Universidade Federal da Paraíba

Coordenação do Curso de Ciência de Dados e Inteligência Artificial

Viabilidade da aprendizagem

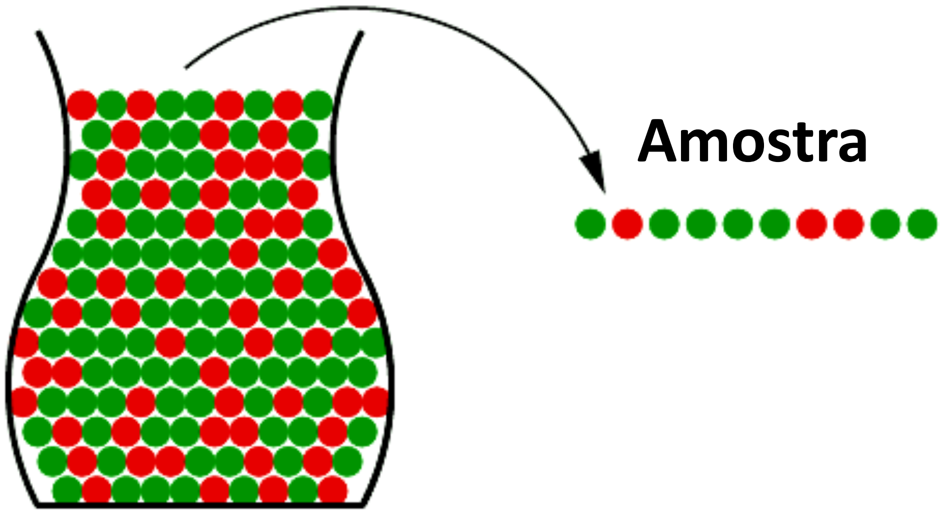
Prof. Dr. Bruno Pessoa

Roteiro

- Relação entre amostra e população
- Desigualdade de Hoeffding
- Conexão com Aprendizagem de Máquina
- Hoeffding para múltiplas hipóteses
- Modelos baseados em distâncias
- Algoritmo k-NN

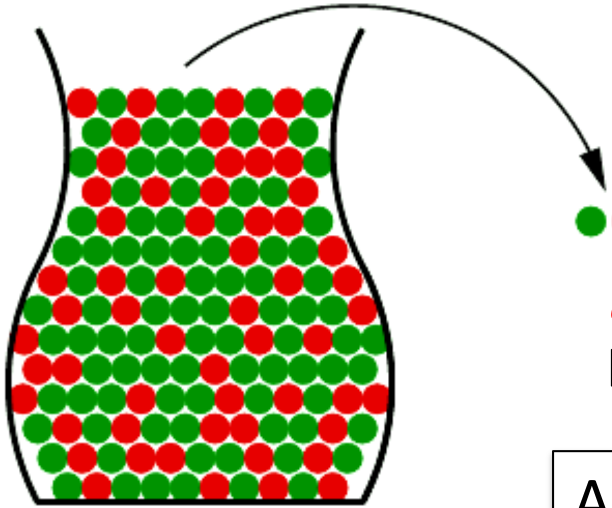
Relação entre amostra e população

- Considere um pote com bolas verdes e vermelhas, onde:
  - $P(\textcolor{red}{pegar\ uma\ vermelha}) = L$
  - $P(\textcolor{green}{pegar\ uma\ verde}) = 1 - L$



- O valor de $\boldsymbol{L}$ é desconhecido.
- A fração de vermelhas na amostra é l.
- O tamanho da amostra é $\boldsymbol{N}$.

Relação entre amostra e população



Amostra

$\boldsymbol{l}$ = **fração de bolas vermelhas**

$\boldsymbol{L}$ = **probabilidade de pegar uma bola vermelha**

A amostra determina o que ocorre na população ou ocorre o inverso?

A amostra dá alguma pista sobre o que ocorre na população?

Teoria da Probabilidade

- Em uma amostra grande o suficiente, l provavelmente se aproxima de L.

