Pontifícia Universidade Católica
do Rio de Janeiro

INF 1771 – Inteligência Artificial

Aula 13 – K-Nearest Neighbor (KNN)

2016.1

Prof. Augusto Baffa
<abaffa@inf.puc-rio.br>

Formas de Aprendizado

- **Aprendizado Supervisionado**
  - Árvores de Decisão.
  - **K-Nearest Neighbor (KNN).**
  - Support Vector Machines (SVM).
  - Redes Neurais.

- Aprendizado Não Supervisionado

- Aprendizado Por Reforço

Aprendizado Supervisionado

- Observa-se alguns pares de **exemplos de entrada e saída**, de forma a aprender uma **função que mapeia a entrada para a saída**.

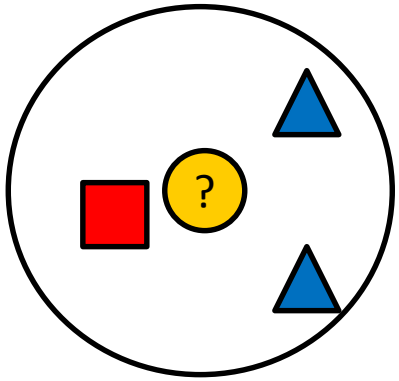
- Damos ao sistema a **resposta correta** durante o processo de treinamento.

- É eficiente pois o sistema pode trabalhar diretamente com informações corretas.

K-Nearest Neighbor

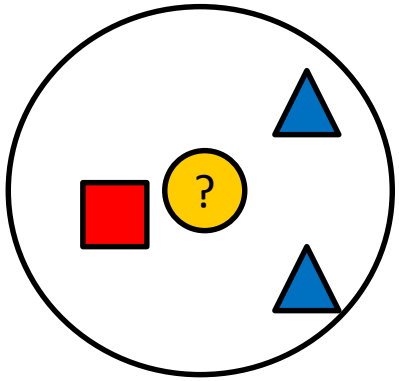
- É um dos algoritmos de classificação clássicos e bem **simples**.

- Usado para classificar objetos com base em **exemplos de treinamento** que estão mais próximos no espaço de características.



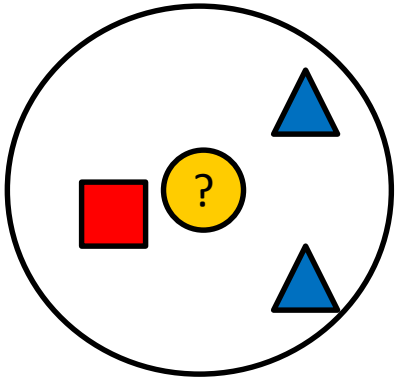
K-Nearest Neighbor

- Significado de k:
  - Classificar x atribuindo a ele o rótulo representado mais frequentemente dentre as k amostras mais próximas.
  - Contagem de votos.



K-Nearest Neighbor

- Para utilizar o KNN é necessário:
  - **(1)** Um conjunto de exemplos de treinamento.
  - **(2)** Definir uma métrica para calcular a distância entre os exemplos de treinamento.
  - **(3)** Definir o valor de K (o número de vizinhos mais próximos que serão considerados pelo algoritmo).



K-Nearest Neighbor

- Calculando a distancia entre dois pontos:
  - Existem varias formas diferentes de calcular essa distancia. A mais simples é a distância euclidiana:

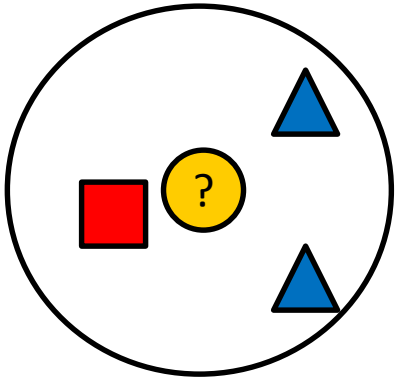
$$d(p,q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

  - É importante normalizar os dados.

