Aprendizado de Máquina

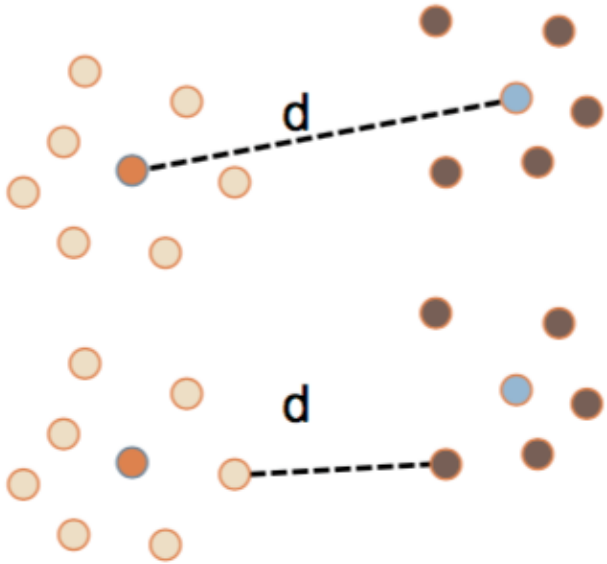
Classificação baseada em Instâncias

Luiz Eduardo S. Oliveira

Universidade Federal do Paraná
Departamento de Informática
http://web.inf.ufpr.br/luizoliveira

Medidas de Distância

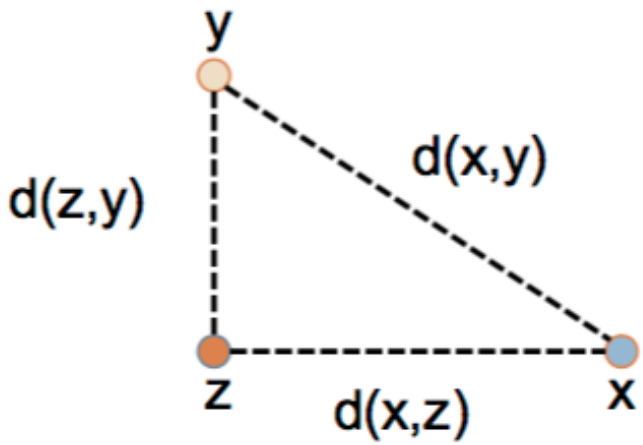
- Entende-se por distância a media de separação de dois objetos
- Em aprendizagem de máquina e reconhecimento de padrões, a distância indica a dissimilaridade ou afastamento entre dois atributos ou vetores de atributos.



Métrica

Formalização do conceito de distância

- Um espaço onde exista uma métrica definida é chamado um espaço métrico
- Para uma função ser considerada uma distância, ou metrica, entre dois vetores de atributos, ela deve seguir alguns aximomas
  - $d(x, y) = d(y, x)$
  - $d(x, y) \geqslant 0$
  - $d(x, x) = 0$
- Além dessas três propriedades, também valem
  - $d(x, y) = 0$, sse $x = y$
  - $d(x, x) \leqslant d(x, z) + d(z, y)$ (desigualdade do triângulo)



Métrica de Minkowski

Métrica geral para padrões d-dimensionais.

$$L_k(a, b) = (\sum_{i=1}^{d} |a_i - b_i|^k)^{\frac{1}{k}}$$

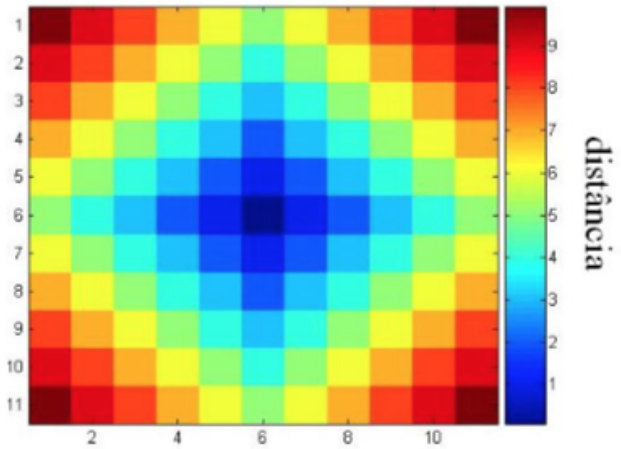
- Também conhecida como norma L_k
- As distâncias Euclidiana e Manhattan (city block) podem ser calculadas usando $k = 2$ e $k = 1$, respectivamente.

