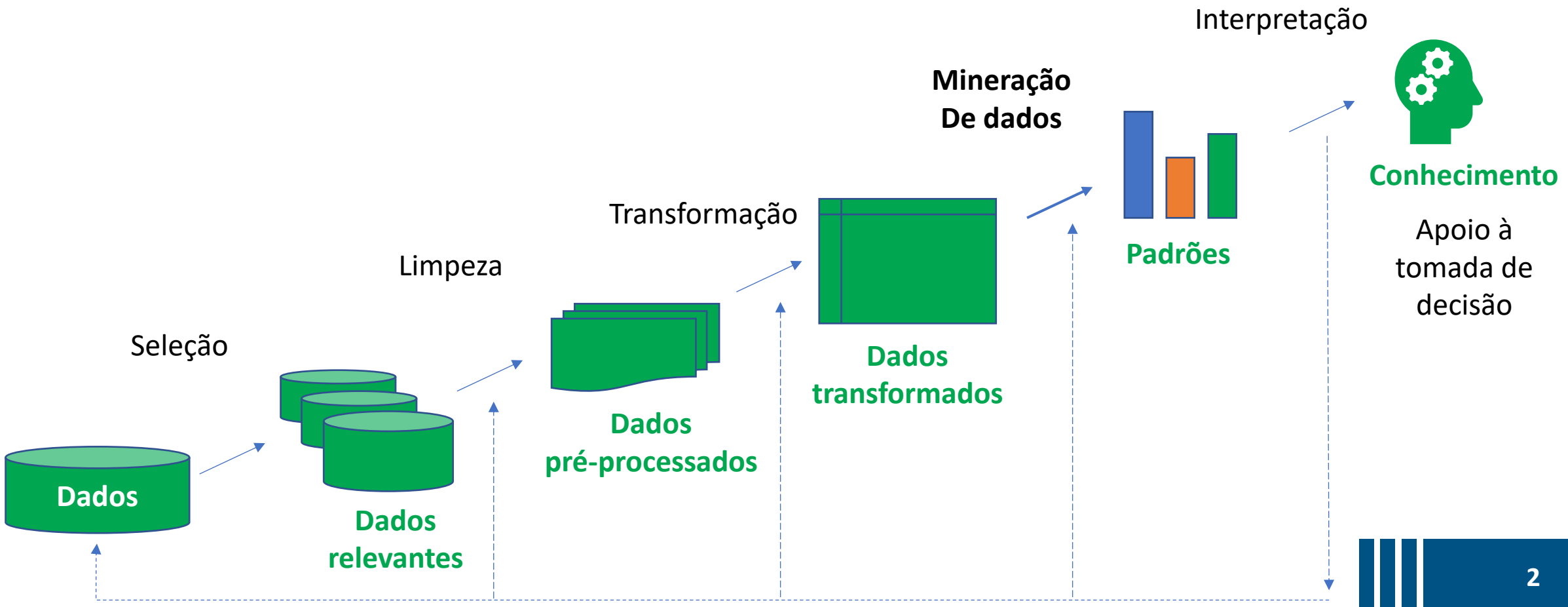
KNN e Naive Bayes

Ciência de Dados II

Professor: Gabriel Machado Lunardi
gabriel.lunardi@ufsm.br

O processo de KDD

"É um processo de várias etapas, não trivial, **interativo** e **iterativo**, para a identificação de **padrões** válidos, novos e potencialmente úteis a partir de um grande conjunto de dados" (FAYYAD, 1996).



1 Seleção

2 Limpeza

3 Transformação

4 **Mineração**

5 Interpretação

- ✓ Selecionar a(s) **tarefa(s)** de mineração de acordo com o problema levantado dentro do projeto de análise de dados:
  - ✓ Associação
  - ✓ Classificação
  - ✓ Agrupamento
  - ✓ **Predição**
  - ✓ ....
- ✓ Escolha do algoritmo dentro da tarefa adotada.
- ✓ Aplicação do algoritmo para construir o modelo.

Principais métodos preditivos

- **Métodos baseados em distâncias**
  - **Algoritmo K-NN**
- Métodos probabilísticos
  - Naive Bayes
  - Redes Bayesianas
- Métodos baseados em procura
  - Árvores de decisão e regressão
- Métodos baseados em otimização
  - Redes neurais artificiais
  - Máquinas de vetores de suporte (SVM)

Métodos baseados em distâncias

- Ideia central
  - Dados similares tendem a estar **próximos**.
  - Dados que não são similares estarão **distantes** entre si.

- Principal algoritmo que segue esta ideia é o **K-NN**
  - K Nearst Neighbors ou K vizinhos mais próximos
  - É o algoritmo mais simples de todos em AM
  - Pode ser utilizado facilmente em problemas de classificação e regressão
  - "Memoriza" os dados de treinamento para classificar uma nova observação
    - Por isso, é chamado de "preguiçoso"
    - Não aprende um **modelo** compacto para os dados
    - O que isso quer dizer?

Métodos baseados em distâncias

Indução
"memorização" dos dados de treinamento

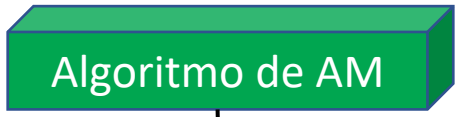
Conjunto de treinamento

Dedução
Comparação da distância

Conjunto de teste

Comparação da nova instância se dá com os dados de treinamento em memória

Outros métodos



Indução
"aprender o modelo"

Modelo

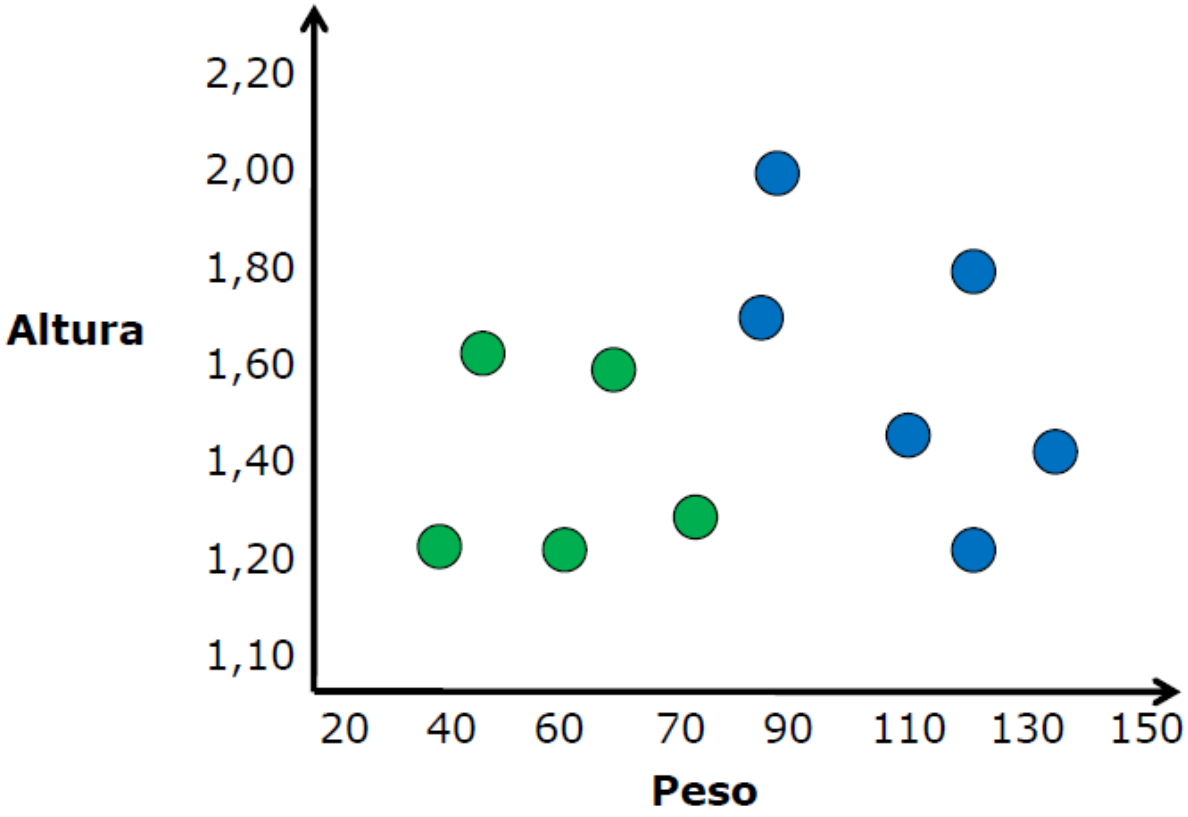
- Uma rede neural
- Árvore decisão
- Rede bayesiana
- ...