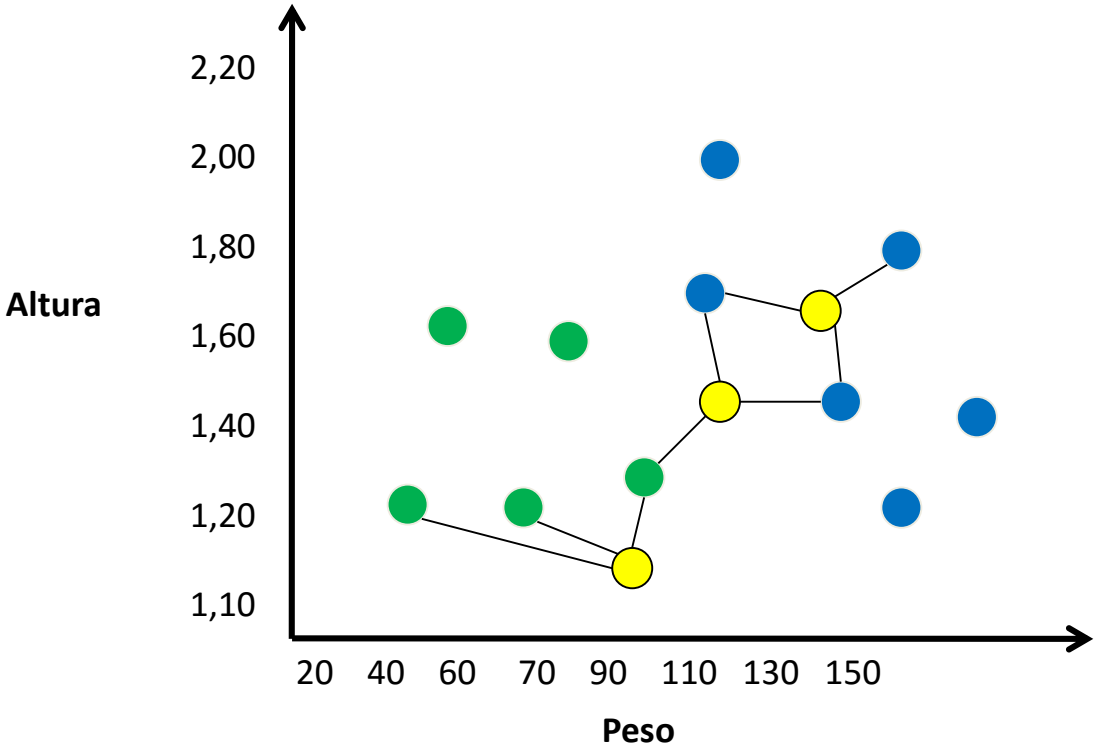
  - Outras formas de mediar a distancia:
    - Distância de Mahalanobis.
    - Distância de Minkowsky.
    - Hamming Distance.
    - ...

K-Nearest Neighbor

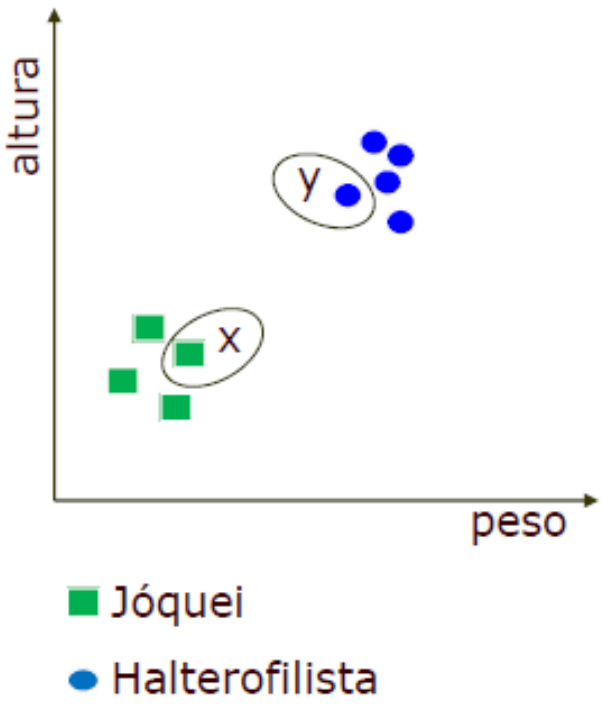
- Classificar um exemplo desconhecido com o algoritmo KNN consiste em:
  - **(1)** Calcular a distância entre o exemplo desconhecido e o outros exemplos do conjunto de treinamento.
  - **(2)** Identificar os K vizinhos mais próximos.
  - **(3)** Utilizar o rotulo da classe dos vizinhos mais próximos para determinar o rótulo de classe do exemplo desconhecido (votação majoritária).



