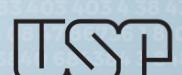


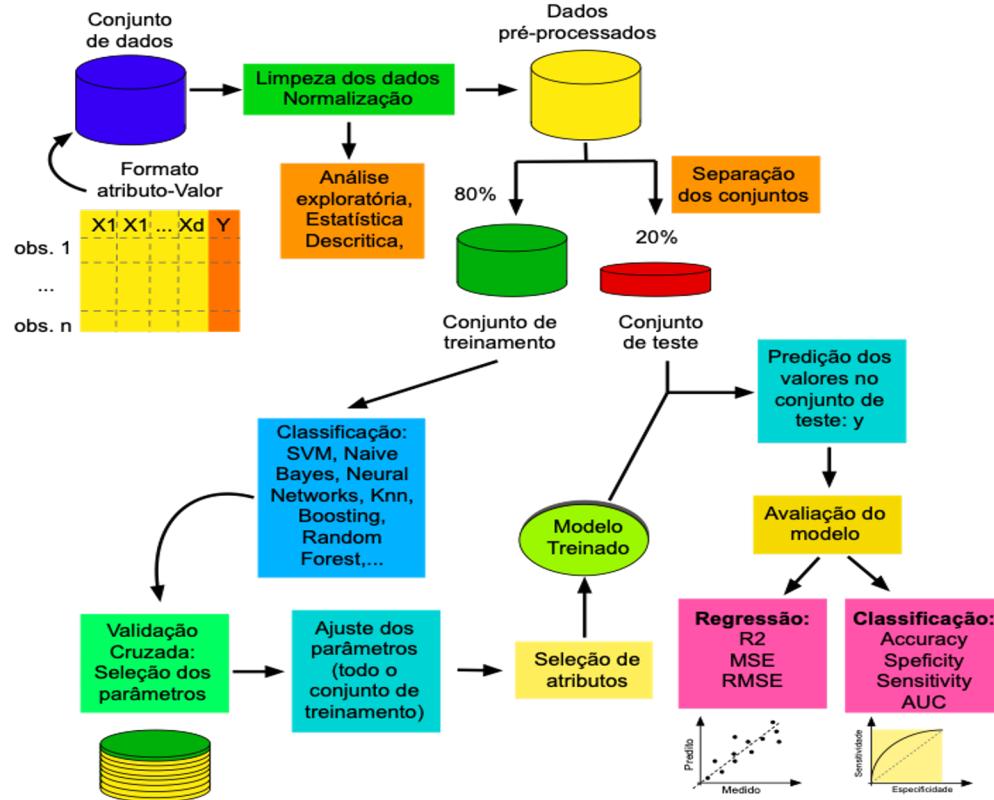
Introdução a Ciências de Dados

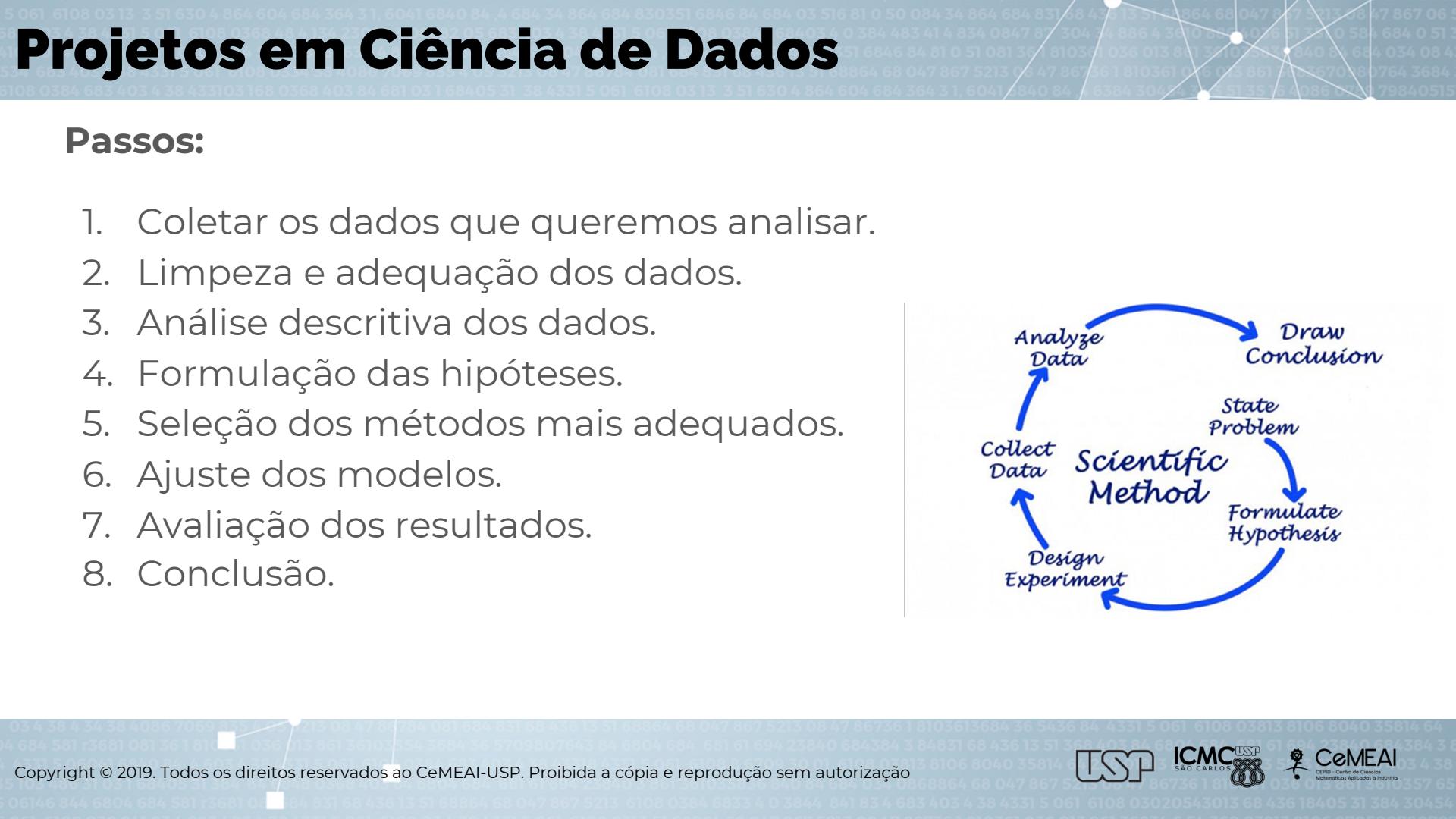
Aula de revisão

Francisco A. Rodrigues
ICMC/USP
francisco@icmc.usp.br



Projetos em Ciência de Dados

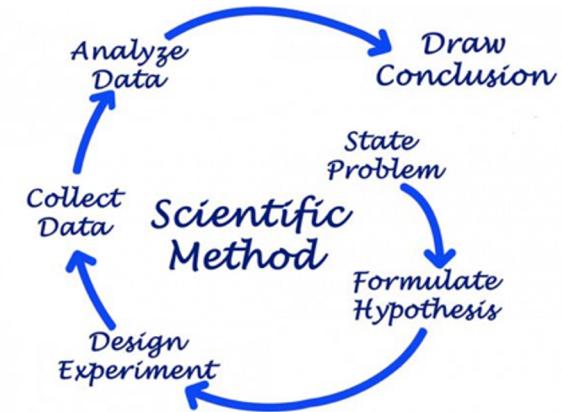




Projetos em Ciência de Dados

Passos:

1. Coletar os dados que queremos analisar.
2. Limpeza e adequação dos dados.
3. Análise descritiva dos dados.
4. Formulação das hipóteses.
5. Seleção dos métodos mais adequados.
6. Ajuste dos modelos.
7. Avaliação dos resultados.
8. Conclusão.



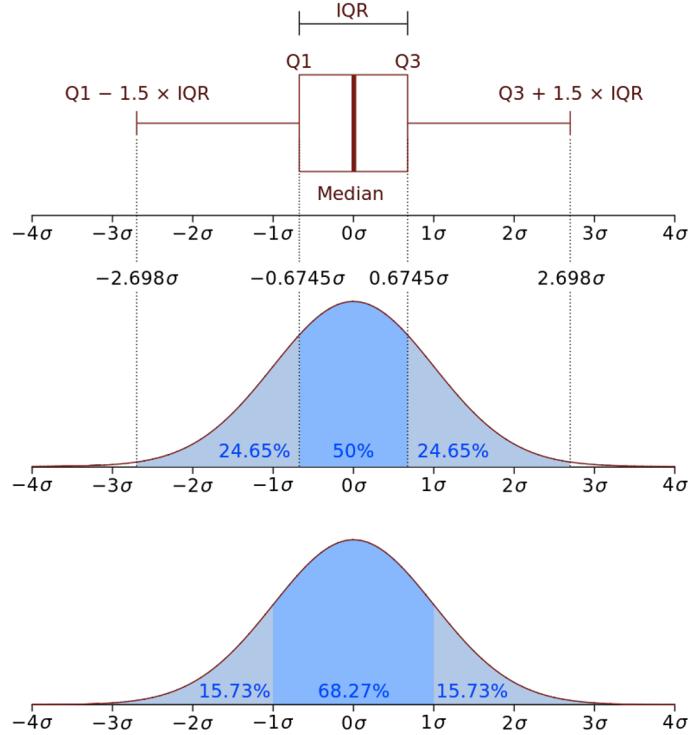
Tratamento e Transformação de Dados

- Tratamento e Transformação de Dados

- Eliminação manual de atributos,
- Integração de dados,
- Amostragem de dados,
- Redução de dimensionalidade,
- Balanceamento de dados,
- Limpeza de dados,
- Transformação de dados.

