Big Data e Machine Learning com Hadoop e Spark

Conteúdo

CONTEÚDO PROGRAMÁTICO

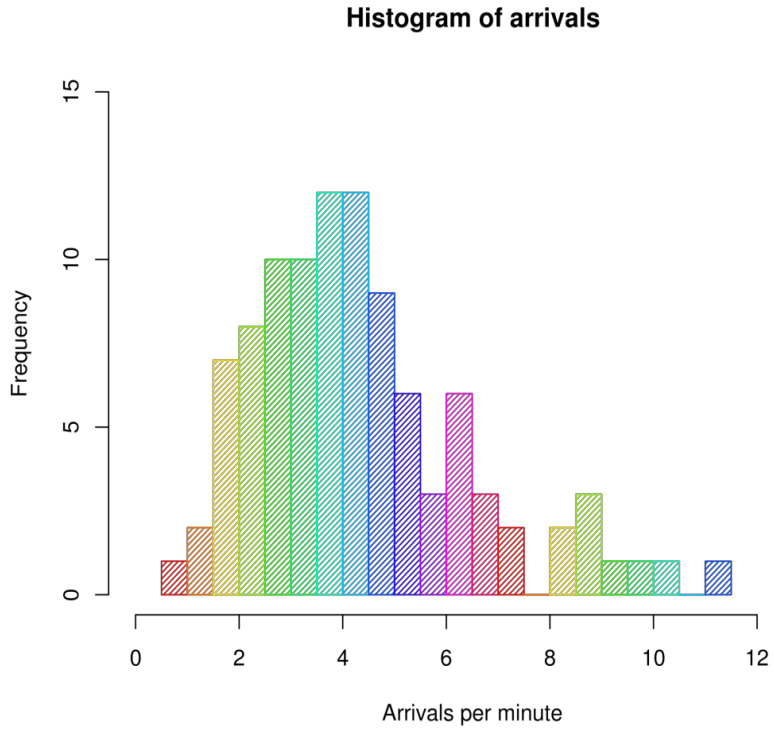
- Visão geral da ciência de dados e aprendizado de máquina em escala
- Visão geral do ecossistema do Hadoop
- Instalação de um Cluster Hadoop
- Trabalhando com dados do HDFS e tabelas do Hive usando o Hue
- Visão geral do Python
- Visão geral do R
- Visão geral do Apache Spark 2
- Leitura e gravação de dados
- Inspeção da qualidade dos dados
- Limpeza e transformação de dados
- Resumindo e agrupando dados
- Combinando, dividindo e remodelando dados
- Explorando dados
- Configuração, monitoramento e solução de problemas de aplicativos Spark
- Visão geral do aprendizado de máquina no Spark MLlib
- Extraindo, transformando e selecionando recursos
- Construindo e avaliando modelos de regressão
- Construindo e avaliando modelos de classificação
- Construindo e avaliando modelos de cluster
- Validação cruzada de modelos e hiperparâmetros de ajuste
- Construção de pipelines de aprendizado de máquina
- Implantando modelos de aprendizado de máquina

MATERIAL DIDÁTICO

- Slides do treinamento em PDF
- GitHub com exercícios e códigos exemplo
- Máquinas virtuais para exercícios simulados
- Gravação das aulas disponível durante 3 meses

Desvio Padrão e Variância

Exemplo de um Histograma



Variância mede quão “espalhados” os dados são.

- Variância (σ^2)é simplesmente a **média das diferenças quadradas da média**
- Exemplo: Qual a variância deste dataset (1, 4, 5, 4, 8)?
  - Calcule a Média: (1+4+5+4+8)/5 = 4.4
  - Agora encontre as diferenças da média: (-3.4, -0.4, 0.6, -0.4, 3.6)
  - Encontre o quadrado das diferenças: (11.56, 0.16, 0.36, 0.16, 12.96)
  - Calcule a média do quadrado das diferenças:

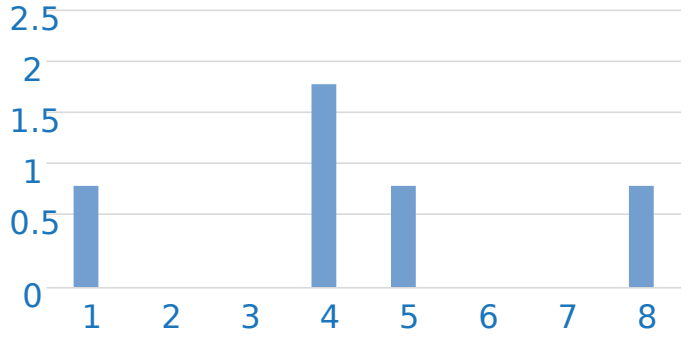
$$\sigma^2 = (11.56 + 0.16 + 0.36 + 0.16 + 12.96) / 5 = 5.04$$

Desvio Padrão é a raiz quadrada da Variância

$$\sigma 2 = 5.04$$

$$\sigma = \sqrt{5.04} = 2.24$$

Então o Desvio Padrão de (1, 4, 5, 4, 8) é 2.24.



Isso é normalmente usado para identificar outliers. Pontos que ficam a mais de um Desvio Padrão da Média podem ser considerados não usuais.

Você pode se referei a quão extremo é um ponto de dados dizendo “quantos sigmas” longe da média ele está.

População vs. Amostra

- Se você está trabalhando com uma Amostra dos dados ao invés de Um dataset completo de dados (a *População* inteira)...
  - Então você vai querer usar a “variância da amostra” ao invés da “variância da população”
  - Para N amostras, você divide a variância quadrada por N-1 ao invés de N.
  - Então, no nosso exemplo, calculamos a variância da população assim:
    - $\sigma^2 = (11.56 + 0.16 + 0.36 + 0.16 + 12.96) / 5 = 5.04$
  - A variância da amostra deve ser:
    - $S^2 = (11.56 + 0.16 + 0.36 + 0.16 + 12.96) / 4 = 6.3$

Fórmulas

- Variância da População:

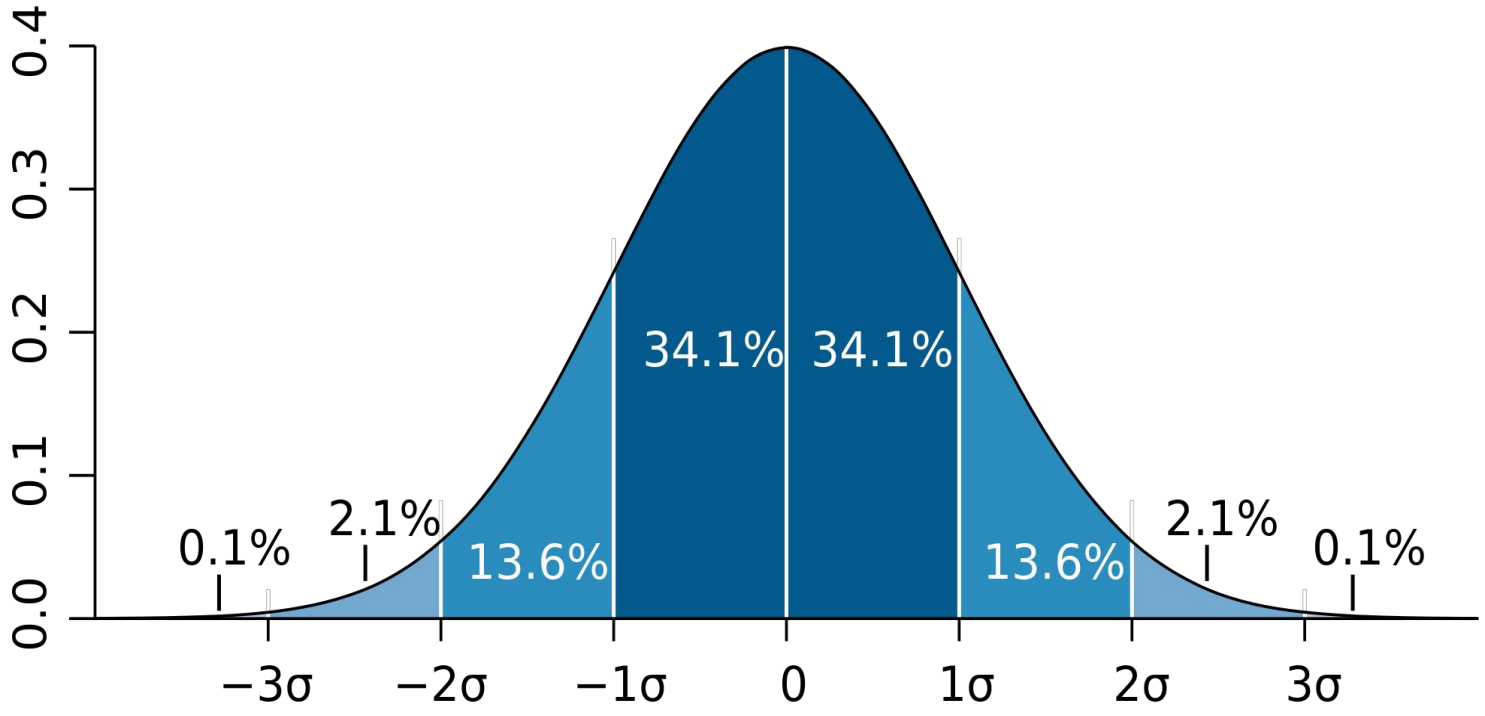
$$\sigma^2 = \frac{\sum(X-\mu)^2}{N}$$

- Variância da Amostra:

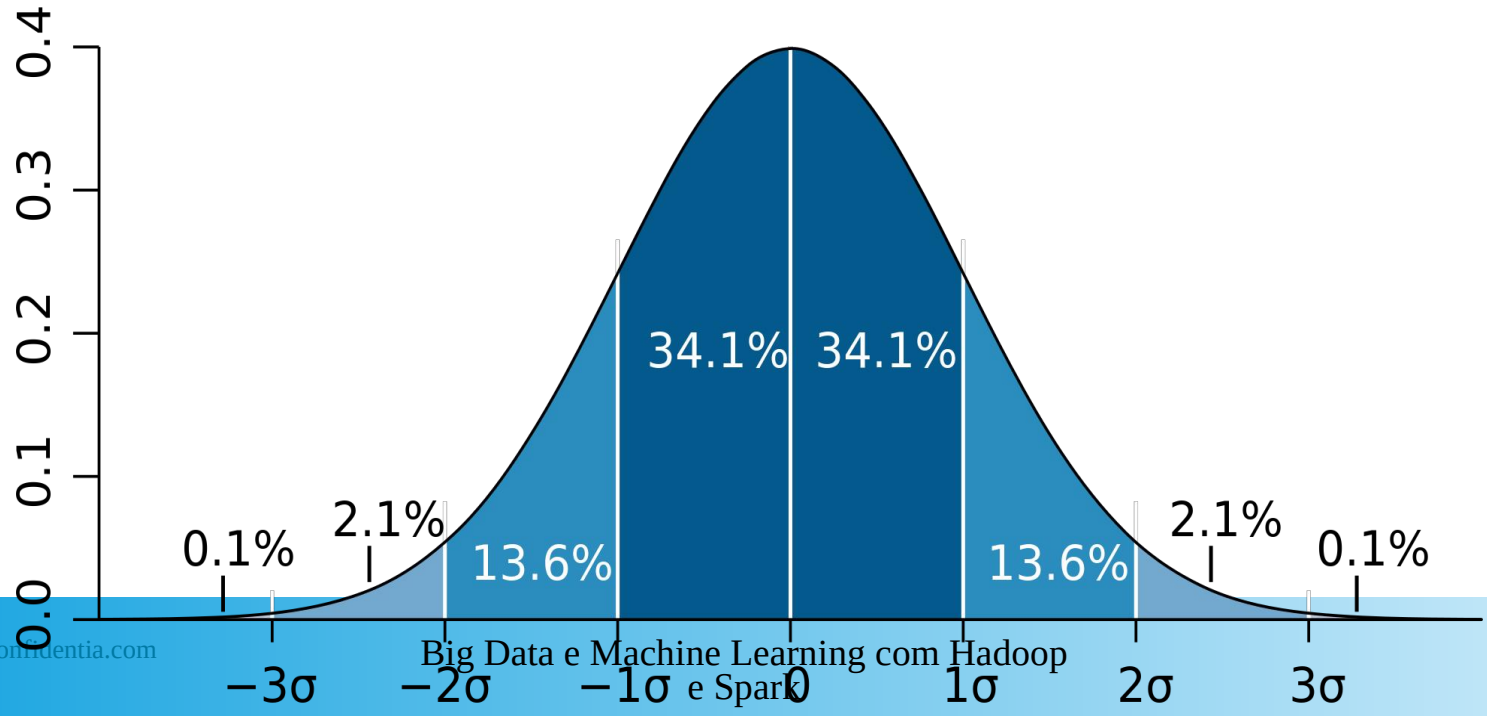
$$s^2 = \frac{\sum(X-M)^2}{N-1}$$

Funções de densidade de probabilidade

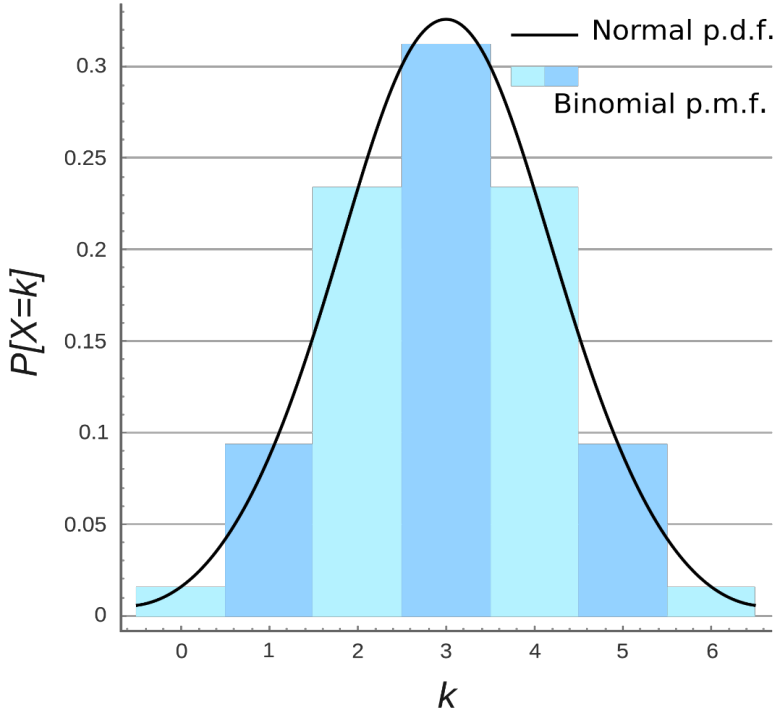
Exemplo: uma “distribuição normal



Dá a probabilidade de um ponto de dados cair dentro de um dado intervalo de um dado valor.



