Guia de estudos *Data Science*

Fundamentos Matemáticos

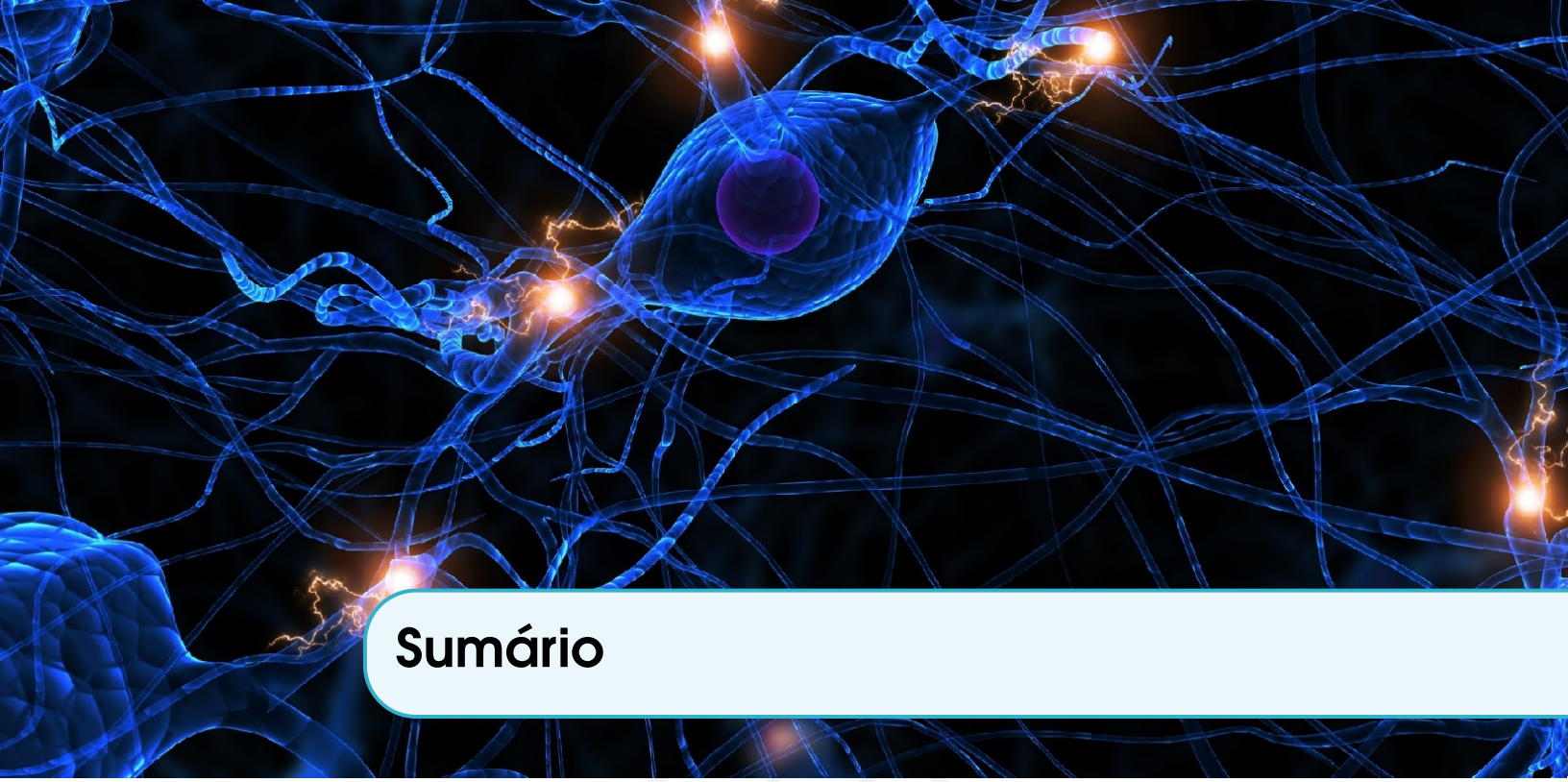
Neuron/DSAI

GRUPO DE ESTUDOS EM *Data Science e Inteligencia Artificial* NEURON/DSAI

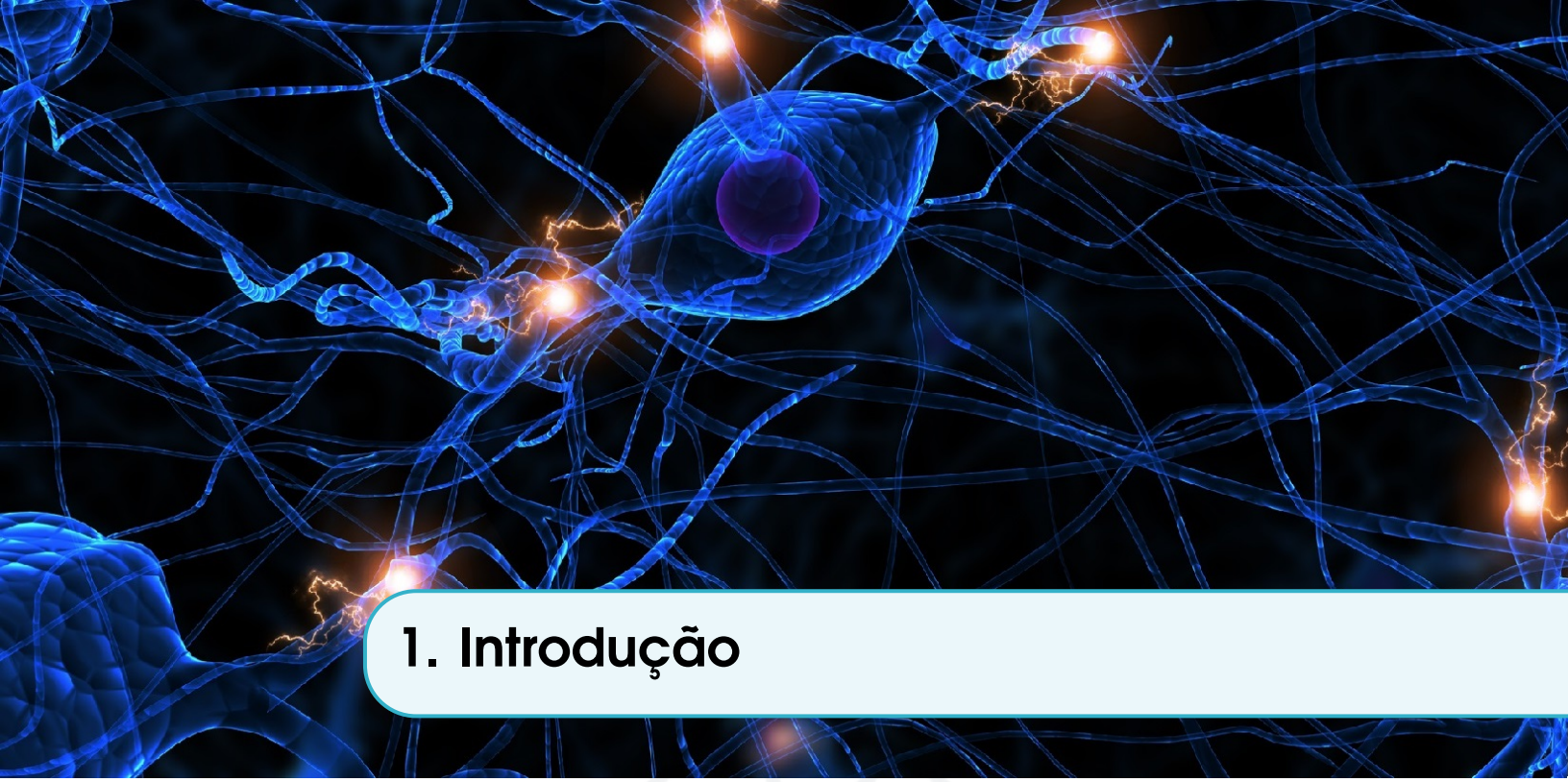
FACEBOOK.COM/NEURONDSAI

Este material, em conjunto com os Notebooks do Jupyter que o acompanham, é de propriedade do grupo Neuron/USP e, legalmente, não pode ser reproduzido para fins comerciais (sujeito a processos judiciais). Qualquer dúvida ou sugestão entre em contato conosco pela nossa página do Facebook ou e-mail: *neuron.conteudo@gmail.com*.

Autores: Leonardo Ernesto, Samuel Henrique, Felipe Maia Polo, Matheus Faleiros, Pablo Leonardo, Murilo Henrique Soave,Carlos Henrique Lavieri, Mateus Padua e Thiago M. Carvalho
Edição: 27 de agosto de 2018



Sumário



1. Introdução

1.1 O grupo

O grupo Neuron surge com o principal intuito de cobrir um *gap* existente na realidade dos brasileiros e até mesmo das principais universidades brasileiras: o estudo, discussão e desenvolvimentos de projetos relacionados ao *Data Science*, que engloba programação, análise de dados e o emprego das tecnologias de Aprendizado de Máquina com fins práticos ligados tanto à pesquisa quanto aos negócios. Como atividade principal, o grupo Neuron abrigará um grupo de estudos em *Data Science* a partir do ano de 2018 localizado no campus da Universidade de São Paulo (USP) em Ribeirão Preto, São Paulo. No que tange às atividades extras, traremos periodicamente profissionais e estudiosos da área para conversar com nossos membros e daremos apoio a alunos que queiram desenvolver projetos próprios. Além disso, fechamos parceria com o grupo HAIT, mais importante grupo de estudos em Inteligência Artificial formado por alunos das melhores universidades japonesas. Essa parceria possibilitará grande troca de experiência e *networking* entre os membros dos dois grupos, sendo que o desenvolvimento de projetos entre alunos de diferentes nacionalidades será o ponto alto. Espero que aproveite seu tempo no Neuron. Grande abraço,

Felipe Maia Polo - Presidente 2018

1.2 O conteúdo

O conteúdo do Neuron foi desenvolvido com a finalidade de maximizar o seu aprendizado em *Data Science*, utilizando a linguagem de programação Python com o enfoque em Aprendizado de Máquina. Os assuntos que serão abordados foram divididos em quatro grandes tópicos: Introdução a Programação em Python, Fundamentos de Matemática, *Machine Learning* e *Deep Learning*. Esses tópicos serão divididos em subtópicos, para melhor explicação dos mesmos. Os conteúdos serão abordados em aulas expositivas com exercícios práticos, que serão disponibilizados antes de cada aula. Também contaremos com palestras de profissionais das diversas áreas, com o intuito de propiciar aos membros do grupo um conhecimento amplo sobre as aplicações das tecnologia no estudos em *Data Science*. Todo o conteúdo foi feito com muito esforço e dedicação com a

colaboração do Felipe Maia Polo e de toda equipe de conteudo. Espero que o conteúdo seja completo e de fácil compreensão. Atenciosamente,

Leonardo Ernesto – Vice-Presidente 2018

1.3 Os autores

Este espaço é dedicado para que você conheça um pouco dos nossos autores.