Estatística descritiva

- Medidas centrais: moda, mediana, média, quantil, percentil
- Medidas de dispersão: variância, desvio padrão
- Correlação: Person, Spearman
- Box-plots
- Histogramas
- Análise dos componentes principais



Classificação

Definição formal: Dado um conjunto de observações:

$$D = \{\mathbf{X}, \mathbf{y}, i = 1, \dots, N\}$$

- **f** representa uma função desconhecida (função objetivo).

$$y_i = f(\mathbf{X}_i, \theta) + \epsilon_i$$

- Essa função mapeia as entradas nas saídas correspondentes.
- O algoritmo preditivo aprende a aproximação, que permite estimar valores de **f** para novos valores de **X**.

Classificação

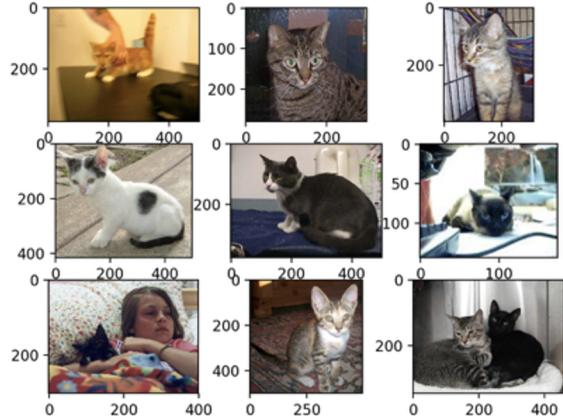
$$y_i \in \{C_1, C_2, \dots, C_n\}$$



K-vizinhos

- Uma maneira simples de definirmos a classificação de objetos é através de distância entre eles:

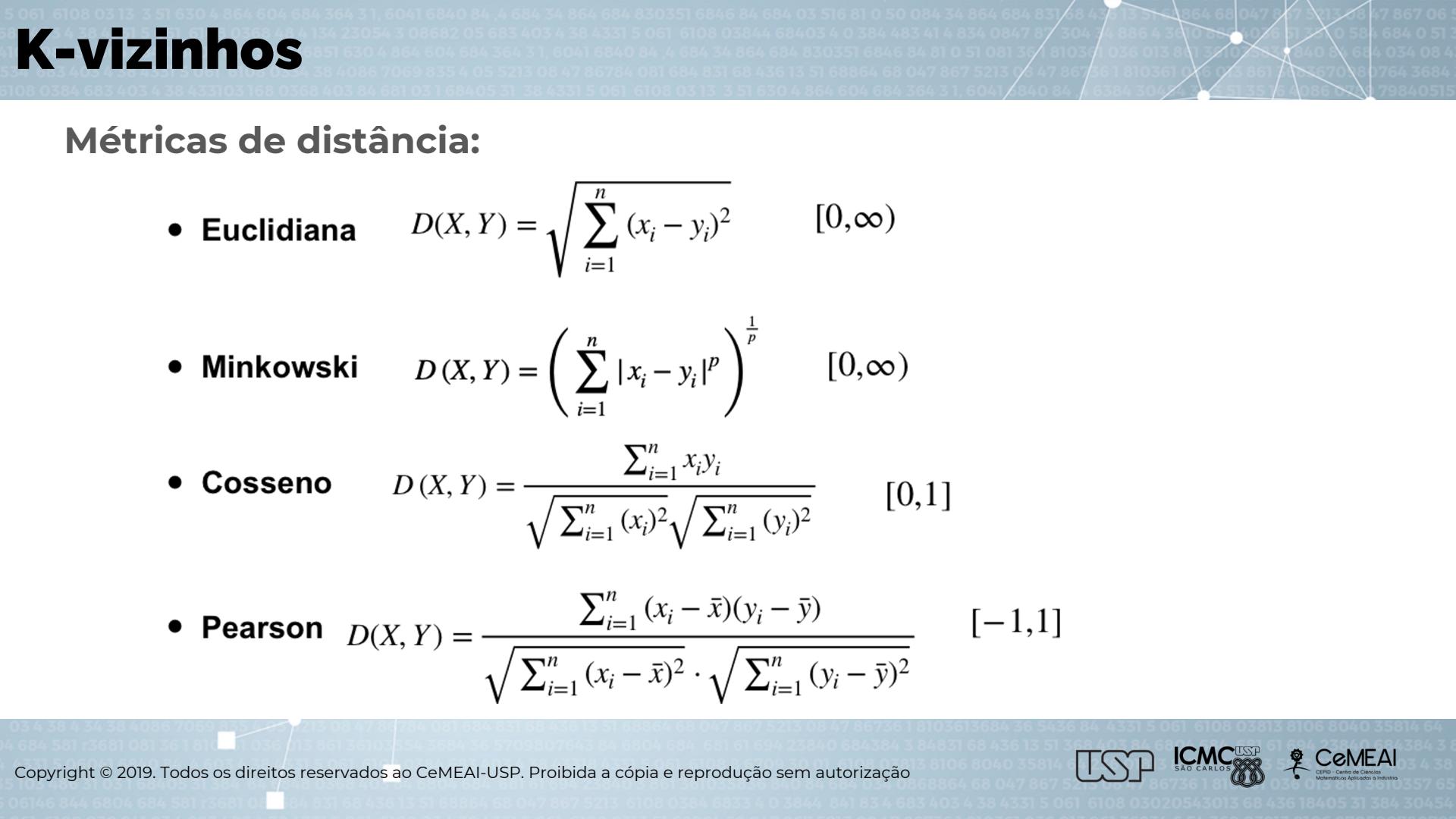
Classe: gatos

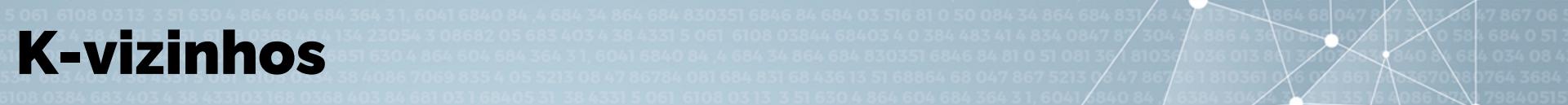


?

Classe: Cachorros





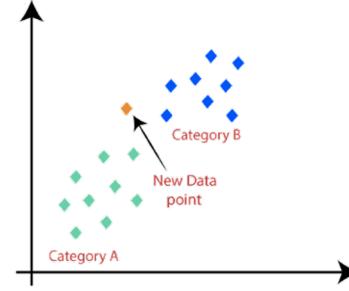


K-vizinhos

Algoritmo:

1. Identifique os k-vizinhos mais próximos do vetor de atributos \mathbf{X} que se quer classificar.
2. Determine o número de vizinhos em cada classe.
3. Classifique \mathbf{X} com pertencente à classe que resultou em um maior número de vizinhos (a moda entre o número de classes).

$$p(y = j | \mathbf{x}_*) = \frac{1}{k} \sum_{i \in R_*} \mathbb{I}\{y_i = j\}$$





K-vizinhos

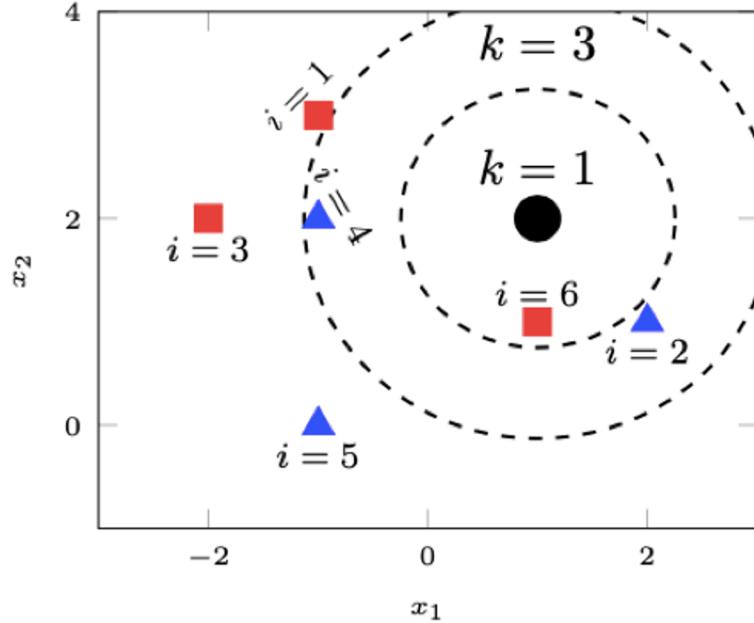
Exemplo:

Observação: $\mathbf{x}_* = [1 \ 2]^\top$

| i | x_1 | x_2 | y |
|-----|-------|-------|------|
| 1 | -1 | 3 | Red |
| 2 | 2 | 1 | Blue |
| 3 | -2 | 2 | Red |
| 4 | -1 | 2 | Blue |
| 5 | -1 | 0 | Blue |
| 6 | 1 | 1 | Red |

Distâncias

| i | $\ \mathbf{x}_i - \mathbf{x}_*\ $ | y_i |
|-----|-----------------------------------|-------|
| 6 | $\sqrt{1}$ | Red |
| 2 | $\sqrt{2}$ | Blue |
| 4 | $\sqrt{4}$ | Blue |
| 1 | $\sqrt{5}$ | Red |
| 5 | $\sqrt{8}$ | Blue |
| 3 | $\sqrt{9}$ | Red |



Regressão Logística

- Vimos que modelos de regressão linear são dados por:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_d X_d = \beta^T \mathbf{x}$$

onde $\mathbf{x}^T = [1, x_1, x_2, \dots, x_d]$

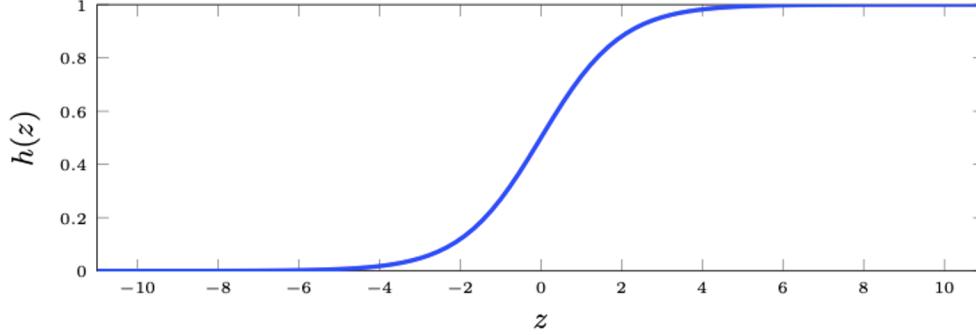
- Se considerarmos a saída \mathbf{Y} como um valor inteiro, podemos usar o modelo de regressão para realizarmos a classificação de dados.
- Vamos definir as probabilidades para o caso de duas classes:

$$p(y=1|\mathbf{x}) \text{ e } p(y=0|\mathbf{x})$$

Regressão Logística

- Vamos considerar a função logística:

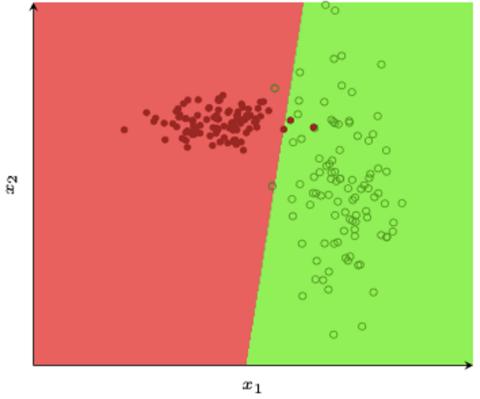
$$h(z) = \frac{e^z}{1 + e^z}$$



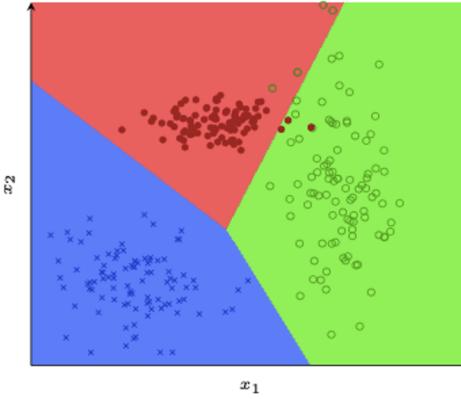
- Essa função retorna valores no intervalo [0,1].

Regressão Logística

Duas classes



Três classes



- Para mais de duas classes, usamos one-hot-encoding e repetimos o mesmo procedimento.
http://www.it.uu.se/edu/course/homepage/sml/literature/lecture_notes.pdf

Naive Bayes

Teoria da decisão Bayesiana:

- Dada M classes $\omega_1, \omega_2, \dots, \omega_M$ e um padrão desconhecido x , determinar a probabilidade condicional $p(\omega_i|x)$ do padrão pertencer a cada classe i .

$$P(\omega_i|x) = \frac{p(x|\omega_i)P(\omega_i)}{p(x)}$$

- Classificar de acordo com a classe mais provável.

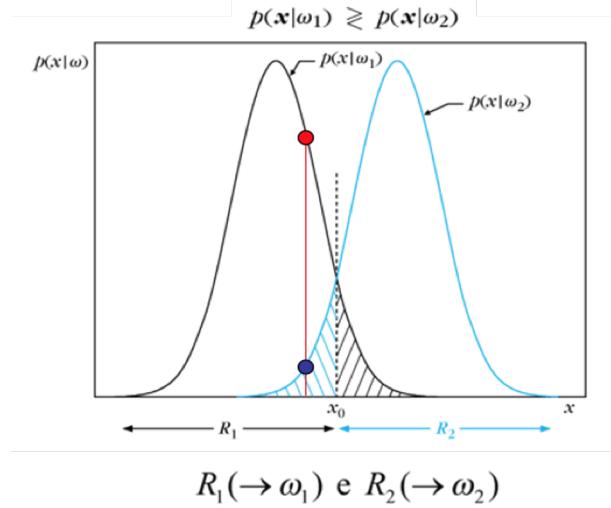
$$\hat{y} = \operatorname{argmax}_{i \in \{1, \dots, M\}} p(\omega_i | x)$$



Naive Bayes

Teoria da decisão Bayesiana:

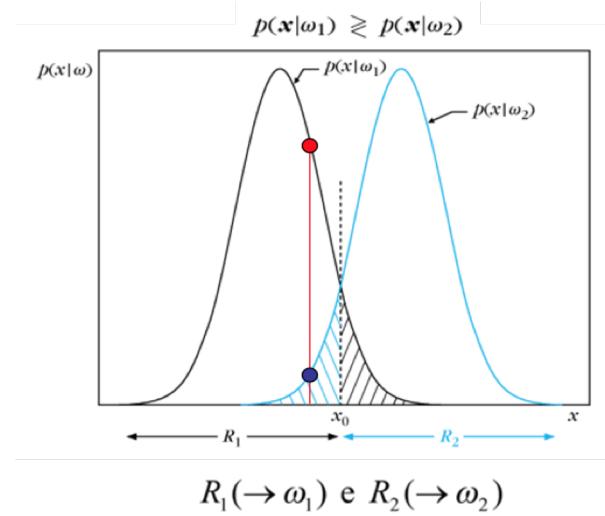
- Se a probabilidade condicional for conhecida para cada classe, o erro obtido é o menor possível (classificador ótimo).



Naive Bayes

Teoria da decisão Bayesiana:

- **Problema:** na maioria das vezes, não sabemos a distribuição de probabilidade conjunta e sua estimativa é bastante complicada.



$$R_1(\rightarrow \omega_1) \text{ e } R_2(\rightarrow \omega_2)$$

Naive Bayes

Algoritmo (para atributos continuos):

- Calcule a média e variância de cada atributo para cada classe.
- Calcule a verossimilhança para cada atributo dentro de cada classe.

$$p(x_j | \omega_i) = \frac{1}{\sqrt{2\pi\sigma_{\omega_i}}} \exp\left(-\frac{(x_j - \mu_{\omega_i})^2}{2\sigma_{\omega_i}^2}\right)$$

- Assuma independência e calcule a distribuição conjunta.

$$p(\mathbf{x} | \omega_i) = \prod_{j=1}^d p(x_j | \omega_i) \quad i = 1, 2, \dots, M$$