Distância Manhattan (city-block)

- Dado dois vetores X e Y, esta distância é dada pelo somatório dos módulos das diferenças

$$d(X, Y) = \sum_{i=1} |X_i - Y_i|$$

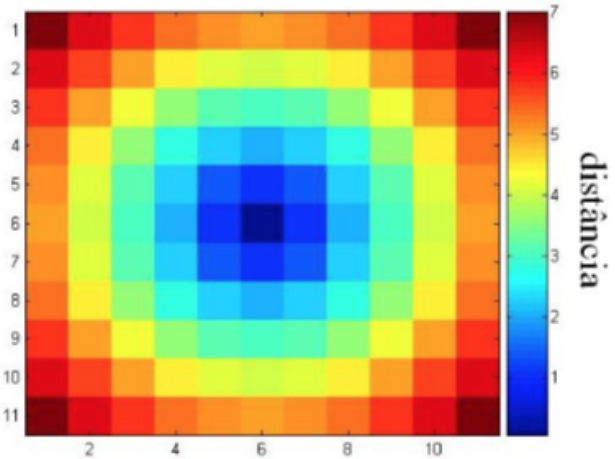
- Trata-se de uma distância que depende da rotação do sistema de coordenadas, mas não depende de sua reflexão em torno de um eixo ou suas translações.



Distância Euclidiana

- Trata-se da distância mais comum entre dois pontos (aquela media com uma régua)
- A distância Euclidiana é invariante a
  - rotação do sistema de coordenadas
  - a sua reflexão em trono de um eixo
  - translações
- É dada pela raiz quadrada do quadrado das diferenças dos vetores X e Y

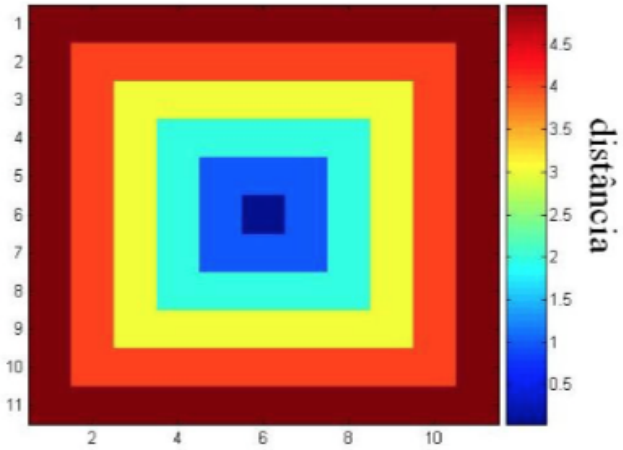
$$d(X, Y) = (\sum_{i=1} |X_i - Y_i|^2)^{\frac{1}{2}}$$



Distância Chebyshev (Chessboard)

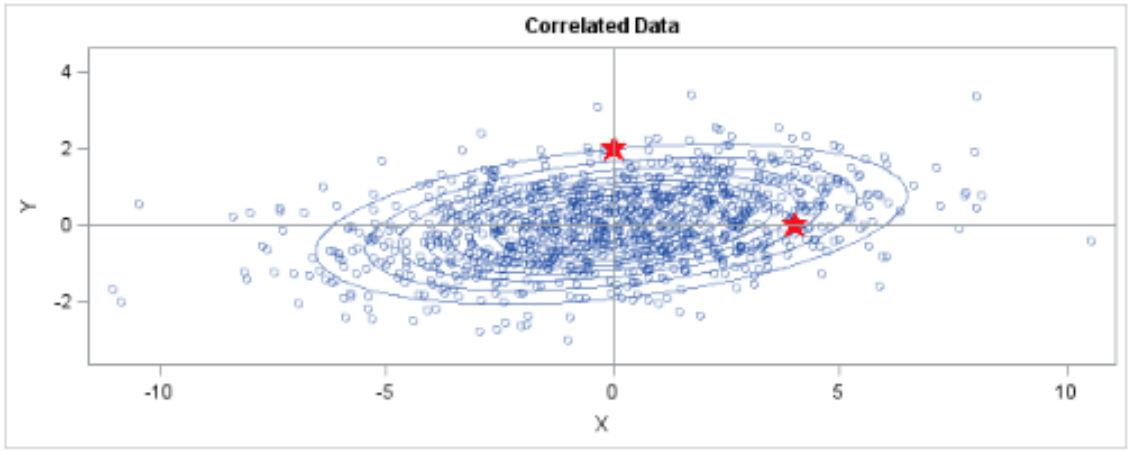
- Assemelha-se a distância Manhattan.
- Considera o valor máximo dos módulos das diferenças dos pontos em respectivas posições