Felipe Maia Polo

Estudante de Economia pela Faculdade de Economia, Administração e Contabilidade de Ribeirão Preto da Universidade de São Paulo. Tem experiência com Estatística, Econometria e Aprendizado de Máquina. Trabalhou no Laboratório de Estudos e Pesquisas em Economia Social (LEPES) focando majoritariamente em avaliação de políticas públicas na área da educação. Concluiu parte da graduação na Universidade de Tóquio, onde teve contato com grupos voltados ao ensino e treinamento de pessoas em *Data Science*, estes que foram grandes inspirações para a concretização do Grupo de Estudos em *Data Science* Neuron/USP.

Leonardo Ernesto

Estudante de Informática Biomédica pela Faculdade de Medicina de Ribeirão Preto e pela Faculdade de Filosofia Ciências e Letras de Ribeirão Preto na Universidade de São Paulo. Exerce o cargo de diretor de assuntos acadêmicos no Centro Estudantil da Informática Biomédica - CEIB, o qual está sob sua responsabilidade desde 2016 e já foi representante Discente na Comissão Coordenadora da IBM em 2016. É aluno de iniciação cientifica desde 2014 no Laboratório de Tráfego Intracelular de Proteínas, onde desenvolve pesquisa de interação de proteínas. Possui amplo conhecimento nas áreas de Bioinformática, Inteligência Artificial, Reconhecimento de Padrões e *Data Mining*.

Matheus Faleiros

Formado em Informática Biomédica e mestrando do programa Interunidades em Bioengenharia da Universidade de São Paulo. Trabalha com classificação e reconhecimento de padrões em imagens médicas. Possui conhecimento em técnicas de *Machine Learning* e *Deep Learning*.

Samuel Henrique

Estudante de Informática Biomédica pela Faculdade de Medicina de Ribeirão Preto e pela Faculdade de Filosofia Ciências e Letras de Ribeirão Preto na Universidade de São Paulo. Conhecimento avançado na linguagem Python 3.6, algumas de suas bibliotecas e Programação Orientada a Objetos.

Pablo Leonardo

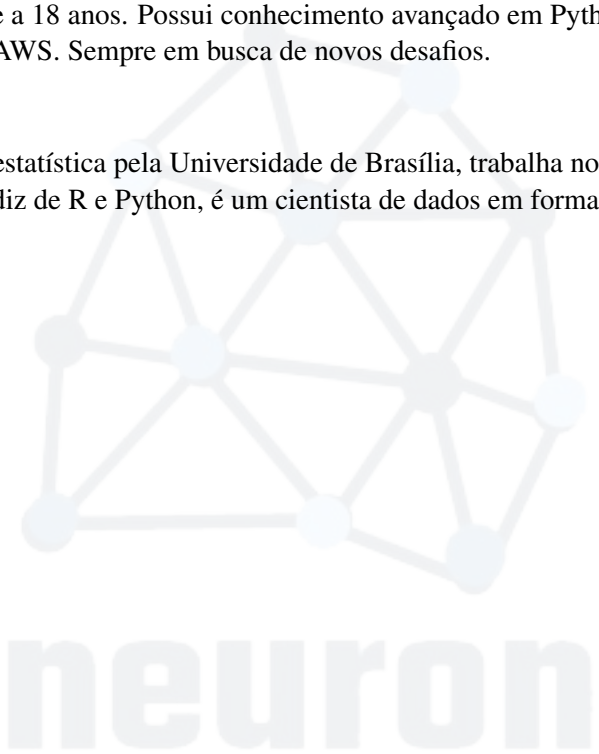
Graduando em engenharia de produção pela faculdade Pitágoras.Graduando em licenciatura plena em matemática pela Universidade Federal do Maranhão. Amante de neuromatemática e matemática computacional.

Murilo Henrique Soave
Estudante de Economia Empresarial e Controladoria na Faculdade de Economia, Administração e Contabilidade da Universidade de São Paulo. Possui conhecimento avançado em Python e C, além de experiência na programação de sistemas embarcados. Atualmente, desenvolve trabalho de iniciação científica vinculado ao Instituto Federal de São Paulo.

Carlos Henrique Lavieri Marcos Garcia
Estudante de Economia na Faculdade de Economia, Administração e Contabilidade de Ribeirão Preto (FEA-RP/USP), com interesse em ciência de dados e Aprendizado de Máquina.

Mateus Padua
Bacharel em Ciência da Computação pela UniSEB COC. Trabalha como desenvolvedor WEB e Engenheiro de Software a 18 anos. Possui conhecimento avançado em Python, Django, JavaScript, OOP, Banco de dados, AWS. Sempre em busca de novos desafios.

Thiago M. Carvalho
Graduado e mestre em estatística pela Universidade de Brasília, trabalha no Banco do Brasil com a linguagem SAS. Aprendiz de R e Python, é um cientista de dados em formação. Atualmente estuda estatística bayesiana.



1.4 Gestão 2018

Presidente	**Vice-Presidente**
Felipe Maia Polo	Leonardo Ernesto

Conteúdo

Diretor:	Samuel Henrique	*Membro:*	Matheus Faleiros
Membro:	Pablo Leonardo	*Membro:*	Murilo Soave
Membro:	Carlos Henrique	*Membro:*	Charles Chen
Membro:	Thiago Carvalho	*Membro:*	Bruno Comitre
Membro:	Mateus Padua		

Projeto

Diretor:	Leonardo Ernesto	*Membro:*	Eduardo Junqueira

Desenvolvimento de pessoas

Diretor:	Gustavo Ribeiro	*Membro:*	Edvaldo Santos
Membro:	Henrique Nogueira		

Financeiro

Diretora:	Natasha Freitas	*Membro:*	Beatriz Machado
Membro:	Murilo Soave		

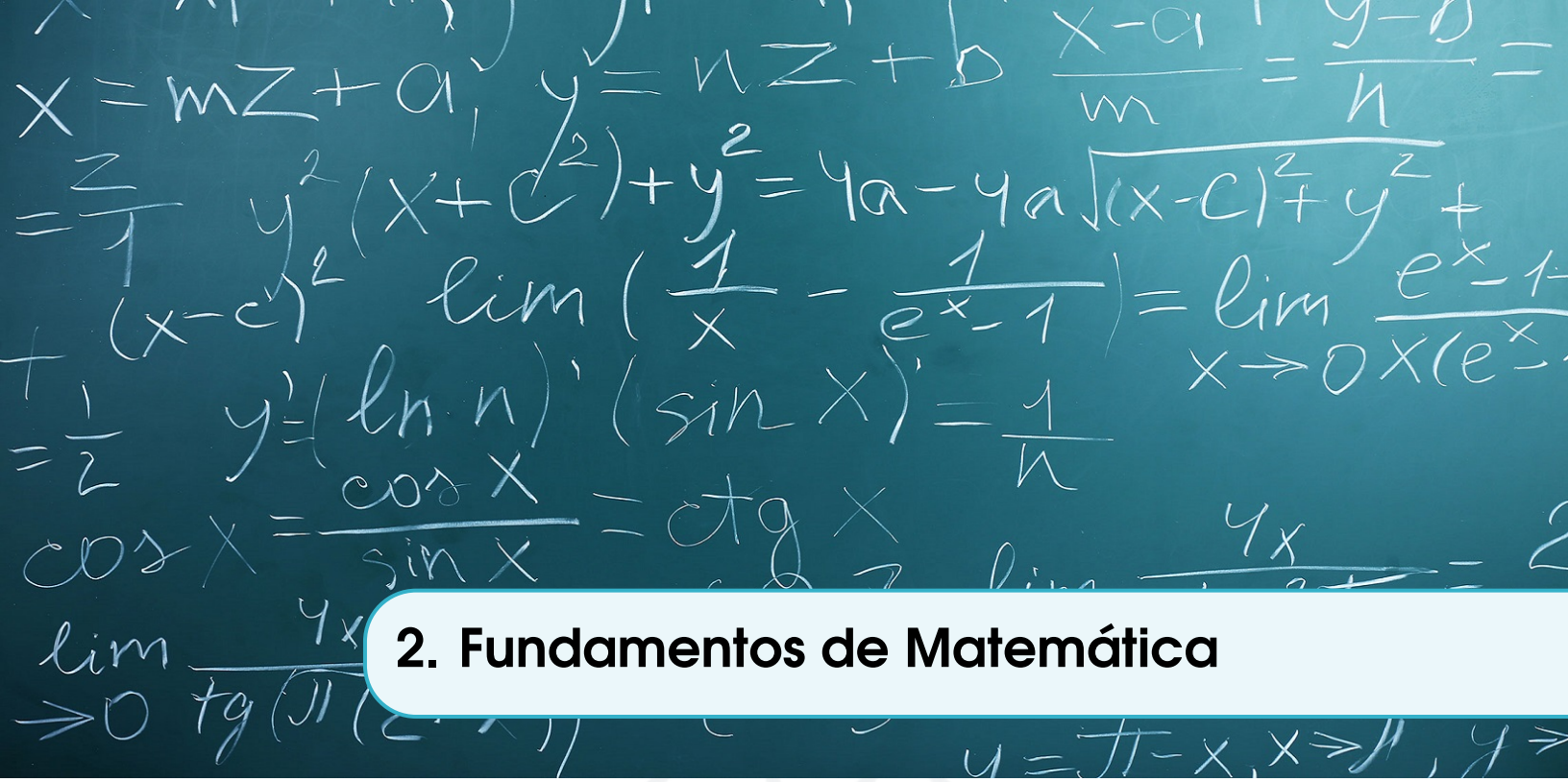
Marketing

Diretor:	Jonathan Batista	*Membro:*	Cassio Vilela
Membro:	Willy Oliveira	*Membro:*	Gustavo Santos