Função de massa de probabilidade

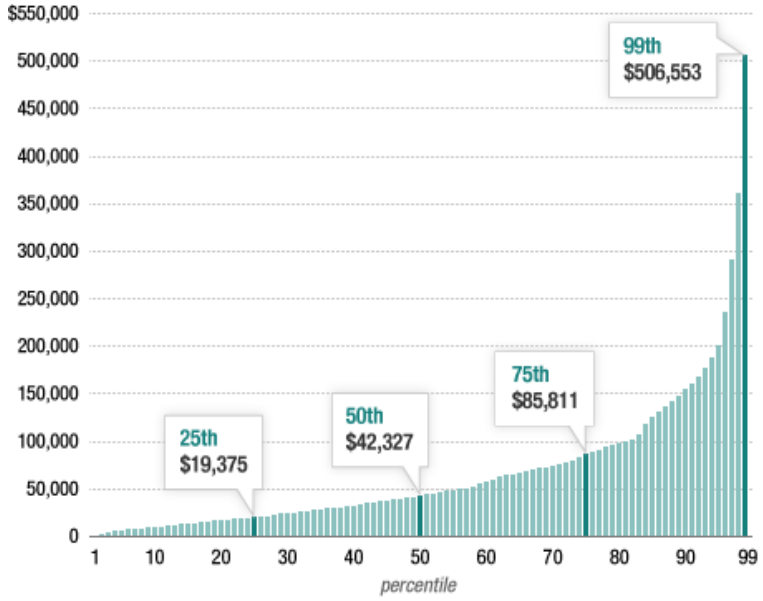


Vamos ver alguns exemplos

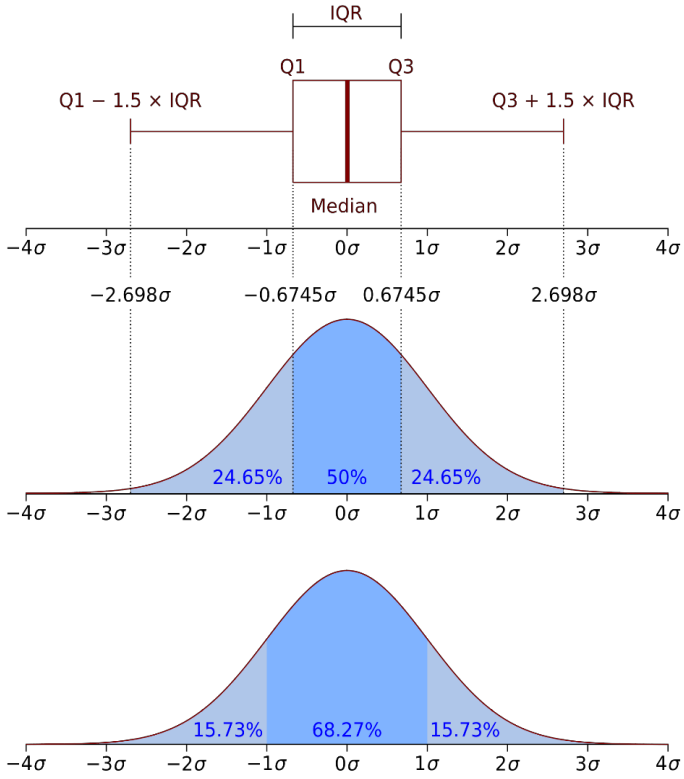
Percentis e Momentos

Percentis

- Em um conjunto de dados, qual é o ponto em que X% dos valores são menores que esse valor?
- Exemplo: distribuição de renda



Percentis em uma distribuição normal



Vamos ver alguns exemplos

Momentos

Medidas quantitativas da forma de uma função de densidade de probabilidade Matematicamente elas são um pouco difíceis de entender:

$$\mu_n = \int_{-\infty}^{\infty} (x - c)^n f(x) dx$$ (para um momento *n* em torno do valor *c*).

Mas intuitivamente, é muito mais simples em estatística.

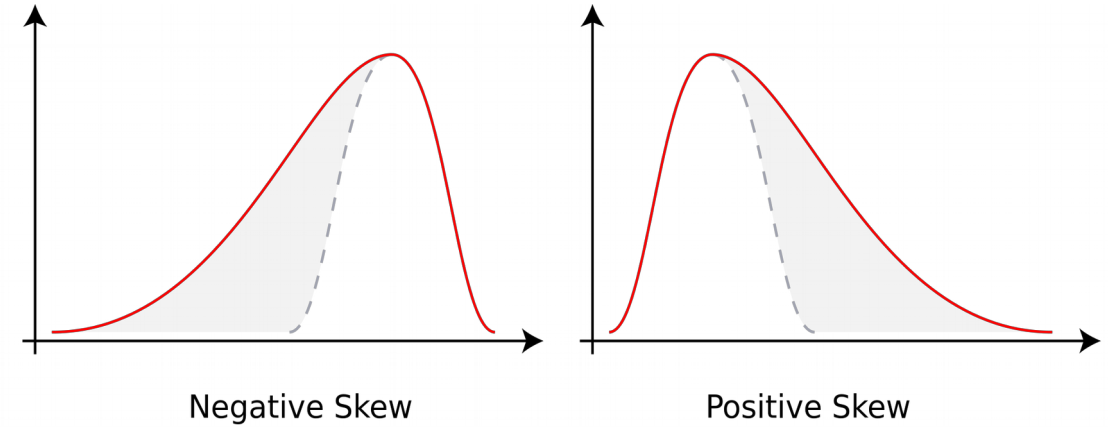
O primeiro momento é a média

O segundo momento é a variância

Simples assim...