$$d(X, Y) = \max(|X_1 - Y_1|, |X_2 - Y_2|, \ldots, |X_n - Y_n|)$$



Distância de Mahalanobis

- As distâncias Euclidiana assume que dos dados tem uma distribuição Gaussiana.
- Entretanto, em alguns casos os dados podem ter uma outra distribuição.
- Na figura abaixo, os dois pontos têm a mesma distância Euclidiana para o centro da elipse.



- Entretanto, um deles é claramente "mais diferente" da população do que o outro.

Distância de Mahalanobis

- A distância de Mahalanobis leva em consideração a variância de cada atributo, bem como a covariância entre eles.
- Para isso, é necessário normalizar os dados (diminuindo do vetor médio), calcular a matriz de covariância e a sua inversa.
- A matriz de covariância ($\sum$) é uma medida de como as variáveis são dispersas em torno da média (elementos da diagonal) e também da variância com as outras variáveis (fora da diagonal).
- A inversa da matriz de covariância ($\sum^{-1}$), também conhecida como matriz de precisão, contém informação sobre a correlação parcial entre as variáveis

Distância de Mahalanobis

- A distância de Mahalanobis é dada por

$$d(X, Y) = [(X - Y)^T A(X - Y)]^{\frac{1}{2}}$$

- Se $A = \sum^{-1}$, temos a distância de Mahalanobis
- Se $A = I$ (matriz identidade), temos a distância Euclidiana

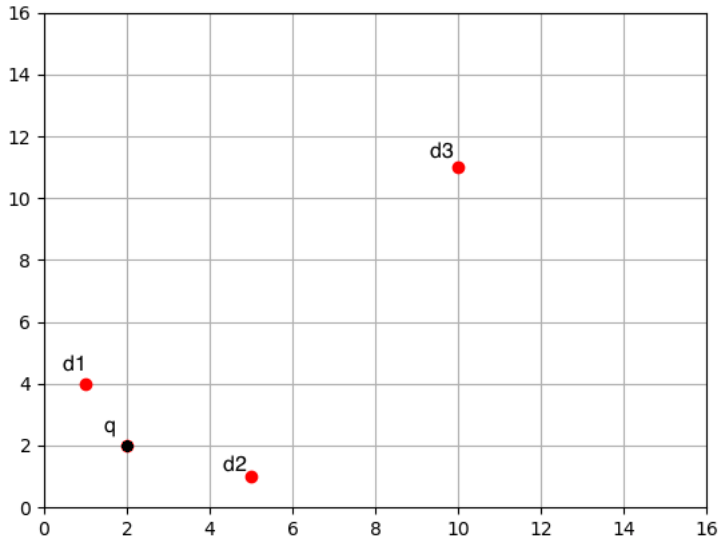
Distância Cosseno (Similaridade Cosseno)

- Bastante utilizada para calcular similaridade.
- Considera a direção do vetor e não a sua magnitude.
- É dado produto escalar dos vetores dividido pela magnitude dos mesmos

$$Cos(\theta) = \frac{A \cdot B}{||A||||B||} = \frac{\sum A_i B_i}{\sqrt{\sum A_i^2}\sqrt{\sum B_i^2}}$$

