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

Configuração

Checklist

Instalar o Enthought Canopy (versão >=1.6.2!)

Abra uma janela de edição e vá para o prompt de comando interativo:

```
!pip install pydot2
```

Abra o package manager, e instale:
scikit_learn, numpy, pandas, stastmodels, xlrd, pydotplus

Python Basics

Vamos ver algum código.

Tipos de Dados

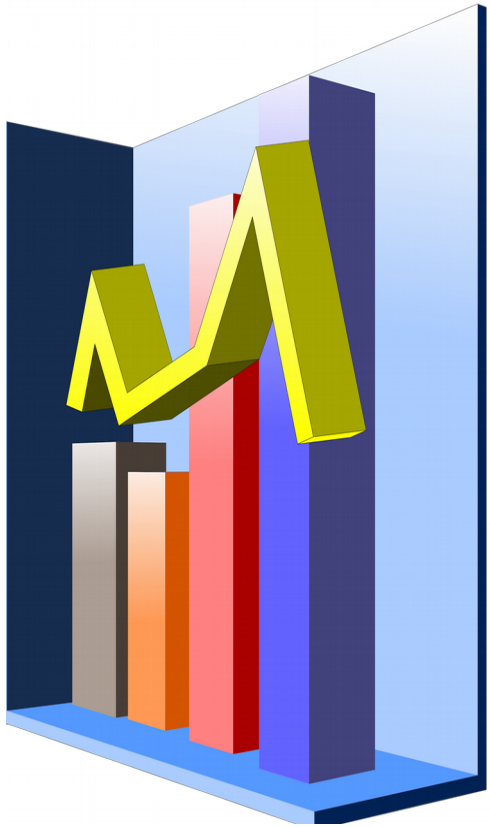
Muitos sabores de dados

Principais Tipos de Dados

- Numéricos
- Nominais (ou Categóricos)
- Ordinais

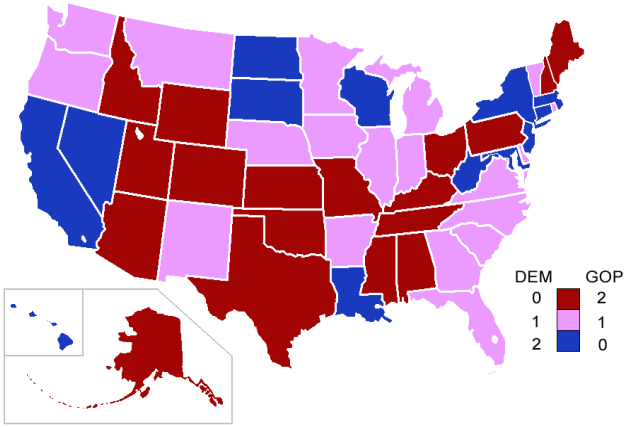
Numéricos

- Representam algum tipo de medida quantitativa
  - Altura da população, tempo de carga de páginas
    Preço de ações, etc...
- Dados Discretos
  - Baseado em inteiros; usualmente contam algo.
    - Quantas compras um cliente faz ao ano?
    - Quantas vezes eu virei a cabeça?
- Dados Contínuos
  - Tem um número infinito de valores possíveis
    - Quanto tempo um usuário gasta no check out?
    - Quanta chuva cai em um determinado dia?



Nominais

- Dados qualitativos que não têm significado matemático inerente
  - Sexo, Sim/Não (dados binários), Raça, Estado De Residência, Categoria de Produto, Partido, etc.
- Você pode assinalar números para as categoria Para representá-las mais compactadamente, mas os números não tem sigificado matemático.

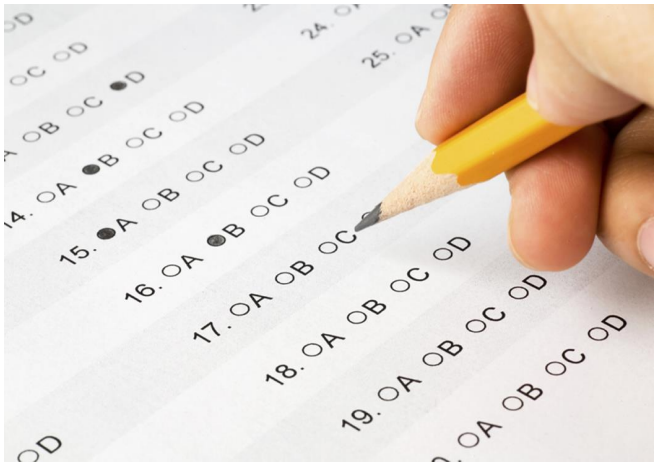


Ordinais

- Mistura de numéricos e nominais
- Dados Nominais não tem significado matemático
- Exemplo: escala de 1-5 para ratings.
  - Ratings devem ser 1, 2, 3, 4, ou 5
  - Mas este valores tem significado matemático; 1 significa um filme pior do que um 2.

