para nosso parâmetro - não ótimo - para que o algoritmo atualize a cada iteração. Vamos supor a taxa de aprendizagem (*learning rate*) $\alpha = 0,1$ e $\theta = 4$ para facilitar. Ou seja, $j(\theta) = 4^2 = 16$. Vamos atualizar os parâmetros:

$\theta := \theta - \alpha.\frac{d}{d\theta}j(\theta)$derivando a função $j(\theta) = \theta^2$ e substituindo:$\theta := \theta - \alpha.2\theta$substituindo os valores de α e $\theta : \theta := 4 - 0,1$

Na iteração obtemos $\theta = 3,2$. Se subsituirmos em $j(\theta)$ novamente, iremos obter $j(\theta) = (3,2)^2 = 10,24$. Agora atualizando novamente para a próxima iteração:

$$\theta := \theta - \alpha.2\theta\theta := 3,2 - 0,1\ .\ 2\ .\ 3,2\theta := 2,56$$

Portanto, $j(\theta) = (2,56)^2 = 6,55$. Sucessivamente, vamos fazendo as iterações até convergir:

θ	$j(\theta)$
4	16
3,2	10,24
2,56	6,55
2,04	4,19
1,632	2,663
.	.
.	.
.	.
0	0

Da mesma forma, se iniciarmos o algoritmo com -4:

θ	$j(\theta)$
-4	16
-3,2	10,24
-2,56	6,55
-2,04	4,19
-1,632	2,663

θ	$j(\theta)$
.	.
.	.
.	.
0	0

Note que conforme θ diminui, o custo também. Conforme mais iterações são aplicadas, mais "ótimo" será. Graficamente para -4 em vermelho e +4 em azul:

feitos/Livro ML/bookdown-demo-master/Figuras/gdx2.bb

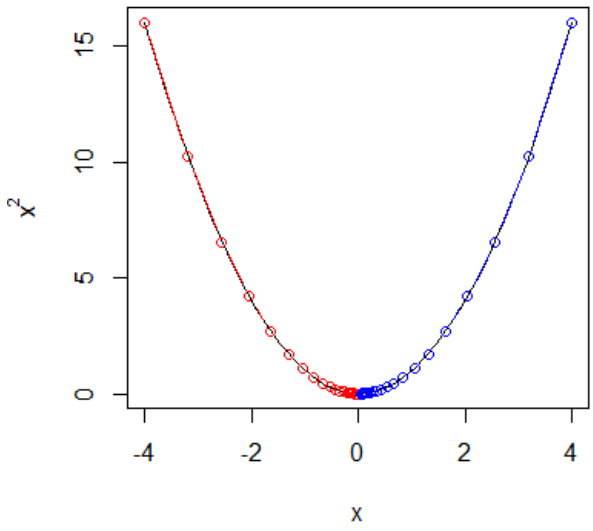


Figure 6.10: Função X^2 com valores de entrada -4 e +4.

2. **Duas variáveis:** Vamos supor a seguinte função de custo com $\alpha = 0,1$, $\theta_1 = 1$ e $\theta_2 = 2$:

$$j(\theta_1, \theta_2) = \theta_1^2 + \theta_2^2 j(\theta_1, \theta_2) = 1^2 + 2^2 = 5$$

Queremos $min\ j(\theta_1, \theta_2)$Como explicado, ao caso de haver mais de um parâmetro precisamos separar atualizar cada um simultâneamente e aplicar derivada parcial em sua função:

$\theta_1 := \theta_1 - \alpha \frac{d}{d\theta_1} j(\theta_1, \theta_2)$ e $\theta_2 := \theta_2 - \alpha \frac{d}{d\theta_2} j(\theta_1, \theta_2)$calculando as derivadas parciais de $j(\theta_1, \theta_2) = \theta_1^2 + \theta_2^2$ ol

substituindo:$\theta_1 := \theta_1 - \alpha.\, 2\theta_1$ e $\theta_2 := \theta_2 - \alpha.\, 2\theta_2$inserindo os valores:$\theta_1 := 1 - 0,1.\, 2.\, 1$ e $\theta_2 := 2 - 0,1.\, 2.\, 2\theta_1 := 0,8$ e θ_2

Portanto após a iteração, temos que $j(\theta_1, \theta_2) = 0,8^2 + 1,6^2 = 3,2$. Da mesma forma, para a próxima iteração temos:

$$\theta_1 := 0,8 - 0,1.\, 2.\, 0,8 \text{ e } \theta_2 := 1,6 - 0,1.\, 2.\, 1,6\theta_1 := 0,64 \text{ e } \theta_2 = 1,28$$

Portanto teremos $j(\theta_1, \theta_2) = 0,64^2 + 1,28^2 = 2,048$. Assim sucessivamente:

θ_1	θ_2	$j(\theta_1, \theta_2)$
1	2	5
0,8	1,6	3,2
0,64	1,28	2,48
.	.	.
.	.	.
.	.	.
0		0

3. **Erro quadrado médio (Regressão Linear Simples:)** Observe a função de regressão linear:

$$f_\theta(X) = \theta_0 + \theta_1 * X$$

A função de custo:

$$j(\theta) = \frac{1}{m}\sum_{i=1}^{m}(f_\theta(x^i) - y^i)^2$$

Primeiramente vamos encontrar a derivada parcial de $j(\theta_0, \theta_1)$:

$$\frac{d}{d\theta_0} j(\theta_0, \theta_1 = \frac{d}{d\theta_0}(\frac{1}{m}\sum_{i=1}^{m}(f_\theta(x^i) - y^i)^2) \rightarrow \frac{2}{m}\sum_{i=1}^{m}(f_\theta(x^i) - y^i)\frac{d}{d\theta_1} j(\theta_0, \theta_1 = \frac{d}{d\theta_1}(\frac{1}{m}\sum_{i=1}^{m}(f_\theta(x^i) - y^i)^2) \rightarrow \frac{2}{m}\sum_{i=1}^{m}(f_\theta(x^i) -$$