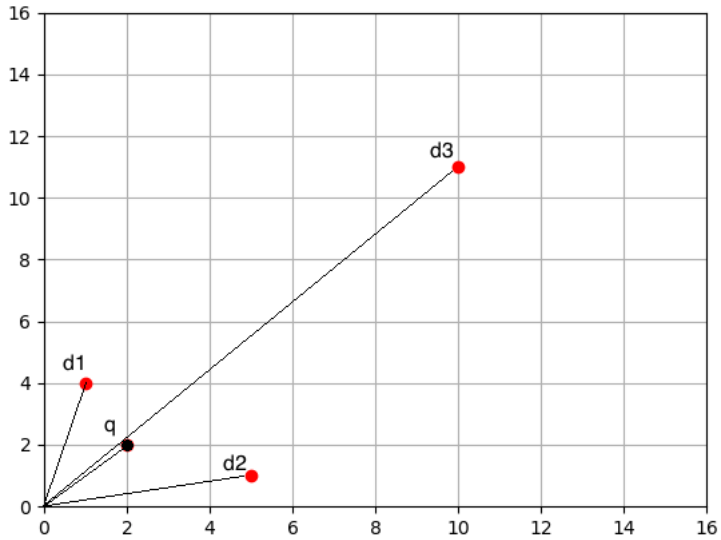
Distância Cosseno (Similaridade Cosseno)

- Considere os seguintes vetores de características, os quais representam a frequência de duas palavras (futebol e eleições)
- Por exemplo, no documento 1 (d1) a palavra futebol aparece quatro vezes e eleições apenas uma. Já o documento 2 (d2) parece ser diferente de d1 pois o termo eleições aparece cinco vezes e futebol apenas uma.
- Já d3 mistura futebol e politica pois os dois termos aparecem 10 vezes cada.
- O documento q é mais similar a qual dos documentos?



Distância Cosseno (Similaridade Cosseno)

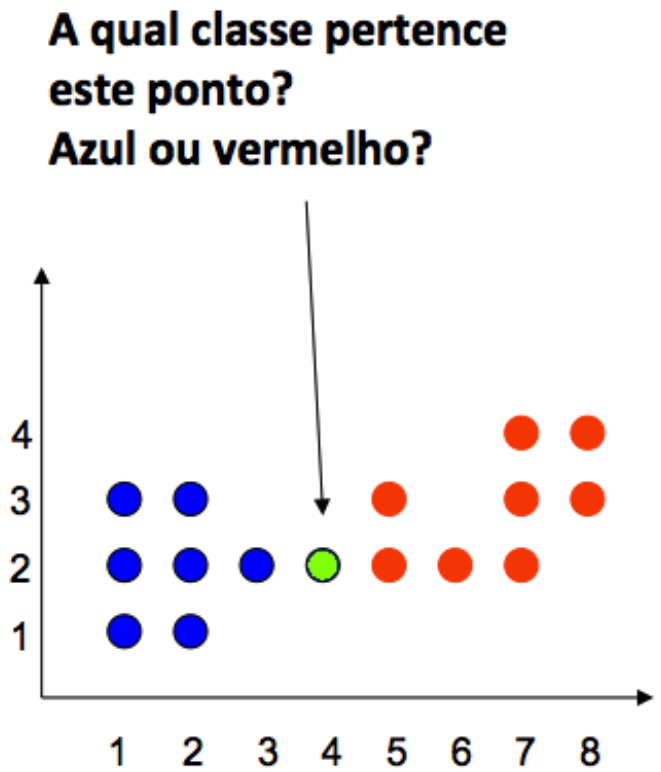
- Se consideramos a distância Euclidiana por exemplo, q está mais próximo de d_1 e d_2 do que de d_3
- Entretanto, q e d_3, possuem exatamente a mesma distribuição de palavras (vetores com a mesma orientação), porém d_3 parece ser um documento mais longo. Ou seja d_3 tem uma magnitude maior do que q.
- Nesse caso distance.cosine$(q, d_3) \approx 0$



Classificação baseada em Instâncias

kNN (k Nearest Neighbors)