Quiz time!

- Estes tipos de dados são numéricos, ordinais ou nominais?
  - Quanta gasolina tem no tanque do seu carro?
  - Um rating de sua saúde geral, onde as opções são 1, 2, 3, ou 4, correspodendo a “ruim”, “moderada”, “boa”, e “excelente”
  - As raças de seus colegas de classe
  - Idade em anos
  - Dinheiro gasto em uma loja



Média, mediana e moda

Média

- Soma / número de amostras
- Exemplo:
  - Numero de crianças em cada casa da minha rua:

0, 2, 3, 2, 1, 0, 0, 2, 0

A **MÉDIA** é (0+2+3+2+1+0+0+2+0) / 9 = **1.11**

Mediana

- Ordene os valores e pegue o do meio da lista.
- Exemplo:

0, 2, 3, 2, 1, 0, 0, 2, 0

Ordene:

0, 0, 0, 0, 1, 2, 2, 2, 3

Mediana

- Se você tiver um número par de amostras, Tire a média dos dois do meio.
- Mediana é menos suscetível outliers do que the mean
  - Exemplo: renda familiar média nos EUA é $72,641, mas a mediana é apenas $51,939 – Porque a média é distorcida por um punhado De bilionários.
  - Mediana representa melhor o Americano “típico” Nesse exemplo.

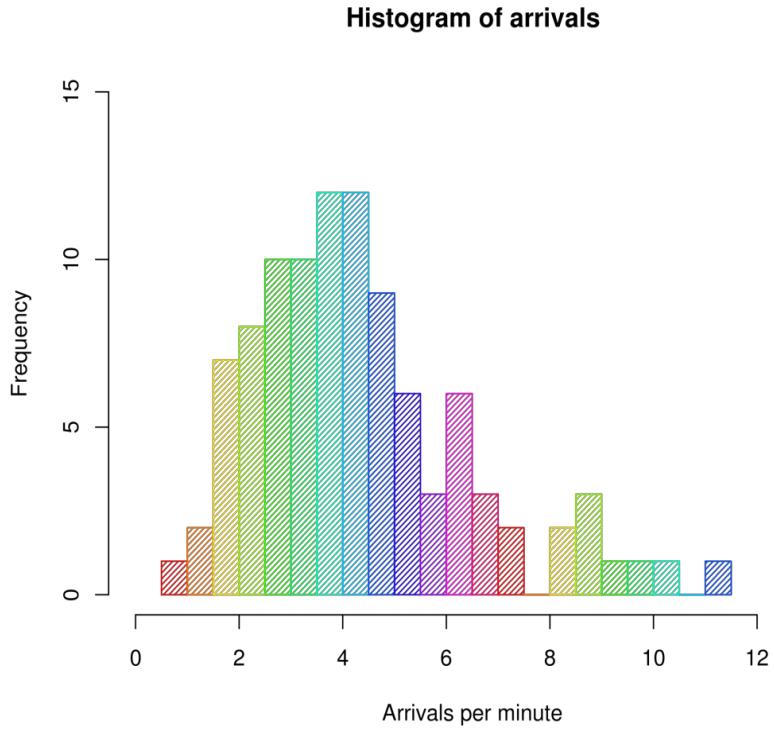
Moda

- O valor mais comum em um dataset
  - Não relevante para dados numéricos contínuos
- De volta ao exemplo do número de crianças:

0, 2, 3, 2, 1, 0, 0, 2, 0
Quantos de cada valor temos?
0: 4, 1: 1, 2: 3, 3: 1
A MODA é 0

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:

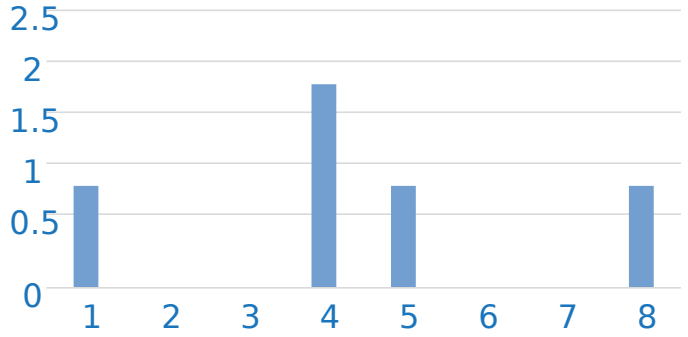
$$\sigma^2 = (11.56 + 0.16 + 0.36 + 0.16 + 12.96) / 5 = 5.04$$