Espaço de Características



Exemplo Atributo Contínuo



- e = [peso(kg); altura(m)]
- x = [70; 1,63] = ?
- y = [83; 1,77] = ?

- Conjunto de Treino:
  - j1 = [50; 1,60] = Jóquei
  - j2 = [53; 1,65] = Jóquei
  - j3 = [60; 1,58] = Jóquei
  - j4 = [62; 1,62] = Jóquei
  - h1 = [91; 1,75] = Halterofilista
  - h2 = [102; 1,85] = Halterofilista
  - h3 = [105; 1,82] = Halterofilista
  - h4 = [103; 1,77] = Halterofilista
  - h5 = [87; 1,73] = Halterofilista

Exemplo Atributo Contínuo

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

- Sabendo que:
  - $a = [a_1, a_2, \ldots, a_n]$;
  - $b = [b_1, b_2, \ldots, b_n]$;

d(x,j1) = 20
d(x,j2) = 17 *
d(x,j3) = 10
d(x,j4) = 8
d(x,h1) = 21
d(x,h2) = 32
d(x,h3) = 35
d(x,h4) = 33
d(x,h5) = 17*

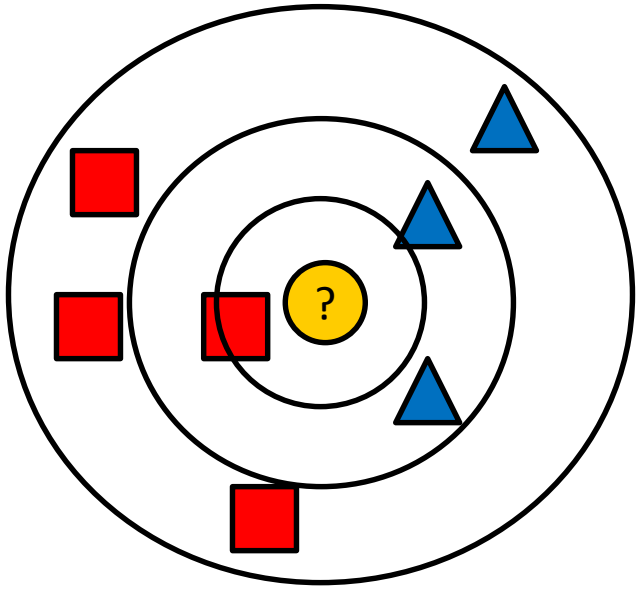
d(Y,j1) = 33
d(Y,j2) = 30
d(Y,j3) = 23
d(Y,j4) = 21
d(Y,h1) = 8
d(Y,h2) = 19
d(Y,h3) = 22
d(Y,h4) = 20
d(Y,h5) = 4

K-Nearest Neighbor

- Determinando a classe do exemplo desconhecido a partir da lista de vizinhos mais próximos:
  - Considera-se o voto majoritário entre os rótulos de classe dos K vizinhos mais próximos.
  - Como escolher o valor de K?

K-Nearest Neighbor

- **K = 1**
  - Pertence a classe de quadrados.
- **K = 3**
  - Pertence a classe de triângulos.
- **K = 7**
  - Pertence a classe de quadrados.



- Como determinar o melhor valor de K (=número de vizinhos) ?
  - Repetindo-se os experimentos.

K-Nearest Neighbor

- Como escolher o valor de K?
  - Se K for muito pequeno, a classificação fica sensível a pontos de ruído.
  - Se k é muito grande, a vizinhança pode incluir elementos de outras classes.
- Além disso, é necessário sempre escolher um valor ímpar para K, assim se evita empates na votação.