- Matematicamente,

$$\boldsymbol{P(|l - L| > \epsilon) \leq 2e^{-2\epsilon^2 N}}$$

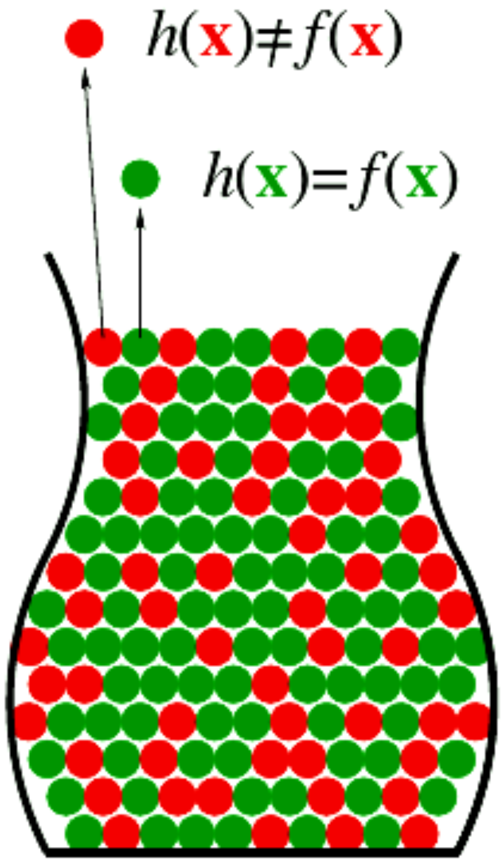
- Conhecida como **desigualdade de Hoeffding**.

Desigualdade de Hoeffding

$$P(|l - L| > \epsilon) \leq 2e^{-2\epsilon^2 N}$$

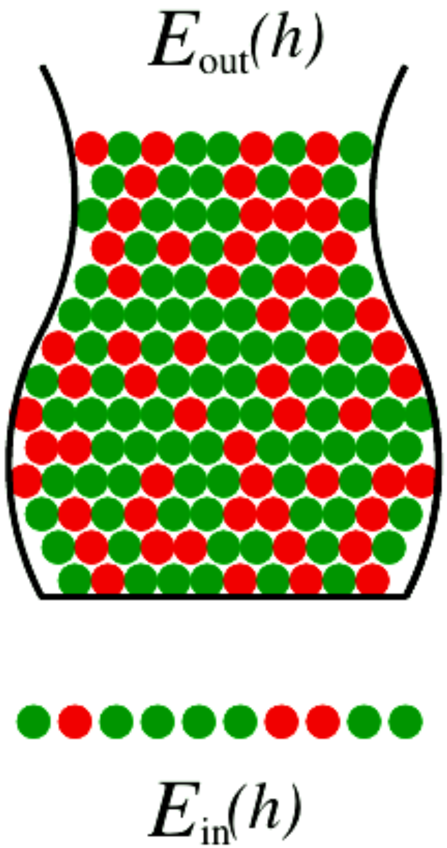
- **Considerações:**
  - O lado direito da inequação, que atua como um **limitante**, não depende de L.
  - l é um componente aleatório da desigualdade.
  - Há claramente um tradeoff entre ϵ e N.
  - Necessidade de aumentar o tamanho da amostra.