Pode-se também multiplicar a função de custo por $\frac{1}{2}$ para que quando faz-se a derivada, facilite no cálculo e multiplicar a função de custo por um escalar não irá afetar a localização do mínimo.

$$j(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(f_\theta(x^i) - y^i)^2$$

Com isso em foco de minimizarmos, basta aplicarmos o banco de dados de X e Y em seu modelo e de seus dois $\theta' s$ de entrada. Repetindo as iterações para atualizar seus valores até a convergência e identificando os parâmetros que se aproximam.

Chapter 7

Algoritmos de Aprendizagem - Parte II

7.1 SVM

7.2 Árvores de Decisão

7.3 Elastic Net

7.4 KNN

7.5 K-means

7.6 Análise de Componentes Principais

A Análise de Componentens Principais, popularmente conhecida como ACP ou PCA (*Principal Component Analysis*), em inglês, foi introduzida por Pearson (1901) e fundamentada no artigo de Hotelling (1933). É uma **análise multivariada** que tem como objetivo explicar a estrutura de variância e covariância de um vetor aleatório, composto por p-variáveis aleatórias, através da construção de combinações lineares das variáveis originais que são chamadas de componentes principais e não correlacionadas entre si (Mingoti, 2007). É uma técnica bastante utilizada em diversas áreas do conhecimento, como a biologia, a agronomia, a zootécnica, a ecologia, a engenharia florestal, a medicina, a economia, entre outras áreas. Muitos sugerem o seu uso quando o volume de dados ou variáveis é grande possibilitando reduzir a dimensão da matriz de dados que compõem o conjunto de variáveis resposta com apenas poucos componentes, ou seja, p variáveis originais substituídas por k (sendo $k < p$) componentes principais não

correlacionadas.

Vamos supor um conjunto de dados em apenas duas dimensões (x, y) e que pode ser plotado em um plano cartesiano. Podemos verificar pelo seu comportamento que possuem alta correlação positiva.

feitos/Livro ML/bookdown-demo-master/Figuras/pca1.bb

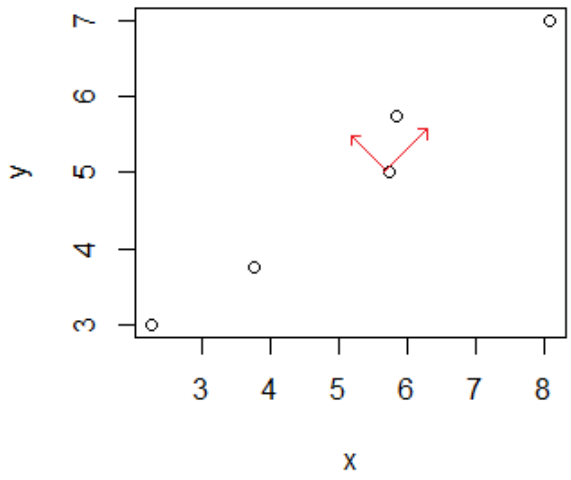


Figure 7.1: Gráfico bidimensional x por y.

Mas se quisermos descobrir a variação do conjunto de dados, o ACP busca encontrar um novo sistema de coordenadas em que cada ponto tem um novo valor (x, y). Os eixos não representam algo físico, mas representam combinações de x e y que denominamos **"componentes principais"**, escolhidas para analisar a variação do eixo. Observe que rotacionamos o gráfico na Figura 7.2 e que após a ACP, podemos verificar a possibilidade de dercartar a componente referente ao eixo y, visto que a componente do eixo x explica 99,30% da variação total dos dados, ou seja, o primeiro componente tem uma maior dispersão (variância). Possibilitando pela componente principal do eixo x, analisar e até mesmo classificar as observações, como por exemplo, a observação 1 e 2 como um conjunto e a 3, 4 e 5 como um segundo conjunto.

Com mais dimensões, o ACP torna-se ainda mais útil pois possibilita observarmos o conjunto de dados num melhor ângulo.

Portanto, a ACP assume que os dados originais estão representados por características (variáveis) correlacionadas com o objetivo de transformar essas var-

feitos/Livro ML/bookdown-demo-master/Figuras/pca2.bb

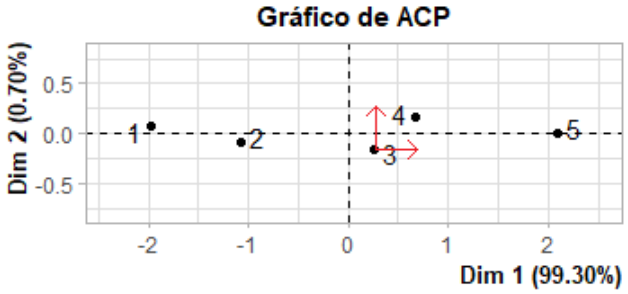


Figure 7.2: Gráfico de x por y rotacionado.

feitos/Livro ML/bookdown-demo-master/Figuras/pca3.bb

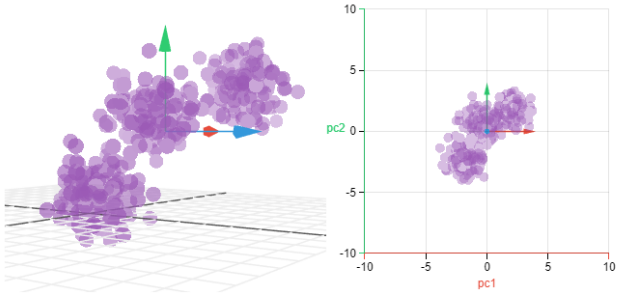


Figure 7.3: Gráfico tridimensional, em Powell, Victor and Lehe, Lewis (2014).

iáveis em novas (componentes principais) por meio de mudança de base do espaço vetorial que não sejam correlacionadas entre si e que estas novas variáveis (menores que as originais) retenha a maior parte da variação apresentada pelas originais, tornando possível a classificação.