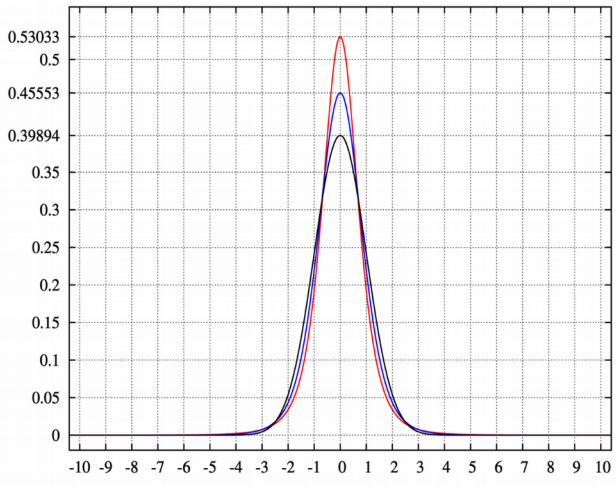
O terceiro momento “inclinação”

Quão “desequilibrada” é a distribuição? Uma distribuição com uma cauda mais longa à esquerda ficará inclinada para a esquerda e terá uma inclinação negativa.



O quarto momento é "curtose"

Quão espessa é a cauda e quão nítido é o pico, comparado a uma distribuição normal? Exemplo: picos mais altos têm maior curtose

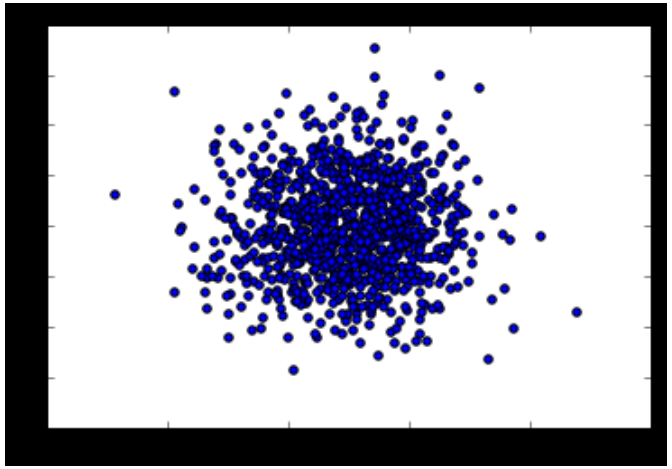


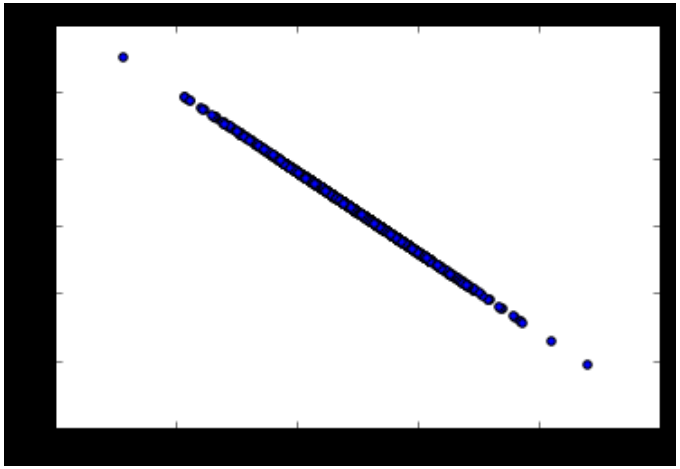
Vamos computar os 4 momentos com Python

Covariância e Correlação

Covariância

Mede como duas variáveis variam em conjunto a partir de suas médias.





