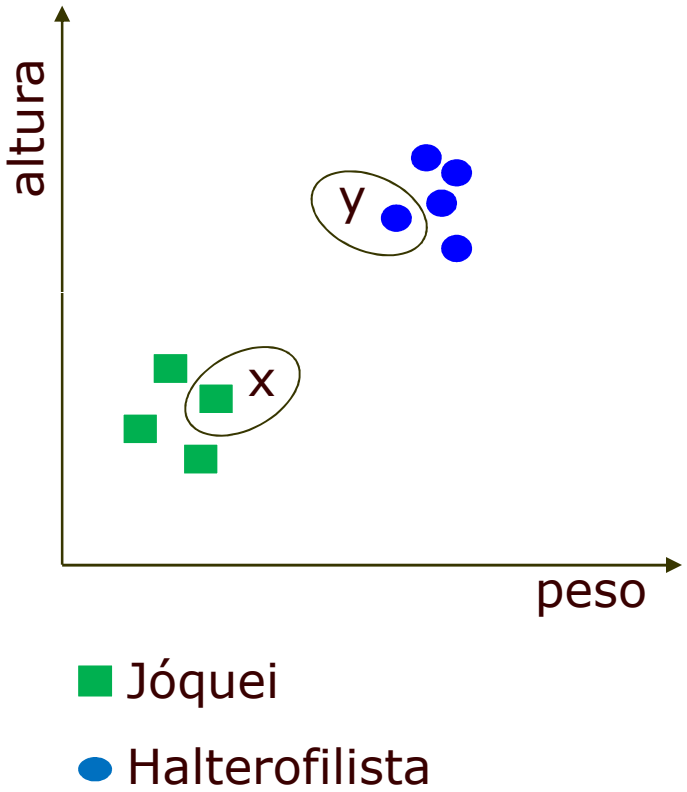
k-Nearest Neighbor

George Darmiton da Cunha Cavalcanti
(gdcc@cin.ufpe.br)
CIn/UFPE

Introdução

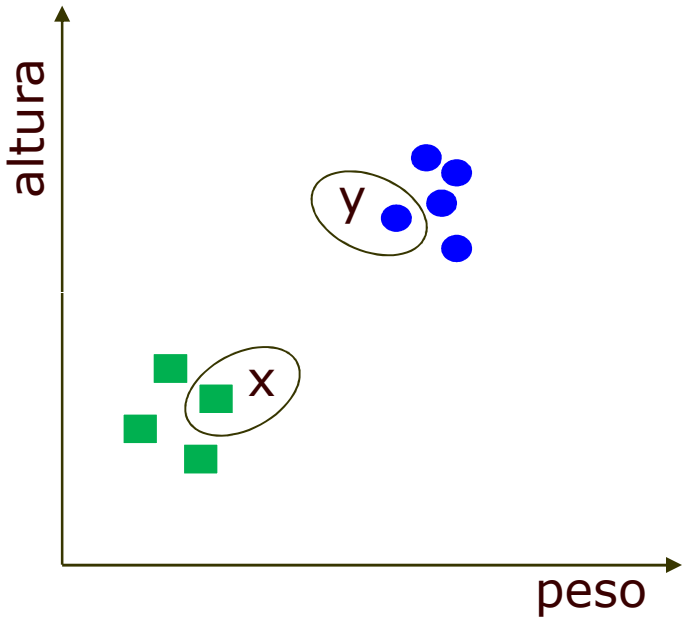
- **Problema:**
  - Como classificar um dado padrão?
  - Comparando com outros
- **k-NN (*k-Nearest Neighbors*)**
  - K-vizinhos mais próximos
  - Compara um padrão *X* de classe desconhecida com um conjunto de outros padrões cujas classes são conhecidas e infere a classe de *X* a partir dos mais semelhantes
    - Mais semelhantes = mais próximos = menor distância

NN ou 1-NN



- **Classifica-se um dado padrão associando a ele a classe do elemento de treinamento mais próximo (que tem a menor distância)**
- **Exemplo:**
  - X está mais próximo de um elemento da classe Jóquei, logo X será classificado como Jóquei
  - Y está mais próximo de um elemento da classe Halterofilista, logo Y será classificado como tal