Dedução
"aplicar o modelo"

Comparação da nova instância se dá com o modelo

Métodos baseados em distâncias

- ✓Os métodos aqui enquadrados podem assumir que os dados são representados como pontos em um espaço Euclidiano.
- ✓Cada eixo representa um atributo (uma dimensão).



Métodos baseados em distâncias

- ✓ Vejamos uma ideia generalista

- ✓ A aprendizagem consiste em armazenar os dados de treinamento.
  - ✓ $< X_1, C_1 >, \ < X_2, C_2 >, \ \ldots \ < X_n, C_n >$

Vetor de atributos — Valor da classe

- ✓ Após a aprendizagem, para o encontrar o valor da classe associado a uma instância desconhecida (teste) $< X_t, ? >$, um conjunto de instâncias similares são **buscadas na memória** e utilizadas para classificar a nova instância.

Métodos baseados em distâncias

Dados de treino

$< X_1, C_1 >$
$< X_2, C_2 >$
.
.
.
$< X_n, C_n >$

Dado a ser classificado (testado)

$< X_t ??? >$

$d(x_1, x_t)$
$d(x_2, x_t)$
.
.
.
$d(x_n, x_t)$

Lista de distâncias

Distância euclidiana

$$d(p,q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

Métodos baseados em distâncias

Dados de treino

$< X_1, C_1 >$
$< X_2, C_2 >$
.
.
.
$< X_n, C_n >$

Dado a ser classificado (testado)

$< X_t ? ? ? >$

$d(x_1, x_t)$
$\boldsymbol{d(x_2, x_t)}$
.
.
.
$d(x_n, x_t)$

Lista de distâncias

Distância euclidiana

$$d(p,q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

Supondo $\boldsymbol{d(x_2, x_t)}$ como a menor distância, classificaríamos o dado a ser classificado como pertencente à classe C_2!

O algoritmo K-NN

- ✓ Assume que todos os dados correspondem a pontos de um espaço n-dimensional. Cada dimensão corresponde a um atributo.

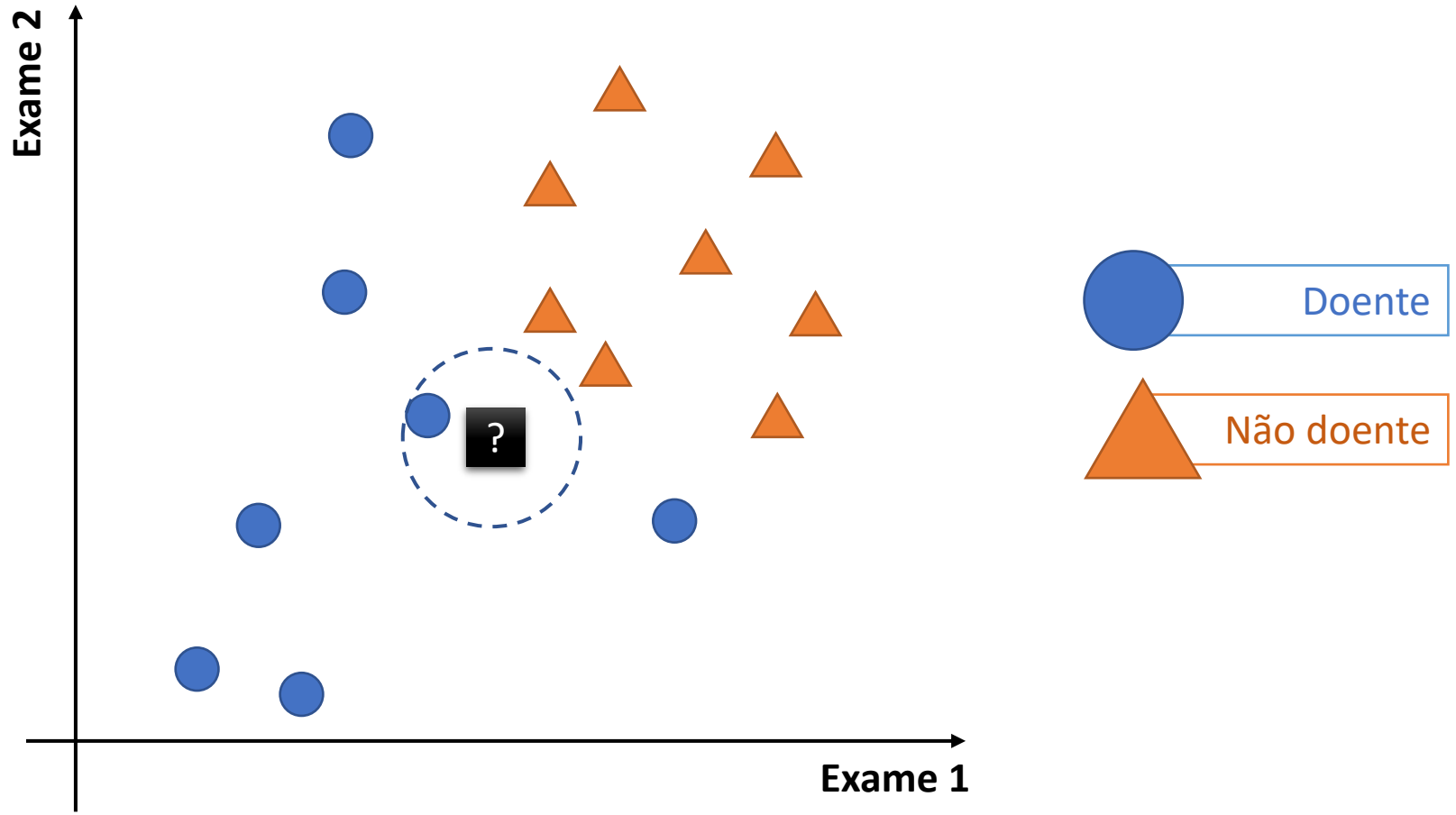
- ✓ Para utilizar o algoritmo é necessário:
  - ✓ 1. Um conjunto de dados de treinamento
  - ✓ 2. Definir uma métrica para calcular a distância entre dois pontos
    - ✓ A mais simples é a distância Euclidiana
    - ✓ Mas existem outras como a distância de Hamming, Minkowsky, etc.
  - ✓ 3. Definir o valor de K
    - ✓ K = número de vizinhos mais próximos que serão considerados pelo algoritmo

O algoritmo K-NN

- Para classificar um exemplo desconhecido é preciso:
  - 1. Calcular a distância entre o exemplo desconhecido e os outros exemplos do conjunto de treinamento.
  - 2. Identificar os K vizinhos mais próximos.
  - 3. Utilizar o rótulo da classe dos vizinhos mais próximos para determinar o rótulo de classe do exemplo desconhecido **(votação majoritária)**.
  - Vamos supor a seguir dados sobre pacientes envolvendo resultados de dois exames classificados em duas classes: doente e não doente. Vamos classificar um exemplo desconhecido a partir de diferentes valores de K.

