PUC Minas

DIRETORIA DE
EDUCAÇÃO CONTINUADA

Pós Graduação *Lato Sensu*

Inteligência Artificial e Aprendizado de Máquina

Técnicas Estatísticas de Predição

Programa

Calendário	Conteúdo	Referencial Teórico
07/02/2019	Regressão Linear (Simples e Multipla)	An Introduction to Statistical Learning with Applications in R, 2013
14/02/2019	Regressão Logística (Simples e Multipla)	
21/02/2019	Modelo Multinomial	Categorical Data Analysis, 2013
28/02/2019	**Aula cancelada**	
07/03/2019	Árvores de decisão e Florestas Aleatórias	An Introduction to Statistical Learning with Applications in R, 2013
14/03/2019	KNN, PCA e K Means Clustering	
21/03/2019	**Avaliação**	-

KNN, PCA e K Means Clustering

Livro texto: *An Introduction to Statistical Learning*

Cap. 4 - Classification

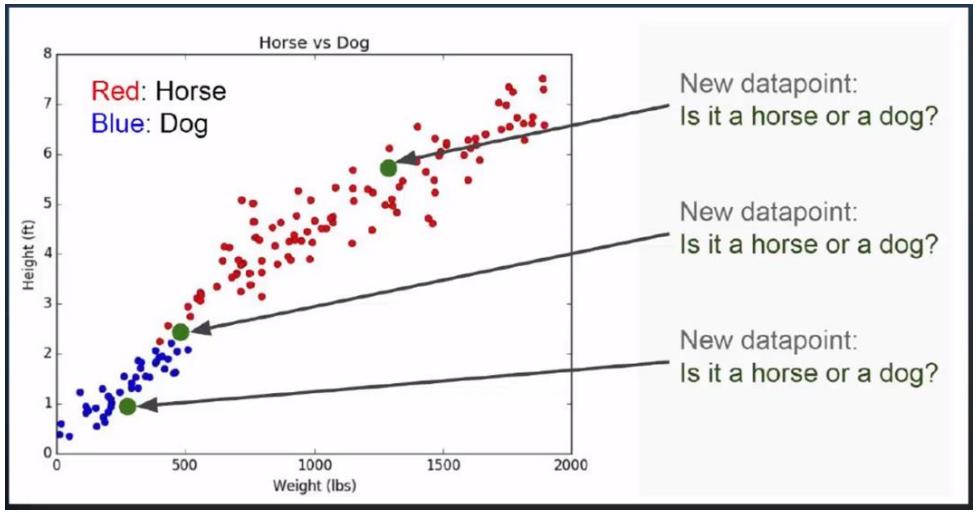
4.6.5 K-Nearest Neighbors

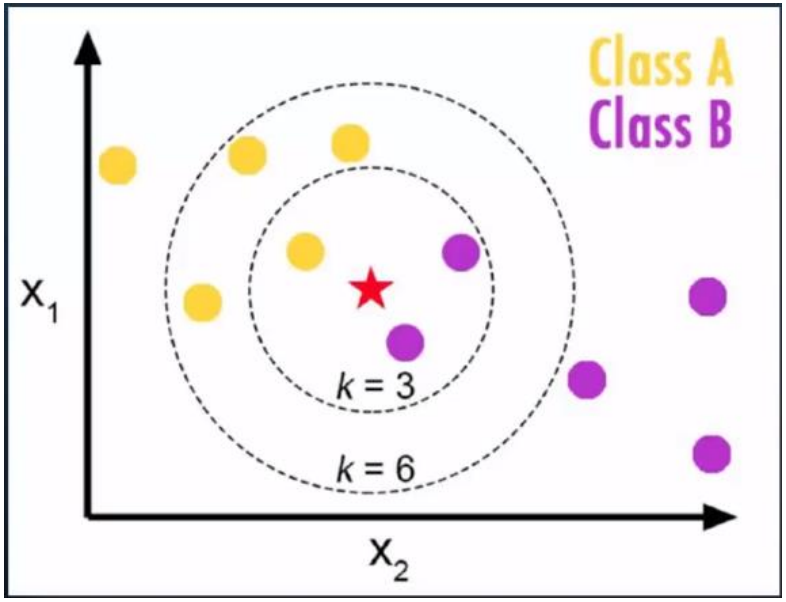
Cap. 10 - Unsupervised Learning

10.2 Principal Components Analysis

10.3 Clustering Methods

K- Nearest Neighbors (KNN)