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:
    - $\sigma^2 = (11.56 + 0.16 + 0.36 + 0.16 + 12.96) / 5 = 5.04$
  - But the sample variance would be:
    - $S^2 = (11.56 + 0.16 + 0.36 + 0.16 + 12.96) / 4 = 6.3$

Fórmulas

- Variância da População:

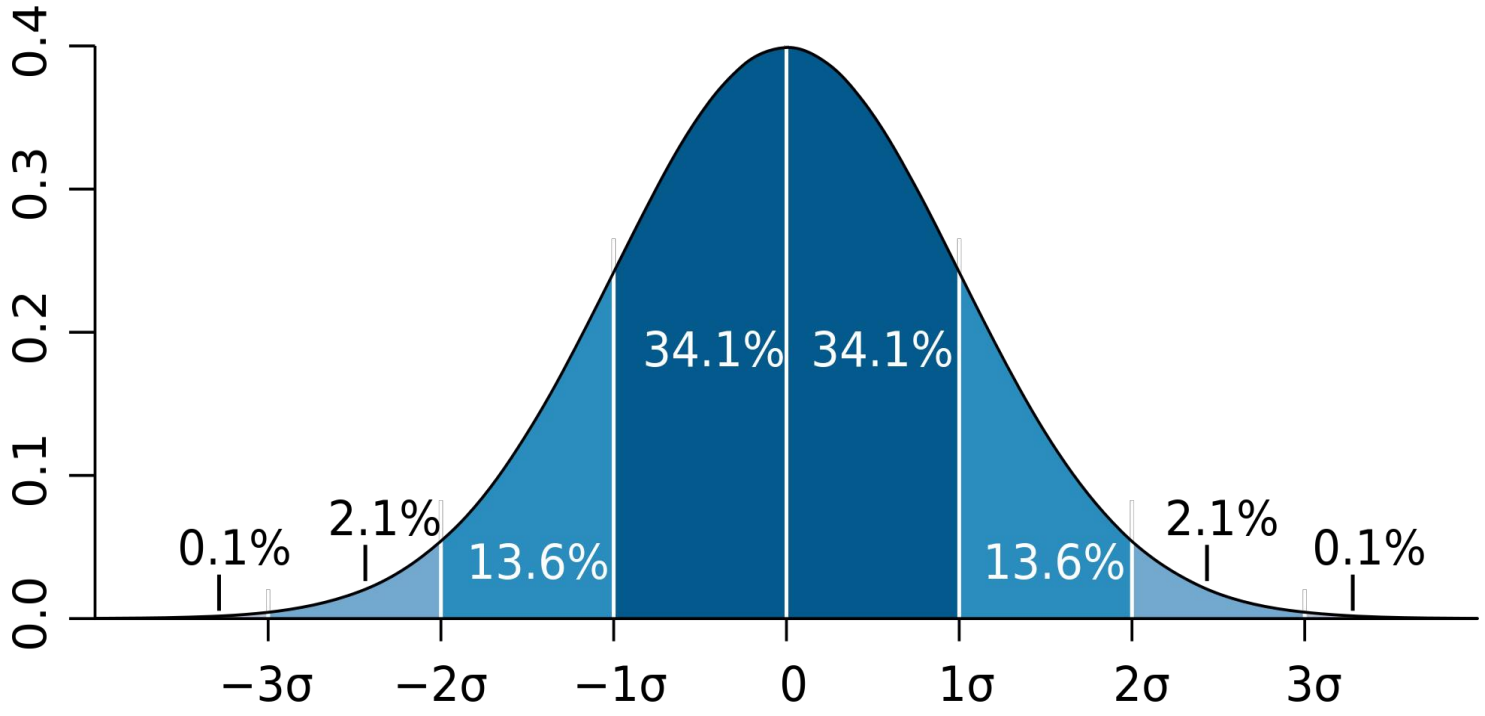
$$\sigma^2 = \frac{\sum(X-\mu)^2}{N}$$

- Variância da Amostra:

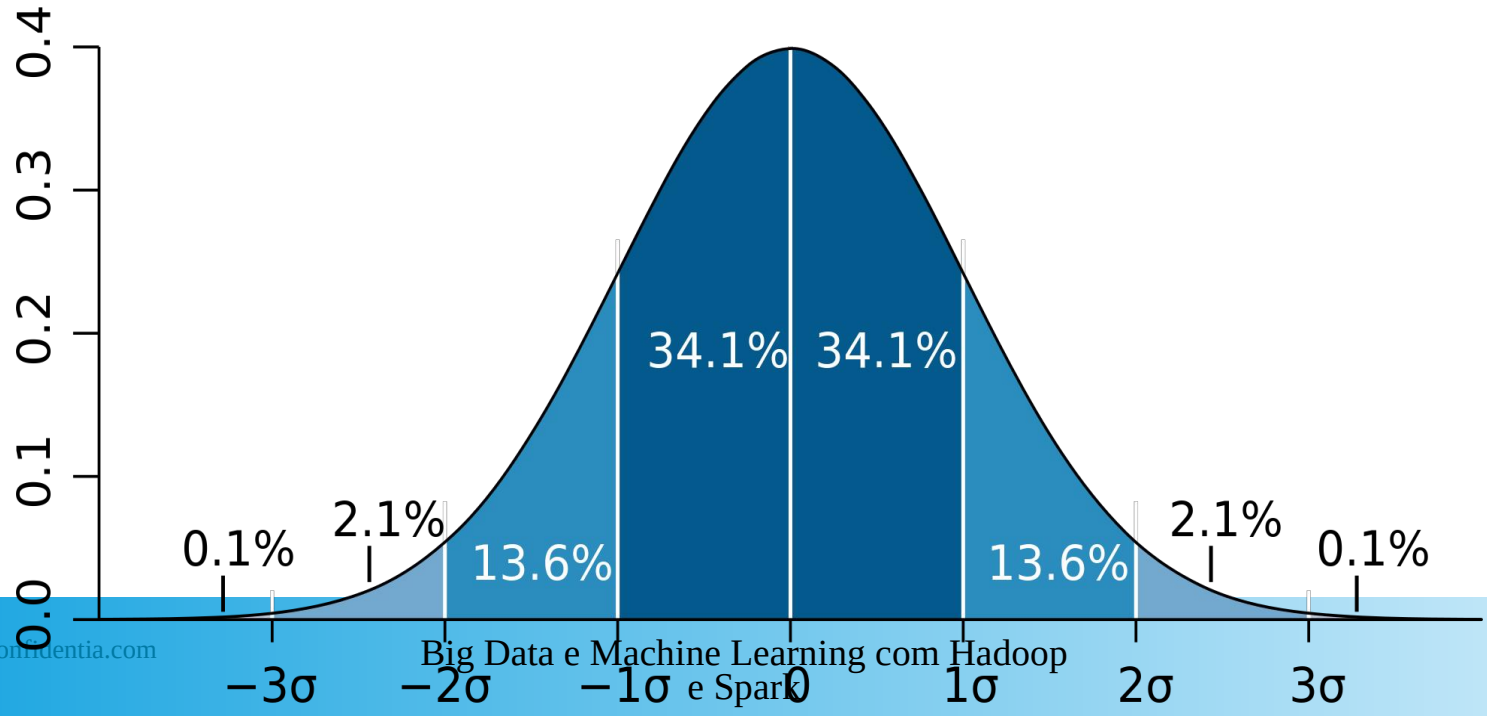
$$s^2 = \frac{\sum(X-M)^2}{N-1}$$

Funções de densidade de probabilidade

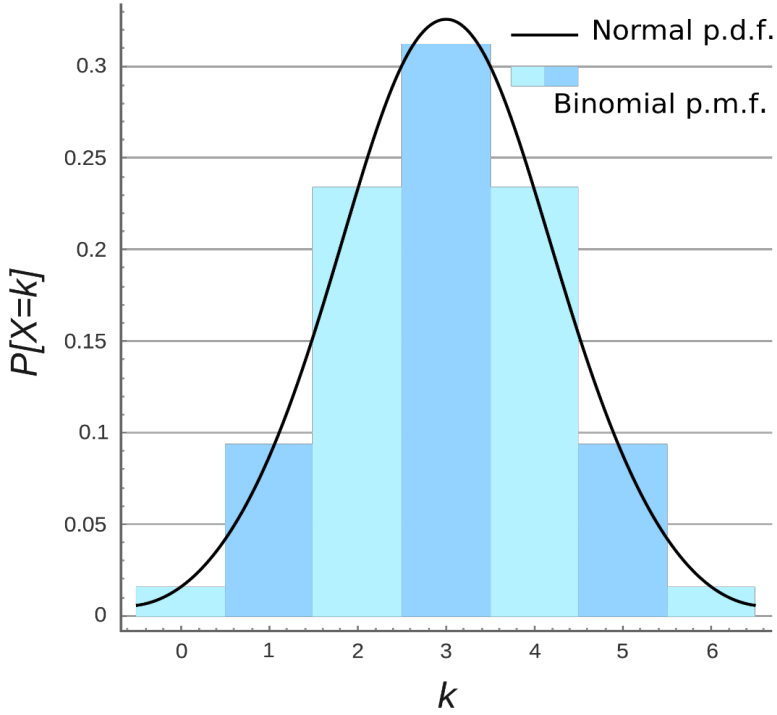
Exemplo: uma “distribuição normal



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

Obrigado!!!

Nos vemos amanhã!!!

Bom descanso!