A suposição de normalidade não é requisito para sua técnica, mas ainda sim é conveniente padronizar (5.2.2) cada variável, permitindo que todas as variáveis tenham o mesmo peso para evitarmos viés de escala (Hongyu et al., 2016). A padronização das variáveis do vetor pelas respectivas médias e desvios padrões, gera novas variáveis centradas em zero e com variâncias iguais a 1. Assim, as componentes principais são determinadas a partir da matriz de covariâncias das variáveis originais padronizadas (Mingoti, 2007).

Agora que sabemos o que é ACP, vamos apresentar alguns conceitos de Álgebra Linear e Estatísticas para compreendermos como é aplicado este método.

7.6.1 Autovalores e Autovetores

Caso ainda não tenha muito contato com a Álgebra Linear, recomendo buscar algumas literaturas a respeito. Em 1.2 encontra-se sobre Escalar, Vetores, Espaço Vetorial e Transformação Linear que serão tratadas neste tópico.

Dado uma matriz A_{mxn} que define uma transformação linear (não muda sua dimensão), existem vetores onde sua orientação não é afetada por esta transformação, os **autovetores**.

feitos/Livro ML/bookdown-demo-master/Figuras/autovetor.bb

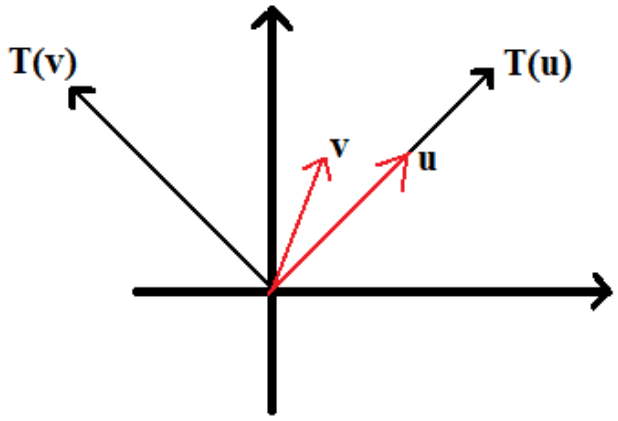


Figure 7.4: u é um autovetor de T, porém v não.

Um vetor é dito ser autovetor da matriz A_{mxn} se a transformação linear deste vetor $T(u)$ é colinear a este vetor, ou seja, $A_{mxn}\vec{u} = \lambda\vec{u}$. Sendo que λ é um escalar e chamado de autovalor da matriz correspondente ao autovetor. Para encontrarmos o autovetor:

$$A_{mxn}\vec{u} = \lambda\vec{u} A_{mxn}\vec{u} - \lambda\vec{u} = 0(A_{mxn} - \lambda l)\vec{u} = 0 \tag{7.1}$$

esta equação tem solução trivial, ou seja, diferentes da nula ($\vec{v} \neq 0$) se e somente se, seu determinante é zero. Conhecido como **Equação caracterísica** e sua solução são os **autovalores**:

$$\text{Eq. Característica } \; det(A_{mxn} - \lambda l) = 0 \tag{7.2}$$

Note também que toda transformação linear (matriz) em um espaço vetorial complexo (números imaginários) tem, pelo menos, um autovetor (real ou complexo).

7.6.1.1 Exemplo

1. Vamos considerar um operador linear $T : R^2 \rightarrow R^2$. Com $T(x, y) = (4x + 5y, 2x + 2y)$. Quais são os autovalores a matriz $A = \begin{bmatrix} 4 & 5 \\ 2 & 2 \end{bmatrix}$?

Vamos resolver a equação característica $det(A_{mxn} - \lambda l) = 0$.

$$det(A_{mxn} - \lambda l) = \begin{bmatrix} 4 & 5 \\ 2 & 2 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 4-\lambda & 5 \\ 2 & 2-\lambda \end{bmatrix}$$

Com $det(A_{mxn} - \lambda l) = 0$:

$(4-\lambda)(2-\lambda)-10 = 0\lambda^2-6\lambda-2 = 0$resolvendo a equação: $\lambda_1 \approx 6,32$ e $\lambda_2 \approx -0,32$

7.6.2 Estatísticas

Alguns conceitos de Estatísticas são fundamentais para que se entenda a ACP:

- **Covariância x Correlação:** como apresentado em 1.2, a covariância é semelhante à correlação (ver 5.2.2) entre duas variáveis, no entanto, elas diferem que os coeficientes de correlação são padronizados. Isso faz com que um relacionamento linear varie entre $-1 \leq \rho \leq 1$. A correlação mede tanto a força como a direção da relação linear entre duas variáveis. Ao caso da covariância os valores não são padronizados. Assim, a covariância pode variar de $-\infty \leq Cov(x, y) \leq \infty$ demonstrando quanto x e y mudam juntas. Portanto o valor para uma relação linear ideal depende muito dos dados. Como os dados não são padronizados, é difícil determinar a força da relação entre as variáveis. Note que o coeficiente de correlação é uma função da covariância:

$$\rho_{x,y} = \frac{cov(x, y)}{\sigma_x \sigma_y}$$

Uma covariância positiva sempre resulta em uma correlação positiva e uma covariância negativa sempre resulta em uma correlação negativa.