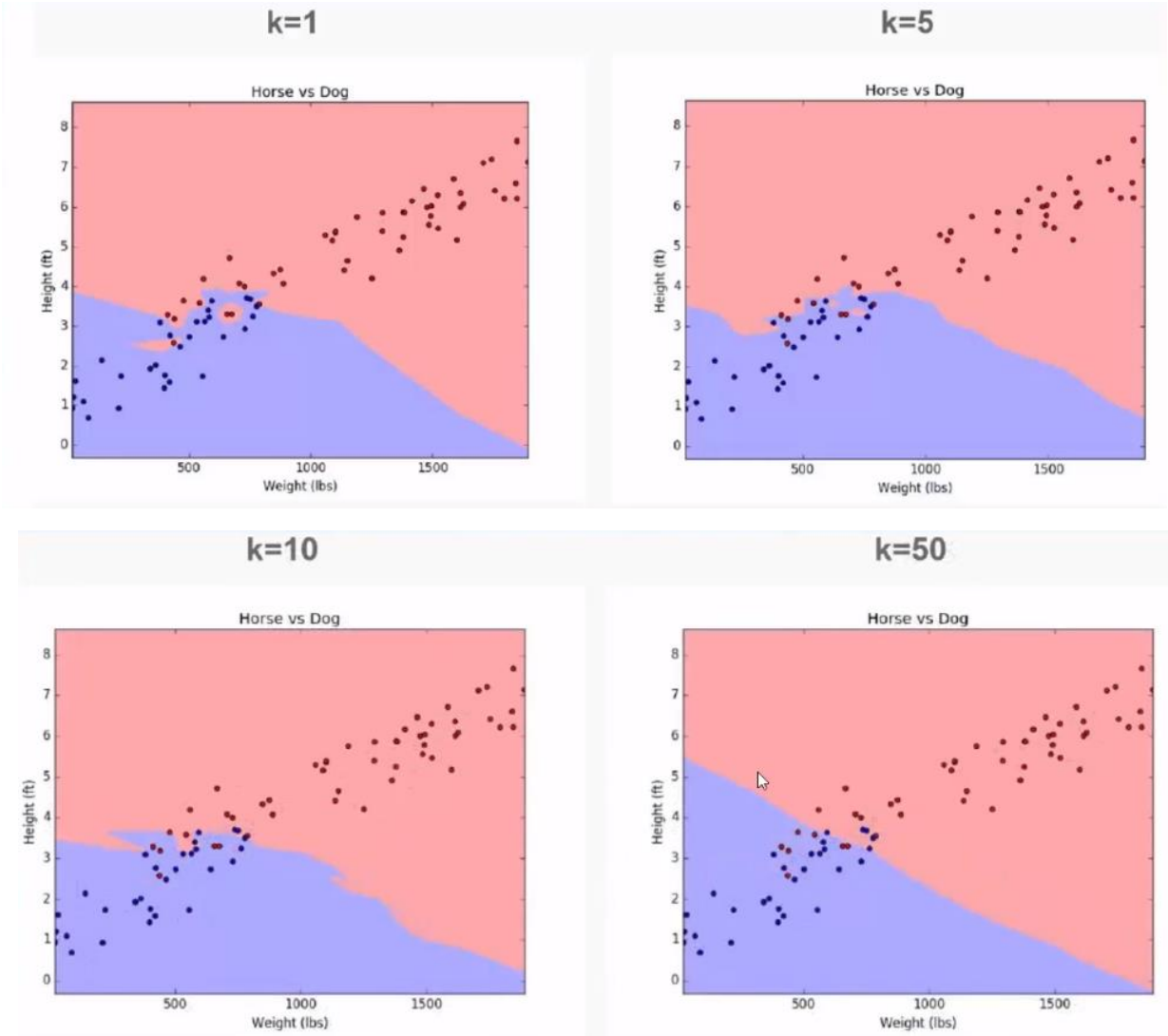




- Algoritmo de treino
  1. Guarda os dados

- Algoritmo de teste/preditor
  1. Calcula as distâncias de X até os demais pontos
  2. Organiza os dados em ordem crescente de distância
  3. Classifica a classe de acordo com a maioria dos "*k*" primeiros valores

O parâmetro “k” do modelo pode afetar as classificações do mesmo



Vantagens

- Simplicidade de aplicação;
- Processo de treino é trivial;
- Funciona muito bem com grande número de classes;
- Fácil de se adicionar mais dados;
- Poucos parâmetros (K e Métrica de distância)

Desvantagens

- Elevado custo computacional para predição (piora para grandes conjuntos de dados);
- Não performa bem em dados com múltiplos níveis;
- Não performa bem para *features* categóricas.