K-Nearest Neighbor

- A **precisão** da classificação utilizando o algoritmo KNN depende fortemente do modelo de dados.

- Na maioria das vezes os atributos precisam ser **normalizados** para evitar que as medidas de distância sejam dominado por um único atributo. Exemplos:
  - Altura de uma pessoa pode variar de 1,20 a 2,10.
  - Peso de uma pessoa pode variar de 40 kg a 150 kg.
  - O salário de uma pessoa podem variar de R$ 800 a R$ 20.000.

Algumas questões

- Como calcular a distância entre duas tuplas ?
  - Para atributos contínuos : distância Euclidiana
  - Para atributos categóricos?
    Se $x_i = y_i$ então $x_i - y_i = 0$
    Se x_i e y_i são distintos: $x_i - y_i = 1$
- Como lidar com valores incompletos (ausentes) ao calcular a distância entre duas tuplas X e Y ?
  - Se x_i e y_i são ausentes: $x_i - y_i = 1$
  - Se x_i é ausente e y_i não: $x_i - y_i = \max \{ |1 - y_i|, |0 - y_i| \}$

Exemplo Atributo Categórico

Compra-computador

ID	IDADE	RENDA	ESTUDANTE	CREDITO	CLASSE
1	≤ 30	Alta	Não	Bom	Não
2	≤ 30	Alta	Sim	Bom	Não
3	31...40	Alta	Não	Bom	Sim
4	> 40	Média	Não	Bom	Sim
5	> 40	Baixa	Sim	Bom	Sim
6	> 40	Baixa	Sim	Excelente	Não
7	31...40	Baixa	Sim	Excelente	Sim
8	≤ 30	Média	Não	Bom	Não
9	≤ 30	Baixa	Sim	Bom	Sim
10	> 40	Média	Sim	Bom	Sim
11	≤ 30	Média	Sim	Excelente	Sim
12	31...40	Média	Não	Excelente	Sim
13	31...40	Alta	Sim	Bom	Sim
14	> 40	Média	Não	Excelente	Não

X = (≤ 30, Média,Sim,Bom)

Exemplo Atributo Categórico

Distância	VALOR
d(X,1)	1,41
d(X,2)	1
d(X,3)	1,73
d(X,4)	1,41
d(X,5)	1,41
d(X,6)	1,73
d(X,7)	1,73
d(X,8)	1
d(X,9)	1
d(X,10)	1
d(X,11)	1
d(X,12)	1,73
d(X,13)	1,41
d(X,14)	1,73