Quando temos um vetor de n variáveis em vez de apenas duas, iremos obter uma matriz de covariâncias ou correlação. Contendo em sua diagonal a variância σ^2, pois $cov(x_i, x_i) = \sigma^2(x_i)$, por exemplo:

$$\begin{bmatrix} cov_{1,1} & cov_{1,2=2,1} & cov_{1,3=3,1} \\ cov_{1,1=2,1} & cov_{2,2} & cov_{2,3=3,2} \\ cov_{3,1=1,3} & cov_{2,3=3,2} & cov_{3,3} \end{bmatrix} = \begin{bmatrix} var_1 & cov_{1,2=2,1} & cov_{1,3=3,1} \\ cov_{1,1=2,1} & var_2 & cov_{2,3=3,2} \\ cov_{3,1=1,3} & cov_{2,3=3,2} & var_3 \end{bmatrix}$$

7.6.3 A ACP

Agora que compreendemos alguns conceitos importantes, podemos entender melhor a metodologia da ACP. Assumindo que os dados originais estão representados por variáveis correlacionadas (etapa de pré-processamento), ou seja, não independentes. Vamos ao objetivo de transformar essas p variáveis em outras novas k (com $k < p$) de ordem decrescesnte de variabilidade e que não sejam correlacionadas e que as primeiras novas variáveis retenham a maior parte da variação apresentadas pelas originais a fim de podermos classificá-los.

Dado um vetor $\vec{u}$ aleatório com p variáveis originais. O primeiro componente principal y_1, como dito que deve ser ordem decrescente de variabilidade, será uma combinação linear do vetor $\vec{u}$ de forma que a variância $var(y_1) = \sigma^2_{y_1}$ seja a máxima (maior possível), ou melhor, precisamos encontrar um vetor $\vec{\beta^1}$ tal que $y_1 = (\vec{\beta^1})^T\vec{u}$ e $var(y_1 = (\vec{\beta^1})^T\vec{u}$ seja máxima. De mesmo modo para y_2 e um vetor $\vec{\beta^2}$ e assim sucessivamente para p variáveis em seu banco de dados.

Ao caso do Exemplo 7.6.1.1 de Autovalores e Autovetores, foi definida a transformação linear $T(x, y) = (4x + 5y, 2x + 2y)$ com as duas respectivas componentes $4x + 5y$ e $2x + 2y$. 4 e 5 da primeira componente refere-se, por exemplo, como o vetor $\vec{\beta^1}$ que multiplicado pelos vetores originais x e y, temos um novo componente desse novo espaço $4x + 5y$. Da mesma forma à segunda componente $2x + 2y$ com um $\vec{\beta^2}$.

Para facilitar a compreensão, vamos utilizar um exemplo com duas variáveis (R^2): AQUI VOU ARRUMAR E COLOCAR COMO TA EM MINGOTI! - Queremos encontrar a primeira componente principal y_1, de modo que $var(y_1)$ seja máxima, ou seja, encontrar um vetor $\vec{\beta^1}$ tal que $y_1 = (\vec{\beta^1})^T\vec{u}$ e $var(y_1 = (\vec{\beta^1})^T\vec{u}$ seja máxima.

$$var(y_1) = var((\vec{\beta^1})^T\vec{u}) = var(\vec{\beta^1_1}\vec{u}_1 + \vec{\beta^1_2}\vec{u}_2) = (\vec{\beta^1_1})^2 var(\vec{u_1}) + (\vec{\beta^1_2})^2 var(\vec{u_2}) + 2\vec{\beta^1_1}\vec{\beta^1_2} Cov(\vec{u_1}\vec{u_2}) = (\vec{\beta^1})^T K_{\vec{u}}\bar{\beta}$$

Os maiores autovalores são os que orientam o sinal, os demais podem ser descartados. Porém quantos componentes principais devemos utilizar? Precisamos

verificar a proporção da variação total dos dados originais que uma componente pode explicar, a partir disso selecionarmos. Lembrando que cada autovalor λ_i refere-se a $var(y_i)$.

Para calcularmos a variação total, expressa-se pela somatória de todos os autovalores:

$$\sum_j \lambda_j \tag{7.3}$$

Portanto, para analisar cada i componente, ou seja, cada autovalor (variação "explicada" por cada componente):

$$p_i = \frac{\lambda_j}{\sum_j \lambda_j} \tag{7.4}$$

Sendo geralmente escolhido as componentes com seus respectivos autovalores que explicam entre 70%-90% segundo alguns pesquisadores. Outros como Kaiser (1960), propõe aceitar, observando diretamente, somente os autovalores iguais ou superiores à unidade.

Importante: sobre utilizar matriz de covariância ou de correlação depende muito das fundamentações teóricas e recomendaçõesdos pesquisadores. Em geral, utiliza-se a matriz de correlação (quando padronizamos e elaboramos a matriz) ao caso de padronizar escalas distintas que podem viesar, como por exemplo, medidas de distância e de peso.

Caso esteja utilizando software para a análise, dependendo do software utilizado com seu determinado modelo de formulação de componentes principais, pode ocorrer essa troca de sinal que nada mais é do que uma reflexão em relação ao eixo, uma rotação em seu espaço vetorial n-dimensional em torno da origem, poderá ocasionar uma "rotação" em torno do eixo. Tratando de algebra linear e suas combinações lineares, a combinação poderá possuir soluções diferentes que diferem apenas o sinal.