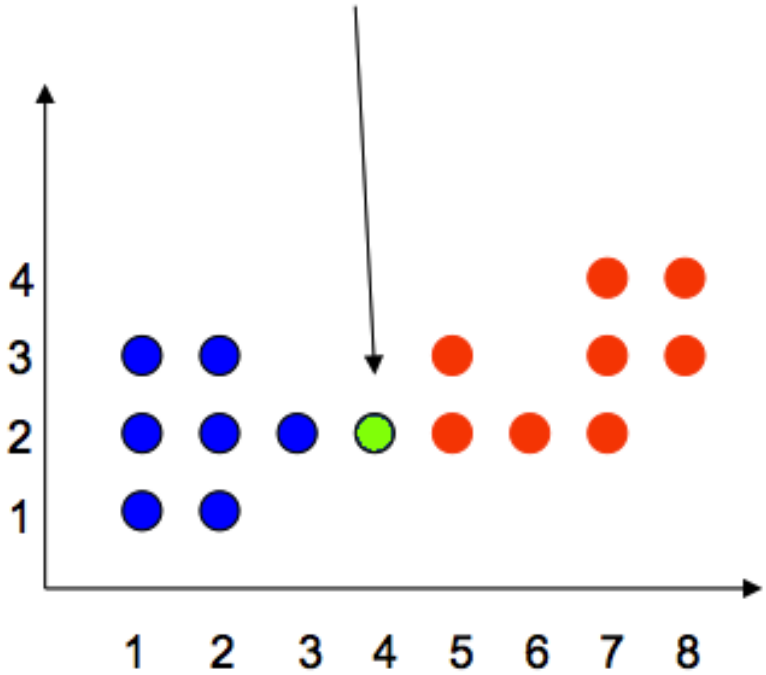
- Para um dados exemplo não rotulado x, encontre os k mais similares a ele na base rotulada e atribua a classe mais frequente para x



kNN

Exemplo

A qual classe pertence este ponto?
Azul ou vermelho?



Calcule para os seguintes valores de k:

k=1 não se pode afirmar

k=3 vermelho – 5,2 - 5,3

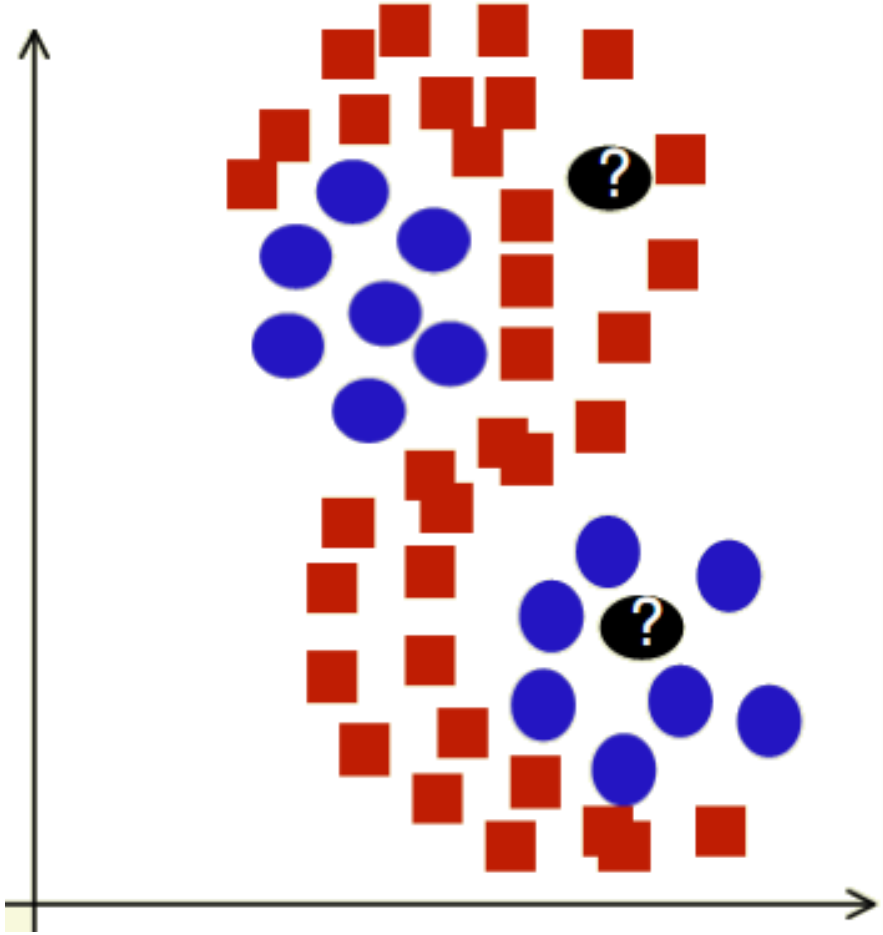
k=5 vermelho – 5,2 - 5,3 - 6,2

k=7 azul – 3,2 - 2,3 - 2,2 - 2,1

A classificação pode mudar de acordo com a escolha de *k*.

kNN

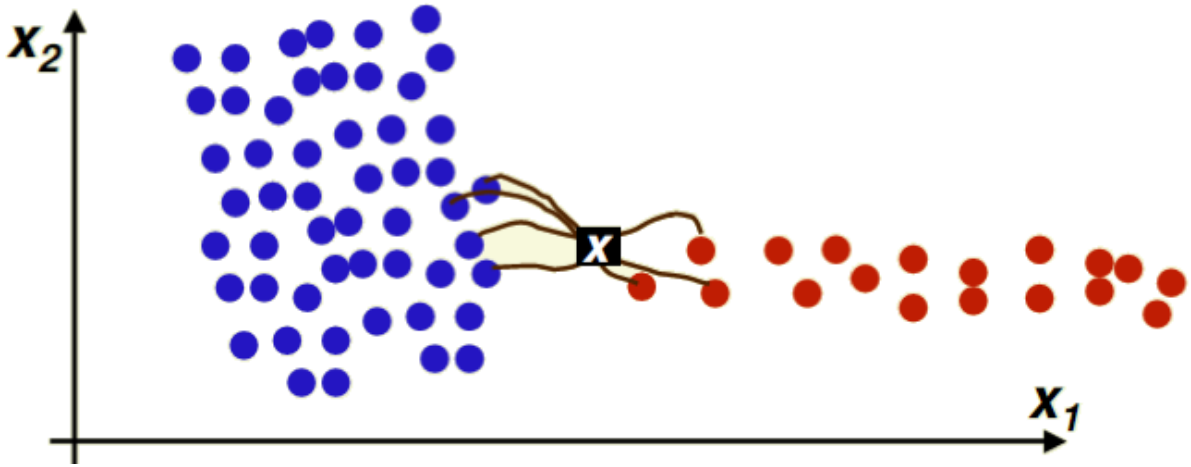
Distribuições multimodais: problemas complexos nos quais o kNN tem bom desempenho.



kNN

Como escolher o valor de k?

- k deve ser grande para minimizar o erro
- Valores pequenos de k levam a fronteiras ruidosas
- Utilizar uma base de validação para definir o valor de k



- Para k $>$ 7 x passa pertencer a classe azul.

kNN

Normalização

- Distância Euclidiana é geralmente utilizada no kNN
- Entretanto, características com diferentes ordens de grandeza impactam no cálculo da distância
- Importante que as características sejam normalizadas, usando por exemplo MinMax (rescaling)

$$z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)}$$

- ou ZScore (Standardizing)

$$z_i = \frac{x_i - \mu}{\sigma}$$

O processo "standardizing" também é conhecido como "z-scaling".

kNN

- Em alguns casos (e.g., vetores binários) a distância Euclidiana pode produzir resultados contra-intuitivos. Por exemplo
- $D([1,1,1,1,0],[0,1,1,1,1]) = 1.41$
- $D([1,0,0,0,0],[0,0,0,0,1]) = 1.41$

kNN

- Em alguns casos (e.g., vetores binários) a distância Euclidiana pode produzir resultados contra-intuitivos. Por exemplo
- $D([1,1,1,1,0],[0,1,1,1,1]) = 1.41$
- $D([1,0,0,0,0],[0,0,0,0,1]) = 1.41$
- Solução
  - Normalizar os vetores para vetores unitários (unit length), ou seja, dividir cada componente pelo seu tamanho. (Normalizing)
  - A magnitude de um vetor $\vec{a} = [3,4]$ é dado por $||\vec{a}|| = \sqrt{(3^2, 4^2)} = 5$
  - O vetor $\hat{u} = [\frac{3}{5}, \frac{4}{5}]$ tem magnitude 1.