Medindo a Covariância

- Pense nos conjuntos de dados para as duas variáveis como vetores de alta dimensionalidade
- Converta-os em vetores de variações a partir da média
- Pegue o produto escalar (cosseno do ângulo entre eles) dos dois vetores
- Divida pelo tamanho da amostra

Interpretar covariância é difícil

- Sabemos que uma pequena covariância, próxima de 0, significa que não há muito correlação entre as duas variáveis.
- E grandes covariâncias - ou seja, longe de 0 (pode ser negativo para inverso relacionamentos) significa que há uma correlação
- Mas quão grande é “grande”?

É aí que entra a correlação!

- Apenas divida a covariância pelos desvios padrão de ambas as variáveis, e isso normaliza as coisas.
- Portanto, uma correlação de -1 significa uma correlação inversa perfeita
- Correlação de 0: sem correlação
- Correlação 1: correlação perfeita

Lembre-se: a correlação não implica causalidade!

- Somente um experimento controlado e randomizado pode fornecer informações sobre causalidade.
- Use a correlação para decidir quais experimentos realizar!

Vamos ver alguns exemplos

Probabilidade Condicional

Probabilidade Condicional

- Se eu tiver dois eventos que dependem um do outro, qual é a probabilidade que ambos irão ocorrer?
- Notação: P (A, B) é a probabilidade de A e B ocorrerem ambos
- P (B | A): Probabilidade de B, dado que A ocorreu
- Nós sabemos:

$$P(B|A) = \frac{P(A,B)}{P(A)}$$

Por exemplo

- Eu passo aos meus alunos dois testes. 60% dos meus alunos passaram nos dois testes, mas o primeiro teste foi mais fácil - 80% foi aprovado. Qual porcentagem de os alunos que passaram no primeiro teste também passaram o segundo?
- A = passando no primeiro teste, B = passando no segundo teste
- Então, estamos pedindo P (B | A) - a probabilidade de B dado A

$$P(B|A) = \frac{P(A,B)}{P(A)} = \frac{0.6}{0.8} = 0.75$$

- 75% dos alunos que passaram no primeiro teste passaram no segundo.

Vamos ver um exemplo

Teorema de Bayes

Teorema de Bayes

• Agora que você entende a probabilidade condicional, você pode entender o Teorema de Bayes:

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

Descrição - a probabilidade de A dado B, é a probabilidade de A vezes o probabilidade de B dado A sobre a probabilidade de B.
O principal insight é que a probabilidade de algo que depende de B depende muito sobre a probabilidade básica de B e A. As pessoas ignoram isso o tempo todo.

Caso do Teorema de Bayes

• O teste de drogas é um exemplo comum. Mesmo um “altamente preciso” teste de drogas pode produzir mais falso-positivos do que verdadeiro-positivos.
• Digamos que tenhamos um teste de drogas que possa determinar com precisão identificar usuários de uma droga 99% do tempo, e com precisão tem um resultado negativo para 99% de não usuários. Mas apenas 0,3% da população total realmente usa essa droga.

Teorema de Bayes

• Evento A = É um usuário do medicamento, Evento B = testado positivamente para o medicamento.
• Podemos calcular a partir dessa informação que P (B) é de 1,3% (0,99 * 0,003 + 0,01 * 0,997 - a probabilidade de teste positivo se você usar, mais o probabilidade de teste positivo se você não fizer isso.