Relações Públicas

Diretor:	Felipe Maia Polo	*Membro:*	Eduardo Heitor

neuron



2. Fundamentos de Matemática

Por que é importante voltar aos contextos matemáticos para entender *Machine Learning* e outras tecnologias de nosso interesse? A resposta é simples e direta: os algoritmos *Machine Learning* são construídos em cima desses conceitos. Os campos mais importantes e que vamos ver rapidamente são os da Estatística, Álgebra Linear e Cálculo. É importante ressaltar que na apostila você encontrará uma visão superficial sobre as técnicas matemáticas necessárias para avançarmos, no entanto, se você decidir se aprofundar seus estudos, recomendamos fortemente que também estude bastante matemática, pois esta será necessária.

2.1 Estatística

Nesta parte do conteúdo veremos conceitos de estatística, a importância e suas aplicações na análise de dados, para auxiliar na interpretação de dados. Trataremos de medidas de tendência centrais e de dispersão e também formulação de gráficos no Python.

2.1.1 O que é Estatística?

Basicamente, a estatística é um braço da matématica que trabalha com dados para a geração de conhecimento e tomada de decisões.

É um tema grandioso que pode ser dividido em duas grandes áreas:

1. Análise Exploratória de Dados
2. Análise Confirmatória de Dados (Inferência Estatística).

Vamos começar nossa jornada do mais simples, a abordagem 1.

Na análise exploratória de dados, como o nome sugere, o interesse é a familiarização com os dados. Queremos conhecê-los, entender sua estrutura, sua grandeza. Dessa forma, saberemos que metodologia aplicar para fazermos nossas análises de forma correta.

Neste vasto campo de análise inicial dos dados, temos como principais ferramentas as medidas-resumo (média, mediana, variância...), também conhecidas como medidas de posição e variabilidade, e as técnicas gráficas cuja entrega de valor para o leitor (estudante...) é mais rápida e fácil, dado que as pessoas têm maior familiaridade com imagens, gráficos do que com tabelas.

Dado este contexto, o melhor cenário na construção de uma boa análise exploratória se dá pela combinação entre **medidas de posição e variabilidade** e **técnicas gráficas**.

Show!!! mas antes de sair jogando um monte de fórmulas nos dados precisamos entender o conceito de variável e seus tipos:

2.1.2 Variáveis e Seus Tipos

Variável é o termo utilizado para se referir a alguma característica dos dados que se está analisando. Para ficar mais claro, vamos utilizar um exemplo. Suponha o conjunto de dados abaixo:

Tabela 2.1: Pessoas aleatórias

Id	Nome	Idade (Anos)	Peso (Kg)	Grau de Instrução
1	Thiago	31	80	Superior
2	Natália	43	53	Superior
3	Augusto	60	74,5	Médio
4	Aline	27	63	Fundamental
5	Joaquim	80	63	Médio

Antes de qualquer coisa, vamos tentar entender esse conjunto de dados. Sabemos pelo título da tabela que os dados se referem a pessoas e suas características. Temos 5 variáveis, ou seja, 5 características que definem essas pessoas. Portanto, cada característica, cada variável é uma coluna da tabela. Vamos começar nossa análise coluna-por-coluna. De imediato percebemos que 3 das 5 variáveis são numéricas e as demais caracteres.

Com relação às variáveis numéricas, a primeira delas **Id**, é apenas uma variável identificadora. Neste momento não é uma variável útil para nossa análise.

A segunda variável numérica é **Idade** (em anos). Se refere a dados de contagem de tempo em números inteiros de anos. Portanto, esta variável é do tipo quantitativa discreta. Gostaria de acrescentar que essa variável poderia também ser classificada como uma variável quantitativa contínua, se assumíssemos que seus valores pudessem ser frações de ano, e não apenas números inteiros.

A terceira variável numérica é **Peso** (em Kg). É uma variável que admite valores fracionados, pertencentes ao conjunto dos números reais. Portanto, trata-se de uma variável quantitativa contínua.

Com relação às variáveis não numéricas (Nome e Grau de Instrução), podemos classificá-las de duas formas: As variáveis podem ser qualitativas nominais ou qualitativas ordinais.

As nominais tratam de categorias sem uma ordem pré-definida entre elas. É o caso da variável **Nome**, que expressa uma característica das pessoas que estão relacionadas na tabela, mas não ordena estas de nenhuma forma. Não há uma relação entre os nomes.

Já as variáveis qualitativas ordinais são aquelas onde pode-se observar uma relação entre seus valores. É o caso da variável **Grau de Instrução**, onde temos valores que seguem uma ordem entre si. É de conhecimento geral que primeiro vem o ensino fundamental, depois o ensino médio e depois o superior. A ordem é esta!

Bem, agora podemos iniciar nossos trabalhos na análise exploratória desses dados e, para isso, vamos conhecer algumas medidas-resumo.

2.1.3 Medidas de posição

Média

Vamos começar pela mais famosa das medidas de posição: a média. Esta é uma estatística, uma medida, que resume os dados em um apenas um único valor, e é definida como a soma das observações de uma série dividida pela quantidade de elementos dessa série. Confira a expressão:

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{2.1}$$

Pra ficar mais claro, vamos calcular a média das variáveis númericas (idade e peso) que temos na [**tabela 1**].

$$\begin{aligned} \bar{X}_{idade} &= \frac{\sum_{i=1}^{5} x_i}{5} \\ &= \frac{31+43+60+27+80}{5} \\ &= 48,2 \end{aligned} \tag{2.2}$$

$$\begin{aligned} \bar{X}_{peso} &= \frac{\sum_{i=1}^{5} x_i}{5} \\ &= \frac{80+53+74,5+63+63}{5} \\ &= 66,7 \end{aligned} \tag{2.3}$$

Portanto, temos que, em média, os indivíduos da tabela [**tabela 1**] têm 48,2 anos de idade e pesam, em média, 66,7Kg. Veja que esses dois números resumem essas duas séries de dados. Vale ressaltar aqui que a média é uma medida altamente influenciada por valores extremos, ou seja, valores discrepantes influenciam no cálculo da média e, dessa forma, as vezes a média pode não ser uma boa alternativa, mas deixemos essa discussão para o futuro. Em breve vamos entrar nesse assunto novamente.

Mediana

A segunda medida de posição que vamos utilizar é a mediana. Trata-se da observação central de uma série de dados **ordenada**. Perceba que a mediana somente será exatamente o termo central de uma série de dados quando essa série tiver um número ímpar de elementos. Quando a série de dados tiver um número par, a mediana será definida como a média entre os dois termos centrais. Ou seja:

$$md(X) = \begin{cases} x_{(\frac{n+1}{2})}, & \textit{se n impar} \\ \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2}, & \textit{se n par.} \end{cases}$$

Dado o método de cálculo da mediana, que leva em conta apenas os termos centrais, temos que essa medida, diferentemente da média, é robusta com relação à valores extremos, isto é, estes não influenciam em seu cálculo. Para clarificar o conceito de mediana, vamos calcular a mediana das variáveis Idade e Peso, da tabela [**tabela 1**].

Idade (ordenada): 27, 31, **43**, 60, 80

Peso (ordenado): 53, 63, **63**, 74.5, 80

Conforme destacado nas séries acima, os termos centrais das séries Idade e Peso são, respectivamente 43 e 63. Portantos, estes são os valores das medianas dessas duas variáveis.

Moda

Outra medida resumo (de posição) que temos é a Moda. Como o próprio nome já diz, ela busca o(s) elemento(s) que está(ão) na moda, ou seja, os elementos que mais se repetem, que aparecem com maior frequencia numa série de dados. Vamos ver como essa medida funciona por meio de um exemplo. Devemos encontrar as modas das variáveis Idade e Peso, da [**tabela 1**].

Para a variável idade, observe que na tabela abaixo as frequências de todos os elementos da série são iguais a 1. Portanto, não temos uma moda nessa série, dado que todos aparecem com a mesma frequência.

Tabela 2.2: Variável Idade

Elementos	Frequência
31	1
43	1
60	1
27	1
80	1

Para a variável Peso, temos uma situação diferente. Podemos observar pela tabela abaixo que o valor **63** aparece com uma frequencia superior aos demais elementos. É o único que aparece duas vezes na série, e isso o diferencia dos demais elementos. Portanto, a seria de dados relativa ao Peso tem moda e o valor dela é 63.

Tabela 2.3: Variável Peso

Elementos	Frequência
80	1
53	1
74,5	1
63	2

2.2 Medidas de Dispersão

Agora podemos dar início ao estudo das medidas de dispersão (ou variabilidade).

Às vezes apenas as medidas de posição (média, moda e mediana) não são suficientes para entendermos minimamente um conjunto de dados numa análise exploratória. Portanto, temos que utilizar as medidas de dispersão para entender melhor.