Conexão com Aprendizagem de Máquina



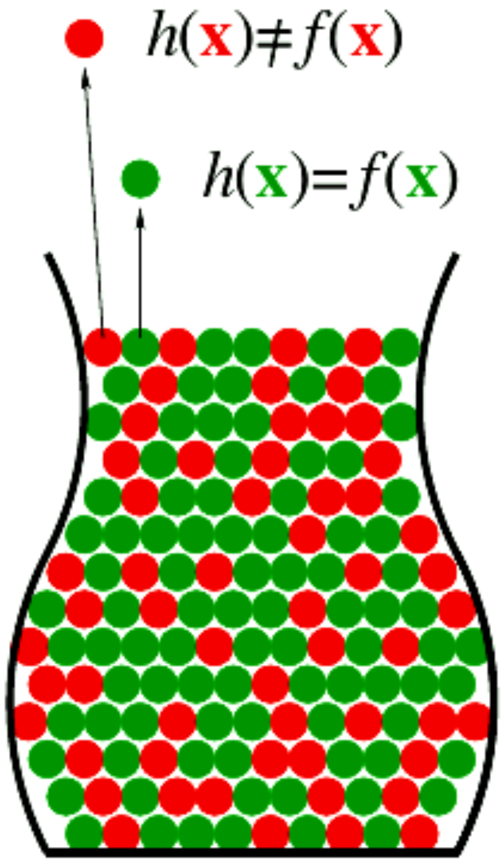
- O ato de pegar uma amostra no pote equivale a escolher um hipótese aleatória para classificar os dados.
- $\boldsymbol{L}$ é desconhecida assim como a função alvo $f: X \rightarrow Y$.
- Cada bola diz respeito a um ponto $x \in X$.
- Uma **bola verde** significa que o ponto foi classificado de forma correta por uma hipótese $\boldsymbol{h}(\boldsymbol{x})$, e uma bola vermelha o contrátrio.
- $\boldsymbol{L}$ e $\boldsymbol{l}$ correspondem à fração de bolas classificadas incorretamente na população e na amostra, respectivamente.
- A amostra consiste nos dados usados no treinamento.

Conexão com Aprendizagem de Máquina



- $\boldsymbol{l}$ é será substituído por $\boldsymbol{E_{in}(h)}$, denotando o **erro dentro da amostra** para uma **hipótese $\boldsymbol{h}$**.
- $\boldsymbol{L}$ é será substituído por $\boldsymbol{E_{out}(h)}$, denotando o **erro fora da amostra** para uma **hipótese $\boldsymbol{h}$**.

Conexão com aprendizagem de máquina

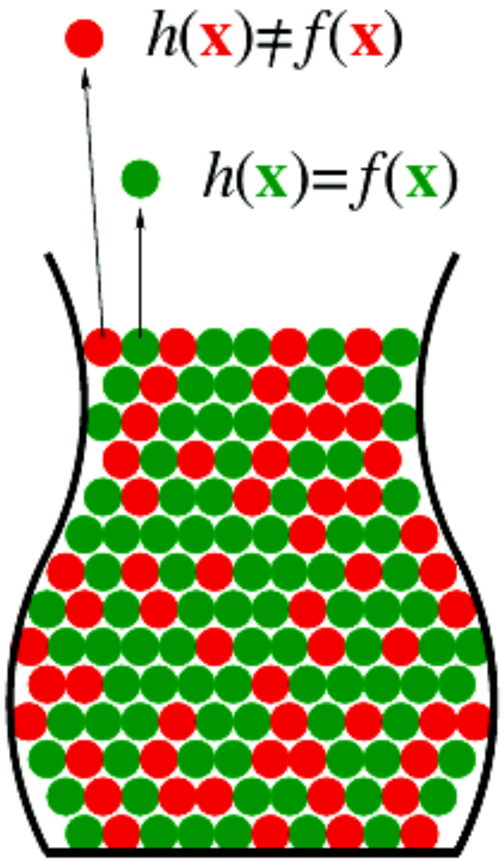


Como resultado:

$$P(|E_{in}(h) - E_{out}(h)| > \epsilon) \leq 2e^{-2\epsilon^2 N}$$

Encontramos uma alicerce matemático para a viabilidade da aprendizagem?