7.6.4 Exemplos

Tomando como base exemplos de Mingoti (2007).

1. **Matriz de covariância amostral**

   A Tabela apresenta dados relativos as 12 empresas no que se refere a 3 variáveis (medidas em unidades monetárias): ganho bruto ($X1$), ganho líquido ($X2$) e o patrimônio acumulado (X_3):

Empresas	**Ganho Bruto (X_1)**	**Ganho Líquido (X_2)**	**Patrimônio Líquido(X_3)**
E1	9893	564	17689
E2	8776	389	17359

Empresas	**Ganho Bruto** (X_1)	**Ganho Líquido** (X_2)	**Patrimônio Líquido**(X_3)
E3	13572	1103	18597
E4	6455	743	8745
E5	5129	203	14397
E6	5432	215	3467
E7	3807	385	4679
E8	3423	187	6754
E9	3708	127	2275
E10	3294	297	6754
E11	5433	432	5589
E12	6287	451	8972

Após calcularmos suas covariâncias (recomendo o leitor calcular e verificar e atentar que por ser exemplificação, passível de ocorrência de arrendondamento dos valores), obtemos a matriz de covariância amostral:

	Ganho Bruto (X_1)	**Ganho Líquido** (X_2)	**Patrimônio Líquido**
Ganho Bruto (X_1)	9550608,6	706121,1	14978232,5
Ganho Líquido (X_2)	706121,1	76269,5	933915,1
Patrimônio Líquido (X_3)	14978232,5	933915,1	34408113,0

Para calcularmos os autovalores:

$$det(A_{mxn} - \lambda l) = 0 : \begin{bmatrix} 9550608,6 - \lambda & 706121,1 & 14978232,5 \\ 706121,1 & 76269,5 - \lambda & 933915,1 \\ 14978232,5 & 933915,1 & 34408113,0 - \lambda \end{bmatrix} = 0$$

Resolvendo o sistema, obtemos os seguintes autovalores das componentes principais:

$$\lambda_1 = 38018192,2 \quad \lambda_2 = 2327881,5 \quad \lambda_3 = 19334,8$$

Para encontrarmos a porcentagem da variância explicada por cada auto valor:

$$\%\lambda_1 = \frac{38018192,2}{38018192,2 + 2327881,5 + 19334,8}.100\% = 94,2\%\%\lambda_2 = \frac{2327881,5}{38018192,2 + 2327881,5 + 19334,8}.10$$

Portanto, podemos descartar o segundo e o terceiro componente principal, pois o primeiro explica cerca de $94,2\%$.

Por fim os autovetores podem sem calculados:

$$A_{mxn}\vec{u} = \lambda\vec{u}$$

Com A_{mxn} a matriz de covariância amostral, $\vec{u}$ o autovetor e λ os respectivos autovalores dos autovetores.

$$\begin{bmatrix}\vec{u_1}\\ \vec{u_2}\\ \vec{u_3}\end{bmatrix}\begin{bmatrix}9550608,6 & 706121,1 & 14978232,5\\ 706121,1 & 76269,5 & 933915,1\\ 14978232,5 & 933915,1 & 34408113,0\end{bmatrix} = \lambda_i \begin{bmatrix}\vec{u_1}\\ \vec{u_2}\\ \vec{u_3}\end{bmatrix}$$

substituindo os autovalores: $\begin{bmatrix}\vec{u_1}\\ \vec{u_2}\\ \vec{u_3}\end{bmatrix}\begin{bmatrix}9550608,6 & 706121,1 & 14978232,5\\ 706121,1 & 76269,5 & 933915,1\\ 14978232,5 & 933915,1 & 34408113,0\end{bmatrix} = \begin{bmatrix}0,942 & 0 & 0\\ 0 & 0,0577 & 0\\ 0 & 0 & 0,0048\end{bmatrix}\begin{bmatrix}\vec{u_1}\\ \vec{u_2}\\ \vec{u_3}\end{bmatrix}$

Teremos os autovetores: | | **Autovetor Ganho Bruto ($\vec{u_1}$)** | **Autovetor Ganho Líquido ($\vec{u_2}$)** | **Autovetor Patrimônio Líquido ($\vec{u_3}$)** | |:-:|:-:|:-:|:-:| | **Autovetor Ganho Bruto ($\vec{u_1}$)** | 0,425 | 0,900 | -0,099 | | **Autovetor Ganho Líquido ($\vec{u_2}$)** | 0,028 | 0,096 | 0,995 | | **Autovetor Patrimônio Líquido ($\vec{u_3}$)** | 0,905 | -0,426 | 0,016 |

Com os autovetores, podemos elaborar as três componentes principais:

$\hat{y_1} = 0,425(GanhoBruto)+0,028(GanhoLquido)+0,905(PatrimnioLquido)\hat{y_2} = 0,900(GanhoBruto)+0,096(Ganho$

por meio da observação de seus resultados podemos analisar que:

- A primeira componente possui alta correlação-positiva com todas as três variáveis, podemos analisar como um índice de desempenho global da empresa. Pelo autovetor, podemos ver que o patrimônio possui o maior peso e de menor o ganho líquido. Podemos verificar que quanto maior for os valores das variáveis, maior será dessa componente, ou melhor, maior será o desempenho global da empresa. Esta ocupa, observando pelos autovalores, 94,\20% de toda variação explicada, dependendo da pesquisa pode-se descartar as outras componentes.
- A segunda componente que ocupa 5,77% de toda variação explicada (autovalor), possui o ganho bruto e patrimônio de maior variância amostral (analisando o tabela de covariância amostra). Pelos autovetores, podemos verificar que o ganho bruto é a variável dominante com segunda maior variância amostral. Com a componente próximo a zero, entende-se que haverá um certo equilíbrio entre ganho bruto e patrimônio acumulado, o que na verdade o aumento do ganho bruto eleva-se esta componente e o patrimônio contrário. Note que há correlação bem menor entre elas.
- A terceira componente com pouca variância total explicada, referente ao ganho líquido de menor variância amostral, possui pouca importância. Apena o ganho líquido possui alta correlação, visto que às outras duas são próximas de zero.