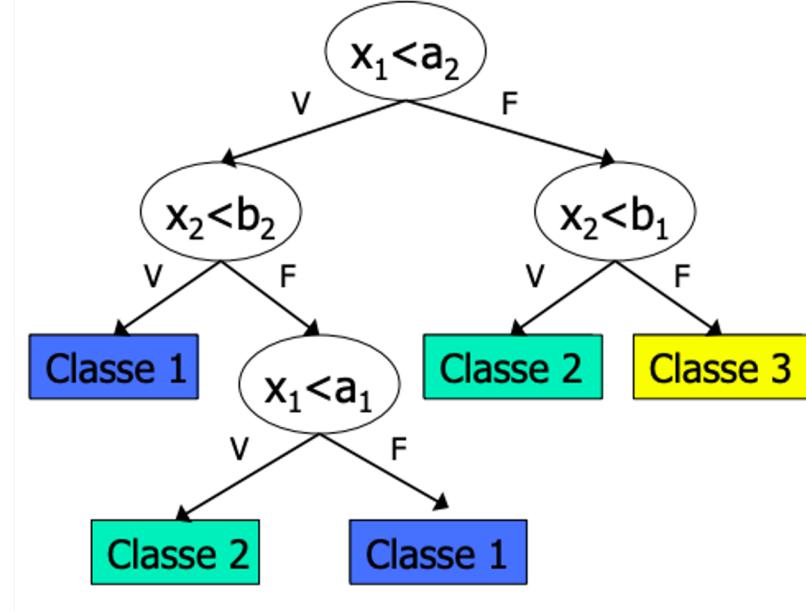
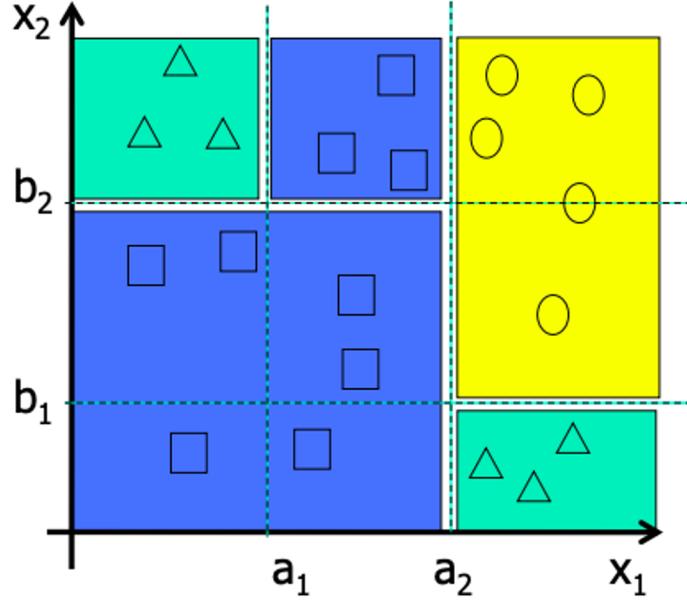
- Classifique de acordo com a classe mais provável.

$$\omega_m = \arg \max_{\omega_i} \prod_{j=1}^d p(x_j | \omega_i), \quad i = 1, 2, \dots, M$$



Árvore de decisão

Como construir a árvore?





Árvores de decisão

Como construir a árvore?

- Medidas de impureza:
 - Índice de Gini
 - Entropia
 - Erro na classificação

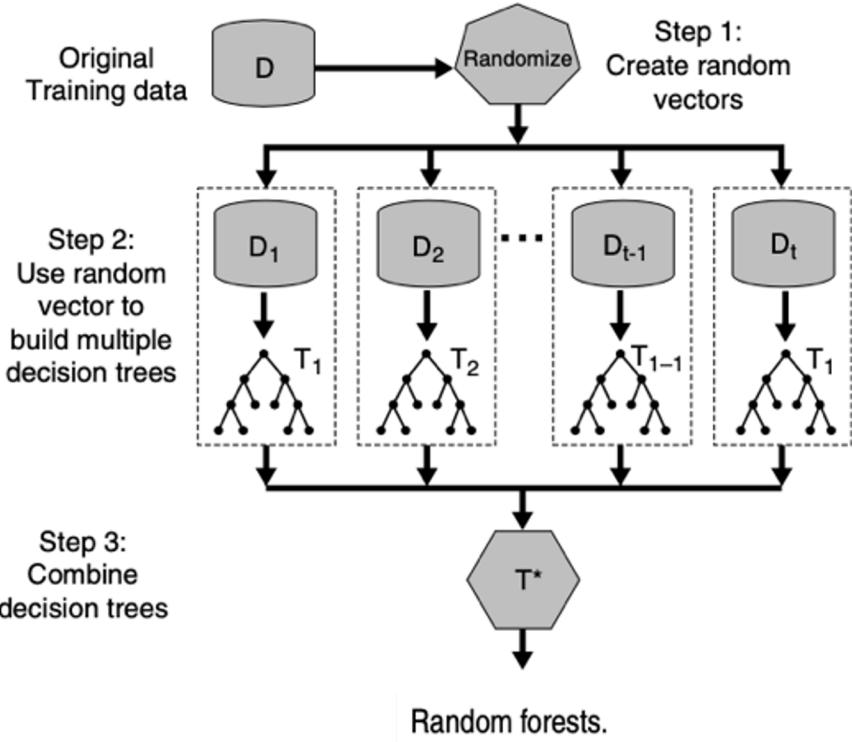




Florestas aleatórias

Florestas aleatórias

- Usa amostragem dos dados para produzir diversas árvores.
- Amostramos as observações e os atributos.

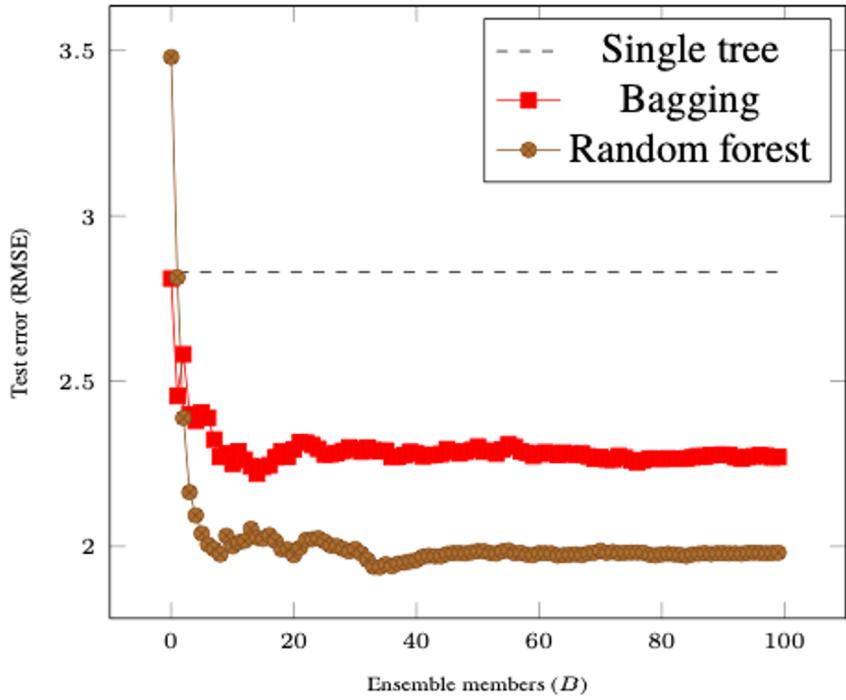




Forests aleatórias

Florestas aleatórias

- A amostragem dos atributos permite que as árvores geradas não sejam dominadas por um atributo com alto poder de discriminação.



Avaliação de modelos

Taxa de erro de um classificador:

$$E(f) = \frac{1}{n} \sum_{i=1}^n I(y_i \neq f(\mathbf{x}_i))$$

- Onde $I(\cdot)$ é a função indicadora, sendo igual a 1 se a entrada for verdadeira.
- $E(f)$ varia entre zero e um, sendo melhor quando for próximo de zero.

Avaliação de modelos

Matriz de confusão

- Linhas representam classes verdadeiras.
- Colunas representam classes preditas.
 - Elemento A_{ij} : número de exemplos da classe c_i classificados como pertencentes à classe c_j .
- Diagonal da matriz: acertos do classificador.
- Elementos fora da diagonal: erros cometidos.