Exemplo Atributo Categórico

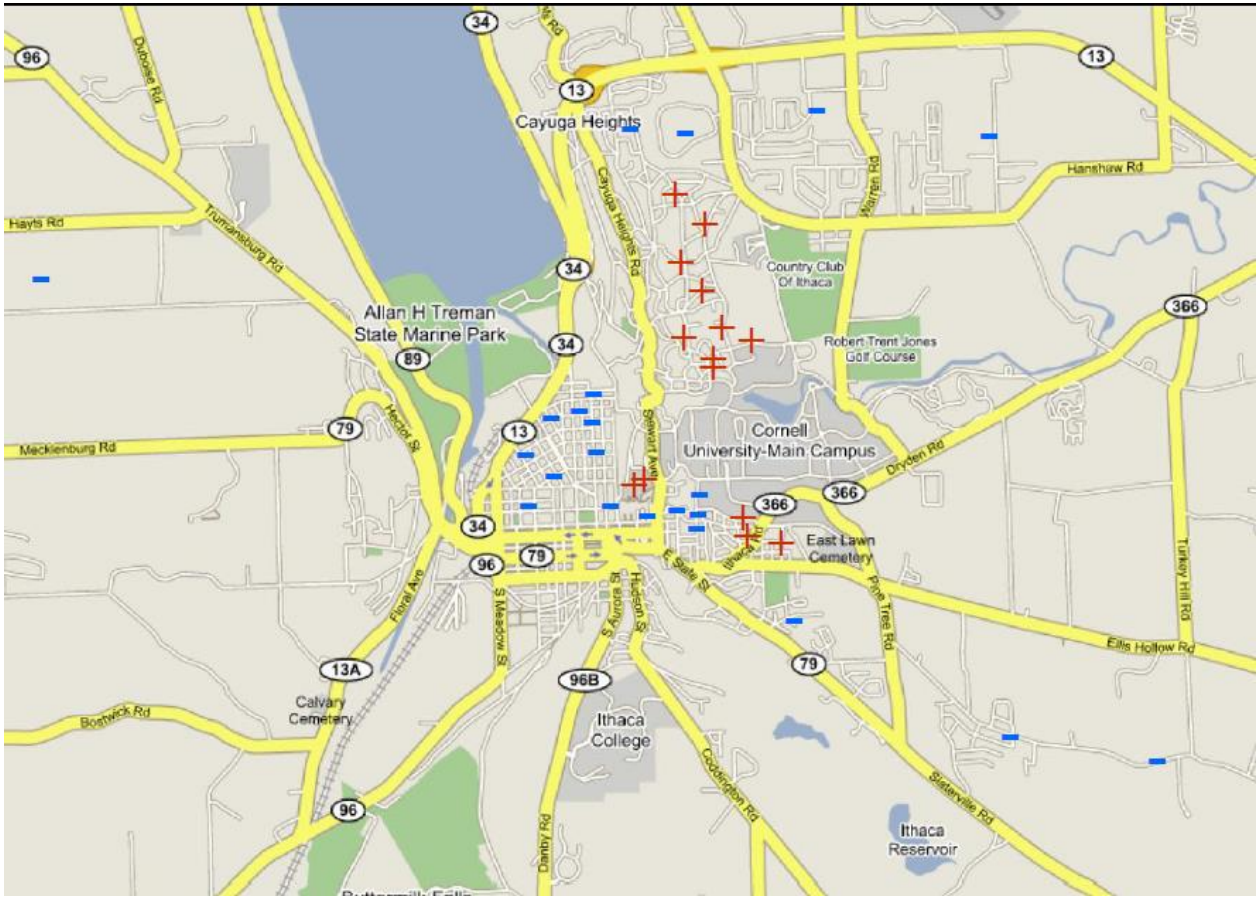
- K = 5

- Os 5 vizinhos mais próximos são:
  - **X1 = (≤ 30 Alta Sim Bom) Classe = Não**
  - **X2 = (≤ 30 Média Não Bom) Classe = Não**
  - **X3 = (≤ 30 Baixa Sim Bom) Classe = Sim**
  - **X4 = (> 40 Média Sim Bom) Classe = Sim**
  - **X5 = (≤ 30 Média Sim Exc.) Clase = Sim**