Determinada as componentes principais, podemos obter seus valores numéricos (**escores**) para cada elemento amostral. Basicamente substituímos os valores originais na funções encontradas de componentes principais (y_1, y_2 e y_3):

Empresas	CP_1	CP_2	CP_3
E1	8857,59	-165,27	-90,18
E2	8079,36	-1046,65	-158,93
E3	11257,93	2810,25	96,18
E4	-690,80	566,19	284,23
E5	3844,09	-3084,94	-30,40
E6	-5915,42	1841,62	-224,93
E7	-5504,97	-119,93	124,81
E8	-3796,38	-1367,83	-0,64
E9	-7729,15	789,46	-160,88
E10	-3848,18	-1473,28	121,59
E11	-3989,16	960,15	25,13
E12	-564,92	290,23	14,02

Podemos observar que a empresa E9 possui o menor desempenho, e as E1, E2 e E3 os melhores. Entenda que não necessariamente o sinal de negativo é sempre ser um pior valor, isso depende da pesquisa e da interpretação do sinal ou como em caso de autovetores, indica a rotação. Para analisarmos por gráfico não é recomendável utilizar neste caso, devido que são valores bem grandes para serem inseridos. No caso de Matriz de correlação, que serão padronizados os dados, podemos visualizar melhor.

7.7 Clusters

7.8 AOC e ROC

7.9 modelos nivel III

7.10 grad boosting -> estudar boosting e bagging dentro de emseamble

7.11 Redes Neurais

Chapter 8

Validação de um modelo

8.1 *Overfitting, Underfitting*

Sendo **muito importantes** nesta área, o Underfitting (sub-ajustado) e Overfitting (sobre-ajustado) são dois termos que temos que estar sempre atentos. Um bom modelo não deve sofrer de nenhum deles (Silver, 2013).

- **Overfitting**: Um cenário de overfitting ocorre quando, nos dados de treino, o seu modelo ML tem um desempenho excelente, porém quando utilizamos os dados em novos bancos de dados, seu resultado é ruim. Nesta situação, seu modelo aprendeu tão bem as relações existentes dos conjuntos de dados para treino que acabou apenas decorando esses dados. Portanto ao receber as informações das variáveis preditoras aos novos dados, o modelo tenta aplicar as mesmas regras decoradas, porém com estes novos dados (diferentes do treino) esta regra não tem validade e seu desempenho é afetado.

  As principais causas e soluções de um Overfitting são:

  1. Algoritmo muito complexo para os dados: caso for possível, pode-se simplificar o modelo utilizado por um algoritmo mais simples, com menos parâmetros. Permitindo reduzir as chances do modelo sofrer overfitting.

  2. Poucos dados para treinar: dependendo da quantidade de dados utilizados para treinar, pode ser que seja uma amostra pequena, com isso recomenda-se aumentar seu tamanho coletando mais dados.

  3. Ruídos nos dados de treinamento: é comum dentro do banco de dados existir algum tipo de ruído, isto é, *outlier* (valores extremos ou até mesmo valores incorretos nos dados). Esses ruídos podem fazer com que o modelo aprenda sobre ele, levando ao overfitting. Seria

recomendado pré-processamento adequado para tratar essa interferência.

8.1.1 Underfitting: No cenário underfitting, o desempenho já é ruim no próprio treinamento de seu algoritmo.

As principais causas e soluções de um Underfitting são:

1. Algoritmo inadequado: bem provável que o modelo estatístico proposto pelo pesquisa pode não ter sido adequado ao comportamento dos dados. Por exemplo aplicar um algoritmo para funções de primeiro grau (linear) em um conjunto de dados com comportamento exponencial (função de segundo grau). Recomendável o pesquisador substituir o algoritmo escolhendo outro com outros parâmetros para solucionar o underfitting.

2. Características não representativas: há possibilidade de que as características que estamos utilizando para treinar o modelo não sejam representativas, ou seja, não possuem relação entre si ou não sejam importantes para o modelo aplicado.

3. Modelo com muitos parâmetros de restrição: o modelo torna-se inflexível, restrito, e não consegue se ajustar de forma adequada aos dados.

Segue abaixo a Figura 8.1 demonstrando os dois casos anteriores e um modelo adequado.

feitos/Livro ML/bookdown-demo-master/Figuras/graficofit.bb

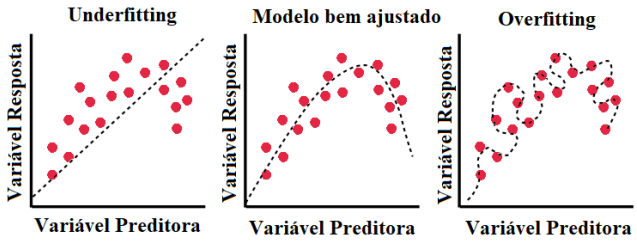


Figure 8.1: "Gráfico representando um *Underfitting*, um Modelo bem ajustado e um *Overfitting* respectivamente."