Dados da Iris

| Clase Predita | Classe Correcta | | |
|------------------|-----------------|------------|-----------|
| | Setosa | Versicolor | Virginica |
| Setosa | 15 | 0 | 0 |
| Versicolor | 0 | 14 | 1 |
| Virginica | 0 | 1 | 4 |

Avaliação de modelos

Problema de duas classes:

- **Exemplo:**

- **Acurácia:**

$$Ac(f) = \frac{VP + VN}{n} = \frac{70 + 60}{200} = 0.65$$

- **Precisão:**

$$Prec(f) = \frac{VP}{VP + FP} = \frac{70}{70 + 40} = 0.64$$

- **Sensitividade:**

$$S(f) = \frac{VP}{VP + FN} = \frac{70}{70 + 30} = 0.7$$

- **Especificidade:**

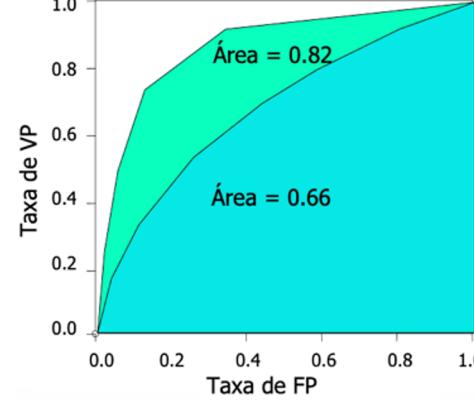
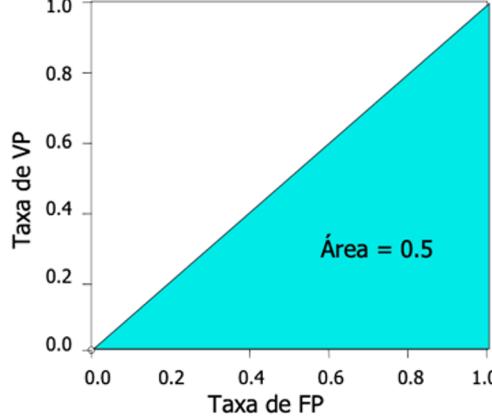
$$Esp(f) = \frac{VN}{VN + FP} = \frac{60}{60 + 40} = 0.60$$

| | | Preditivo | |
|------------|---|-----------|----|
| | | p | n |
| Verdadeiro | p | VP | FN |
| | n | FP | VN |
| Preditivo | p | 70 | 30 |
| | n | 40 | 60 |

Avaliação de modelos

Problema de duas classes:

- Curva ROC (Receiving Operating Characteristics):
 - Área sob a curva ROC:
 - Produz valores no intervalo [0,1]
 - Valores mais próximos de 1 são considerados melhores.



Exemplo de um projeto

- Classificaçã dos dados da base Titanic do site Kaggle...



Regressão

Definição formal: Dado um conjunto de observações:

$$D = \{\mathbf{X}, \mathbf{y}, i = 1, \dots, N\}$$

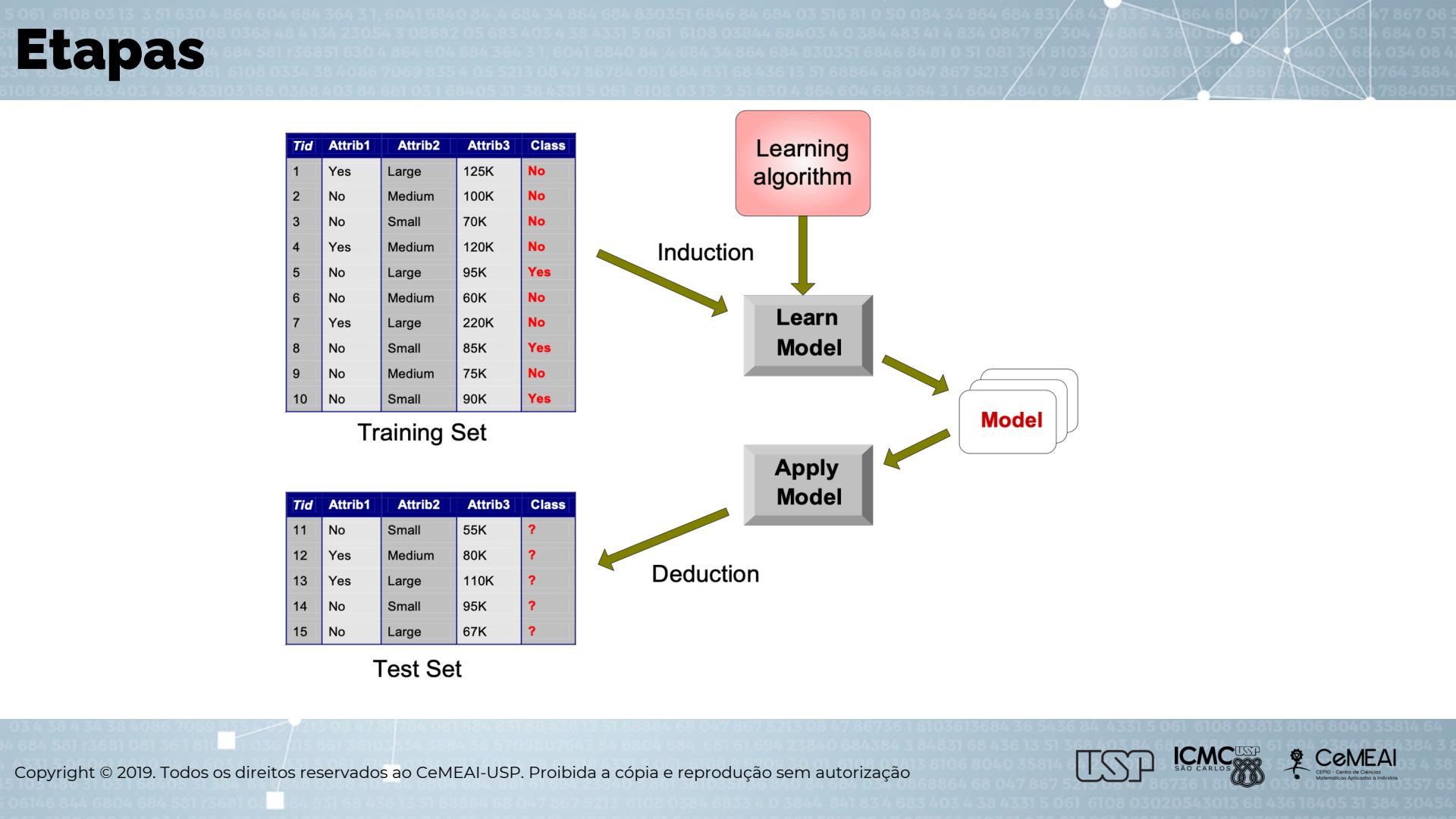
- **f** representa uma função desconhecida (função objetivo).

$$y_i = f(X_i, \theta) + \epsilon_i$$

- Essa função mapeia as entradas nas saídas correspondentes.
- O algoritmo preditivo aprende a aproximação, que permite estimar valores de **f** para novos valores de **X**.

Regressão

$$y_i \in \mathbb{R}$$



Regressão linear

Modelos de regressão podem ser usados principalmente para duas tarefas:

- Prever dados desconhecidos a partir do modelo treinado.
- Determinar a importância de cada variável independente na previsão.

O modelo de regressão linear é um dos mais usados na literatura:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_d X_d$$

Variáveis independentes

↑ ↑ ↑ ↑

Variável resposta (saída) Parâmetros

Regressão linear simples

Para estimar os parâmetros (coeficientes) do modelo, usamos o conjunto de treinamento.

O modelo ajustado é representado por:

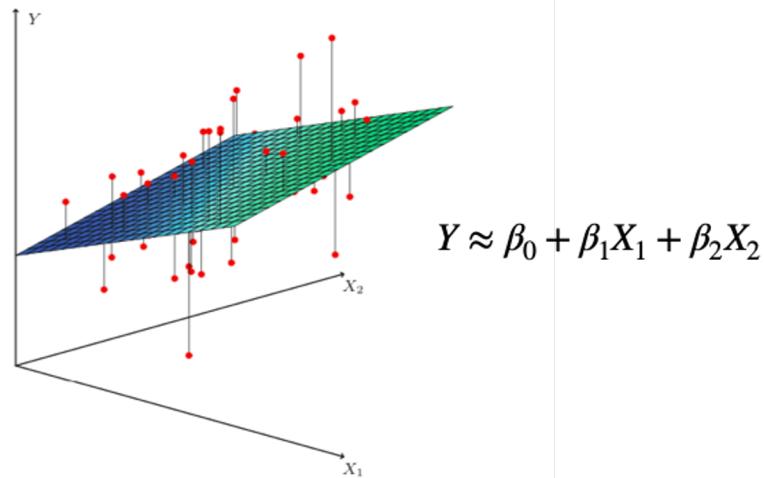
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Que permite determinar estimativas do valor de Y.

A acurácia do modelo é verificada no conjunto de teste.

Regressão linear múltipla

Na maioria dos casos, estamos interessados na influência de várias variáveis em uma variável alvo.



Por exemplo, podemos analisar o efeito do investimento em TV, rádio e jornal na quantidade de itens vendidos, ao invés de considerar apenas TV.

Regressão linear múltipla

Assim:

$$\beta = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

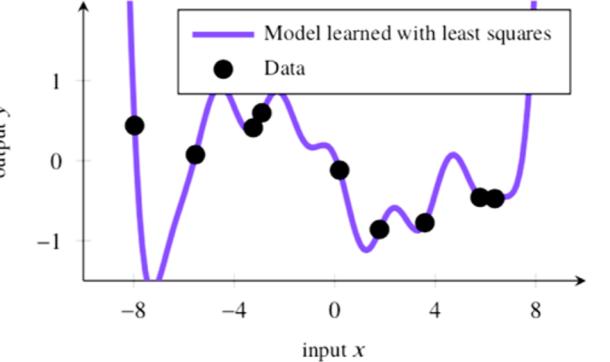
- $\mathbf{X}^T \mathbf{X}$ deve ser positiva definida.