- **Logo, X é classificada na classe = Sim**

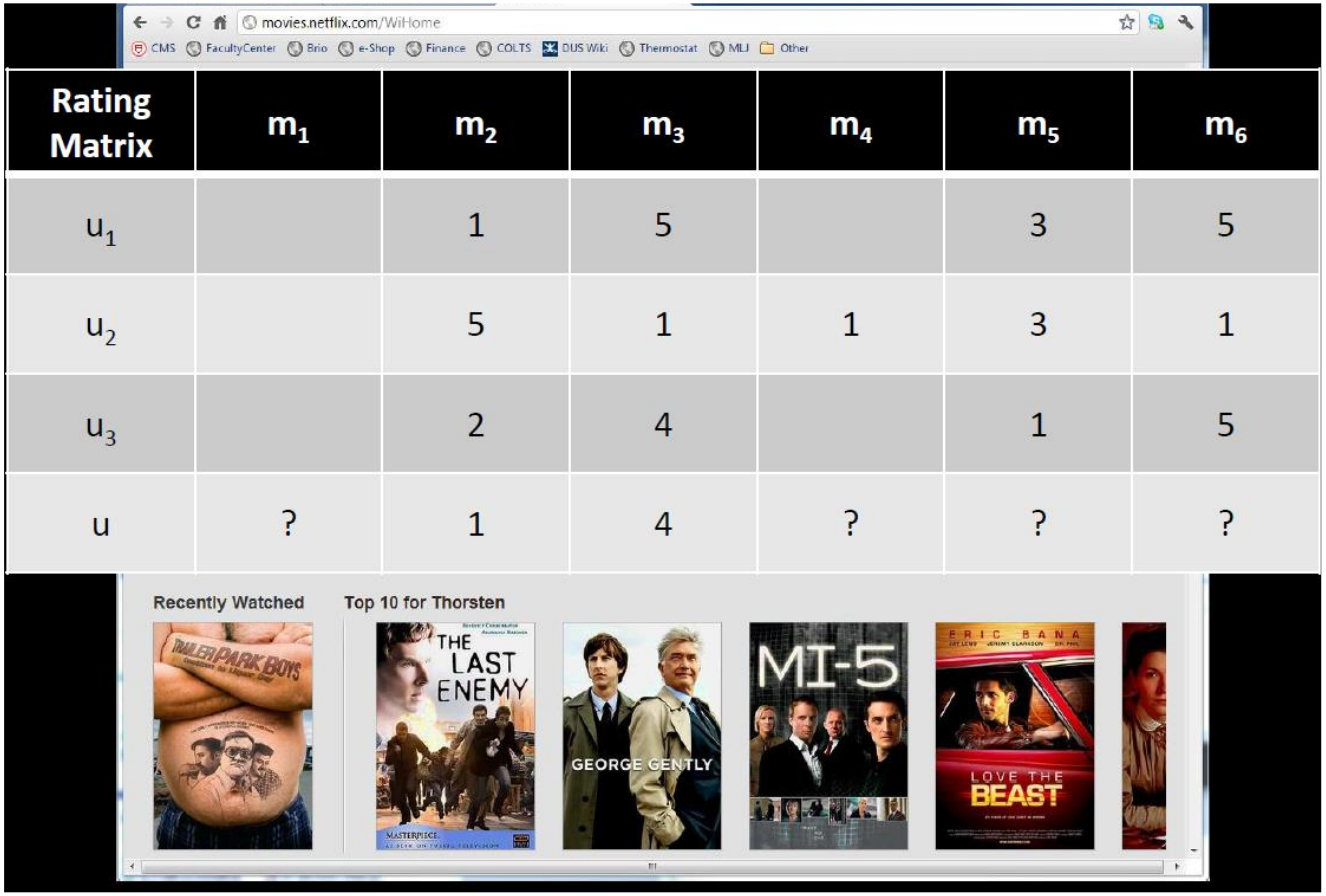
Exemplos de Aplicações

- Preços de Empreendimentos Imobiliários



Exemplos de Aplicações

- Sistemas de Sugestão baseados em Rating



Rating Matrix	m_1	m_2	m_3	m_4	m_5	m_6
u_1		1	5		3	5
u_2		5	1	1	3	1
u_3		2	4		1	5
u	?	1	4	?	?	?

K-Nearest Neighbor

- **Vantagens:**
  - Técnica simples e facilmente implementada.
  - Não constrói um modelo de classificação.
  - Bastante flexível.
  - Em alguns casos apresenta ótimos resultados.
- **Desvantagens:**
  - Classificar um exemplo desconhecido pode ser um processo computacionalmente complexo. Requer um calculo de distancia para cada exemplo de treinamento.
    - Pode consumir muito tempo quando o conjunto de treinamento é muito grande.
  - A precisão da classificação pode ser severamente degradada pela presença de ruído ou características irrelevantes.
  - KNN faz predição baseando-se em informações locais à tupla sendo classificada.
  - Arvores de decisão, redes neurais e bayesianas encontram modelo global que se leva em conta todo o banco de dados de treinamento.
- No Weka, a implementação do K-NN chama-se IBk.

Leitura Complementar

- Mitchell, T. **Machine Learning**, McGraw–Hill Science/Engineering/Math, 1997.

- Duda, R.; Hart, P.; Stork, D. **Pattern Classification**, John Wiley & Sons, 2000

- Bishop, C. **Pattern Recognition and Machine Learning**. Springer. 2006