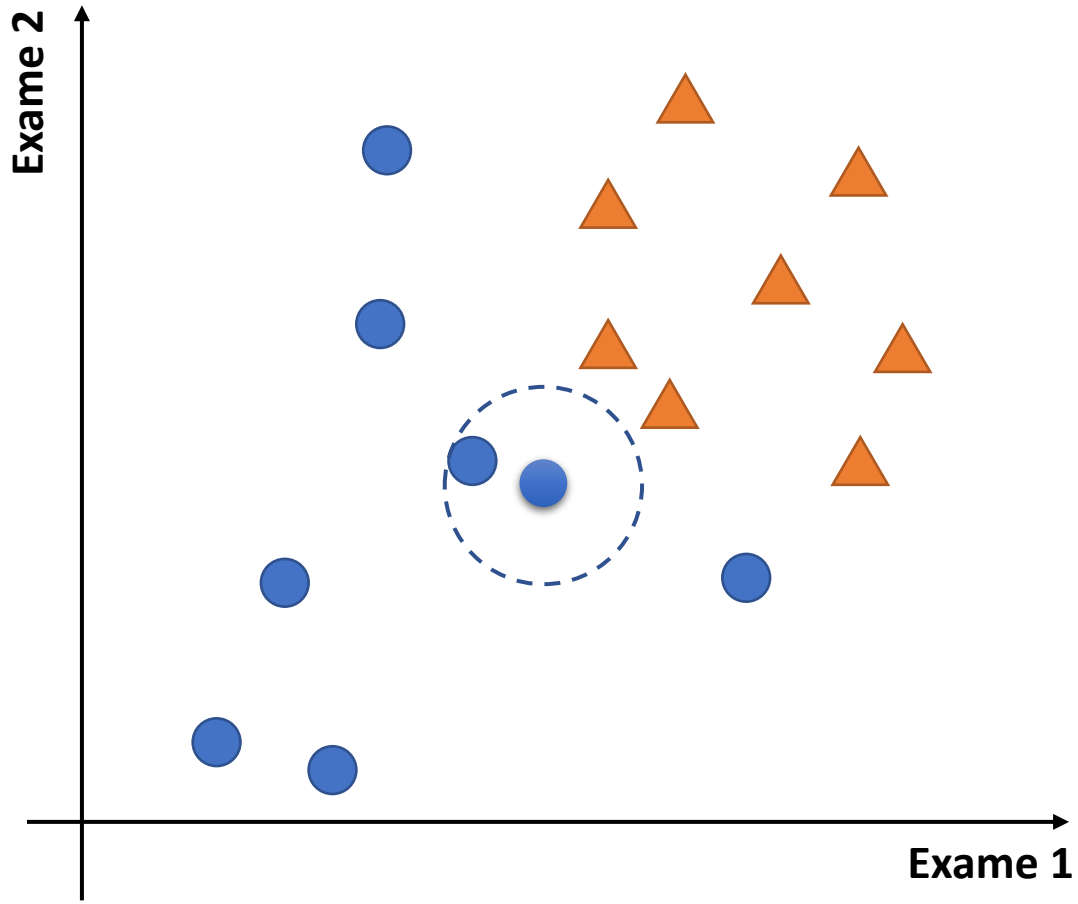
O algoritmo K-NN

- ✓Exemplo de teste K = 1



O algoritmo K-NN

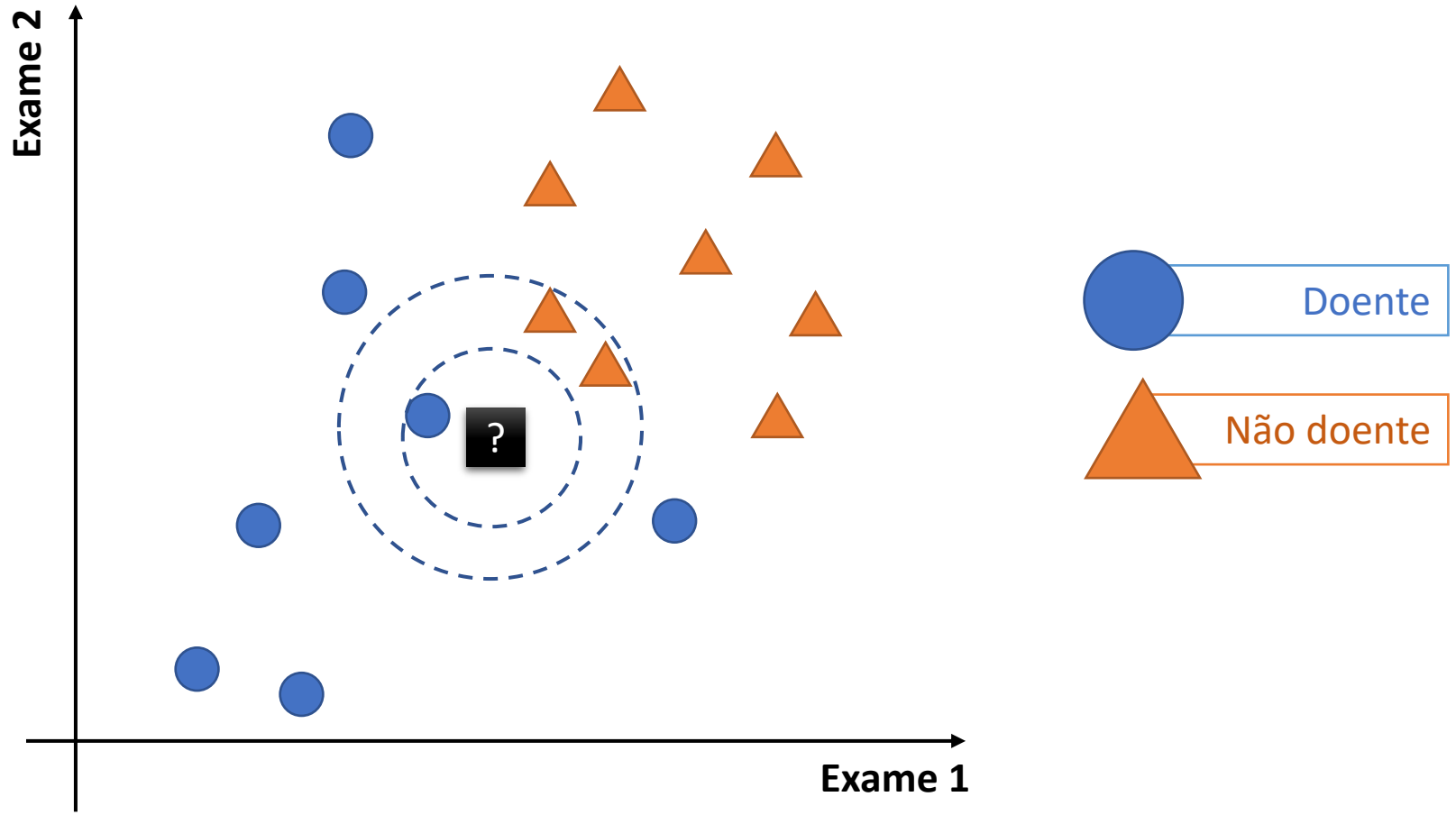
- ✓Exemplo de teste K = 1



Classificado como Doente!