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1d} \\ 1 & x_{21} & x_{22} & \dots & x_{2d} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nd} \end{pmatrix} \quad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_d \end{pmatrix}$$

Inversa de Moore-Penrose $(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$

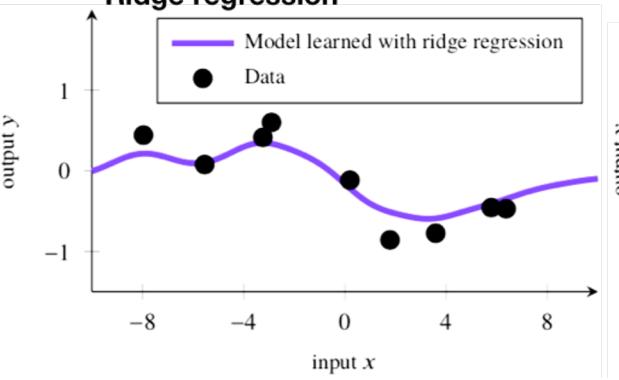
Regularização

Regressão linear

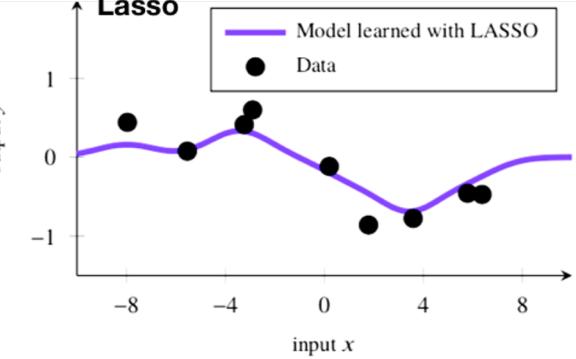


$$\underset{\beta}{\text{minimize}} \underbrace{V(\beta, X, y)}_{\text{data fit}} + \gamma \underbrace{R(\beta)}_{\text{model flexibility penalty}}.$$

Ridge regression



Lasso



Regressão linear múltipla

Suponha que temos d preditores distintos para a variável \mathbf{Y} . Então, o modelo de regressão linear múltipla é:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_d X_d$$

No exemplo:

$$\text{Vendas} = \beta_0 + \beta_1 \text{TV} + \beta_2 \text{Radio} + \beta_3 \text{Jornal}$$

Importante: o modelo é linear nos parâmetros e não nas variáveis, que podem ser não lineares (ex. X^2 , $\text{sen}(X)$, $\ln(X)$).

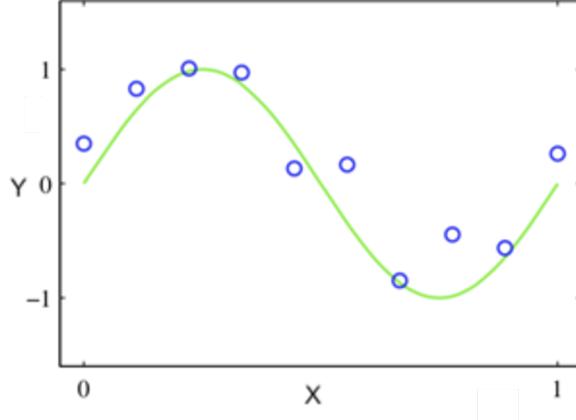
Aprendizado de máquina

- Muitos problemas em aprendizado de máquina consideram os mesmos ingredientes:
 1. O primeiro ingrediente é um conjunto de dados $D=(\mathbf{X}, \mathbf{y})$, onde \mathbf{X} é uma matriz de variáveis independentes e \mathbf{y} é o vetor de variáveis dependentes.
 2. O segundo ingrediente é modelo $f(x, \theta)$, onde f é uma função dos parâmetros θ .
 3. O terceiro ingrediente é a função custo $C(y, f(x, \theta))$ que permite determinar o quanto o modelo f é adequado para predizer \mathbf{y} .

Aprendizado de máquina

Exemplo:

1. $D(\mathbf{X}, y)$



2. Modelo

$$f(\mathbf{x}, \theta) = \sum_{j=0}^M \theta_j x^j$$

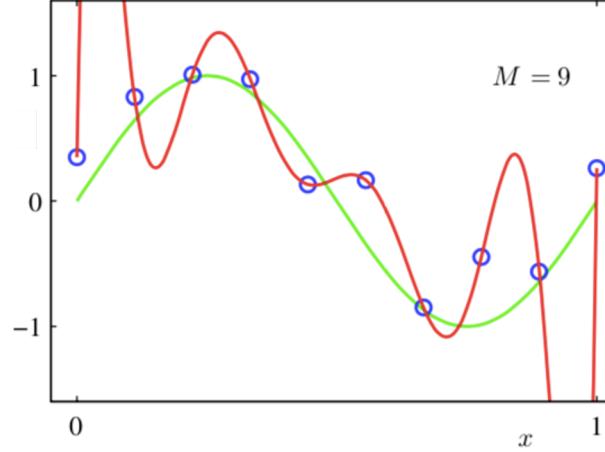
3 Função Custo

$$E(\theta) = \sqrt{\frac{1}{N} \sum_{n=1}^N \{f(x_n, \theta) - y_n\}^2}$$

Overfitting

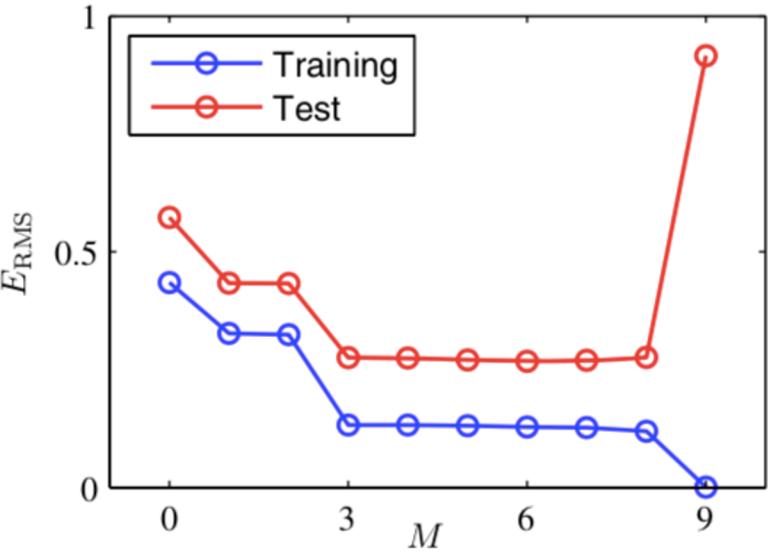
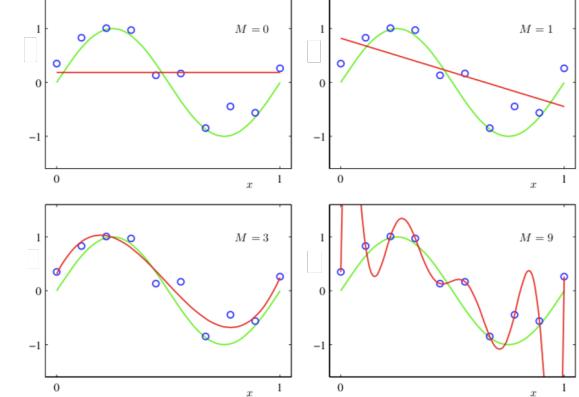
Overfitting: Ocorre quando um modelo se ajusta muito bem ao conjunto de dados anteriormente observado, mas se mostra ineficaz para prever novos resultados.

- Nesse exemplo, temos 10 pontos e um polinômios de grau 9.
- O modelo está super adaptado aos dados de treinamento.



