INF623

2024/1

UFV

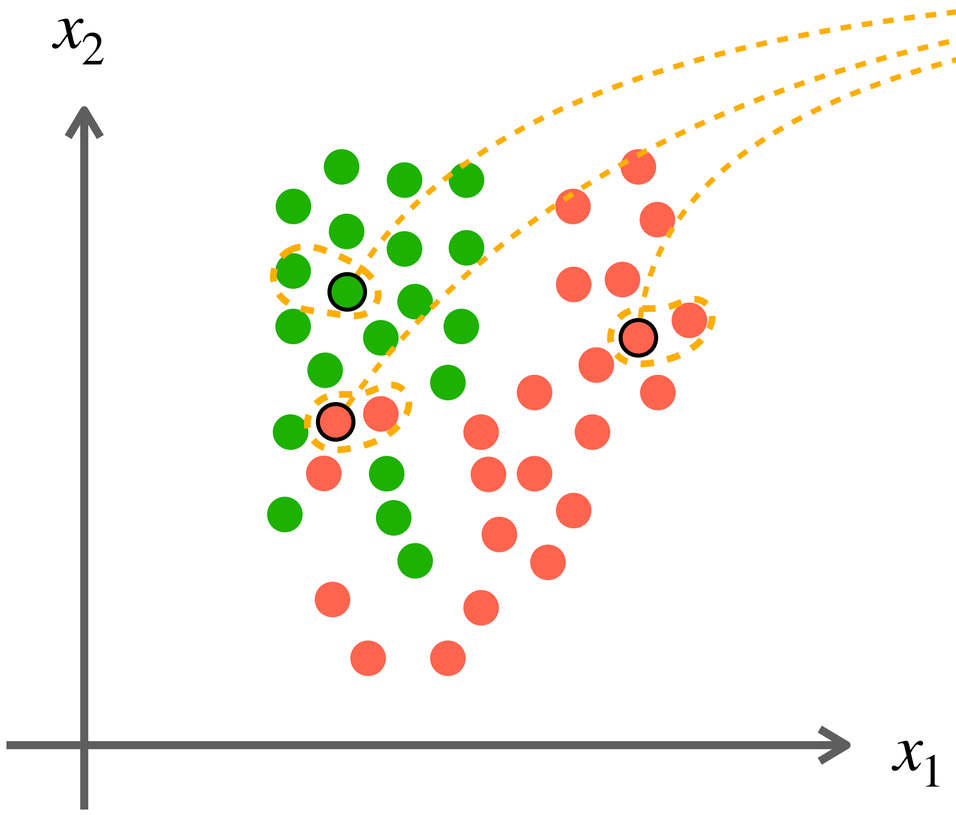
Inteligência Artificial

A25: Aprendizado supervisionado II

Plano de aula

- k-Nearest Neighbors
  - Hipótese
  - Métricas de distância
  - Convergência
  - A maldição da dimensionalidade
  - Implementação

k-Nearest Neighbors (k-NN)



Qual é a classe desse vetor de caratetísticas $\mathbf{x}$?