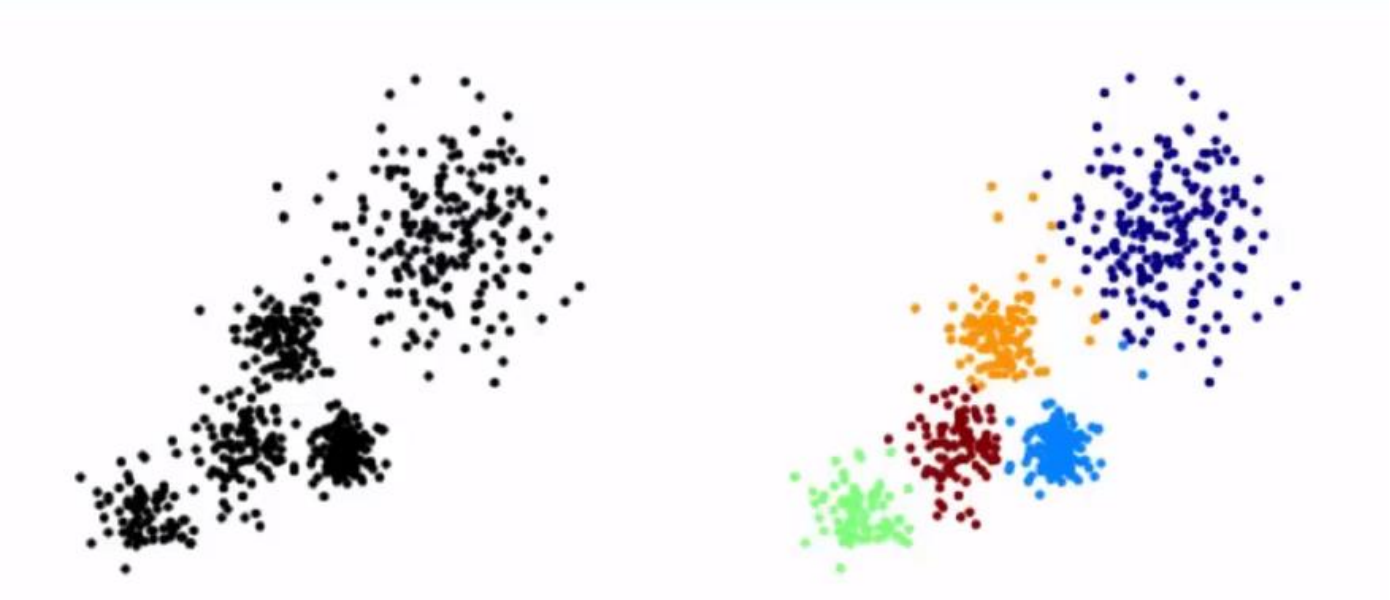
K- Means Clustering

K- Means Clustering é um método de ML baseado em aprendizado não supervisionado que tentará agrupar seus dados em grupos de características similares.

Exemplo de utilização:

- Agrupamento automático de documentos;
- Agrupamento de clientes;
- Segmentação de mercado;
- Geoestatística