Suponha as seguintes séries de dados:

Perceba que apesar de serem séries diferentes de dados, as médias são iguais. Portanto, uma maneira de aumentar o conhecimento sobre esses dados, ou seja, de obter mais informações é conhecendo sua variabilidade e isso pode ser feito de várias maneiras, dentre as mais comuns temos as medidas de dispersão variância e desvio-padrão.

Tabela 2.4: Séries de dados diferentes, mas com valores iguais para a média

Série	Elementos	Média
A	9, 10, 11, 12, 13	11
B	7, 9, 11, 13, 15	11
C	11, 11, 11, 11, 11	11
D	9, 11, 11, 13	11

2.2.1 Variância

Por definição, a variância corresponde à relação entre o somatório dos quadrados das diferenças entre os elementos da série e sua média, e o número de observações da série. Para ficar mais didático, vamos à sua expressão:

$$Var(X) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} \tag{2.4}$$

Vamos aplicar essa expressão da variância às séries de dados da [**tabela 2.4**]:

$$\begin{aligned} Var(A) &= \frac{\sum_{i=1}^{5}(a_i - \bar{a})^2}{5} \\ &= \frac{(9-11)^2 + (10-11)^2 + (11-11)^2 + (12-11)^2 + (13-11)^2}{5} \\ &= \frac{4+1+0+1+4}{5} \\ &= \frac{10}{5} \\ &= 2 \end{aligned}$$

Faça você mesmo!

Replique a aplicação da fórmula para as demais séries da [**tabela 2.4**]. Verifique que a variância da série **D** é zero e explique porque isso acontece. Responda: existe variância negativa? pq?

Desvio-Padrão

Apesar de a variância ser uma boa medida de dispersão dos dados, o seu resultado apresenta sempre o quadrado de sua dimensão, isto é, de sua unidade de medida. Por exemplo, se a variância for calculada sobre uma séries de dados de peso de pessoas, o resultado será acompanhado da unidade Kg^2 e isso pode gerar um pouco de dificuldade de interpretação dos resultados e de possíveis comparações.

Para contornar essa situação, podemos utilizar outra medida de dispersão, o **desvio-padrão**, que é exatamente a raiz quadrada da variância. Dessa forma, o problema da dimensão fica resolvido. Segue a expressão do desvio-padrão:

$$\begin{aligned} Dp(X) &= \sqrt{Var(X)} \\ &= \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}} \end{aligned} \tag{2.5}$$

Faça você mesmo!
Calcule os desvios-padrão das séries contidas na [**tabela 2.4**].

Coeficiente de Variação

Além de todas essas medidas, temos outra que combina duas das que já vimos aqui e que nos permite fazer comparações entre séries distintas, haja vista que está livre de qualquer unidade/dimensão. Trata-se de uma medida absoluta, o **coeficiente de variação**, que é definido da seguinte maneira:

$$CV = \frac{dp(X)}{\bar{X}} \tag{2.6}$$

Ou seja, o coeficiente de variação (CV) corresponde à razão entre o desvio-padrão e a média. E porque é uma medida absoluta?

Imagine uma série de dados sobre o peso (Kg) de alguns indivíduos. Suponha que dessa série temos uma média $\bar{X} = 85kg$ e um desvio-padrão $dp(X) = 5kg$. Calculando o CV, temos:

$$\begin{aligned} CV(X) &= \frac{dp(X)}{\bar{X}} \\ &= \frac{85Kg}{5Kg} \\ &= 17. \end{aligned}$$

Perceba que na expressão acima, havia a unidade Kg tanto no numerador quanto no denominador. Se as duas grandezas tinham essa mesma unidade, ao final da divisão o resultado se apresenta de forma adimensional.

Faça você mesmo!
Calcule os CV's das séries de dados da [**tabela 2.4**]

2.2.2 Variáveis aleatórias

Dada essa paequena introdução sobre medidas-resumo é importante entrar um pouco no campo da teoria da probabilidade. A probabilidade é uma medida relativa a um Evento específico, que está inserido em um Espaço Amostral[1]. Em outras palavras, a probabilidade, em uma perspectiva clássica, mede o quão provável um evento é dadas certas circunstâncias. Para clarear essa ideia, vamos voltar ao exemplo das alturas: suponha que em nossa sala de aula tenhamos 10 alunos, sendo que 3 deles tenham mais de 1,80m. A probabilidade de se tirar aleatóriamente um aluno que tenha mais de 1,80m é $\frac{3}{10}$. Nesse caso, nosso Evento é tirar um aluno com mais de 1,80m e o Espaço Amostral são os 10 alunos que estão na sala, sendo que apenas 3 têm mais de 1,80m. Se, por exemplo, um desses 3 alunos for substituído por um de 1,70m, o Espaço Amostral muda, alterando também a probabilidade do nosso Evento.

O conceito de Variável Aleatória[2] vem e é fundamental no campo da Probabilidade. Uma Variável Aleatória é uma função que conecta um Evento contido no Espaço Amostral aos números reais. Voltando ao exemplo da sala de aula, considere a variável aleatória X[3] como sendo uma variável

[1] `https://en.wikipedia.org/wiki/Sample_space`
[2] `https://en.wikipedia.org/wiki/Random_variable`
[3] Geralmente são denotadas por letras maiúsculas.

binária que assume 1 caso se tire uma pessoa com altura maior do que 1,80m e 0 caso contrário. Então dizemos que a probabilidade de tirarmos uma pessoa maior que 1,80m no caso em que temos 3 pessoas com altura superior a 1,80m entre as 10 pessoas que estão na sala é $P(X = 1) = \frac{3}{10}$. Um outro exemplo de variável aleatória seria Y, que representa o número de pessoas com mais de 1,80m escolhidas sucessivamentes em 2 tentativas. Se consideramos a situação com reposição[4], a probabilidade de se tirar 2 pessoas com mais de 1,80m de altura é igual a $P(Y = 2) = \frac{9}{100}$. Na situação sem reposição, teríamos que $P(Y = 2) = \frac{3}{10}\frac{2}{9} = \frac{1}{15}$.

Em um contexto mais prático, podemos pensar em alguns eventos que podem ser representados por variáveis aleatórias e que são interessantes no campo do *Data Science*: o salário do brasileiro, gênero das pessoas, idade, nível educacional, uma peça de uma máquina estragar ou não, o números de gols em uma partida de futebol etc. Nos exemplos de eventos citados acima, podemos considerar que o contradomínio da função que é uma variável aleatória são os números reais em todos os casos. No entanto, a imagem dessa função varia de acordo com as especificidades de cada caso. Por exemplo, a imagem da função que liga os salários aos números reais são, teoricamente, os reais não-negativos e a imagem da função que liga os gols de uma partida aos reais é, teoricamente, os números naturais.

2.2.3 Distribuição de uma variável aleatória

Como vimos anterioemente, é possível atribuir probabilidades aos eventos contidos em um Espaço Amostral e, assim, também é possível dar probabilidades aos diversos valores, ou intervalos, que uma variável aleatória pode assumir. Antes de proseguirmos é necessário fazer a distinção entre uma variável aleatória discreta[5] e uma variável aleatória contínua[6]: basicamente a diferença entre esses dois tipos de variáveis está em seu contradomínio, sendo que se esse conjunto for um conjunto de "pontos isolados"temos uma variável discreta e se esse conjunto é contínuo, temos uma variável contínua. Exemplos de variáveis discretas são: gênero das pessoas, nível educacional, uma peça de uma máquina estragar ou não, o números de gols em uma partida de futebol. Exemplos de variáveis contínuas são: o salário do brasileiro e a altura das pessoas. É possível observar que variáveis contínuas podem ser "discretizadas", mas o contrário não é verdade.

Quando falamos em variáveis aleatórias discretas, consequentemente falamos de uma *função probabilidade*[7] que é uma função que liga um valor assumido pela variável aleatória a uma probabilidade. Se X é a face de um dado jogado (não viciado), sua função probabilidade pode ser expressa como $P(X = x)$ ou $p(x)$ sendo que X é a variável aleatória e x são possíveis valores da mesma, ou seja, elementos de seu domínio. No caso, o domínio é {1,2,3,4,5,6}:

Por outro lado, quando tratamos de variáveis aleatórias contínuas, trocamos a função probabilidade pela *função densidade*[8]. Ao contrário da função probabilidade, a função densidade não retorna a probabilidade associada a um valor específico, mas sim a densidade associada àquele valor. Na prática, isso que dizer que quando usamos variáveis aleatórias contínuas e o conceito de funções densidade, a probabilidade de uma variável contínua Y de assumir um valor específico é sempre igual a zero e, por isso, trabalhamos com intervalos. Ou seja, ao invés de perguntarmos qual a probabilidade de uma pessoa ter exatamente 1,80m (que é igual a zero), perguntaríamos "qual a probabilidade de uma pessoa ter entra 1,80m e 1,85m de altura?". Se, por exemplo, dizemos que a

[4] Ou seja, após escolhermos uma pessoa a retornamos na sala de aula.

[5] https://en.wikipedia.org/wiki/Probability_mass_function

[6] http://www.portalaction.com.br/probabilidades/23-variavel-aleatoria-continua

[7] https://en.wikipedia.org/wiki/Probability_mass_function

[8] http://www.portalaction.com.br/probabilidades/23-variavel-aleatoria-continua

x	P(X=x)
1	0,167
2	0,167
3	0,167
4	0,167
5	0,167
6	0,167

variável Z tem distribuição Normal Padrão[9], a densidade $f(z)$ é dada por:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \tag{2.7}$$

Para calcular a probabilidade de Z estar entre -1 e 2, ou seja, $P(-1 \leq Z \leq 2)$ basta calcular a área delimitada pelas retas $z = -1$, $z = 2$, $y = 0$ e pela função densidade $f(z)$[10]. Assim é possível perceber se calcularmos a área entra a função e o eixo z, com z variando de menos a mais infinito, esta vai ser igual a 1. O gráfico da função densidade nesse caso seria:

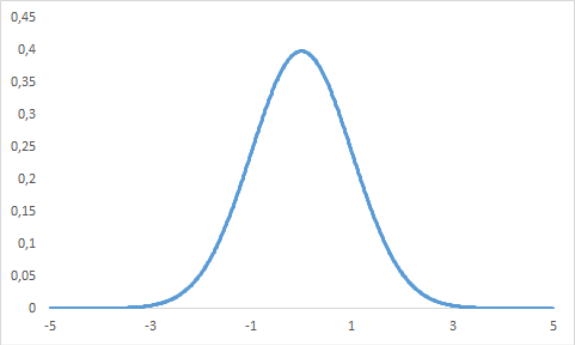


Figura 2.1: Função Densidade

[9] https://pt.wikipedia.org/wiki/Distribuiç~ao_normal

[10] Basta calcular a integral $\int_{-1}^{2} f(z)\mathrm{d}z$.