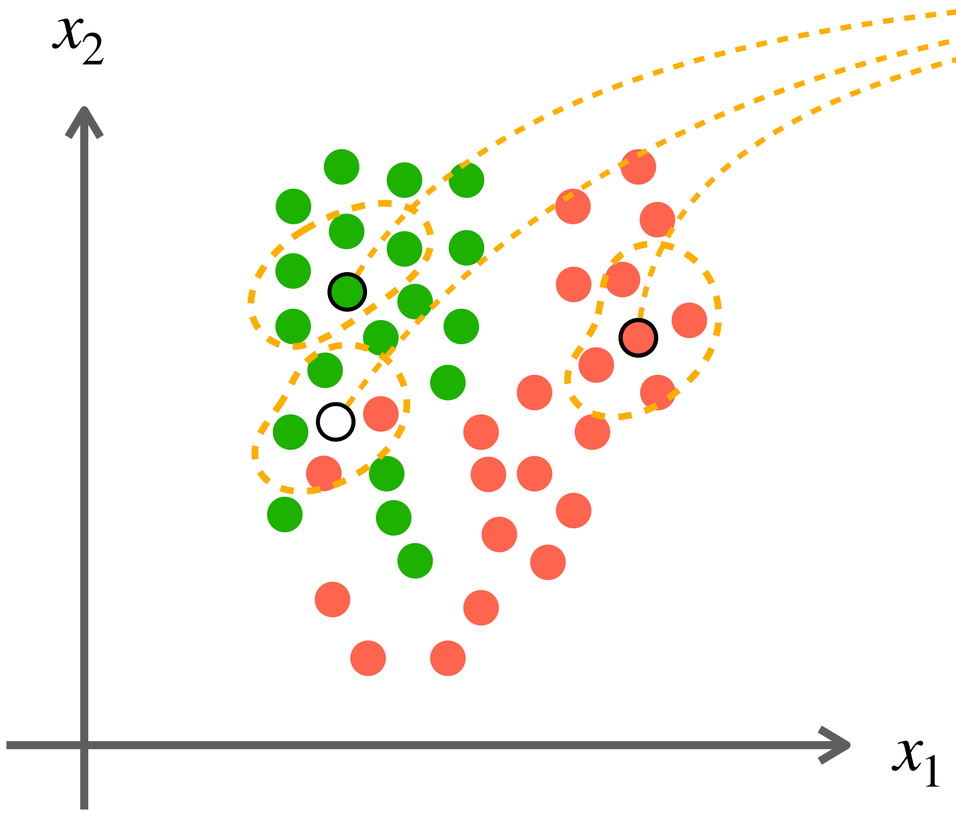
Hipótese

- O k-NN assume que vetores $\mathbf{x}$ que estão próximos no espaço possuem o mesmo rótulo y

Exemplos

- k = 1

k-Nearest Neighbors (k-NN)



Qual é a classe desse vetor de caratetísticas $\mathbf{x}$?

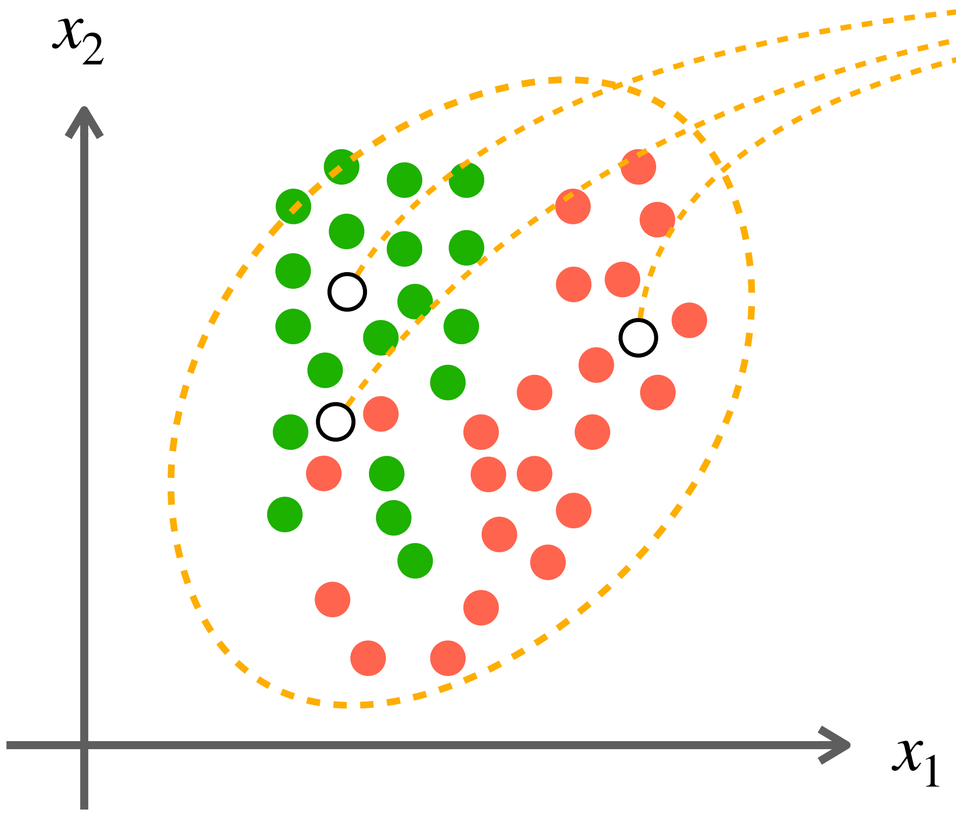
Hipótese

- O k-NN assume que vetores $\mathbf{x}$ que estão próximos no espaço possuem o mesmo rótulo y

Exemplos

- k = 1
- k = 4

k-Nearest Neighbors (k-NN)



Qual é a classe desse vetor de caratetísticas $\mathbf{x}$?