Conexão com aprendizagem de máquina



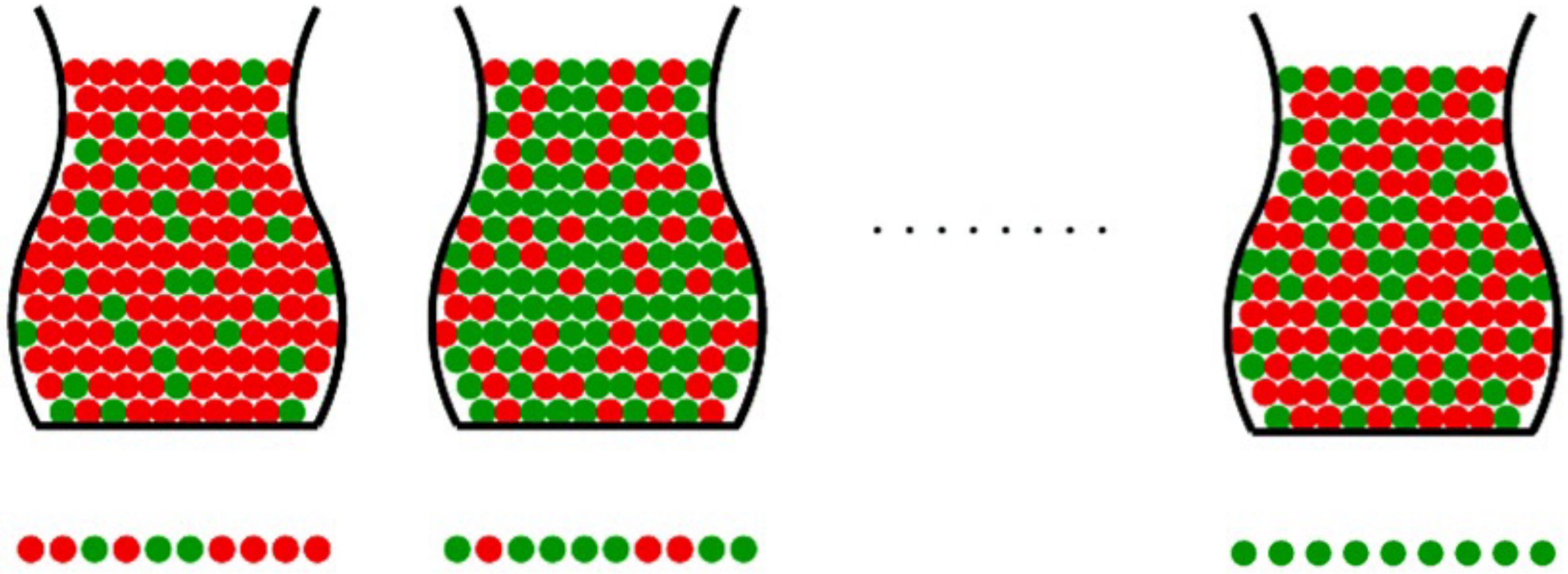
- A desigualdade de Hoeffding dá suporte matemático para a retirada aleatória de apenas uma amostra.

- Portanto, a hipótese $h(x)$ deve ser **fixada** a priori, e não aprendida.

- Para essa hipótese especificamente, $\boldsymbol{l}$ generaliza $\boldsymbol{L}$. Contudo, pode-se generalizar um **aprendizado equivocado**.

- A desigualdade de Hoeffding **não funciona** para um algoritmo que trabalha com **várias hipóteses**.

Hoeffding para múltiplas hipóteses



h_1 h_2 h_i

$i \in \{1, \dots M\}$

Situação de exceção resultando em um aprendizado equivocado.

Um pouco de probabilidade

Problema 1: Se você lançar duas moedas duas vezes, qual a probabilidade de se obter duas caras?

Eventos:

M_1: Obter duas caras com a moeda 1.

M_2: Obter duas caras com a moeda 2.