Regressão Polinomial

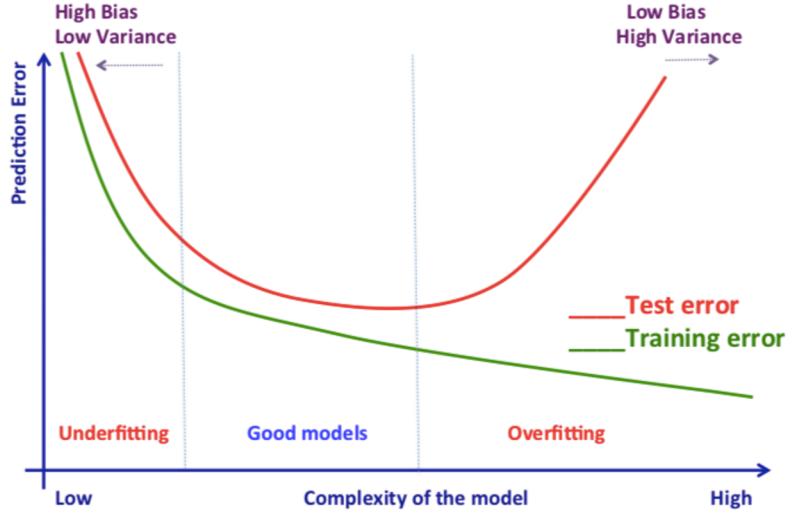
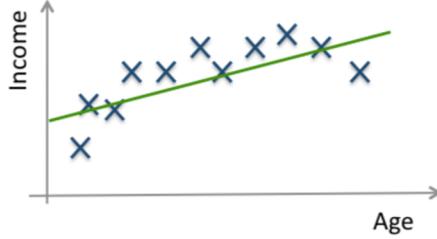
$N = 10$
observações



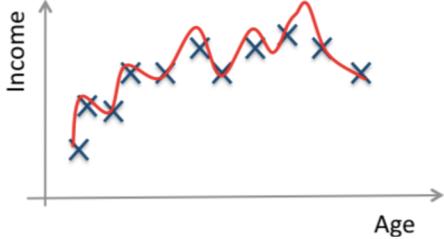
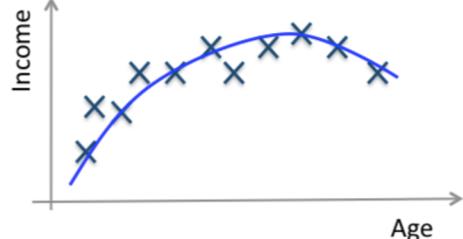
- Quando aumentamos a complexidade do modelo, ocorre um super ajuste.

Viés-variância (bias-variance tradeoff)

Alto viés
Baixa variância
Underfitting



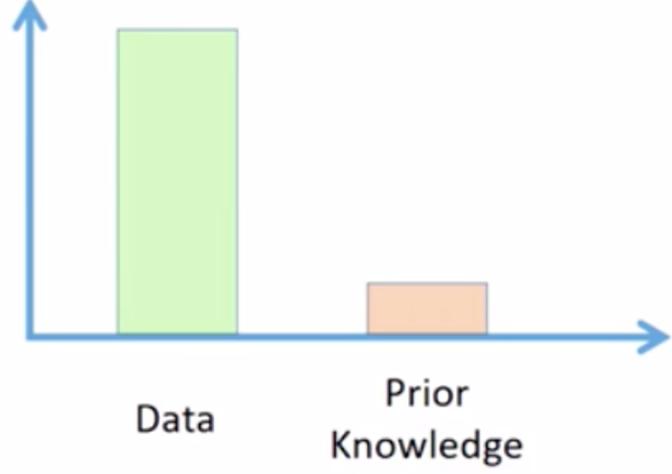
Baixo viés
Alta variância
Overfitting



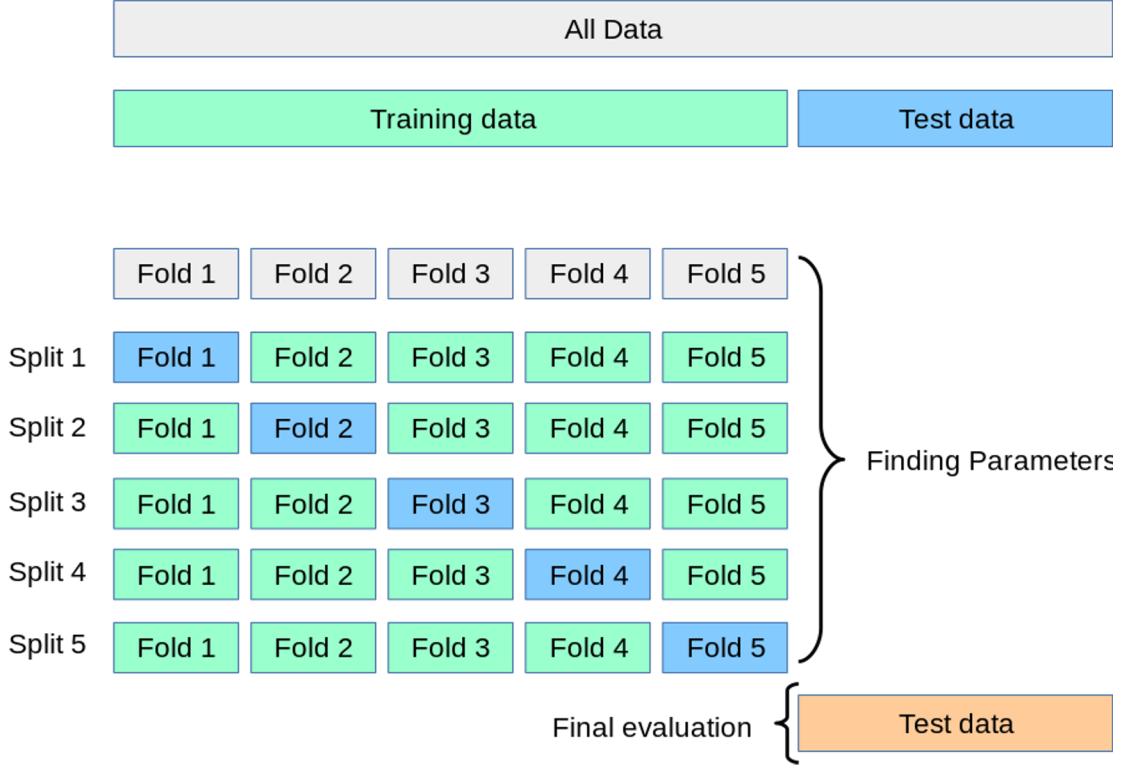


Escolha dos Modelos

Há uma relação entre a quantidade de dados e a complexidade do modelo.



Validação cruzada





Validação cruzada

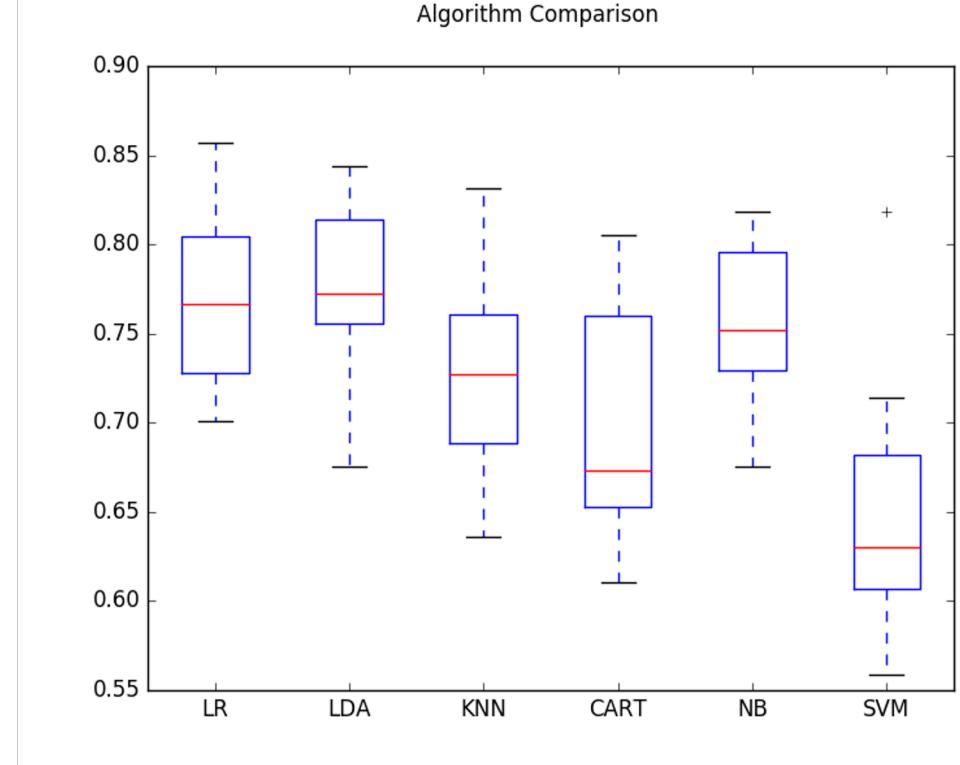
- Na validação cruzada, **todos os dados rotulados são usados**.
- Pode haver **grande variação** nos resultados de cada classificação, pois uma fração $1/k$ dos dados são colocadas no conjunto de teste.
- A média de todas as classificações **reduz a variância** de todo o processo.
- **Validação não serve para determinar a precisão do modelo**, mas para escolher os atributos e modelos.
- Após a validação, usamos **todo o conjunto de dados** para ajustar o método de classificação ou regressão, para aplicar no conjunto de teste.



Validação cruzada

Usamos validação para as seguintes tarefas:

- Comparar modelos.
- Escolha dos parâmetros do modelo (ex. grau do polinômio).