kNN

Complexidade

- O algoritmo básico do kNN armazena todos os exemplos. Suponha que tenhamos n exemplos
  - O(n) é a complexidade para encontrar o vizinho mais próximo
  - O(nk) complexidade para encontrar k exemplos mais próximos
- Considerando que precisamos de um n grande para o kNN funcionar bem, a complexidade torna-se problema

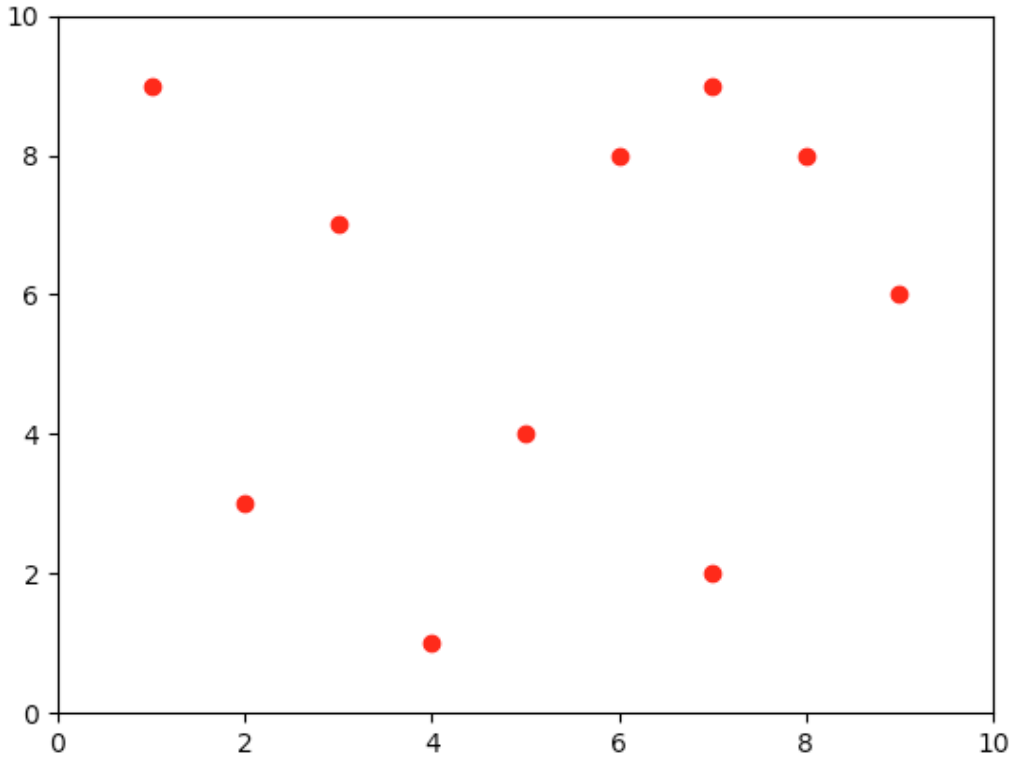
kNN - Space Partition Tree

- Uma solução para reduzir a complexidade do kNN consiste em utilizar uma árvore para particionar o espaço de busca
- Selecione a dimensão com maior variância, encontre a mediana e divida os dados.
- Repita o processo até um critério de parada (por exemplo, um número máximo de vizinhos na folha)