Espaço amostral:

$A = \{\boldsymbol{(K, K)}, (K, C), (C, K), (C, C)\}$

$$\boldsymbol{P(M_1) = P(M_2) = 0,25}$$

Um pouco de probabilidade

Continuação:

Probabilidade da união

$$\boldsymbol{P(M_1 \cup M_2) = P(M_1) + P(M_2) - P(M_1 \cap M_2)}$$

$$= P(M_1) + P(M_2) - P(M_1).P(M_2)$$

$$= \frac{1}{4} + \frac{1}{4} - \frac{1}{16} = \frac{7}{16} \cong \mathbf{0,44}$$

Um pouco de probabilidade

Problema 2: Sabendo-se que a probabilidade de obter-se 10 caras, ao lançar **uma** moeda 10 vezes, é $\frac{1}{2^{10}} \cong \mathbf{0,1\%}$, qual a probabilidade do mesmo evento no lançamento de 1000 moedas?

$$P(M_1 \cup M_2 \ldots \cup M_{1000}) = \left(\sum_{i=1}^{1000} P(M_i)\right) - \boldsymbol{expr}$$

$$\boldsymbol{P(M_1 \cup M_2 \ldots \cup M_{1000}) \leq \sum_{i=1}^{1000} P(M_i)}$$

$\boldsymbol{R \cong 63\%}$

Hoeffding para múltiplas hipóteses

Seja $\boldsymbol{g}$ a **hipótese final,** selecionada por um algoritmo de aprendizagem qualquer, em um **conjunto de hipóteses $\boldsymbol{H}$**, onde $|\boldsymbol{H}| = \boldsymbol{M}$. Ao aplicar Hoeffding ao algoritmo de aprendizagem, obtemos:

$$
\begin{aligned}
P(|E_{in}(g) - E_{out}(g)| > \epsilon) \leq P(\quad & |E_{in}(h_1) - E_{out}(h_1)| > \epsilon \\
& \cup |E_{in}(h_2) - E_{out}(h_2)| > \epsilon \\
& \dots \\
& \cup |E_{in}(g) - E_{out}(g)| > \epsilon \\
& \dots \\
& \cup |E_{in}(h_M) - E_{out}(h_M)| > \epsilon \ )
\end{aligned}
$$

$$
= \sum_{i=1}^{M} \boldsymbol{P}(\ |\boldsymbol{E}_{in}(\boldsymbol{h}_i) - \boldsymbol{E}_{out}(\boldsymbol{h}_i)| > \boldsymbol{\epsilon} \ ) - \boldsymbol{expr}
$$

Hoeffding para múltiplas hipóteses

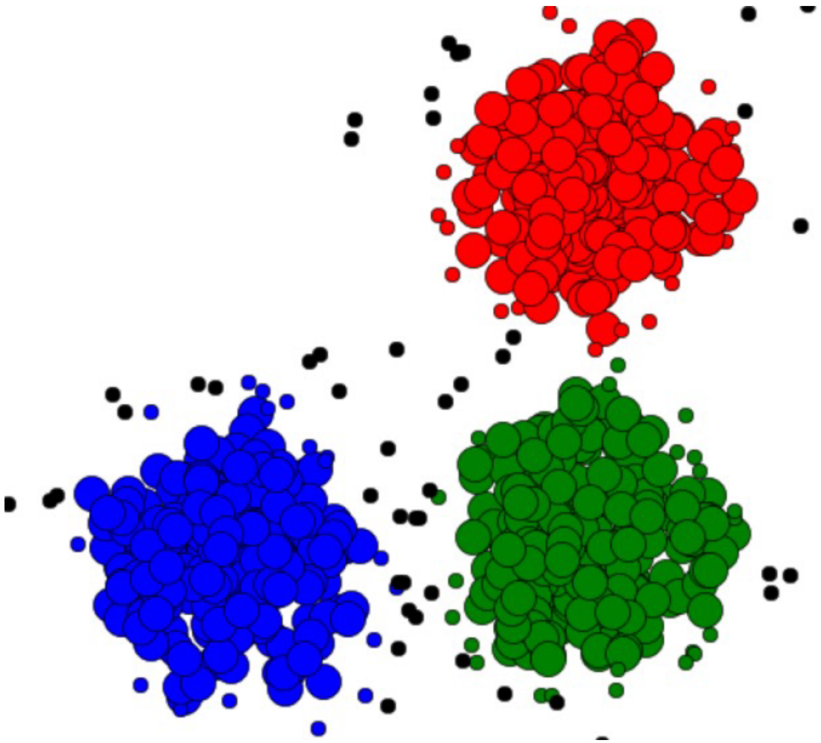
$$P(|E_{in}(g) - E_{out}(g)| > \epsilon) \leq \sum_{i=1}^{M} P(\ |E_{in}(h_i) - E_{out}(h_i)| > \epsilon)$$