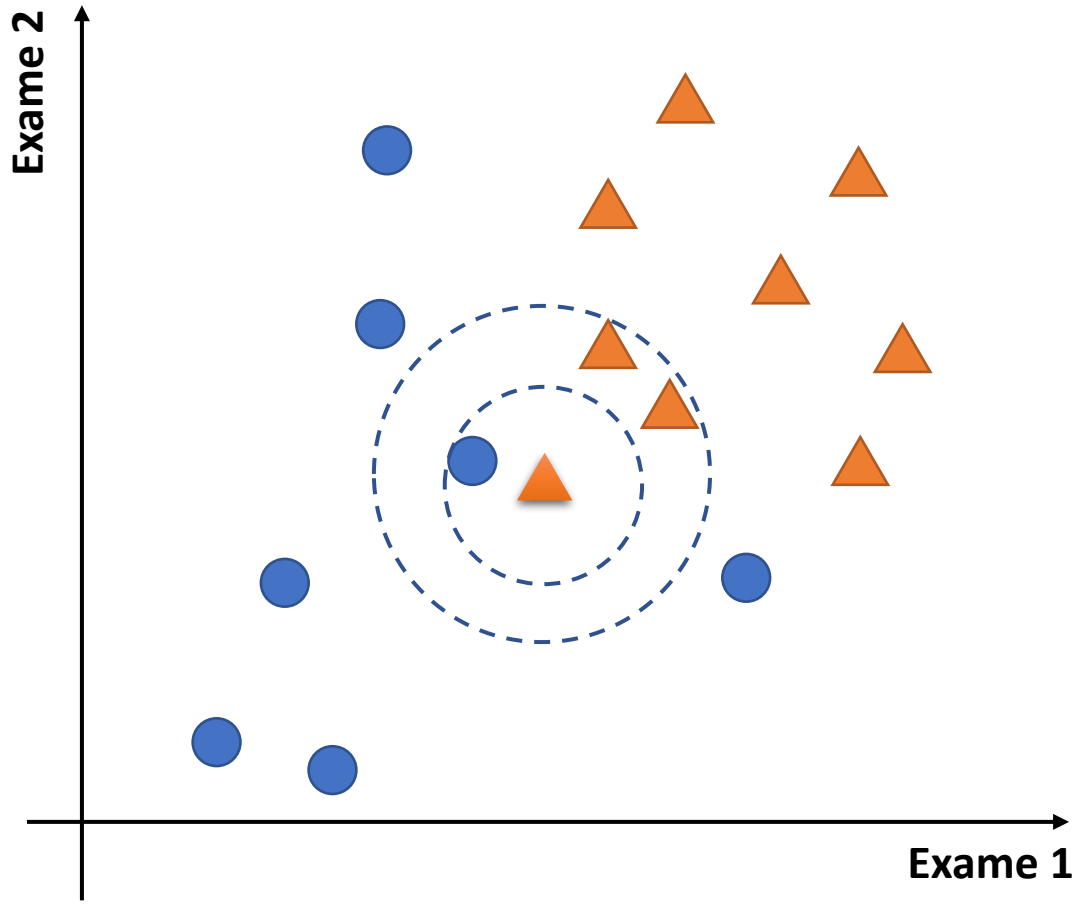
O algoritmo K-NN

- ✓Exemplo de teste K = 3



O algoritmo K-NN

- ✓Exemplo de teste K = 3

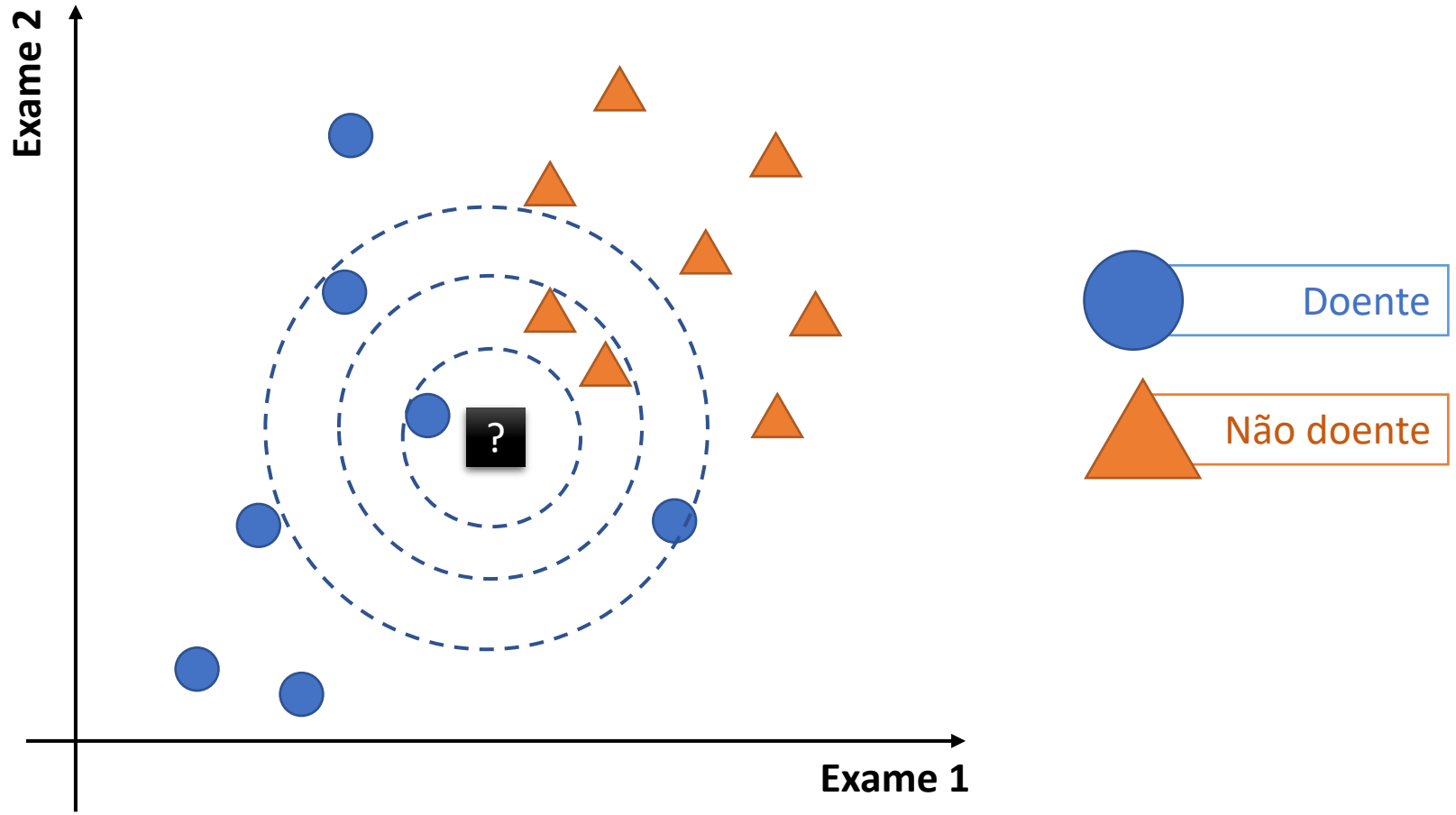


Classificado como Não doente!

Observe como a votação mudou a classe.
Afinal, temos mais vizinhos agora.