Agrupamento de dados

Limitação: Não há uma definição clara sobre o significado de “cluster” e como encontrá-los.

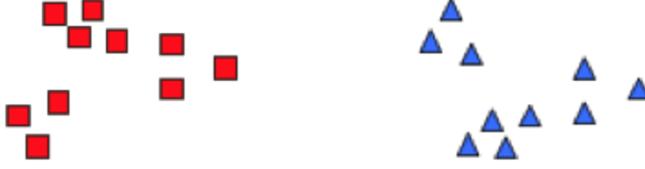
Quantos clusters?



Seis clusters?



Dois clusters?

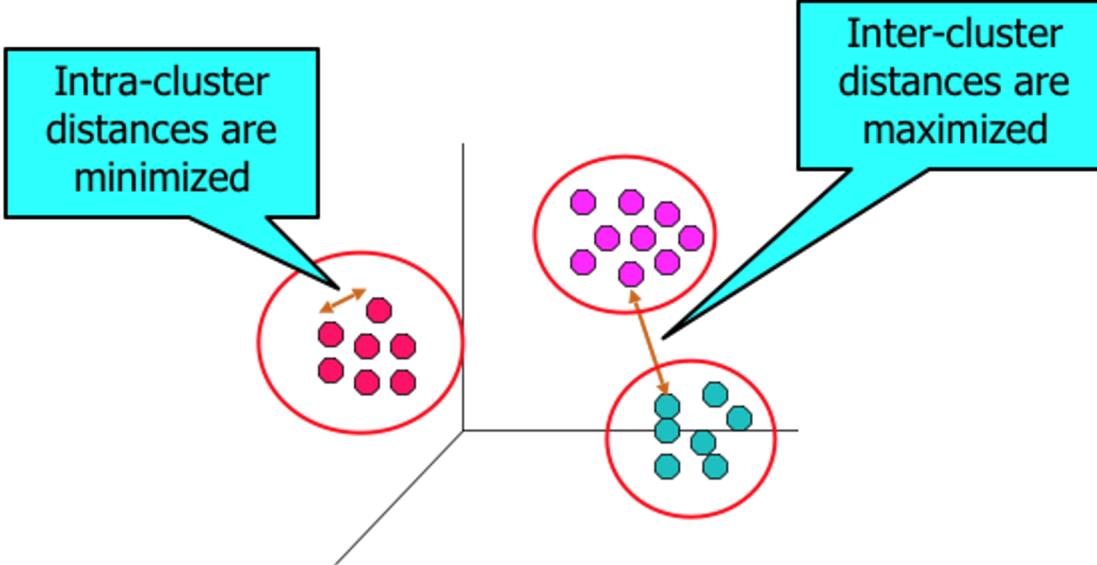


Quatro clusters?



Agrupamento de dados

Encontrar os grupos de objetos tal que objetos no mesmo grupo serão similares (ou relacionados) um ao outro e diferentes (ou não relacionados) a objetos nos outros grupos.





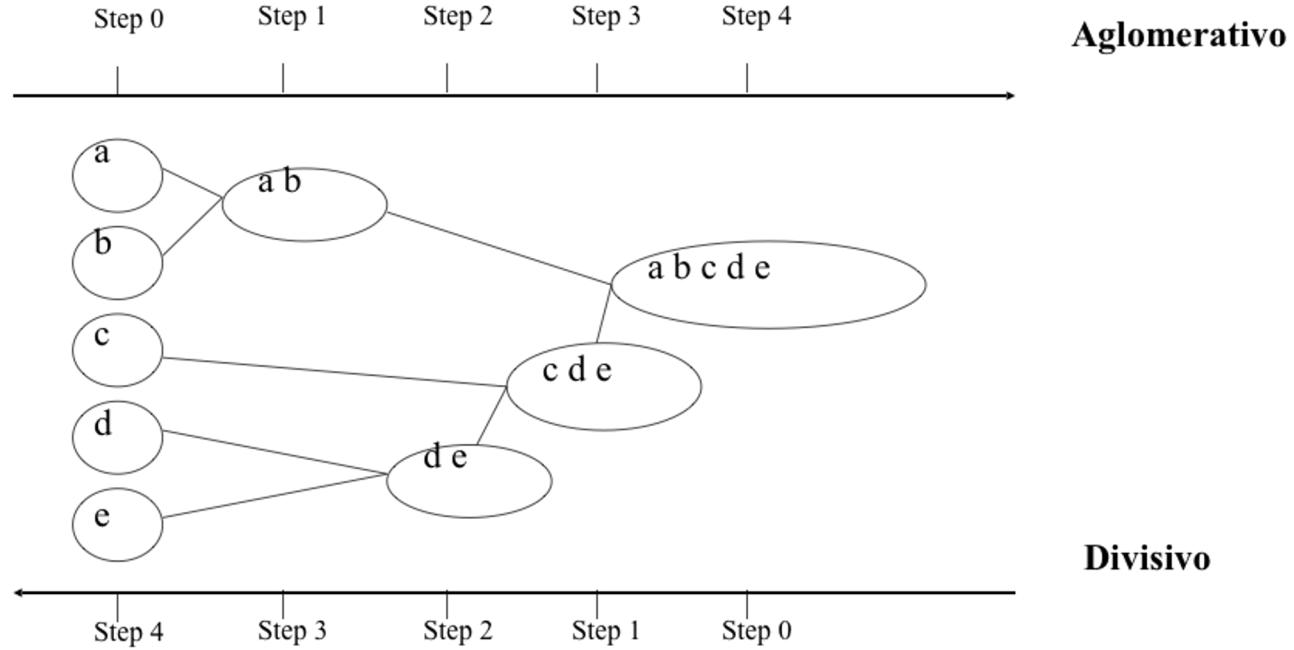
k-means

Algoritmo:

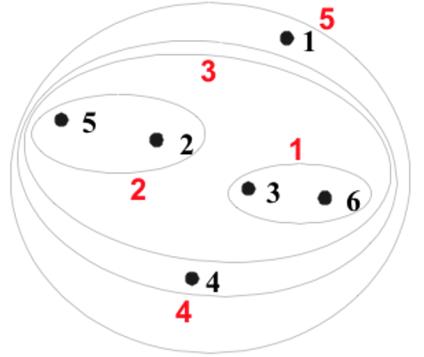
1. Selecione k pontos como centróides iniciais.
2. Repita até que os centróides não mudem.
 - a. Forme k grupos associado todos os pontos aos centróide mais próximos,
 - b. Calcule o centróide de cada grupo obtido.



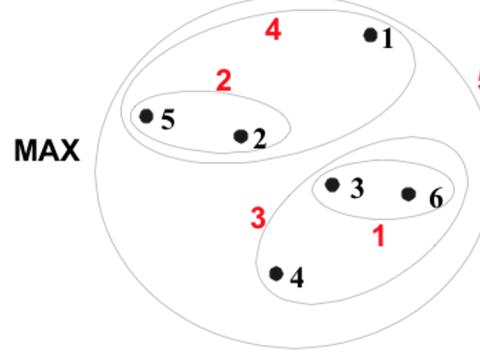
Agrupamento Hierárquico



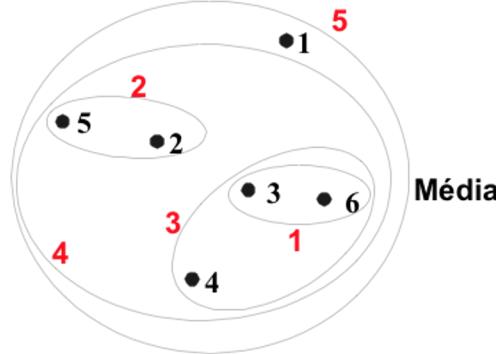
Agrupamento Hierárquico



MIN

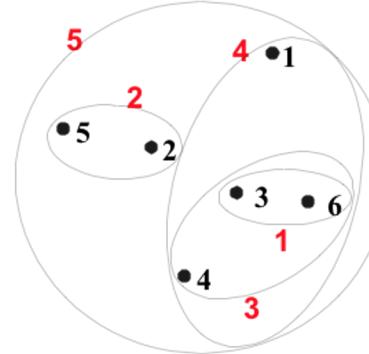


MAX



Média

Método de
Ward's



Avaliando Agrupamentos

- Medidas para avaliar agrupamentos são usadas em três casos:
 1. **Índice externo:** Usado quando os rótulos dos objetos são conhecidos e queremos avaliar se os clusters correspondem aos grupos originais.
 - a. Exemplo: Medidas de entropia.
 2. **Índice interno:** Usado para avaliar um agrupamento sem usar informações externas.
 - a. Exemplo: Soma do erro quadrático
 3. **Índice relativo:** Usado para comparar agrupamentos ou grupos.
 - a. Exemplo: Índices internos ou externos são usados para esse fim.



Dúvidas?