2.3 Álgebra Linear

A Álgebra Linear é uma área fundamental da matemática com uma importância tão grande quanto à do Cálculo. Os objetos de estudo da Álgebra Linear são os Espaços Vetoriais, Vetores, Matrizes, Transformações Lineares entre outras coisas. Mas qual a importância disso tudo para a análise de dados e *Machine Learning*? Grande parte dos modelos estatísticos e matemáticos que veremos daqui para frente usam os conceitos de Vetores e Matrizes em sua construção, sem falar que muitas vezes conhecê-los pode facilitar a vida um bocado. É importante salientar que o conteúdo que será exposto nesse capítulo tem caráter extremamente prático, então deixaremos as partes mais teóricas para os livros didáticos. Mesmo que já tenha cursado essa disciplina anteriormente, acompanhe o capítulo e faça os exercícios propostos. Se ainda sim não ficar satisfeito, acompanhe nossas sugestões de leitura. Vamos lá!

2.3.1 Matrizes

Matrize[11] são uma coleção[12] de números em forma de "tabela":

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \end{pmatrix} \tag{2.8}$$

A matriz A tem ordem $3x4$ (3 linhas e 4 colunas), logo dizemos que $A \in \mathbb{R}^{3x4}$. A título de curiosidade, dizemos que $A \in \mathbb{R}^{3x4}$ pertence a um espaço com 12 dimensões, pois os elementos que o habitam tem 12 entradas compostas por números reais. A um exemplo mais palpável de matriz A seria;

$$A = \begin{pmatrix} 0 & 4 & 1 & 7 \\ 9 & 3 & 4 & 1 \\ 6 & 3 & 1 & 0 \end{pmatrix} \tag{2.9}$$

2.3.2 Vetores

Os vetores são objetos que são encontrados nos Espaços Vetorias[13], que são estruturas matemáticas - se você for curioso(a) vale a pena conferir, mas se não tiver paciência, não há grandes problemas. Para nossos fins, vamos assumir que o espaço no qual estaremos trabalhando é o $\mathbb{R}^n$, sendo que n é um número Natural diferente de zero. Ou seja, os elementos que habitam o $\mathbb{R}^n$ são pontos, como os pares ordenados, mas com n coordenadas. Devido a uma série de definições e resultados matemáticos[14] cada um desses pontos representa um objeto geométrico chamado Vetor[15] e devido a esses resultados matemáticos já citados, trabalhar com os pontos no $\mathbb{R}^n$ é equivalente a trabalhar com os vetores diretamente. Então a partir de agora passaremos a chamar pontos da forma $\mathbf{x} = (x_1, x_2, ..., x_n)$ de

[11] Para mais detalhes: `https://en.wikipedia.org/wiki/Matrix_(mathematics)`
[12] Em computação, essas coleções se chamam *arrays*.
[13] `https://en.wikipedia.org/wiki/Vector_space`
[14] Base ortonormal, coordenadas e espaços isomorfos.
[15] Mais tarde veremos os vetores na forma matricial.

vetores. Além dessa forma, o vetor $\mathbf{x}$ também pode ser representado na forma matricial[16]:

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{pmatrix} \tag{2.10}$$

2.3.3 Operações com matrizes e vetores

Adição e subtração

Quando queremos somar duas matrizes (incluindo vetores na forma matricial) basta somarmos elemento por elemento. Logo, se $A+B=C$, então $a_{ij}+b_{ij}=c_{ij}$. por outro lado, se $A-B=C$, então $a_{ij}-b_{ij}=c_{ij}$. Pela definição de adição e subtração, temos que as matrizes A e B devem ter a mesma ordem nXm.

Produto por um escalar

Quando queremos multiplicar matrizes (incluindo vetores na forma matricial) por um escalar α, basta multiplicarmos todos os elementos da matriz pelo escalar. Ou seja, se $B=\alpha A$ então, $b_{ij}=\alpha a_{ij}$.

Multiplicação de duas matrizes

Quando queremos multiplicar uma matriz (incluindo vetores na forma matricial) por outra matriz temos que aplicar uma fórmula específica. Vamos supor que $A \in \mathrm{R}^{nxk}$ e $B \in \mathrm{R}^{kxm}$ então se $C=AB$, logo $c_{ij}=a_{i1}b_{1j}+a_{i2}b_{2j}+\dots+a_{ik}b_{kj}=\sum_{p=1}^{k}a_{ip}b_{pj}$. Da nossa definição seguem as seguintes conclusões: (i) o número de colunas da primeira matriz (A) deve ser igual ao número de linhas da segunda (B) e (ii) e matriz produto C tem o número de linhas da primeira e o número de colunas da segunda.

Produto elemento a elemento

Nesse caso necessitamos que as duas matrizes que estão sendo multiplicadas tenham a mesma ordem. Se $C=A\bigodot B$ então $c_{ij}=a_{ij}b_{ij}$.

Transposição

A transposição é uma operação que inverte as colunas e as linhas de uma matrix, ou seja, se $B=A^T$ é a transposta de A, então $b_{ij}=a_{ji}$.

Links externos

Para ver com mais detalhes como funcionam as operações com matrizes, basta acessar os seguintes links:

- `https://www.georgebrown.ca/uploadedFiles/TLC/_documents/Basic%20Matrix%20Operations.pdf`
- `https://www.infoescola.com/matematica/operacoes-com-matrizes-1/`
- `https://en.wikipedia.org/wiki/Matrix_(mathematics)`

2.3.4 Matrizes especiais

1. Matriz simétrica: a matriz simétrica é aquela que satisfaz $A=A^T$. Deve ser quadrada, ou seja, o número de linhas deve ser igual ao número de colunas;

[16] A convenção é de adotar vetores coluna - com n linhas e 1 coluna.

2. Matriz identidade: a matriz identidade é tem zeros em todas suas entradas, exceto pela sua diagonal principal, que é composta exclusivamente por uns. Deve ser quadrada e é denotada pela letra "i"maiúscula - I. A matriz I_n, por exemplo, é uma matriz identidade de ordem nXn. Uma propriedade de I é que se $AI_m = I_nA = A$, para qualquer matriz $A \in \mathrm{R}^{nxm}$;
3. Matriz inversa: se A^{-1} é a inversa de A, então $A^{-1}A = AA^{-1} = I$.
4. Matriz de co-ocorrência são matrizes quadradas, geralmente utilizadas em imagens, onde são utilizadas para o cálculo de descritores. Nessas matrizes são armazenadas os valores de níveis de cinza da imagem nos seguintes ângulos: $0, 45, 90, 135$, os demais ângulos são calculados por simetria [17].
5. Matriz de covariância: são as matrizes que contém as variâncias das variáveis em sua diagonal e nos demais pontos da matriz possui a covariância de todas as combinações dos pares das variáveis. São caracterizadas por serem matrizes quadradas. Geralmente são empregadas em aplicações estatísticas.

2.3.5 Produto escalar e norma euclidiana

Se ambos $\mathbf{x} = \begin{pmatrix} x_1 & x_2 & \dots & x_n \end{pmatrix}$ e $\mathbf{y} = \begin{pmatrix} y_1 & y_2 & \dots & y_n \end{pmatrix}$ são elementos do conjunto R^n então o produto escalar, ou muitas vezes referenciado como produto interno ou *inner product*, desses dois vetores é definido como $\sum_{i=1}^{n} x_i y_i = \mathbf{x}^T\mathbf{y}$. A norma euclidiana também conhecida como "módulo"do vetor pode ser calculada com a seguinte fórmula $\|\mathbf{x}\| = \sqrt{x_1^2 + x_2^2 + \dots + x_n^2} = \sqrt{\sum_{i=1}^{n} x_i^2} = \sqrt{\mathbf{x}^T\mathbf{x}}$.

2.3.6 Distância entre vetores

Se ambos $\mathbf{x} = \begin{pmatrix} x_1 & x_2 & \dots & x_n \end{pmatrix}$ e $\mathbf{y} = \begin{pmatrix} y_1 & y_2 & \dots & y_n \end{pmatrix}$ são elementos do conjunto R^n então uma distância entre $\mathbf{x}$ e $\mathbf{y}$ é uma função $d : \mathrm{R}^n \; X \; \mathrm{R}^n \to \mathrm{R}$ tal que (i) $d(\mathbf{x},\mathbf{y}) \geq 0$, (ii) $d(\mathbf{x},\mathbf{y}) = d(\mathbf{y},\mathbf{x})$, (iii) $d(\mathbf{x},\mathbf{y}) \leq d(\mathbf{x},\mathbf{z}) + d(\mathbf{z},\mathbf{y})$ para $\mathbf{z} \in \mathrm{R}^n$ e (iv) $d(\mathbf{x},\mathbf{y}) = 0$ se e somente se $\mathbf{x} = \mathbf{y}$. Observe que dizer $\mathbf{x} = \mathbf{y}$ é igual a dizer que $x_i = y_i$ para qualquer i. Vamos agora introduzir as distâncias mais importantes para a nossa evolução no conteúdo:

Distância *Manhattan* ou *Taxicab*

Dados $\mathbf{x},\ \mathbf{y} \in \mathrm{R}^n$, a Distância *Manhattan*[18] é $L = \sum_{i=1}^{n} |x_i - y_i|$.