8.2 Validação Cruzada

A fim de que não haja previsões desastrosas geradas pelo modelo, para medirmos o desempenho real do modelo criado, é necessário que realizemos testes com ele, utilizando dados diferentes dos que foram apresentados no início. Portanto uma das técnicas mais utilizadas é a **Cross-validation (Validação Cruzada)**.

Após a realização do pré-processamento (analisar), iremos separar a totalidade dos dados históricos existentes em dois grupos, sendo o primeiro responsável pelo aprendizado do modelo, e o segundo por realizar os testes. Seguindo o mesmo exemplo de bons ou mau pagadores, usualmente separamos o conjunto de dados dos clientes em duas amostras. Uma com

8.3 Como escolher um bom modelo?

Bibliography

AQUARELA (2017). Otimizando agendamentos médicos com inteligência artificial. *AQUARELA*.

Assunção, F. (2012). *Estratégias para tratamento de variáveis com dados faltantes durante o desenvolvimento de modelos preditivos*. PhD thesis, Universidade de São Paulo.

Banzatto, D. A. and Kronka, S. d. N. (1992). Experimentação agrícola. *Jaboticabal: Funep*, 2.

Batista, G. E. d. A. P. et al. (2003). *Pré-processamento de dados em aprendizado de máquina supervisionado*. PhD thesis, Universidade de São Paulo.

Bobrow, D. G. (1967). Problems in natural language communication with computers. *IEEE Transactions on Human Factors in Electronics*, (1):52–55.

Bolfarine, H. and Sandoval, M. C. (2001). *Introdução à inferência estatística*, volume 2. SBM.

Box, G. E. and Jenkins, G. M. (1976). Time series analysis: Forecasting and control san francisco. *Calif: Holden-Day*.

Buchanan, B., Sutherland, G., and Feigenbaum, E. (1969). Heuristic dendral: A program for generating explanatory hypotheses. *Organic Chemistry*.

Buchanan, B. G. and Shortliffe, E. H. (1984). Rule-based expert systems: the mycin experiments of the stanford heuristic programming project.

CARVALHO, A., Faceli, K., LORENA, A., and Gama, J. (2011). Inteligência artificial–uma abordagem de aprendizado de máquina. *Rio de Janeiro: LTC*.

Casella, G. and Berger, R. L. (2010). Inferência estatística. *São Paulo: Cengage Learning*.

Cordeiro, G. M. (1999). *Introduçao a teoria assintótica*. IMPA.

Covões, T. F. (2010). *Seleção de atributos via agrupamento*. PhD thesis, Universidade de São Paulo.

Cox, D. (1970). Analysis of binary data london: Methuen &co.

Cross, S. E. and Walker, E. (1994). Dart: applying knowledge-based planning and scheduling to crisis action planning. *Intelligent Scheduling. Morgan Kaufmann*.

da Silveira, J. A. P. (2013). Searle e dennett: duas perspectivas de estudo da mente. *Problemata: Revista Internacional de Filosofía*, 4(2):238–258.

de Andrade, D. F., Borgatto, A. F., Araujo, P. H., and Schmitt, J. (2019). *Caderno de Pesquisa 1: Técnicas de imputação de dados na análise de questionários contextuais*. Cebraspe, Brasília. ISBN 978-85-5656-010-0.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

Dennett, D. C. (2009). The part of cognitive science that is philosophy. *Topics in Cognitive Science*, 1(2):231–236.

Dougherty, J., Kohavi, R., and Sahami, M. (1995). Supervised and unsupervised discretization of continuous features. In *Machine learning proceedings 1995*, pages 194–202. Elsevier.

Enders, C. K. (2010). *Applied missing data analysis*. Guilford press.

Evans, T. G. (1964). A program for the solution of a class of geometric-analogy intelligence-test questions. Technical report, AIR FORCE CAMBRIDGE RESEARCH LABS LG HANSCOM FIELD MASS.

Fahlman, S. E. (1974). A planning system for robot construction tasks. *Artificial intelligence*, 5(1):1–49.

Fisher, R. A. (1912). On an absolute criterion for fitting frequency curves. *Messenger of Mathematics*, 41:155–156.

Freund, J. E. (2009). *Estatística Aplicada-: Economia, Administração e Contabilidade*. Bookman Editora.

Garcia, S., Luengo, J., Sáez, J. A., Lopez, V., and Herrera, F. (2012). A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning. *IEEE Transactions on Knowledge and Data Engineering*, 25(4):734–750.

Guimarães, R. R. C. (2019). A inteligência artificial e a disputa por diferentes caminhos em sua utilização preditiva no processo penal. *Revista Brasileira de Direito Processual Penal*, 5(3):1555–1588.

Gujarati, D. N. and Porter, D. C. (2011). *Econometria básica-5*. Amgh Editora.

Hartley, R. V. (1928). Transmission of information 1. *Bell System technical journal*, 7(3):535–563.

Hebb, D. O. (1949). *The organization of behavior: a neuropsychological theory*. J. Wiley; Chapman & Hall.

Henri, T. (1978). *Introduction to econometrics*. Prentice Hall, Englewood Cliffs, New Jersey.

Hongyu, K., Sandanielo, V. L. M., and de Oliveira Junior, G. J. (2016). Análise de componentes principais: resumo teórico, aplicação e interpretação. *E&S Engineering and Science*, 5(1):83–90.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417.

Huffman, D. A. (1971). Impossible object as nonsense sentences. *Machine intelligence*, 6:295–324.

John, G. H., Kohavi, R., and Pfleger, K. (1994). Irrelevant features and the subset selection problem. In *Machine Learning Proceedings 1994*, pages 121–129. Elsevier.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and psychological measurement*, 20(1):141–151.