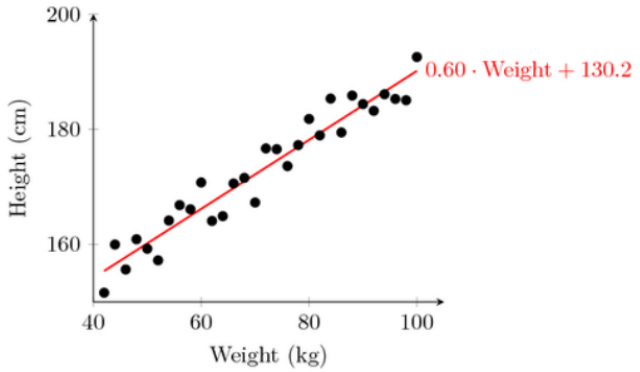
$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} = \frac{0.003 * 0.99}{0.013} = 22.8\%$$

• Então, as chances de alguém ser um usuário real da droga, dado que eles testado positivo é de apenas 22,8%!
• Embora P (B | A) seja alto (99%), não significa que P (A | B) esteja alto.

Regressão Linear

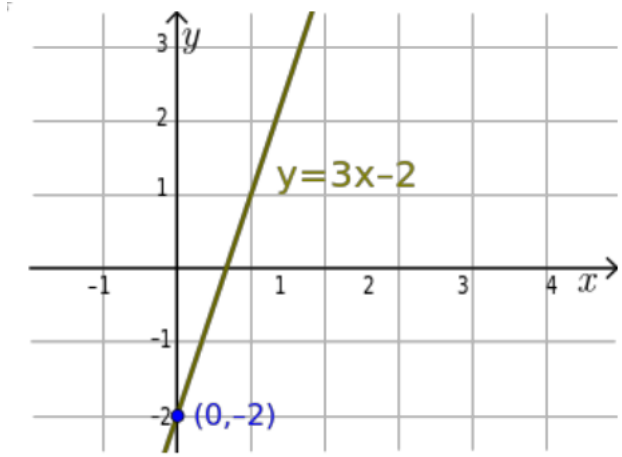
Regressão Linear

• Ajustar uma linha a um conjunto de dados de observações
• Use esta linha para prever valores não observados
• Eu não sei por que eles chamam de "regressão".É realmente enganador. Você pode usá-lo para prever pontos no futuro, o passado, tanto faz. Na verdade, o tempo geralmente não tem nada a ver com isso.



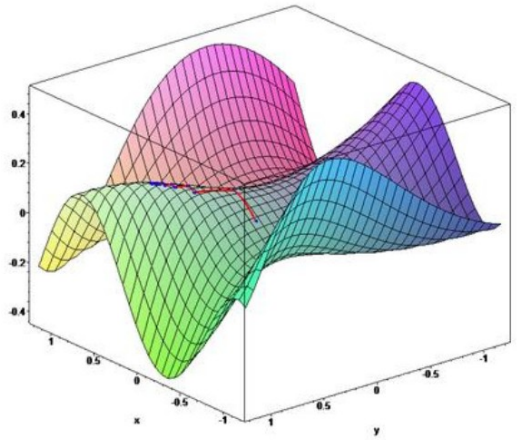
Regressão Linear: como funciona?

• “Mínimos quadrados” minimiza a soma dos erros quadrados.
• Isto é o mesmo que maximizar a probabilidade dos dados observados se você começar a pensar no problema em termos de probabilidades e probabilidades funções de distribuição
• Isso às vezes é chamado de “estimativa de máxima verossimilhança”



Mais de uma maneira de fazer isso

- Gradiente descendente é um método alternativo aos mínimos quadrados.
- Basicamente itera para encontrar a linha que melhor segue os contornos definidos pelos dados.
- Pode fazer sentido quando se lida com dados 3D
- Fácil de experimentar em Python e apenas comparar resultados para mínimos quadrados
  - Mas geralmente os mínimos quadrados são perfeitamente boas escolha.



Medição de Erro com R-Quadrado

- Como medimos quão bem nossa linha se ajusta aos nossos dados?
- Medidas de R-Quadrado (coeficiente de determinação):

A fração da variação total em Y que é capturado pelo modelo

Computação R-Quadrado

$$1{,}0 - \frac{\text{soma de erros quadrados}}{\text{soma da variação quadrática da média}}$$

Interpretando o R-Quadrado

- Varia de 0 a 1
- 0 é ruim (nenhuma das variações é capturada), 1 é bom (todas as variações são capturadas).

Vamos ver um exemplo

Obrigado!!!

Nos vemos amanhã!!!

Bom descanso!