Intuitivamente se imaginarmos um triangulo retangulo a distancia de manhattan é a soma do tamanho dos catetos.

Distância Euclidiana

Dados $\mathbf{x},\ \mathbf{y} \in \mathrm{R}^n$, a Distância Euclidiana[19] é $L^2 = \|\mathbf{x} - \mathbf{y}\| = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} = \sqrt{(\mathbf{x}-\mathbf{y})^T(\mathbf{x}-\mathbf{y})}$.

Novamente imaginando o triangulo retangulo, a distancia euclidiana seria o comprimento da hipotenusa.

Distância de Mahalanobis

Dados $\mathbf{X},\ \mathbf{Y} \in \mathrm{R}^n$ vetores aleatórias e $\mathbf{x}$ e $\mathbf{y}$ realizações dos mesmos, a Distância de Mahalanobis[20] empírica é dada por $D_M = D_M(\mathbf{x},\mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})^T S^{-1}(\mathbf{x}-\mathbf{y})}$, sendo que S é a matriz estimada de covariância entre $\mathbf{X}$ e $\mathbf{Y}$.

[17] http://iris.sel.eesc.usp.br/wvc/Anais_WVC2012/pdf/97967.pdf

[18] https://en.wiktionary.org/wiki/Manhattan_distance

[19] https://en.wikipedia.org/wiki/Euclidean_distance e http://mathonline.wikidot.com/the-distance-between-two-vectors

[20] https://en.wikipedia.org/wiki/Mahalanobis_distance

Exercício

Se $\mathbf{x} = (1,5)$, $\mathbf{y} = (2,-4)$, $S_X = 2$ (desvio-padrão), $S_Y = 1$ (desvio-padrão) e $S_{XY} = 3$ (covariância), calcule:

- L
- L^2
- D_M

2.3.7 Materiais complementares

Para ter uma melhor noção de Álgebra Linear com aplicações em Python, acesse: http://www.ulaff.net/. Para entender um pouco mais a Lógica do Machine Learning `http://www.deeplearningbook.org/contents/ml.html`. Outros materiais de apoio para entender melhor KNN:

- `https://kevinzakka.github.io/2016/07/13/k-nearest-neighbor/`
- `http://cs231n.github.io/classification/#nn`
- `https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm`
- `https://saravananthirumuruganathan.wordpress.com/2010/05/17/a-detailed-introduction-to-k`

2.4 Otimização de funções

2.4.1 O conceito de Derivadas

Vocês já pararam para se perguntar qual é o significado do conceito de Velocidade e por que esta tem sempre uma unidade de espaço dividida por uma de tempo? Velocidade não é nada mais que uma medida de variação de um corpo no espaço dividida por uma medida na variação de tempo que o corpo tomou para se deslocar. Quando a variação de tempo é uma constante maior do que zero, se aplicarmos a definição de velocidade exposta acima, teremos a velocidade média do corpo durante seu percurso em um determinado intervalo de tempo, logo $V_m = \frac{\Delta y}{\Delta x}$. Para facilitar o raciocínio, estamos assumindo que o corpo viaja em uma linha reta ao longo do eixo y e que o eixo x é o tempo. Na Figura 2.4.1 [21] é possível ver a situação de forma ilustrada.

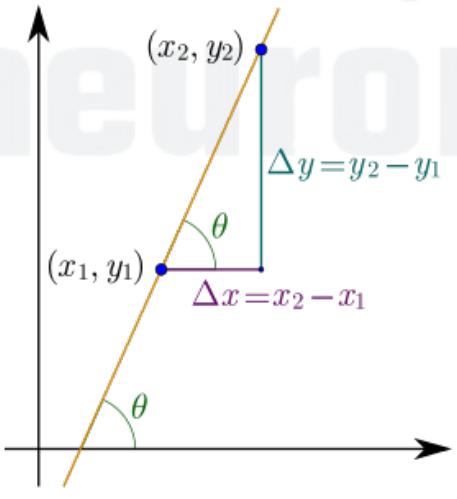


Figura 2.2: Função deslocamento 1

[21] `https://en.wikipedia.org/wiki/Derivative`.

Pela nossa definição de Velocidade Média, esta é dada pela tangente do ângulo θ, ou seja, $V_m = \frac{\Delta y}{\Delta x} = \text{tg}(\theta)$. Nossa função de deslocamento do veículo na Figura 2.4.1 é linear e positivamente inclinada, ou seja, a velocidade com a que o carro se move é constante e positiva $\rightarrow \Delta y$ e Δx têm sempre sinais iguais. O caso da Figura 2.4.1 é um tanto restrito, pois estamos tratando de uma função linear. Por que não partimos para uma função não-linear agora? Vamos lá. na Figura 2.4.1 [22], a curva azul denota nossa função deslocamento e a curva roxa uma reta secante "cortando"a função nos pontos $(x_1, y_1) = (x, f(x))$ e $(x_2, y_2) = (x+\Delta x, f(x+\Delta x)) = (x+h, f(x+h))$. Logo a velocidade média do carro durante esse deslocamento é dada pelo coeficiente angular (*slope*) da reta roxa. Em outras palavras, a tangente do menor ângulo formado entre a reta roxa e o eixo x é a velocidade média no trecho $f(x) \rightarrow f(x+h)$.

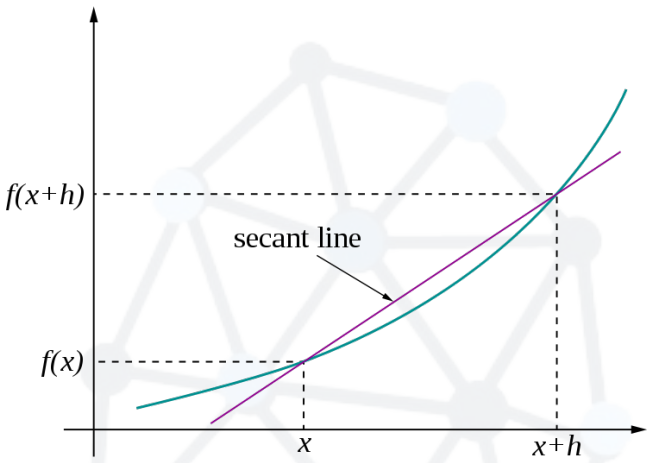


Figura 2.3: Função deslocamento 2

Aí surge uma outra pergunta muito interessante: e quando olhamos para o velocímetro do carro? O que exatamente aquilo está no dizendo? E a resposta é: o velocímetro nos dá a velocidade instantânea do carro e não a velocidade média. Ou seja, o intervalo de tempo (Δx ou h) que o carro utiliza para calcular a velocidade mostrada no velocímetro é tão pequeno que dizemos que este é **infinitesimal** ou infinitesimalmente pequeno[23]. Quando temos um $\Delta x = h$ infinitesimalmente pequeno, que **tende** a zero, a reta, que antes era secante, agora é tangente à função em x (imagem) e seu coeficiente angular dá a velocidade que o velocímetro indica no momento. Veja a Figura 2.4.1:

As notações mais usuais para denotar a inclinação (coeficiente angular) da reta tangente a uma função $f(.)$ quando $x = x_0$ são $f'(x_0)$ e $\frac{\text{d}f(x_0)}{\text{d}x}$. Na segunda notação dx faz o papel do Δx e o $df(x_0)$ faz o papel do $\Delta f(x)$, ambos quando Δx é infinitesimal. Essa inclinação é nada mais do que a derivada no ponto $x = x_0$. Assim sendo, a derivada tem dois grandes significados:

- Geométrico: a inclinação da reta tangente em um certo ponto;
- Abstrato: denota a relação entre duas grandezas. No caso do carro, a pergunta seria "**o que acontece com a posição do carro quando o tempo passa mais um pouquinho (relativamente ao tempo passado)?"**

[22] `https://en.wikipedia.org/wiki/Derivative`.

[23] Essa ideia de número infinitesimal está intimamente ligada ao conceito de Limite - `https://en.wikipedia.org/wiki/Limit_(mathematics)`.

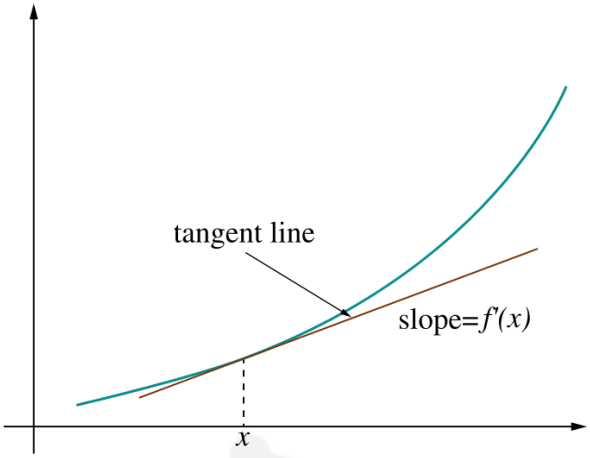


Figura 2.4: Função deslocamento 3

Se a derivada no ponto $x = x_0$ é 4, ou seja, $\frac{\mathrm{d}f(x_0)}{\mathrm{d}x} = 4$, para cada unidadezinha de tempo passada, o carro se move 4 vezes mais, relativamente à essa unidadezinha, para frente. Se $\frac{\mathrm{d}f(x_0)}{\mathrm{d}x} = -4$, a interpretação é a mesma, no entanto, os deltas têm sinal contrários e o carro está indo para trás. Agora, se $\frac{\mathrm{d}f(x_0)}{\mathrm{d}x} = 0$, quer dizer que o carro está momentaneamente parado (provavelmente fazendo uma conversão). Se eu fizer o cálculo da derivada para um x genérico, teremos o que se chama de **função derivada**. Existem técnicas para se derivar uma função $f(x)$ e você **deve** saber as mais básicas. Para isso basta assistir os vídeos 6, 7, 8, 9 e 10 deste curso de derivadas (clique aqui)[24].