NN ou 1-NN



- **e = [peso(kg); altura(m)]**
- **x = [70; 1,63] → ?**
- **y = [83; 1,77] → ?**
- **Conjunto de Treino:**
  - j_1 = [50; 1,60] → Jóquei
  - j_2 = [53; 1,65] → Jóquei
  - j_3 = [60; 1,58] → Jóquei
  - j_4 = [62; 1,62] → Jóquei
  - h_1 = [91; 1,75] → Halterofilista
  - h_2 = [102; 1,85] → Halterofilista
  - h_3 = [105; 1,82] → Halterofilista
  - h_4 = [103; 1,77] → Halterofilista
  - h_5 = [87; 1,73] → Halterofilista

Distância Euclidiana – D(a,b)

$$D(a,b) = \sqrt{(a_1-b_1)^2 + (a_2-b_2)^2 + \ldots + (a_n-b_n)^2}$$

Sabendo que:

$a = [a_1, a_2, \ldots, a_n]$;

$b = [b_1, b_2, \ldots, b_n]$;

$D(x, j_1) = 20$

$D(x, j_2) = 17^*$

$D(x, j_3) = 10$

$D(x, j_4) = 8$

$D(x, h_1) = 21$

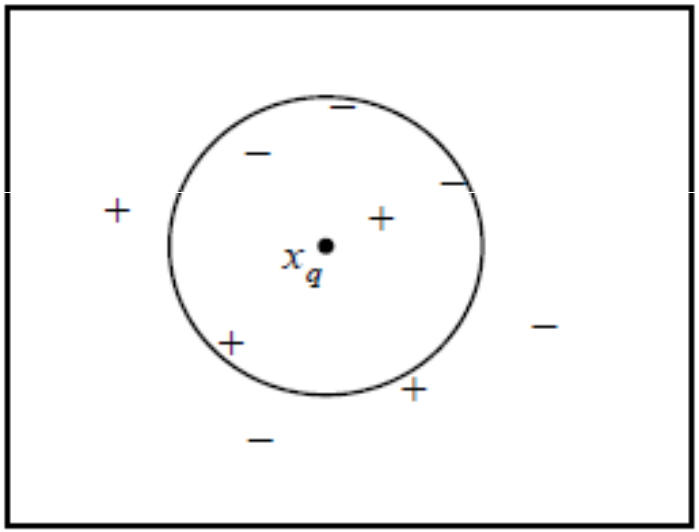
$D(x, h_2) = 32$

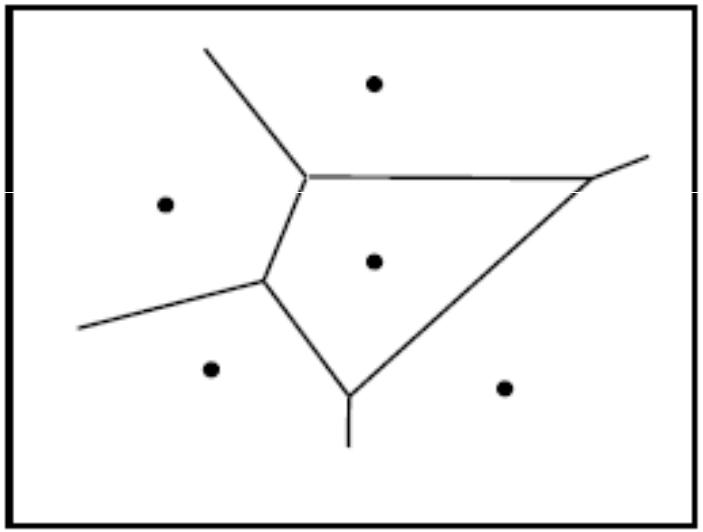
$D(x, h_3) = 35$

$D(x, h_4) = 33$

$D(x, h_5) = 17^*$

Diagrama de Voronoi





Distância Euclidiana Normalizada

$$D_n(a_l, a_k) = \sqrt{\sum_{i=1}^{n} \left(\frac{a_{li} - a_{ki}}{range_i} \right)^2}$$

Sendo:

- $a_j = [a_{j1}, a_{j2}, ..., a_{jn}]$
- $1 <= j <= M$
- M é o número de elementos no conjunto de treinamento
- $max_i = max(a_{ji})$
- $min_i = min(a_{ji})$
- $range_i = max_i - min_i$

Distância Euclidiana:

- $x = [70; 1,63]$
- $j_1 = [50; 1,60]$
- $D(x, j_1)^2 = (70-50)^2 + (1,63 – 1,60)^2$
- $D(x, j_1)^2 = 20^2 + 0,03^2$
- A altura tem influência desprezível no cálculo da distância

Centro de Informática U.F.P.E

Distância Euclidiana Normalizada

$$DN(a_l, a_k) = \sqrt{\sum_{i=1}^{n} \left(\frac{a_{li} - a_{ki}}{range_i} \right)^2}$$

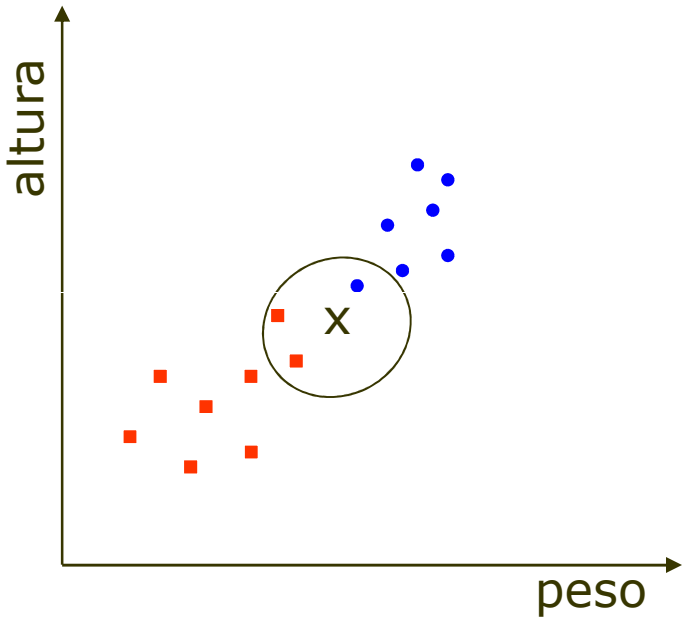
- $D(x, j_1) = 0{,}38$
- $D(x, j_2) = 0{,}32^*$
- $D(x, j_3) = 0{,}26$
- **$D(x, j_4) = 0{,}15$**
- $D(x, h_1) = 0{,}59$
- $D(x, h_2) = 1{,}00$
- $D(x, h_3) = 0{,}95$
- $D(x, h_4) = 0{,}79$
- $D(x, h_5) = 0{,}48^*$

$$D(a_l, a_k) = \sqrt{\sum_{i=1}^{n} (a_{li} - a_{ki})^2}$$

- $D(x, j_1) = 20$
- $D(x, j_2) = 17^*$
- $D(x, j_3) = 10$
- **$D(x, j_4) = 8$**
- $D(x, h_1) = 21$
- $D(x, h_2) = 32$
- $D(x, h_3) = 35$
- $D(x, h_4) = 33$
- $D(x, h_5) = 17^*$