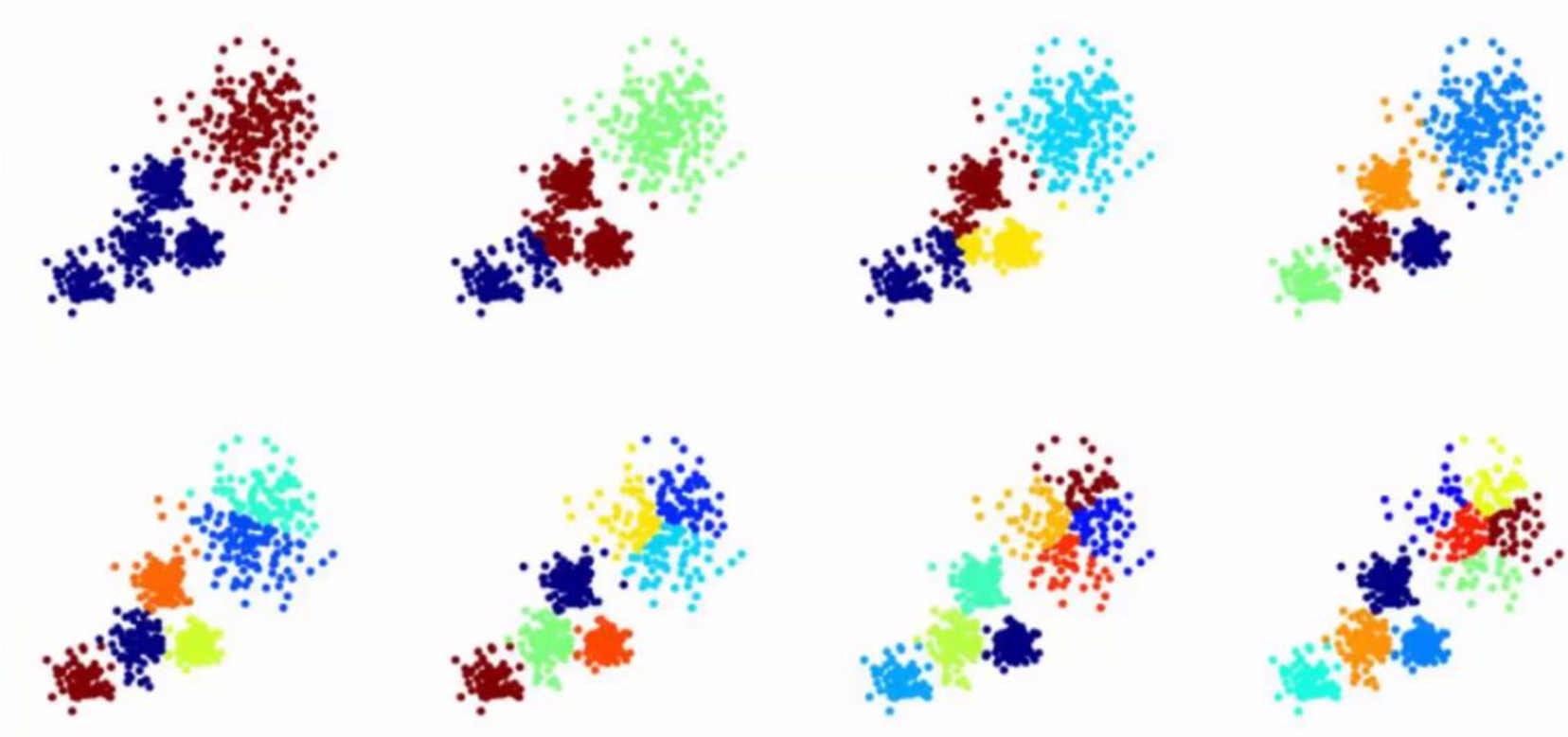
O objetivo do k-Means Clustering é dividir os dados em "k" grupos distintos fora do grupo e similar dentro do grupo.



O algoritmo

- Escolher um número K de grupos (Cluster)
    a. Aleatoriamente definir um centroide para cada grupo
    b. Até os clusters pararem de mudar, faça:
        i. Para cada cluster obter centroide de acordo com a média dos vetores de pontos dos clusters
        ii. Defina cada ponto ao cluster no qual o centroide é o mais próximo.