Hipótese

- O k-NN assume que vetores $\mathbf{x}$ que estão próximos no espaço possuem o mesmo rótulo y

Exemplos

- k = 1
- k = 4
- ...
- k = m (número de exemplos)

Implementação do kNN

```
def kNN(z, k, D, dist):
1. dists = []
2. for (x,y) in D:
3.   d = dist(x, z)
4.   dists.append({'distance': dist, 'class': y})
5. sorted_dists = sort(dists)
6. k_nearest_neighbors = sorted_dists[1:k]
7. predicted_class = majority_vote(k_nearest_neighbors)
```

Métricas de distâncias para o kNN

```
def kNN(z, k, D, dist):
```

O k-NN depende fundamentalmente de uma **métrica de distância**.

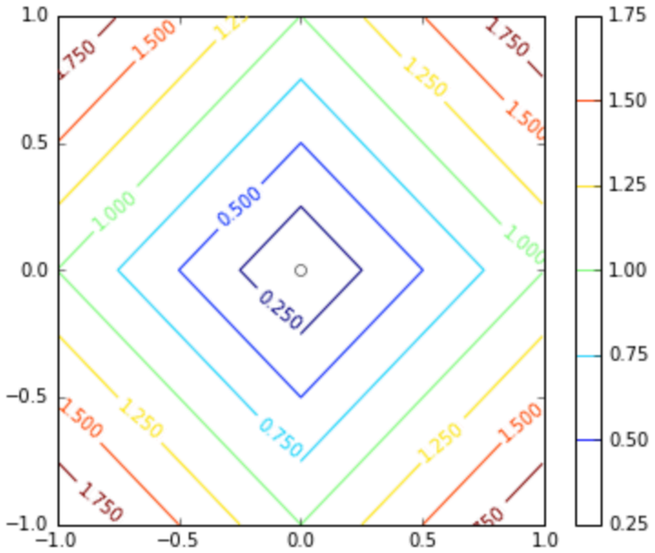
- Quanto melhor essa métrica refletir a similaridade do rótulo, melhor será a classificação.
- A escolha mais comum é a distância de Minkowski:

$$dist(\mathbf{x}, \mathbf{z}) = (\sum_{i=1}^{d} |x_i - z_i|^p)^{\frac{1}{p}}$$

- Essa métrica é bastante genérica e representa diversas métricas de distância conhecidas:
  - $p = 1$?
  - $p = 2$?
  - $p \to \infty$?

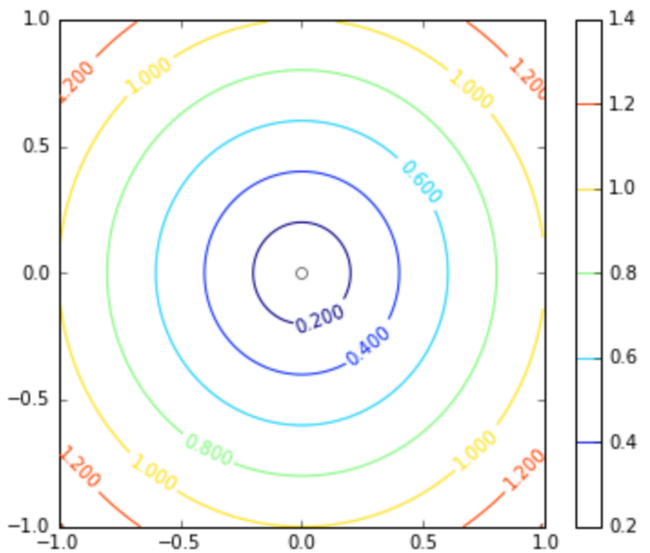
Métricas de distâncias para o kNN

Distância de Minkowski: $dist(\mathbf{x}, \mathbf{z}) = (\sum_{i=1}^{d} |x_i - z_i|^p)^{\frac{1}{p}}$