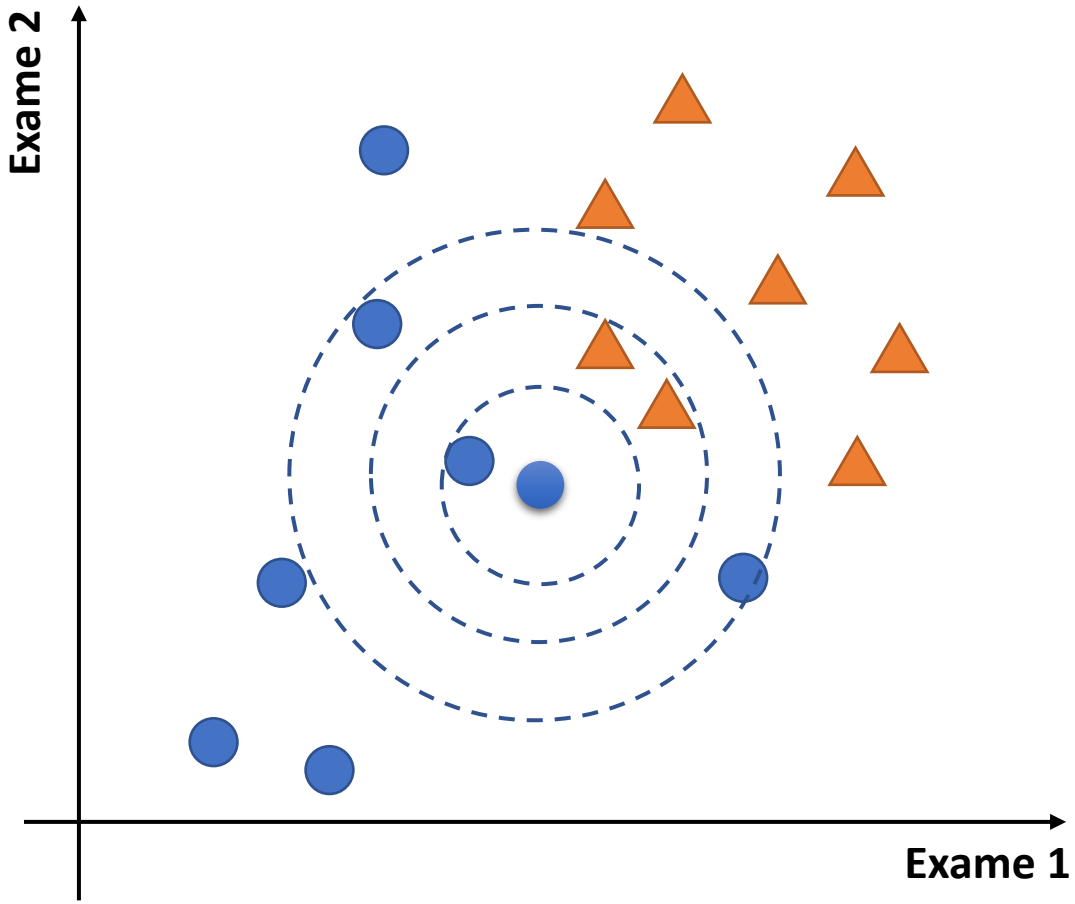
O algoritmo K-NN

- ✓Exemplo de teste K = 5



O algoritmo K-NN

✓Exemplo de teste K = 5



Classificado como Doente!

Observe como a votação mudou a classe.
Afinal, temos mais vizinhos agora.

O algoritmo K-NN

- Na prática, como é determinado o valor do atributo alvo da instância desconhecida sendo testada?

- Em problemas de **classificação**:
  - A **classe** é determinada através da moda de todas as classes dos vizinhos mais próximos ao exemplo desconhecido sendo testado.

- Em problemas de **regressão**:
  - O **valor** é determinado a partir da **média** ou da **mediana** do atributo alvo dos vizinhos mais próximos ao exemplo desconhecido sendo testado

  - Média: se o objetivo for minimizar o erro quadrático
  - Mediana: se o objetivo for minimizar o desvio absoluto.

O algoritmo K-NN

- Como escolher o valor de K?
  - Se K for muito pequeno, a classificação fica sensível a ruídos.
  - Se K for muito grande, a vizinhança pode incluir dados de outras classes.
  - **Para problemas de classificação, escolha um valor pequeno e ímpar para K**, assim evita-se empates na votação.
  - O valor de K também pode ser estimado utilizando validação cruzada.

O algoritmo K-NN

- Outras considerações importantes
  - A precisão da classificação utilizando o K-NN depende fortemente da forma dos dados.
  - Na maioria das vezes, os atributos precisam ser **normalizados** para evitar que as medidas de distância sejam dominadas por um único atributo.
  - Exemplos de um dataset:
    - Altura de um indivíduo pode variar de 1,20 a 2,10.
    - Peso de um indivíduo pode variar de 40Kg a 150Kg.
    - O salário de uma pessoa pode variar de R$ 800,00 a R$ 20.000,00.
    - O salário aqui dominaria a história.

O algoritmo K-NN

✓Vantagens

- ✓Sistemática simples e de fácil implementação
- ✓O treinamento do algoritmo é baseado apenas na memorização dos dados de treino
- ✓Oferece flexibilidade tanto para classificação como para regressão
- ✓Em alguns casos apresenta ótimos resultados
- ✓É um algoritmo naturalmente incremental
  - ✓Assim que obtém-se novos dados de treinamento, basta armazená-los em memória.

O algoritmo K-NN

- **Desvantagens**
  - Classificar um exemplo desconhecido pode ser computacionalmente complexo, pois requer o cálculo da distância para cada exemplo de treinamento.
  - Pode consumir muito tempo quando o conjunto de treino é muito grande.
  - A precisão da classificação pode ser gravemente afetada pela presença de ruído ou atributos irrelevantes ou redundantes.

Principais métodos preditivos

- ✓ Métodos baseados em distâncias
  - ✓ Algoritmo K-NN
- ✓ **Métodos probabilísticos**
  - ✓ **Naive Bayes**
  - ✓ Redes Bayesianas
- ✓ Métodos baseados em procura
  - ✓ Árvores de decisão e regressão
- ✓ Métodos baseados em otimização
  - ✓ Redes neurais artificiais
  - ✓ Máquinas de vetores de suporte (SVM)

Métodos probabilísticos

- Utilizam o **teorema de Bayes** para determinar a probabilidade de **uma instância nunca vista antes** pertencer à cada classe. Em seguida, selecionam a classe mais provável.

- Exemplo: determinar se uma pessoa X pertence à classe doente ou saudável
  - Probabilidade de X estar Doente = 70%
  - Probabilidade de X estar Saudável = 30%
  - Logo, classificaríamos X na classe “Doente” (maior probabilidade)

Métodos probabilísticos

- O Teorema de Bayes para AM probabilístico:

$$P(C,d) = \frac{P(d\,|C)\ \times P(C)}{P(d)}$$

- $P(C,d) =$ probabilidade do exemplo d ser da classe C
  - É o que queremos descobrir

- $P(d,C) =$ probabilidade de gerar o exemplo d dada a classe C
  - Podemos pensar que o exemplo d ser da classe C leva a ter o atributo envolvido com alguma probabilidade

Métodos probabilísticos

- O Teorema de Bayes para AM probabilístico:

$$P(C,d) = \frac{P(d\,|C)\ \times P(C)}{P(d)}$$

- $P(C) =$ probabilidade de ocorrência da classe C
  - Representa apenas a contagem da classe C na nossa base de dados.

- $P(d) =$ probabilidade do exemplo d ocorrer
  - Pode ser ignorado porque é igual para todas as classes.

Métodos probabilísticos

✓Exemplo: Vamos assumir que temos objetos em duas classes:

$$C_1 = homem \qquad C_2 = mulher$$

Vamos supor que temos uma pessoa cujo sexo não saibamos e que se chama Ariel (d). Sendo assim, classificar Ariel como homem ou mulher é igual a se perguntar:

"Ariel" pode ser o nome de um homem ou de uma mulher.

✓Qual é a probabilidade de ser chamado de **Ariel** dado que você é homem?

✓$P(homem \mid Ariel)$

✓Qual é a probabilidade de ser chamado de **Ariel** dado que você é mulher?

✓$P(mulher \mid Ariel)$

Métodos probabilísticos

✓Exemplo: Vamos assumir que temos objetos em duas classes:

$$C_1 = homem \qquad C_2 = mulher$$

Vamos supor que temos uma pessoa cujo sexo não saibamos e que se chama Ariel (d). Sendo assim, classificar Ariel como homem ou mulher é igual a se perguntar:

"Ariel" pode ser o nome de um homem ou de uma mulher.

✓Qual é a probabilidade de ser chamado de **Ariel** dado que você é homem?

Probab. de ser chamado de Ariel dado que é vc é homem

Probab. de ser homem

$$P(homem, Ariel) = \frac{P(Ariel \mid homem) \times P(homem)}{P(Ariel)}$$

Probab. de ser chamado de Ariel (mesmo para ambas as classes)

Métodos probabilísticos

✓Exemplo: Vamos assumir que temos objetos em duas classes:

$$C_1 = homem \qquad C_2 = mulher$$

Vamos supor que temos uma pessoa cujo sexo não saibamos e que se chama Ariel (d). Sendo assim, classificar Ariel como homem ou mulher é igual a se perguntar:

"Ariel" pode ser o nome de um homem ou de uma mulher.

✓Qual é a probabilidade de ser chamado de **Ariel** dado que você é mulher?