Kennedy, P. E. (1981). The "ballentine": a graphical aid for econometrics. *Australian Economic Papers*, 20(37):414–416.

Langley, P. et al. (1994). Selection of relevant features in machine learning. In *Proceedings of the AAAI Fall symposium on relevance*, volume 184, pages 245–271. Citeseer.

Lee, H. D. (2005). *Seleção de atributos importantes para a extração de conhecimento de bases de dados*. PhD thesis, Universidade de São Paulo.

Little, R. J. and Rubin, D. B. (2019). *Statistical analysis with missing data*, volume 793. John Wiley & Sons.

Liu, H. and Motoda, H. (1998). *Feature extraction, construction and selection: A data mining perspective*, volume 453. Springer Science & Business Media.

Liu, H. and Motoda, H. (2008). Computational methods of feature selection (chapman & hall/crc data mining and knowledge discovery series).

Liu, H. and Motoda, H. (2012). *Feature selection for knowledge discovery and data mining*, volume 454. Springer Science & Business Media.

Maroco, J. (2014). Análise estatística com o spss. *Statistics*, 6.

McCarthy, J. (1968). Programs with common sense'in minsky m (ed) semantic information processing.

McCarthy, J., Minsky, M. L., Rochester, N., and Shannon, C. E. (2006). A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955. *AI magazine*, 27(4):12–12.

McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133.

Meehl, P. E. (1954). Clinical versus statistical prediction: A theoretical analysis and a review of the evidence minneapolis: University of minnesota press.[reprinted with new preface. In *In Proceedings of the 1955 Invitational Conference on Testing Problems*. Citeseer.

Mingoti, S. A. (2007). Análise de dados através de métodos estatística multivariada: uma abordagem aplicada. In *Análise de dados através de métodos estatística multivariada: uma abordagem aplicada*, pages 295–295.

Minsky, M. L. and Papert, S. (1969). Perceptrons: an introduction to. *Computational Geometry*.

Montgomery, D. C., Peck, E. A., and Vining, G. G. (2012). *Introduction to linear regression analysis*, volume 821. John Wiley & Sons.

MOSER, J. d. M. (2006). O golem. *Estudos em homenagem a Margarida Llosa*, pages 323–336.

Moser, S. M. and Chen, P.-N. (2012). *A student's guide to coding and information theory*. Cambridge University Press.

Newell, A. and Shaw, J. (1959). A variety op intelligent learning in a general problem solver. *RAND Report P-1742, dated July*, 6.

NG, Andrew Y. (2019). Gradient descent algorithm.

Nyquist, H. (1924). Certain factors affecting telegraph speed. *Transactions of the American Institute of Electrical Engineers*, 43:412–422.

Orgânica Digital (2019). Algoritmo de classificação naive bayes.

Parmezan, A. R. S., Lee, H. D., Spolaôr, N., and Chung, W. F. (2012). Avaliação de métodos para seleção de atributos importantes para aprendizado de máquina supervisionado no processo de mineração de dados.

Paviotti, J. R. and Magossi, C. J. (2019). Considerações sobre o conceito de entropia na teoria da informação.

Pearson, K. (1901). Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572.

Pereira, S. G. (2019). Inserção de dados faltantes não aleatórios para estimativa de variável geometalúrgica.

Powell, Victor and Lehe, Lewis (2014). Análise do componente principal.

Rendle, S. and Schmidt-Thieme, L. (2008). Online-updating regularized kernel matrix factorization models for large-scale recommender systems. In *Proceedings of the 2008 ACM conference on Recommender systems*, pages 251–258.

Rosenthal, R. (1994). Science and ethics in conducting, analyzing, and reporting psychological research. *Psychological Science*, 5(3):127–134.

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons.

Russel, S. and NORVIG, P. (2004). Inteligência artificial. 2ª. edição. *Rio de Janeiro: Campus*.

RUSSEL, S. and Norvig, P. (2013). Inteligência artificial. tradução de regina célia smille. *Rio de Janeiro: Campus Elsevier*.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, 3(3):210–229.

Schafer, J. L. (1999). Multiple imputation: a primer. *Statistical methods in medical research*, 8(1):3–15.

Schafer, J. L. and Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological methods*, 7(2):147.

Searle, J. R. (1980). Minds, brains, and programs, from the behavioral and brain sciences, vol. 3. *Cambridge University Press http://members. aol. com/NeoNoetics/MindsBrainsPrograms. html From*, 23:2004.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Shannon, C. E. (1950). Xxii. programming a computer for playing chess. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 41(314):256–275.

Shelley, M. (1818). Frankenstein or the modem prometheus. *London: Printed for Lackington, Hughes, Harding, Mayor & Jones*.

Silver, N. (2013). *O sinal e o ruído*. Editora Intrinseca.

Simon, P. (2013). *Too big to ignore: the business case for big data*, volume 72. John Wiley & Sons.

Slagle, J. R. (1963). A heuristic program that solves symbolic integration problems in freshman calculus. *Journal of the ACM (JACM)*, 10(4):507–520.

Souza, F. A. d. (2014). *Computational Intelligence Methodologies for Soft Sensors Development in Industrial Processes*. PhD thesis.

TURING, I. B. A. (1950). Computing machinery and intelligence-am turing. *Mind*, 59(236):433.

Waltz, D. (1975). Understanding line drawings of scenes with shadows. In *The psychology of computer vision*. Citeseer.

Winograd, T. (1972). Understanding natural language. *Cognitive psychology*, 3(1):1–191.

Winston, P. H. (1970). Learning structural descriptions from examples.