Centro de Informática UFPE

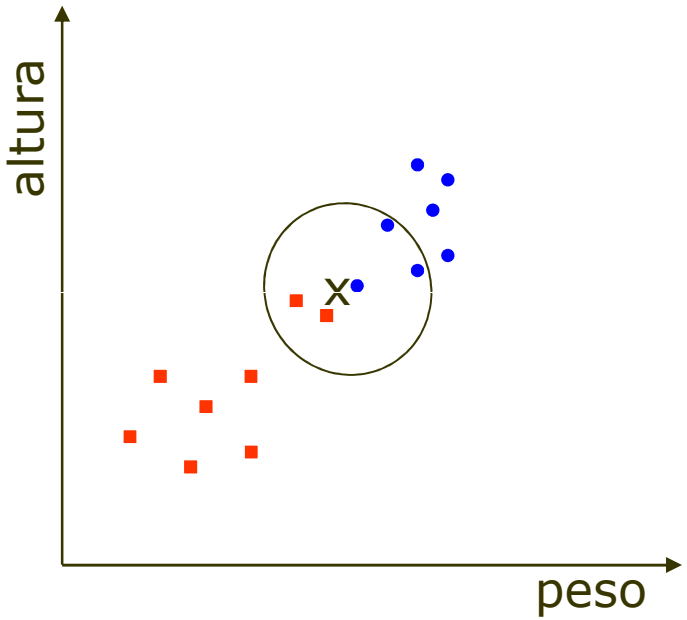
k-NN



- Jóquei
- Halterofilista

- **Classifica-se um dado padrão associando a ele a classe de maior freqüência entre os k vizinhos mais próximos**
- **Exemplo (3-NN):**
  - X está mais próximo de um elemento da classe Halterofilista
  - Mas outros dois elementos da classe Jóquei também estão entre o 3 vizinhos mais próximos
  - X será classificado como Jóquei

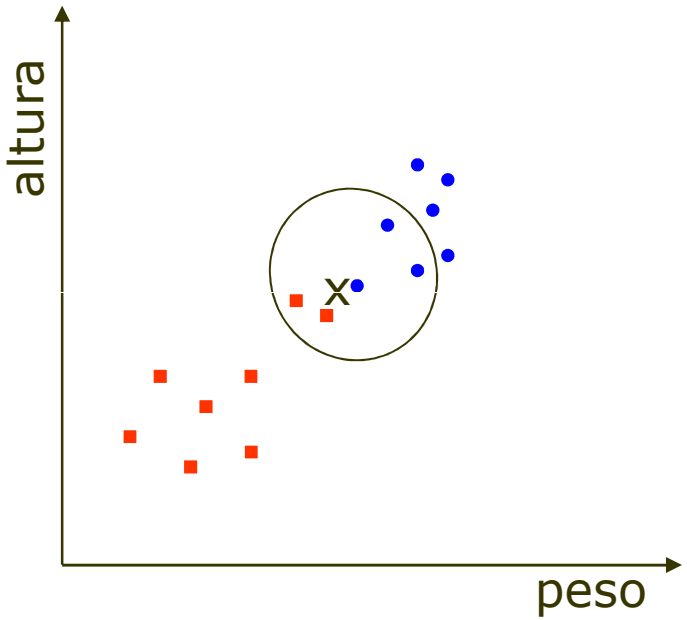
k-NN



- Jóquei
- Halterofilista

- **Classifica-se um dado padrão associando a ele a classe que apresentar a maior soma dos peso entre os k vizinhos mais próximos**
- **Exemplo (5-NN):**
  - X tem 3 vizinhos "Halterofilista" e 2 vizinhos "Jóquei"
  - X será classificado como Halterofilista

k-NN (peso pela distância)



- Jóquei
- Halterofilista

- Calculando o peso
  - Os vizinhos da classe “Jóquei” tem um peso maior.
  - Desta forma, X é associado à classe “Jóquei”

$$w_i = \frac{1}{d(x, e_i)}$$

Vantagens e Desvantagens

- **Vantagens**
  - Rápido treinamento
  - Capaz de aprender funções complexas
  - Não perde/desperdiça informação
- **Desvantagens**
  - Lento para realizar uma consulta
  - Facilmente enganado por um atributo irrelevante

Exemplo

- **Base de Dados *Iris***
  - 3 classes
  - 50 padrões por classe
  - 4 atributos numéricos por padrão
  - Fonte: UCI (http://archive.ics.uci.edu/ml/)

	SEM PESO	COM PESO
k = 1	95,80	95,80
k = 2	95,67	**95,80**
k = 3	95,33	**95,80**
k = 5	95,27	**95,53**
k = 6	**96,67**	95,53
k = 11	**95,53**	95,47
k = 16	**96,27**	95,47
k = 21	95,13	**95,60**
k = 31	95,13	**95,40**

Referências

- Tom Mitchell. **Machine Learning**. McGraw-Hill. 1997.
- S. Theodoridis and K. Koutroumbas. ***Pattern Recognition***. Academic Press. 2006.
- Christopher M. Bishop. **Pattern Recognition and Machine Learning**. Springer. 2006
- Richard O. Duda, Peter E. Hart and David G. Stork. **Pattern Classification**. Wiley-Interscience. 2000