Probab. de ser chamado de Ariel dado que é vc é mulher

Probab. de ser mulher

$$P(mulher, Ariel) = \frac{P(Ariel \mid mulher) \times P(mulher)}{P(Ariel)}$$

Probab. de ser chamado de Ariel (mesmo para ambas as classes)

Métodos probabilísticos

✓ Este(a) é o(a) policial Ariel. Vamos aplicar o Teorema de Bayes para determinar a classe desse(a) policial desconhecido(a).

$$P(C,d) = \frac{P(d\,|C)\ \times P(C)}{P(d)}$$

Base dados

Nome	Classe
Ariel	Homem
Roberta	Mulher
Ariel	Mulher
Ariel	Mulher
Pedro	Homem
Renata	Mulher
Justina	Mulher
Plínio	Homem

$$P(homem, Ariel) = \frac{P(Ariel \mid homem)\ \times P(homem)}{P(Ariel)}$$

$$P(homem, Ariel) = \frac{1/3 \times 3/8}{3/8} \cong 33\%$$

Métodos probabilísticos

✓ Este(a) é o(a) policial Ariel. Vamos aplicar o Teorema de Bayes para determinar a classe desse(a) policial desconhecido(a).

$$P(C,d) = \frac{P(d\,|C)\ \times P(C)}{P(d)}$$

Base dados

Nome	Classe
Ariel	Homem
Roberta	Mulher
Ariel	Mulher
Ariel	Mulher
Pedro	Homem
Renata	Mulher
Justina	Mulher
Plínio	Homem

$$P(mulher, Ariel) = \frac{P(Ariel \mid mulher)\ \times P(mulher)}{P(Ariel)}$$

$$P(mulher, Ariel) = \frac{2/5 \times 5/8}{3/8} \cong 66\%$$

Métodos probabilísticos

✓Este(a) é o(a) policial Ariel. Vamos aplicar o Teorema de Bayes para determinar a classe desse(a) policial desconhecido(a).

$$P(C,d) = \frac{P(d\,|C)\ \times P(C)}{P(d)}$$

Base dados

Nome	Classe
Ariel	Homem
Roberta	Mulher
Ariel	Mulher
Ariel	Mulher
Pedro	Homem
Renata	Mulher
Justina	Mulher
Plínio	Homem

$$P(homem, Ariel) = \cong 33\%$$
$$P(mulher, Ariel) \cong 66\%$$

Naive Bayes

- ✓ Um dos algoritmos de aprendizagem mais práticos e utilizados na literatura.

- ✓ Naive = ingênuo
  - ✓ Assume que os atributos de entrada (preditores) são independentes

- ✓ Apesar dessa premissa, o classificador reporta bom desempenho em diversas tarefas de classificação **onde há dependência**.

- ✓ Aplicações bem sucedidas:
  - ✓ Diagnóstico médico
  - ✓ Classificação de documentos textuais

Naive Bayes

- Exemplo: Vamos supor que possuímos uma base de dados de animais que contém quatro atributos:
  - Tamanho: pequeno, médio ou grande
  - Pele: lisa ou peluda
  - Cor: marrom, verde ou vermelho
  - Carne: macia ou dura
  - Seguro: sim ou não (classe)

- Queremos determinar se é seguro ou perigoso ingerir a carne de certo animal, com base nos atributos preditores.

Naive Bayes

Pele	Cor	Tamanho	Carne	Seguro
Peludo	Marrom	Grande	Dura	Sim
Peludo	Verde	Grande	Dura	Sim
Liso	Vermelho	Grande	Macia	Não
Peludo	Verde	Grande	Macia	Sim
Peludo	Vermelho	Pequeno	Dura	Sim
Liso	Vermelho	Pequeno	Dura	Sim
Liso	Marrom	Pequeno	Dura	Sim
Peludo	Verde	Pequeno	Macia	Não
Liso	Verde	Pequeno	Dura	Não
Peludo	Vermelho	Grande	Dura	Sim
Liso	Marrom	Grande	Macia	Sim
Liso	Verde	Pequeno	Macia	Não
Peludo	Vermelho	Pequeno	Macia	Sim
Liso	Vermelho	Grande	Dura	Não
Liso	Vermelho	Pequeno	Dura	???
Peludo	Verde	Pequeno	Dura	???

Naive Bayes

Pele	Cor	Tamanho	Carne	Seguro
Liso	Vermelho	Pequeno	Dura	Sim? ou Não?

- ✓ Dado um pequeno animal que tem pele lisa de cor vermelha e carne dura. É seguro ingerir a carne?

d = [pequeno, lisa, vermelho, dura]

- ✓ A pergunta condicional fica assim: $P(seguro = sim, \boldsymbol{d})$

- ✓ Deve ser lida como: Qual a probabilidade de um animal ser seguro para ingerir dado que ele é pequeno, tem pele lisa, cor vermelha e carne dura?

Naive Bayes

Pele	Cor	Tamanho	Carne	Seguro
Liso	Vermelho	Pequeno	Dura	Sim? ou Não?

- ✓ Dado um pequeno animal que tem pele lisa de cor vermelha e carne dura. É seguro ingerir a carne?

d = [pequeno, lisa, vermelho, dura]

- ✓ Pelo teorema de Bayes, vamos ter:

$$P(seguro = sim, d) = \frac{P(d \mid seguro = sim) \times P(seguro = sim)}{P(d)}$$

Naive Bayes

- Com a fórmula anterior e a tabela de dados podemos calcular a probabilidade de cada animal não classificado pertencer a uma classe ou outra.
  - (seguro = “sim”)
  - (seguro = “não)

- Para isso, o que o algoritmo aplicará a fórmula:

$$P(seguro = x, d) = \frac{\prod_i P(d_i \mid seguro = x) \times P(seguro = x)}{\prod_i P(d_i)}$$

  - x corresponde a “sim” ou “não” que é o valor de cada classe.

Naive Bayes

Pelo	Cor	Tamanho	Carne	Seguro
Liso	Vermelho	Pequeno	Dura	Sim? ou Não?

- ✓ d = [liso, vermelho, pequeno, dura]
- ✓ $\prod_i P(d_i)$

$$P(pelo = liso) =$$
$$P(cor = vermelho) =$$
$$P(tamanho = pequeno) =$$
$$P(carne = dura) =$$

Naive Bayes

Pelo	Cor	Tamanho	Carne	Seguro
Peludo	Marrom	Grande	Dura	Sim
Peludo	Verde	Grande	Dura	Sim
Liso	Vermelho	Grande	Macia	Não
Peludo	Verde	Grande	Macia	Sim
Peludo	Vermelho	Pequeno	Dura	Sim
Liso	Vermelho	Pequeno	Dura	Sim
Liso	Marrom	Pequeno	Dura	Sim
Peludo	Verde	Pequeno	Macia	Não
Liso	Verde	Pequeno	Dura	Não
Peludo	Vermelho	Grande	Dura	Sim
Liso	Marrom	Grande	Macia	Sim
Liso	Verde	Pequeno	Macia	Não
Peludo	Vermelho	Pequeno	Macia	Sim
Liso	Vermelho	Grande	Dura	Não
Liso	Vermelho	Pequeno	Dura	???
Peludo	Verde	Pequeno	Dura	???

Naive Bayes

Pelo	Cor	Tamanho	Carne	Seguro
Liso	Vermelho	Pequeno	Dura	Sim? ou Não?