$$\leq \sum_{i=1}^{M} 2e^{-2\epsilon^2 N}$$

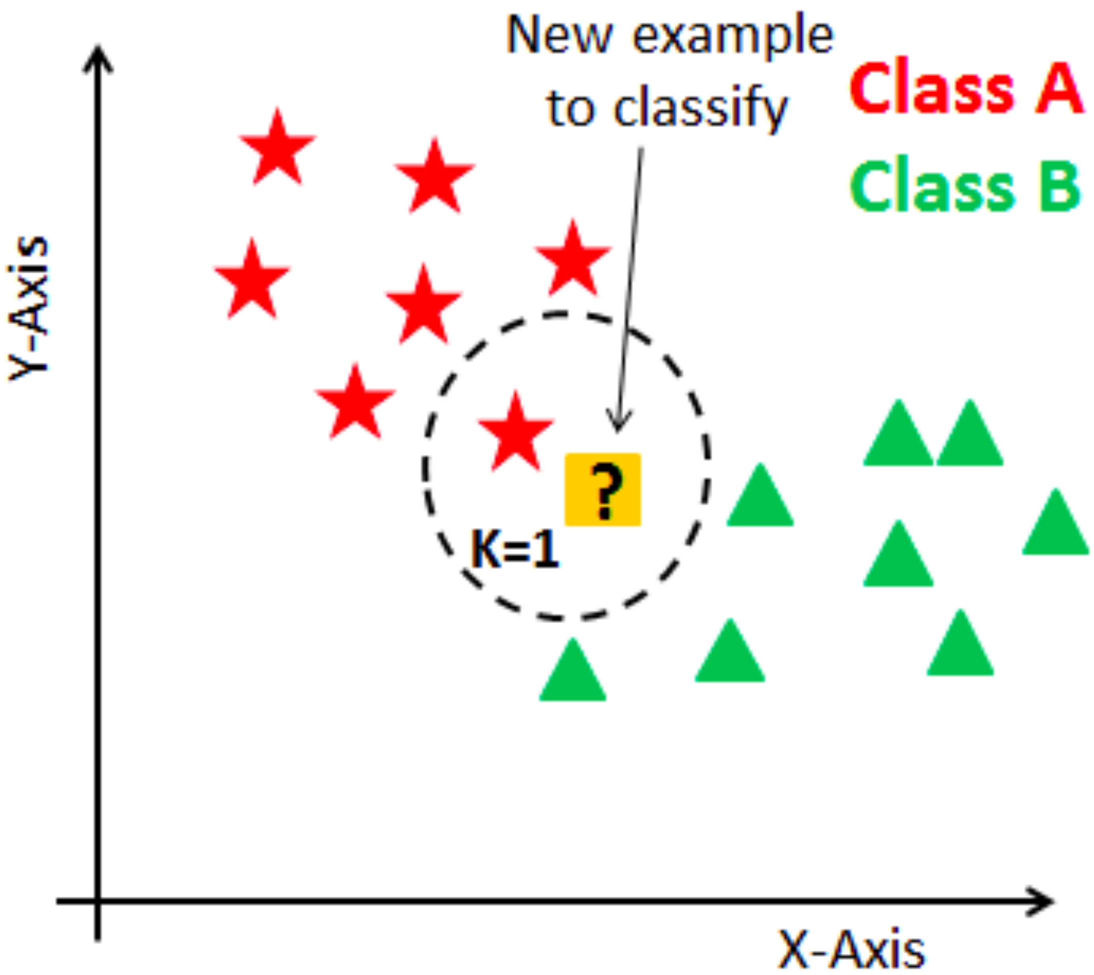
$$\boldsymbol{P(|E_{in}(g) - E_{out}(g)| > \epsilon) \leq 2Me^{-2\epsilon^2 N}}$$

Modelos baseados em distâncias

Algoritmo do vizinho mais próximo (*k-Nearest Neighbors*)



Modelos baseados em distâncias



Modelos baseados em distâncias

- Modelos que consideram a **proximidade de objetos** para a realização de predições.