Um exemplo aqui é muito importante: vamos assumir $f(x) = x^2$. Pela "regra do tombo", nossa função derivada é $f'(x) = \frac{\mathrm{d}f(x)}{\mathrm{d}x} = 2x$. Quando $x = -2$ temos $f'(-2) = -4$, $x = 0$ implica em $f'(0) = 0$ e $x = 2$ implica em $f'(2) = 4$. Ou seja, quando $x = -2$ a função é decrescente (derivada negativa) e quando $x = 2$ a função é crescente (derivada positiva). Por outro lado, quando $x = 0$ a função não é nem decrescente e nem crescente. Nesse caso, dizemos que $x = 0$ é responsável pelo ponto ponto crítico $(0, f(0)) = (0,0)$ (quando a derivada é **nula**). Observe que se $f(x)$ é diferenciável [25], ser ponto crítico é condição necessária para um ponto ser um ponto de máximo ou de mínimo (onde há uma "conversão"[26]). Por exemplo, se uma função diferenciável $g(x)$ é o custo de uma empresa, quando a empresa minimiza seu custo, temos que $g'(x) = 0$. Geralmente, quando queremos encontrar o mínimo ou máximo de uma função diferenciável, a derivamos e a igualamos a zero, caindo em uma equação. Resolvendo a equação, chegamos a possíveis valores de x que otimizam a função, a partir daí, testamos os valores possíveis de x e descobrimos qual estamos procurando.

Exercícios

Descubra $f'(x)$ para as seguintes funções:

- $f(x) = 2x^4$, Resposta: $f'(x) = 8x^3$
- $f(x) = 2xe^x$, Resposta: $f'(x) = 2(e^x + xe^x)$

[24] Você pode baixar a uma tabela de derivadas aqui `https://drive.google.com/file/d/0B1hOECdZUa2xMjA1elMxY1BsbFk/view`

[25] Se tiver curiosidade ver `https://en.wikipedia.org/wiki/Derivative`. Se não tiver, não é muito necessário, dado que as funções vistas no *Machine Learning são diferenciáveis.*

[26] Ou a função está decrescente e começa a ser crescente ou vice-versa

- $f(x) = e^{6x}$, Resposta: $f'(x) = 6e^{6x}$

2.4.2 Otimizando uma função na prática

Suponha que queremos minimizar a função $f(x) = 2 + (x-5)^2$ para qualquer x no conjunto dos Reais. Seguiremos, então, os passos apresentados anteriormente: (i) derivar e igualar a zero, (ii) resolver a equação e testar os possíveis candidatos e ver qual dá o mínimo. Derivando $f(x)$ igualando a zero, temos que $f'(x) = 2(x-5) = 0 \Rightarrow x^* = 5$ é o nosso ponto de mínimo. Como é possível perceber, tivemos apenas um candidato e isso ocorreu pois a função $f(x)$ tem pelo menos um ponto de mínimo e é **convexa**[27] em todo seu domínio. Mas e se tivéssemos que resolver esse problema no Python, como seria?

O Método do *Gradient Descent* para a minimização de funções

Para resolver esse problema no Python, até certa parte do caminho o raciocínio é idêntico. Primeiramente, temos que calcular a derivada usando uma caneta e papel - no caso de $f(x)$, sabemos que sua derivada $f'(x) = 2(x-5)$. Em segundo lugar aplicamos o método do *Gradient Descent* (Gradiente Descendente)[28]. Este método é bem simples e segue o seguinte roteiro: (i) escolher um x_0 inicial[29], (ii) calcular a derivada no ponto $x = x_0$, (iii) fazer o *update* do x com a seguinte fórmula $x_{n+1} = x_n - \gamma f'(x_n)$ e repetir o processo até que a sequência $\{x_n\}$ tenha convergido. Aqui não estamos tão preocupados com a convergência no processo e adotaremos um número de repetições grande para garantir que isso aconteça. O γ na nossa fórmula é um número arbitrário e pequeno (e.g. $0,001$) mais conhecido na linguagem do *Machine Learning* como *Learning Rate*, ou seja, é a taxa que o algoritmo usa para "aprender"os padrões nos dados.

Antes de entrar mais a fundo no exemplo, preste muita atenção na Figura 2.4.2. O gráfico nos mostra que à esquerda do $x = 5$ a função é decrescente e à sua direita a função é crescente. É importante relembrar que a função ser decrescente em certo intervalo é o mesmo que dizer que sua derivada é negativa. Por outro lado, dizer que é crescente dá na mesma que dizer que a função tem derivada positiva.

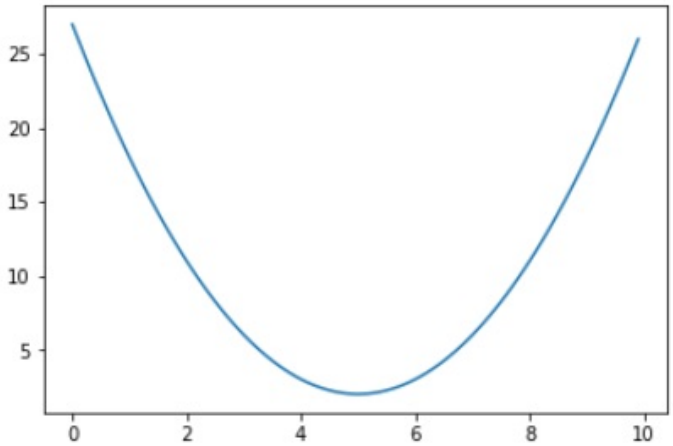


Figura 2.5: $f(x) = 2 + (x-5)^2$

[27] `https://en.wikipedia.org/wiki/Convex_function`
[28] `https://en.wikipedia.org/wiki/Gradient_descent.`
[29] Geralmente randômico e pequeno.

Vamos então seguir os passos para a minimização da função pelo método do *Gradient Descent*: escolhemos um $x_0 = 2$ e uma *Learning Rate* $\gamma = 0,001$ para começar. Em seguida, é necessário aplicar a fórmula para o *update* do nosso x: $x_1 = x_0 - \gamma f'(x_0) = 2 - 0,001(-6) = 2,006$. Seguindo a mesma lógica $x_2 = x_1 - \gamma f'(x_1) = 2,006 - 0,001(-5,988) \approx 2,012$. Perceba que quando vou atualizando o x, este vai cada vez mais se aproximando do nosso x ótimo, ou seja, o 5. Perceba também que conforme o x vai se aproximando de 5, o valor absoluto da derivada $f'(x)$ vai se aproximando do zero. Assim sendo, quando o x estiver perto o suficiente de 5, a atualização pela fórmula $x_{n+1} = x_n - \gamma f'(x_n)$ vai deixando de surtir efeito. Quando isso ocorre, dizemos que o x **convergiu** para seu valor ótimo, ou seja, o valor que nesse caso minimiza a função. É importante que você se convença que se começássemos com um $x > 5$, por exemplo, $x = 10$ a mesma coisa ocorreria. A única diferença é que o x estaria se aproximando de 5 pela direita e não pela esquerda.

Falta decidir apenas uma coisa para o nosso algoritmo (processo) de minimização - quantas vezes faremos esse *update*. O número de repetições é arbitrário e é conhecido na linguagem de *Machine Learning* como *steps* ou *epochs*. Como padrão nos nossos exemplos, vamos trabalhar com *epochs*= 10000. No entanto, isso não é uma regra e ao longo do seu percurso de aprendizado, você irá aprender melhor como decidir que numero usar e algumas implicações dessa escolha. **Vamos ao código!!! :)**

Exemplo/Exercício 1

Faça o download do Jupyter Notebook e leia o Exemplo 1. Por favor, resolva o Exercício 1.

2.4.3 Otimização de funções com mais de uma variável real

Até o momento estamos trabalhando com funções univariadas (com uma variável real) da forma $f(x)$. Agora começaremos a trabalhar com funções de várias variáveis reais, algo da forma $f(x_1, x_2, ..., x_n)$. Um exemplo de função do tipo seria $f(x,y) = x^2 + y^2$ - nesse caso específico, o domínio da nossa função é um conjunto de pares ordenados, como por exemplo o $\mathbb{R}^2$, que é o conjunto de pares ordenados (x,y) sendo que x e y são números reais.

Derivadas Parciais e Vetor Gradiente

O raciocínio da otimização de funções de várias variáveis reais é análogo ao raciocínio que desenvolvemos para funções univariadas, no entanto, devemos fazer adaptações para caso de maior dimensão. Primeiramente, é necessário ver o conceito de **Derivada Parcial**[30] e para isso devemos relembrar do significado abstrato da derivada no caso univariado: no caso univariado a derivada nos diz a resposta de uma variável y a uma pequena variação de uma variável x. No caso das Derivadas Parciais, a interpretação é a mesma: como um exemplo, tome $z = f(x,y) = x^2 + y^2 + xy$ sendo que estamos interessados no efeito que uma pequena variação na variável y tem sobre a variável z, mantendo x constante. Nesse caso, basta derivar parcialmente z em relação a y - as regras de derivação são as mesmas do caso univariado considerando x como uma constante. O resultado[31] desse cálculo seria $\frac{\partial z}{\partial y} = \frac{\partial f(x,y)}{\partial y} = 0 + 2y + 1x = 2y + x$.