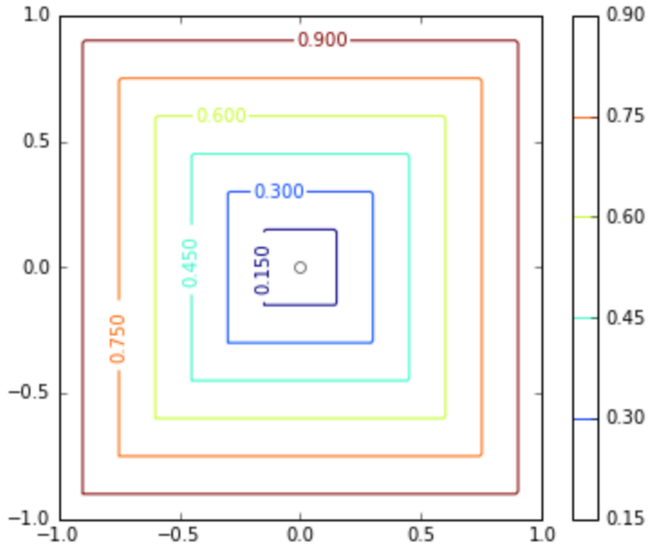
Distância de manhattan $(p = 1)$

$$dist(\mathbf{x}, \mathbf{z}) = \sum_{i=1}^{d} |x_i - z_i|$$



Distância euclidiana $(p = 2)$

$$dist(\mathbf{x}, \mathbf{z}) = \sqrt{\sum_{i=1}^{d} |x_i - z_i|^2}$$



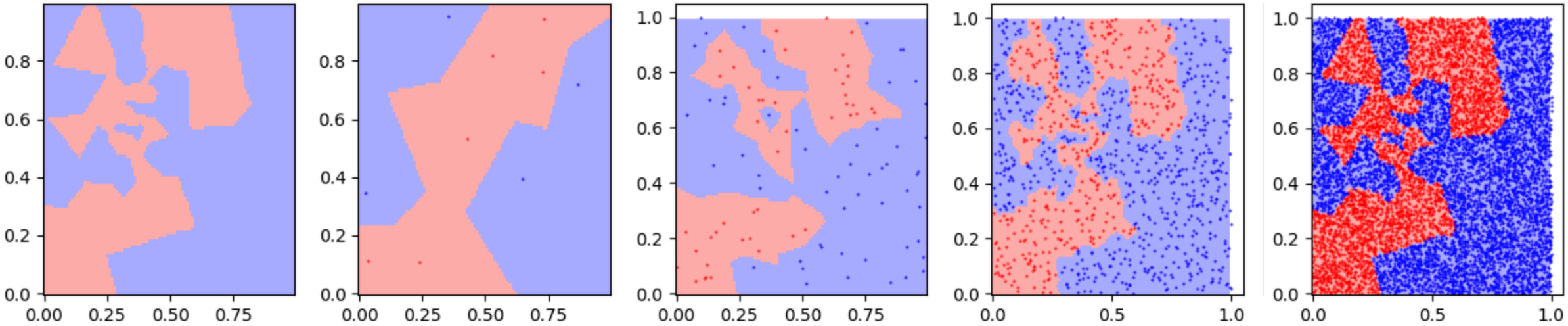
Distância chebyshev $(p \to \infty)$

$$dist(\mathbf{x}, \mathbf{z}) = max_d |x_d - z_d|$$

UFV

Convergência do k-NN

O k-NN converge para a os rótulos reais quando $m \to \infty$



Rótulos reais | $m = 10,\ erro = 0{,}28$ | $m = 100,\ erro = 0{,}14$ | $m = 1000,\ erro = 0{,}05$ | $m = 10000,\ erro = 0{,}02$

Cover, Thomas, and, Hart, Peter. Nearest neighbor pattern classification. IEEE Transactions on Information Theory. 1967, 13(1): 21-27

Demonstração do k-NN

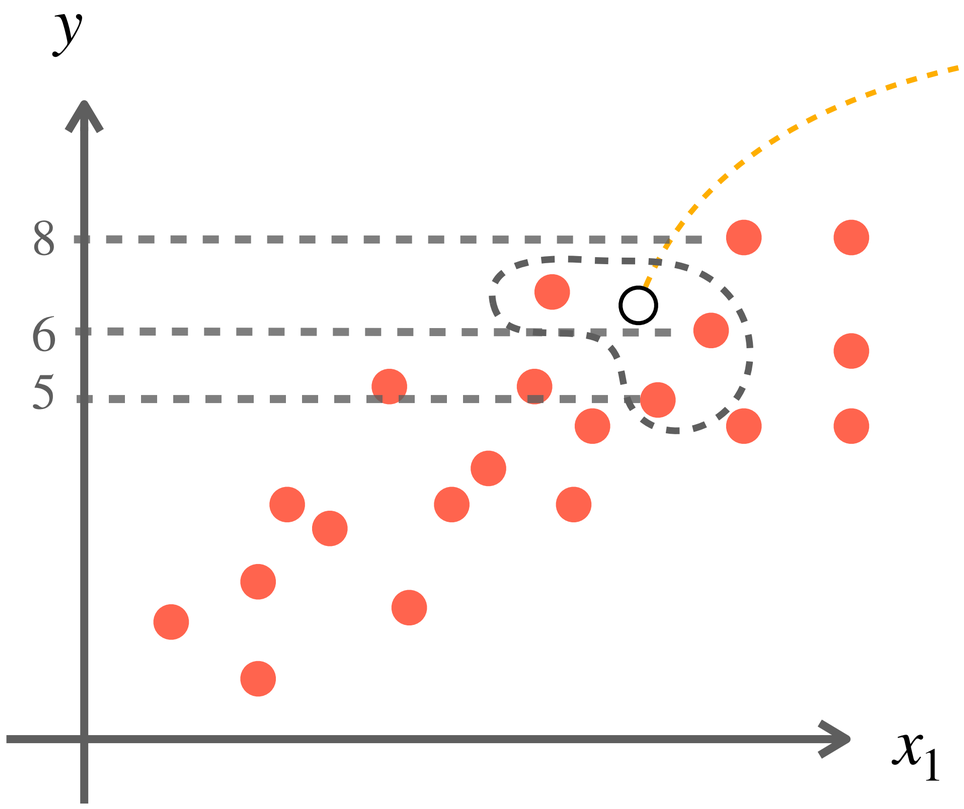
Abrir o seguinte colab:

https://colab.research.google.com/drive/1EDFQ4FTYVTx2XW4iz6FHcO_GCYy2aoUg?usp=sharing

Visualizar:

- Métricas de distância
- Classificação com k-NN
- Convergência do k-NN quando o múmero de exemplos cresce

kNN para regressão



Qual é o valor y desse vetor de caratetísticas $\mathbf{x}$?

Média dos rótulos y_i

$$y = \frac{1}{k}\sum_{i}^{k} y_i$$

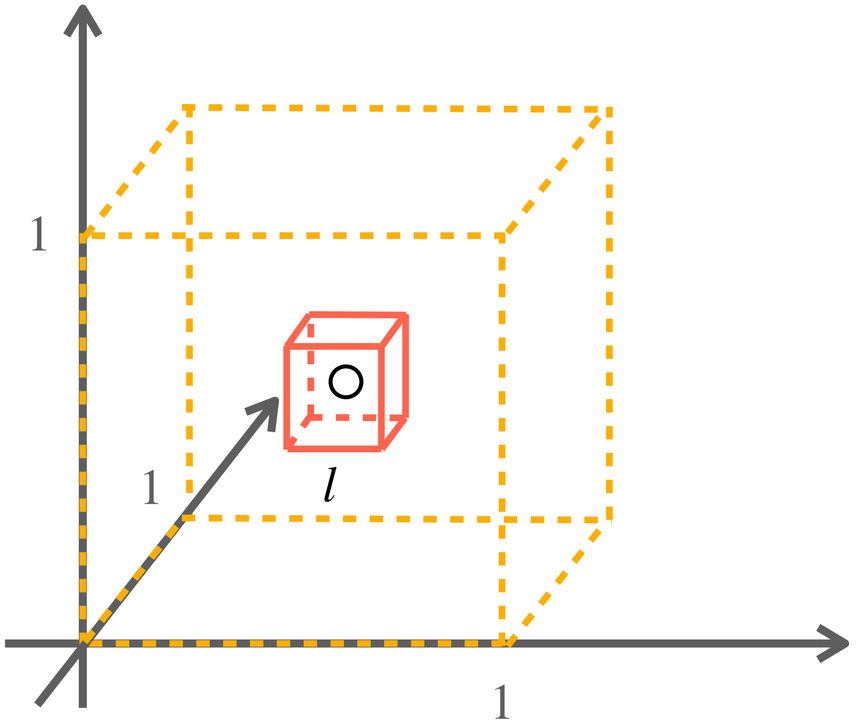
Exemplo

$$y = \frac{8+6+5}{3} = 6.333$$

O k-NN é um algoritmo de aprendizado supervisionado não-paramétrico para classificação e regressão!

A maldição da dimensionalidade

A **maldição da dimensionalidade** no contexto do k-NN mostra que a distância euclidiana é inútil em dimensões altas porque todos os vetores $\mathbf{x}$ são quase equidistantes do vetor de teste $\mathbf{z}$:



- Considere que todos os exemplos de treinamento foram amostrados uniformemente de um **cubo de lado 1**
- Seja l o comprimento do **menor cubo** que contém todos os k vizinhos mais próximos do ponto de teste $\mathbf{z}$

$$l^d \approx \frac{k}{m} \rightarrow l \approx (\frac{k}{m})^{\frac{1}{d}}$$

- Se $m = 1000$, qual o comprimento l?

d	l
2	0,1
10	0,63
100	0,955
1000	0,9954

Próxima aula

A26: Aprendizado supervisionado III

Árvores de decisão