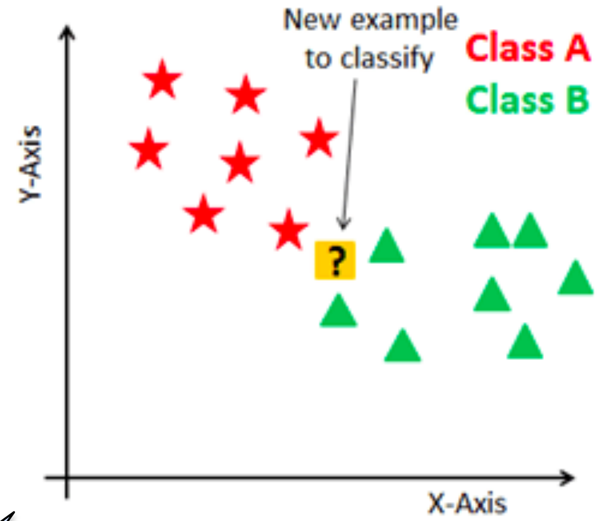
- Supõe-se que objetos similares tendem a se concentrar em uma mesma região do espaço de entrada.

- **Não há aprendizado** no sentido de indução de uma função hipótese.
  - Há uma memorização do conjunto de treinamento.

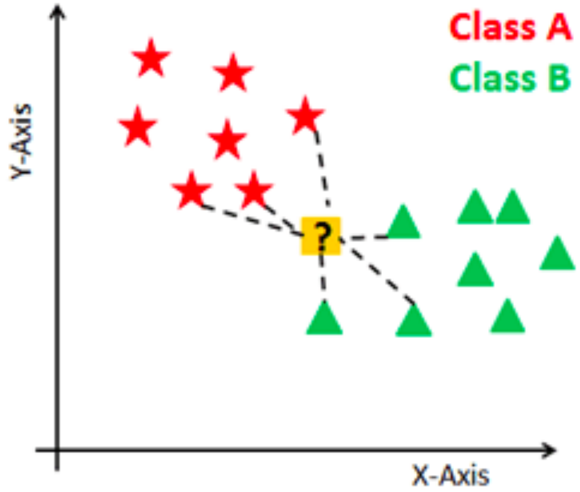
Algoritmo k-NN

- **Pode ser dividido em três etapas:**
  - Calcular as distâncias;
  - Encontrar os K vizinhos mais próximos;
  - Eleger o rótulo mais adequado.

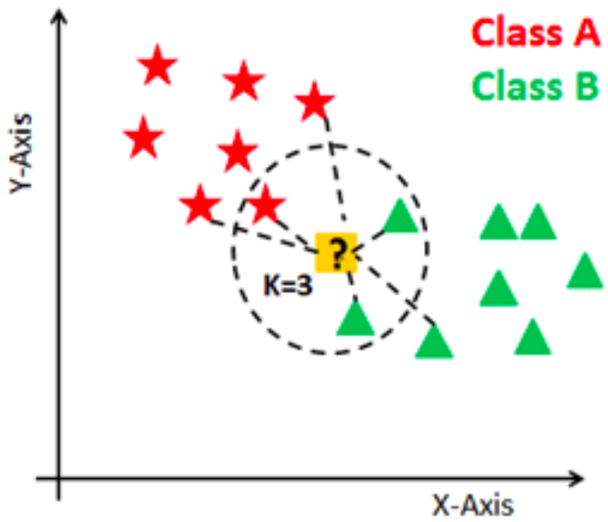
Algoritmo k-NN



Dados
iniciais



Cálculo
das
distâncias



Seleção de vizinhos
mais próximos e
eleição do rótulo.