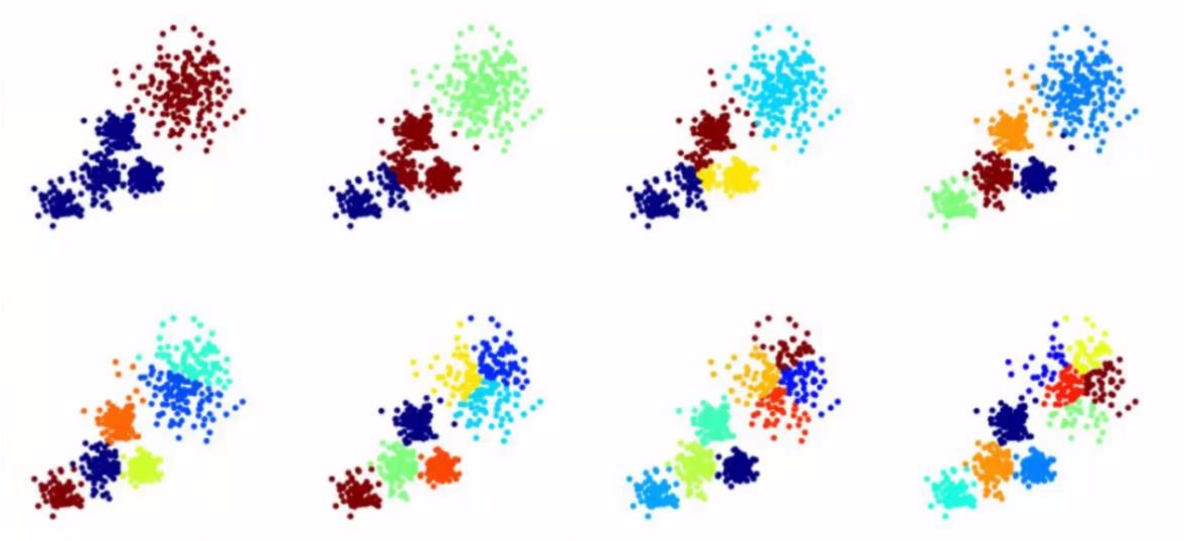
Como escolher o “k” mais adequado ou o “k”correto?



Não existe uma resposta fácil para a escolha do "k" mais adequado.

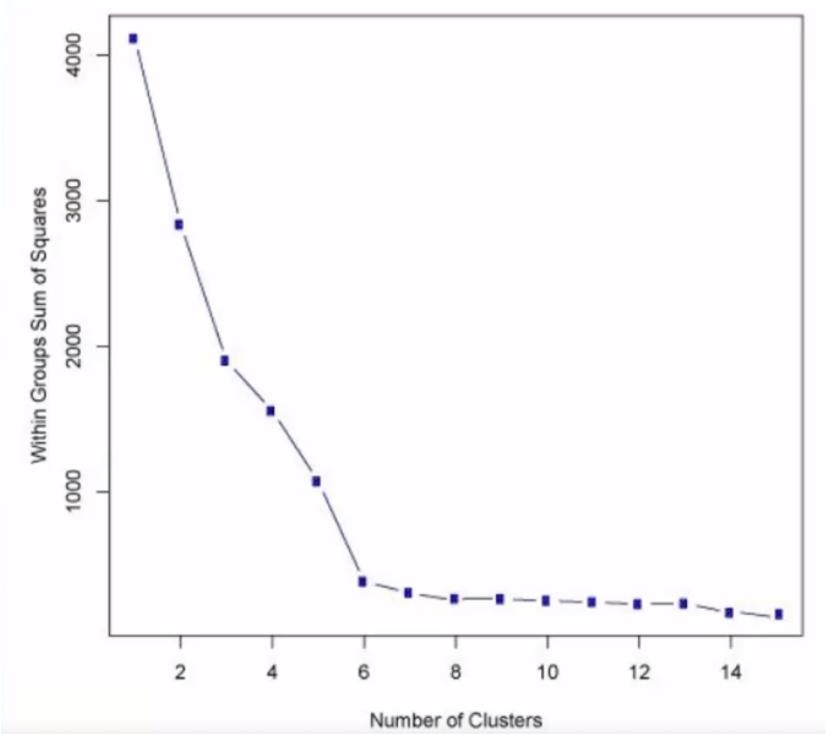
Uma forma sugerida é o método do cotovelo.

- Primeiro, calcula-se a soma dos erros quadráticos (SEQ) para alguns valores de K (Ex.: 2,4,6,...)
- A soma dos quadrados dos erros é definida como o quadrado das distâncias entre cada membro e o seu centroide.



Se plotado K versus SEQ, você verá que o erro diminui à medida em que K aumenta.

A ideia do método é escolher um valor de K no qual a SEQ caia bruscamente. Isso produz um efeito cotovelo no gráfico.



META

- Participação em uma competição de ***Machine Learning*** no *Kaggle*
- Aplicar uma das técnicas de predição utilizadas no curso
- Escrever um artigo sobre modelagem preditiva no *Linkedin*
  - *Causalidade*
  - *Risco*
  - *Identificação*
  - *etc*

Exercício

Jupyter Notebook