Analogamente ao caso univariado, em uma situação de optimalidade, é necessário que todas as derivadas parciais, ou seja, derivadas em relação a todas as variáveis, sejam iguais a zero. No caso em que $f(x,y) = x^2 + y^2$, por exemplo, a situação de optimalidade exige que $\frac{\partial f(x,y)}{\partial x} = \frac{\partial f(x,y)}{\partial y} = 0$, pois a função $f(x,y)$ é diferenciável. Ou seja, $2y = 2x = 0 \Rightarrow x^* = y^* = 0$, logo o resultado ótimo e que dá o valor mínimo da função f é o ponto $(0,0)$. Para facilitar as contas e os algoritmos futuros,

[30] `https://pt.wikipedia.org/wiki/Derivada_parcial`

[31] Se não entendeu o resultado do cálculo, favor voltar aos materiais indicados para as regras de derivação.

introduziremos agora o conceito de Vetor Gradiente, que é o lugar da onde saiu o nome do nosso algoritmo *Gradient Descent*.

O Vetor Gradiente é um vetor que está relacionado a uma função multivariada da forma $f : \mathbb{R}^n \to \mathbb{R}$ e sempre terá a ordem n. Se $f = f(x_1, x_2, ..., x_n)$ então o Vetor Gradiente de f é $\nabla f = \nabla f(x_1, x_2, ..., x_n) = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, ..., \frac{\partial f}{\partial x_n}\right)$ ou na forma matricial $\nabla f = \nabla f(x_1, x_2, ..., x_n) = \left(\frac{\partial f}{\partial x_1} \ \frac{\partial f}{\partial x_2} \ ... \ \frac{\partial f}{\partial x_n}\right)^T$. Se f é definida como $f(x,y) = x^2 + y^2$, logo o Vetor Gradiente de f é $\nabla f = \nabla f(x,y) = \left(\frac{\partial f}{\partial x} \ \frac{\partial f}{\partial y}\right)^T = (2x \ \ 2y)^T$. Logo a condição necessária para a optimalidade, para funções bem comportadas, é de que o vetor gradiente seja nulo.

Exercícios

Descubra ∇f para as seguintes funções:

- $f(x,y) = 2yx^4$
- $f(x,y) = 2xe^{xy}$
- $f(x,y,z) = xyz + ze^{6xy}$

Adaptando o método do *Gradient Descent* para o caso multivariado

A adaptação do método de minimização de funções utilizado até agora é simples. Considere que estamos trabalhando com uma função bivariada $f(x,y)$[32], aí então precisamos escolher um ponto inicial (x_0, y_0) e depois aplicar o *update* sucessivamente para que nosso par (x,y) convirja para o ponto ótimo (x^*, y^*). A regra do *update* no caso multivariado fica $(x_n, y_n) = (x_{n-1}, y_{n-1}) - \gamma \nabla f(x_{n-1}, y_{n-1})$. Se você prestar bem a atenção, o caso univariado não nada mais ou menos do que um caso particular.

Vamos dizer então que $f(x,y) = x^2 + y^2$ e que escolhemos uma *Learning Rate* $\gamma = 0,001$ e um ponto inicial $(x_0, y_0) = (-2, 3)$. Logo $(x_1, y_1) = (x_0, y_0) - \gamma \nabla f(x_0, y_0) = (-2,\ 3) - 0,001(-4,\ 6) = (-1.996,\ 2.994)$ e, seguindo a mesma lógica, $(x_2, y_2) = (x_1, y_1) - \gamma \nabla f(x_1, y_1) = (-1.996,\ 2.994) - 0,001(-3.992,\ 5.988) \approx (-1.992,\ 2.988)$. Como vimos anteriormente, o ponto que minimiza a função nesse caso é $(0,0)$ e conforme os *updates* vão sendo feitos a sequência de pontos $\{(x_n, y_n)\}$ se aproxima do ponto ótimo. É importante salientar que, assim como no caso univariado, conforme vamos nos aproximando do ponto ótimo, o Vetor Gradiente vai se aproximando do vetor nulo, até que os *updates* deixam de surtir efeito e dizemos que houve convergência. **Vamos ao código!**

Exemplo/Exercício 2

Faça o download do Jupyter Notebook e leia o Exemplo 2. Por favor, resolva o Exercício 2.

2.4.4 *Machine Learning* - Ajuste de curvas entre pontos

O ajuste de curvas em nuvens de pontos é um velho problema da matemática e da estatística. A resolução desse problema se dá pelo método dos Mínimos Quadrados que é a base para o conceito de **Regressão Linear**. Suponha que você tenha a seguinte situação: um(a) vendedor(a) de sorvetes percebe que a venda de sorvetes está positivamente relacionada à temperatura média do dia. Após ter coletado os dados de temperatura média (eixo x) e quantidade de sorvetes vendidos (eixo y) chegamos ao seguinte gráfico (2.4.4):

Um(a) cientista de dados quer saber agora qual a função linear da forma $f(x) = wx$ que melhor descreve essa relação. No entanto, a constante w é desconhecida. Para isso, o(a) especialista em análise de dados propõe o método dos Mínimos Quadrados para a estimação do parâmetro w. Nossa

[32] Para se trabalhar com N variáveis é só estender o raciocínio.

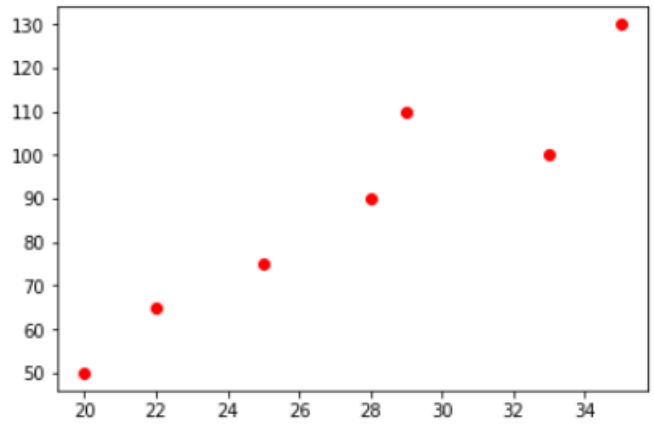


Figura 2.6: Temperatura Média Vs. Venda de sorvetes no dia

amostra é composta pelos pares $(20,50),(22,65),...,(35,130)$, sendo que temos 7 pontos (pares) no total. Logo, nossa amostra pode ser reescrita da seguinte maneira $(x_0,y_0),(x_1,y_1),...,(x_6,y_6)$. O método dos Mínimos Quadrados se consiste em encontrar o valor de w que **minimiza** o Erro Quadrático Médio (EQM). No nosso caso, o EQM é uma função de w e é dado pela equação 2.11:

$$\text{EQM}(w) = \frac{1}{7}(y_0 - wx_0)^2 + ... + \frac{1}{7}(y_6 - wx_6)^2 = \frac{1}{7}\sum_{i=0}^{6}(y_i - wx_i)^2 = \frac{1}{7}(\mathbf{y} - w\mathbf{x})^T(\mathbf{y} - w\mathbf{x}) \quad (2.11)$$

Se **y** e **x** forem vetores que contém as amostras. Pela resolução conseguimos traçar uma reta entre os pontos com uma equação igual a $y = 3.28x$. Veja na Figura 2.4.4:

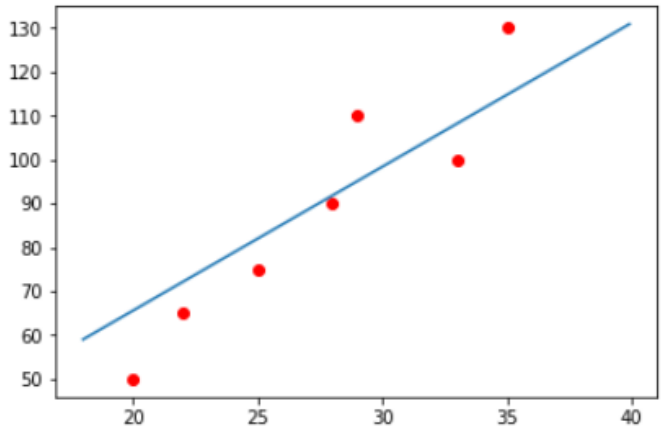


Figura 2.7: Temperatura Média Vs. Venda de sorvetes no dia

Para aplicarmos o Método de *Gradient Descent* e achar o w que minimiza o EQM, basta seguir os passos que repetimos diversas vezes anteriormente!

Exemplo/Exercício 3

Faça o download do Jupyter Notebook e leia o Exemplo 3. Por favor, resolva o Exercício 3.

2.4.5 Observações finais

Nos exemplos apresentados durante essa parte de otimização de funções focamos praticamente na minimização de funções. Fizemos isso pois na maioria das aplicações que temos em *Machine Learning* precisamos minimizar funções ao invés de maximizar. No entanto, se houver a necessidade de se maximizar uma função, é possível fazê-lo com pequenas adaptações no algoritmo do *Gradient Descent* (trocar o sinal de menos pelo sinal de mais). Outra observação importante é que muitas funções que poderíamos querer minimizar não são convexas e isso pode desembocar em situações que há mais de um ponto de mínimo (problema!!!). No entanto, aqui focamos em funções convexas e que, consequentemente, têm somente um ponto de mínimo. Os algoritmos mais simples de *Machine Learning* não sofrem com esse problema e os mais complexos, como as Redes Neurais, podem ou não sofrer com o problema - assunto para o futuro!

2.4.6 Materiais complementares

Se você achou o material muito básico até aqui, sugiro que leia o capítulo `http://www.deeplearningbook.org/contents/numerical.html`. Outros materiais de apoio:

- `https://www.youtube.com/playlist?list=PL2Wzg8U7YXS9IganqhtPS2YSgZ8dtAM3N`
- `https://www.youtube.com/playlist?list=PL2Wzg8U7YXS9HjwRCy_1kcAMHLFUoV9tu`
- `https://en.wikipedia.org/wiki/Limit_(mathematics)`
- `https://en.wikipedia.org/wiki/Derivative`
- `https://en.wikipedia.org/wiki/Partial_derivative`
- `https://en.wikipedia.org/wiki/Gradient`
- `https://en.wikipedia.org/wiki/Gradient_descent`