- ✓ d = [liso, vermelho, pequeno, dura]
- ✓ $\prod_i P(d_i)$

$$P(pelo = liso) = 7/14$$

$$P(cor = vermelho) =$$

$$P(tamanho = pequeno) =$$

$$P(carne = dura) =$$

Naive Bayes

Pelo	Cor	Tamanho	Carne	Seguro
Peludo	Marrom	Grande	Dura	Sim
Peludo	Verde	Grande	Dura	Sim
Liso	Vermelho	Grande	Macia	Não
Peludo	Verde	Grande	Macia	Sim
Peludo	Vermelho	Pequeno	Dura	Sim
Liso	Vermelho	Pequeno	Dura	Sim
Liso	Marrom	Pequeno	Dura	Sim
Peludo	Verde	Pequeno	Macia	Não
Liso	Verde	Pequeno	Dura	Não
Peludo	Vermelho	Grande	Dura	Sim
Liso	Marrom	Grande	Macia	Sim
Liso	Verde	Pequeno	Macia	Não
Peludo	Vermelho	Pequeno	Macia	Sim
Liso	Vermelho	Grande	Dura	Não
Liso	Vermelho	Pequeno	Dura	???
Peludo	Verde	Pequeno	Dura	???

Naive Bayes

Pelo	Cor	Tamanho	Carne	Seguro
Liso	Vermelho	Pequeno	Dura	Sim? ou Não?

- ✓ d = [liso, vermelho, pequeno, dura]
- ✓ $\prod_i P(d_i)$

$$P(pelo = liso) = 7/14$$

$$P(cor = vermelho) = 6/14$$

$$P(tamanho = pequeno) =$$

$$P(carne = dura) =$$

Naive Bayes

Pelo	Cor	Tamanho	Carne	Seguro
Peludo	Marrom	Grande	Dura	Sim
Peludo	Verde	Grande	Dura	Sim
Liso	Vermelho	Grande	Macia	Não
Peludo	Verde	Grande	Macia	Sim
Peludo	Vermelho	Pequeno	Dura	Sim
Liso	Vermelho	Pequeno	Dura	Sim
Liso	Marrom	Pequeno	Dura	Sim
Peludo	Verde	Pequeno	Macia	Não
Liso	Verde	Pequeno	Dura	Não
Peludo	Vermelho	Grande	Dura	Sim
Liso	Marrom	Grande	Macia	Sim
Liso	Verde	Pequeno	Macia	Não
Peludo	Vermelho	Pequeno	Macia	Sim
Liso	Vermelho	Grande	Dura	Não
Liso	Vermelho	Pequeno	Dura	???
Peludo	Verde	Pequeno	Dura	???

Naive Bayes

Pelo	Cor	Tamanho	Carne	Seguro
Liso	Vermelho	Pequeno	Dura	Sim? ou Não?

- ✓ d = [liso, vermelho, pequeno, dura]
- ✓ $\prod_i P(d_i)$

$$P(pelo = liso) = 7/14$$

$$P(cor = vermelho) = 6/14$$

$$P(tamanho = pequeno) = 7/14$$

$$P(carne = dura) =$$

Naive Bayes

Pelo	Cor	Tamanho	Carne	Seguro
Peludo	Marrom	Grande	Dura	Sim
Peludo	Verde	Grande	Dura	Sim
Liso	Vermelho	Grande	Macia	Não
Peludo	Verde	Grande	Macia	Sim
Peludo	Vermelho	Pequeno	Dura	Sim
Liso	Vermelho	Pequeno	Dura	Sim
Liso	Marrom	Pequeno	Dura	Sim
Peludo	Verde	Pequeno	Macia	Não
Liso	Verde	Pequeno	Dura	Não
Peludo	Vermelho	Grande	Dura	Sim
Liso	Marrom	Grande	Macia	Sim
Liso	Verde	Pequeno	Macia	Não
Peludo	Vermelho	Pequeno	Macia	Sim
Liso	Vermelho	Grande	Dura	Não
Liso	Vermelho	Pequeno	Dura	???
Peludo	Verde	Pequeno	Dura	???

Naive Bayes

Pelo	Cor	Tamanho	Carne	Seguro
Liso	Vermelho	Pequeno	Dura	Sim? ou Não?

- ✓ d = [liso, vermelho, pequeno, dura]
- ✓ $\prod_i P(d_i)$

$$P(pelo = liso) = 7/14$$
$$P(cor = vermelho) = 6/14$$
$$P(tamanho = pequeno) = 7/14$$
$$P(carne = dura) = 8/14$$

$$\prod_i P(d_i) \cong 0{,}06 = 6\%$$

É igual a P(d) na fórmula base do teorema de Bayes. Logo, já temos o denominador da fórmula

Naive Bayes

Pelo	Cor	Tamanho	Carne	Seguro
Liso	Vermelho	Pequeno	Dura	Sim? ou Não?

- ✓ d = [liso, vermelho, pequeno, dura]
- ✓ $\prod_i P(d_i \mid seguro = "sim")$

$$P(pelo = liso \mid seguro = sim) =$$
$$P(cor = vermelho \mid seguro = sim) =$$
$$P(tamanho = pequeno \mid seguro = sim) =$$
$$P(carne = dura \mid seguro = sim) =$$

Naive Bayes

Pelo	Cor	Tamanho	Carne	Seguro
Peludo	Marrom	Grande	Dura	Sim*
Peludo	Verde	Grande	Dura	Sim*
Liso	Vermelho	Grande	Macia	Não
Peludo	Verde	Grande	Macia	Sim*
Peludo	Vermelho	Pequeno	Dura	Sim*
Liso	Vermelho	Pequeno	Dura	Sim*
Liso	Marrom	Pequeno	Dura	Sim*
Peludo	Verde	Pequeno	Macia	Não
Liso	Verde	Pequeno	Dura	Não
Peludo	Vermelho	Grande	Dura	Sim*
Liso	Marrom	Grande	Macia	Sim*
Liso	Verde	Pequeno	Macia	Não
Peludo	Vermelho	Pequeno	Macia	Sim*
Liso	Vermelho	Grande	Dura	Não
Liso	Vermelho	Pequeno	Dura	???
Peludo	Verde	Pequeno	Dura	???

Naive Bayes

Pelo	Cor	Tamanho	Carne	Seguro
Liso	Vermelho	Pequeno	Dura	Sim? ou Não?

- ✓ d = [liso, vermelho, pequeno, dura]
- ✓ $\prod_i P(d_i \mid seguro = "sim")$

$$P(pelo = liso \mid seguro = sim) = 3/9$$

$$P(cor = vermelho \mid seguro = sim) =$$

$$P(tamanho = pequeno \mid seguro = sim) =$$

$$P(carne = dura \mid seguro = sim) =$$

Naive Bayes

Pelo	Cor	Tamanho	Carne	Seguro
Peludo	Marrom	Grande	Dura	Sim*
Peludo	Verde	Grande	Dura	Sim*
Liso	Vermelho	Grande	Macia	Não
Peludo	Verde	Grande	Macia	Sim*
Peludo	Vermelho	Pequeno	Dura	Sim*
Liso	Vermelho	Pequeno	Dura	Sim*
Liso	Marrom	Pequeno	Dura	Sim*
Peludo	Verde	Pequeno	Macia	Não
Liso	Verde	Pequeno	Dura	Não
Peludo	Vermelho	Grande	Dura	Sim*
Liso	Marrom	Grande	Macia	Sim*
Liso	Verde	Pequeno	Macia	Não
Peludo	Vermelho	Pequeno	Macia	Sim*
Liso	Vermelho	Grande	Dura	Não
Liso	Vermelho	Pequeno	Dura	???
Peludo	Verde	Pequeno	Dura	???

Naive Bayes

Pelo	Cor	Tamanho	Carne	Seguro
Liso	Vermelho	Pequeno	Dura	Sim? ou Não?