Distâncias

Sejam $\boldsymbol{x_i}$ e $\boldsymbol{x_j}$ dois vetores em $\mathbb{R}^d$, as distâncias entre eles são dadas por:

$$\boldsymbol{d}(x_i, x_j) = \sqrt{\sum_{l=1}^{d} (x_i^l - x_j^l)^2}$$

Euclidiana

$$\boldsymbol{d}(x_i, x_j) = \sum_{l=1}^{d} |x_i^l - x_j^l|$$

Manhattan

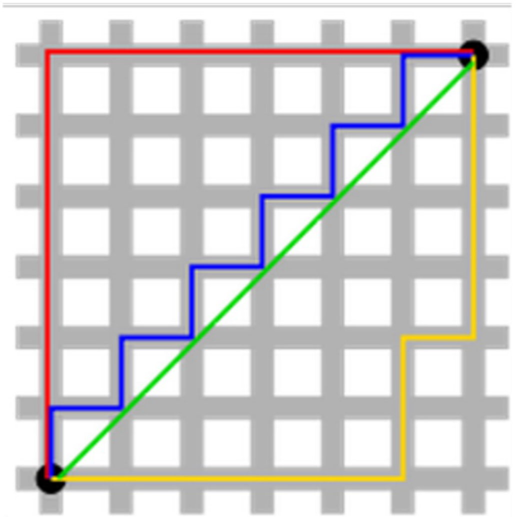
$$\boldsymbol{d}(x_i, x_j) = \left(\sum_{l=1}^{d} |x_i^l - x_j^l|^p\right)^{1/p}$$

Minkowski

Distâncias

- **Generalização das distâncias:**

$$\boldsymbol{d(x_i, x_j) = \left(\sum_{l=1}^{d} |x_i^l - x_j^l|^p\right)^{1/p}}$$



- Para $\boldsymbol{p = 1}$, temos a distância de Manhattan.
- Para $\boldsymbol{p = 2}$, a distância euclidiana.
- Para $\boldsymbol{p = \infty}$, a dimensão dominante é destacada.

Distâncias

- **Considerações:**
  - A distância euclidiana é mais **sensível** a pequenas modificações do que a distância de Manhattan.
    - Não é adequada para dimensões com muito ruído.

  - Um valor de $\boldsymbol{p}$ **alto** pode dar muita ênfase a dimensões com outliers.

Algoritmo k-NN

- **Considerações**
  - A escolha do k não é trivial.
    - Normalmente um valor pequeno e ímpar.
    - A integração com algoritmos evolutivos é uma alternativa.
  - Considerado um algoritmo preguiçoso.
    - Maior parte da computação ocorre no momento da classificação.

Algoritmo k-NN

- **Vantagens**
  – Algoritmo intuitivo e simples de implementar.

  – Consiste em um algoritmo incremental.
    - Novos exemplos são adicionados sem gerar esforço computacional adicional.

  – Poucos parâmetros a serem ajustados.

Algoritmo k-NN

- **Desvantagens**
  - Necessário recalcular as distâncias para cado novo ponto a ser rotulado.
  - Susceptível a atributos redundantes e irrelevantes.
  - Com o aumento no número de dimensões, há um salto na magnitude das distâncias.
    - A distância do vizinho mais próximo aproxima-se da do mais afastado.
  - Necessidade de normalização dos valores das dimensões.

Aplicações

- Reconhecimento facial.

- Identificação de padrões de fraude na utilização de cartões de crédito.

- Identificação de padrões de compra em lojas varejistas.

- Sistemas de recomendação.

- Benchmark para modelos mais sofisticados.

Referências bibliográficas

- Abu-Moustafa, Y.S.; Magdon-Ismail, M.; Lin, H-S. "*Learning from data*". AMLBook, 2012.

- Faceli, K.; Lorena, A.C.; Gama, J.; Carvalho, A.C.P.L.F. "*Inteligência Artificial Uma Abordagem de Aprendizado de Máquina*". LTC, 2011.

- Notas de aula do prof. Abu-Moustafa.