- ✓ d = [liso, vermelho, pequeno, dura]
- ✓ $\prod_i P(d_i \mid seguro = "sim")$

$$P(pelo = liso \mid seguro = sim) = 3/9$$

$$P(cor = vermelho \mid seguro = sim) = 4/9$$

$$P(tamanho = pequeno \mid seguro = sim) =$$

$$P(carne = dura \mid seguro = sim) =$$

Naive Bayes

Pelo	Cor	Tamanho	Carne	Seguro
Peludo	Marrom	Grande	Dura	Sim*
Peludo	Verde	Grande	Dura	Sim*
Liso	Vermelho	Grande	Macia	Não
Peludo	Verde	Grande	Macia	Sim*
Peludo	Vermelho	Pequeno	Dura	Sim*
Liso	Vermelho	Pequeno	Dura	Sim*
Liso	Marrom	Pequeno	Dura	Sim*
Peludo	Verde	Pequeno	Macia	Não
Liso	Verde	Pequeno	Dura	Não
Peludo	Vermelho	Grande	Dura	Sim*
Liso	Marrom	Grande	Macia	Sim*
Liso	Verde	Pequeno	Macia	Não
Peludo	Vermelho	Pequeno	Macia	Sim*
Liso	Vermelho	Grande	Dura	Não
Liso	Vermelho	Pequeno	Dura	???
Peludo	Verde	Pequeno	Dura	???

Naive Bayes

Pelo	Cor	Tamanho	Carne	Seguro
Liso	Vermelho	Pequeno	Dura	Sim? ou Não?

- ✓ d = [liso, vermelho, pequeno, dura]
- ✓ $\prod_i P(d_i \mid seguro = "sim")$

$$P(pelo = liso \mid seguro = sim) = 3/9$$

$$P(cor = vermelho \mid seguro = sim) = 4/9$$

$$P(tamanho = pequeno \mid seguro = sim) = 4/9$$

$$P(carne = dura \mid seguro = sim) =$$

Naive Bayes

Pelo	Cor	Tamanho	Carne	Seguro
Peludo	Marrom	Grande	Dura	Sim*
Peludo	Verde	Grande	Dura	Sim*
Liso	Vermelho	Grande	Macia	Não
Peludo	Verde	Grande	Macia	Sim*
Peludo	Vermelho	Pequeno	Dura	Sim*
Liso	Vermelho	Pequeno	Dura	Sim*
Liso	Marrom	Pequeno	Dura	Sim*
Peludo	Verde	Pequeno	Macia	Não
Liso	Verde	Pequeno	Dura	Não
Peludo	Vermelho	Grande	Dura	Sim*
Liso	Marrom	Grande	Macia	Sim*
Liso	Verde	Pequeno	Macia	Não
Peludo	Vermelho	Pequeno	Macia	Sim*
Liso	Vermelho	Grande	Dura	Não
Liso	Vermelho	Pequeno	Dura	???
Peludo	Verde	Pequeno	Dura	???

Naive Bayes

Pelo	Cor	Tamanho	Carne	Seguro
Liso	Vermelho	Pequeno	Dura	Sim? ou Não?

- ✓ d = [liso, vermelho, pequeno, dura]
- ✓ $\prod_i P(d_i \mid seguro = "sim")$

$$P(pelo = liso \mid seguro = sim) = 3/9$$
$$P(cor = vermelho \mid seguro = sim) = 4/9$$
$$P(tamanho = pequeno \mid seguro = sim) = 4/9$$
$$P(carne = dura \mid seguro = sim) = 6/9$$

- ✓ $\prod_i P(d_i \mid seguro = "sim") \cong 0{,}04 = 4\%$

É igual a P(d | seguro) na fórmula base do teorema de Bayes. Logo, já temos outro componente calculado

Naive Bayes

Pelo	Cor	Tamanho	Carne	Seguro
Liso	Vermelho	Pequeno	Dura	Sim? ou Não?

- ✓ O último componente da fórmula que nos falta calcular é P(seguro = sim)

- ✓ $\prod_i P(seguro = "sim") \cong$

Naive Bayes

Pelo	Cor	Tamanho	Carne	Seguro
Peludo	Marrom	Grande	Dura	Sim
Peludo	Verde	Grande	Dura	Sim
Liso	Vermelho	Grande	Macia	Não
Peludo	Verde	Grande	Macia	Sim
Peludo	Vermelho	Pequeno	Dura	Sim
Liso	Vermelho	Pequeno	Dura	Sim
Liso	Marrom	Pequeno	Dura	Sim
Peludo	Verde	Pequeno	Macia	Não
Liso	Verde	Pequeno	Dura	Não
Peludo	Vermelho	Grande	Dura	Sim
Liso	Marrom	Grande	Macia	Sim
Liso	Verde	Pequeno	Macia	Não
Peludo	Vermelho	Pequeno	Macia	Sim
Liso	Vermelho	Grande	Dura	Não
Liso	Vermelho	Pequeno	Dura	???
Peludo	Verde	Pequeno	Dura	???

Naive Bayes

Pelo	Cor	Tamanho	Carne	Seguro
Liso	Vermelho	Pequeno	Dura	Sim? ou Não?

- ✓ O último componente da fórmula que nos falta calcular é P(seguro = sim)

- ✓ $\prod_i P(seguro = "sim") = \frac{9}{14} \cong 0{,}64 = 64\%$

Agora já temos todas as probabilidades a priori calculadas e, com elas, podemos calcular a probabilidade a posteriori que é justamente a probabilidade da carne ser segura para comer

Naive Bayes

Pelo	Cor	Tamanho	Carne	Seguro
Liso	Vermelho	Pequeno	Dura	Sim? ou Não?

- ✓ Assim, podemos calcular:

$$P(seguro = sim, d) = \frac{\prod_i P(d_i \mid seguro = sim) \times P(seguro = sim)}{\prod_i P(d_i)}$$

$$P(\boldsymbol{seguro = sim}, d) = \frac{0{,}04 \times 0{,}64}{0{,}06} \cong 0{,}42 = 42\%$$

Naive Bayes

Pelo	Cor	Tamanho	Carne	Seguro
Liso	Vermelho	Pequeno	Dura	Sim? ou Não?

- ✓ Fazendo todo o processo anterior para P(seguro=“não”), ou seja, a probabilidade da carne não ser segura para comer...

$$P(seguro = \text{não}, d) = \frac{\prod_i P(d_i \mid seguro = \text{não}) \times P(seguro = \text{não})}{\prod_i P(d_i)}$$

Naive Bayes

Pelo	Cor	Tamanho	Carne	Seguro
Liso	Vermelho	Pequeno	Dura	Sim? ou Não?

- ✓ d = [liso, vermelho, pequeno, dura]
- ✓ $\prod_i P(d_i \mid seguro = \text{"não"})$

$$P(pelo = liso \mid seguro = não) = 4/5$$
$$P(cor = vermelho \mid seguro = não) = 2/5$$
$$P(tamanho = pequeno \mid seguro = não) = 3/5$$
$$P(carne = dura \mid seguro = não) = 2/5$$

- ✓ $\prod_i P(d_i \mid seguro = \text{"não"}) \cong 0{,}08 = 8\%$

É igual a P(d | seguro) na fórmula base do teorema de Bayes. Logo, já temos outro componente calculado

Naive Bayes

Pelo	Cor	Tamanho	Carne	Seguro
Liso	Vermelho	Pequeno	Dura	Sim? ou Não?

- ✓ Fazendo todo o processo anterior para P(seguro="não"), ou seja, a probabilidade da carne não ser segura para comer...

$$P(seguro = \text{não}, d) = \frac{\prod_i P(d_i \mid seguro = \text{não}) \times P(seguro = \text{não})}{\prod_i P(d_i)}$$

$$P(\boldsymbol{seguro = não}, d) = \frac{0{,}08 \times 0{,}36}{0{,}06} \cong 0{,}48 = 48\%$$

Naive Bayes

- ✓ Agora que possuímos as probabilidade a posteriori para as duas classes, podemos decidir a qual classe pertence a observação desconhecida

$$P(\boldsymbol{seguro} = sim, d) = 42\%$$

$$P(\boldsymbol{seguro} = \boldsymbol{não}, d) = 48\%$$

Pelo	Cor	Tamanho	Carne	Seguro
Liso	Vermelho	Pequeno	Dura	**Não**

Observe que os valores de probabilidade não são complementares.
P(seguro = “sim” | d) é diferente de 1 – P(seguro = “não” | d)

Dúvidas?

Ciência de Dados II

Professor: Gabriel Machado Lunardi
gabriel.lunardi@ufsm.br