kNN - Space Partition Tree

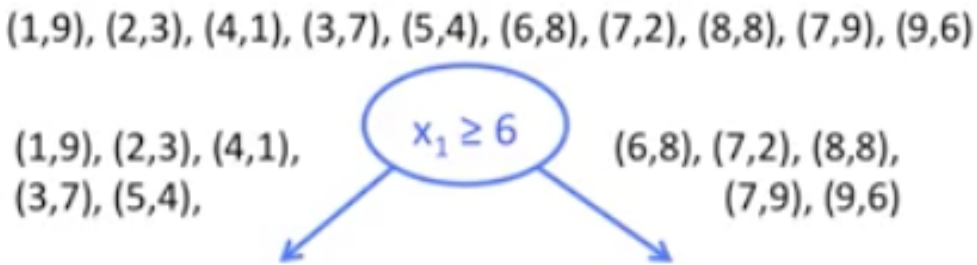
Considere o seguinte conjunto de dados

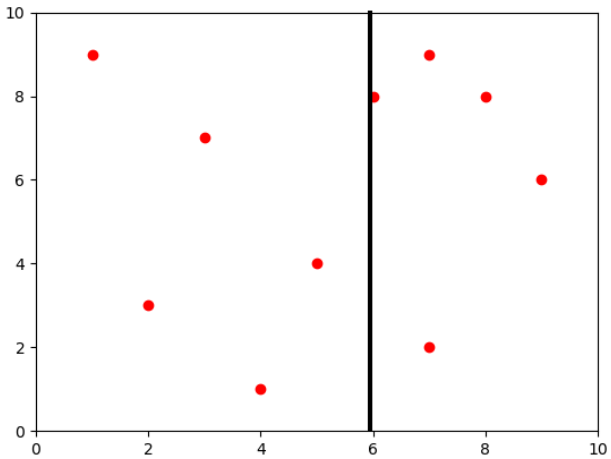
[(1,9), (2,3), (4,1), (3,7), (5,4), (6,8), (7,2), (8,8), (7,9), (9,6)]



kNN - Space Partition Tree

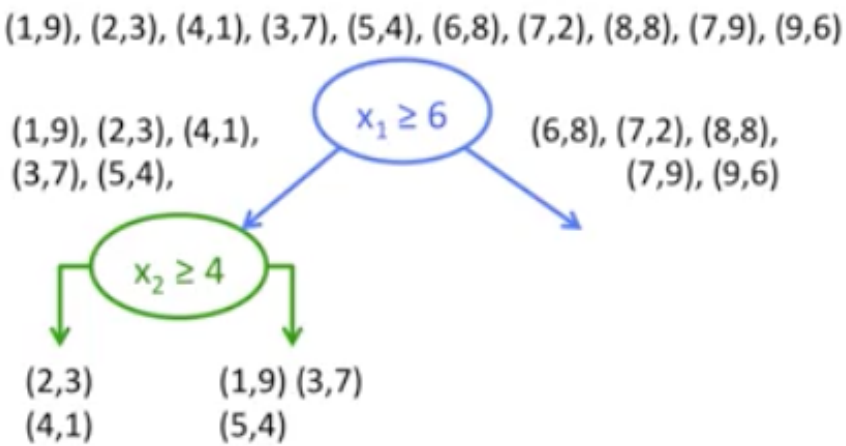
- Escolhendo a primeira dimensão do vetor, temos a mediana = 6
- Desta forma temos o primeiro nó da árvore

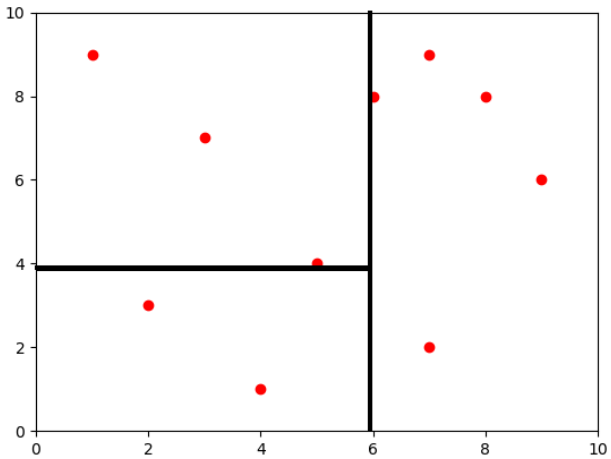




kNN - Space Partition Tree

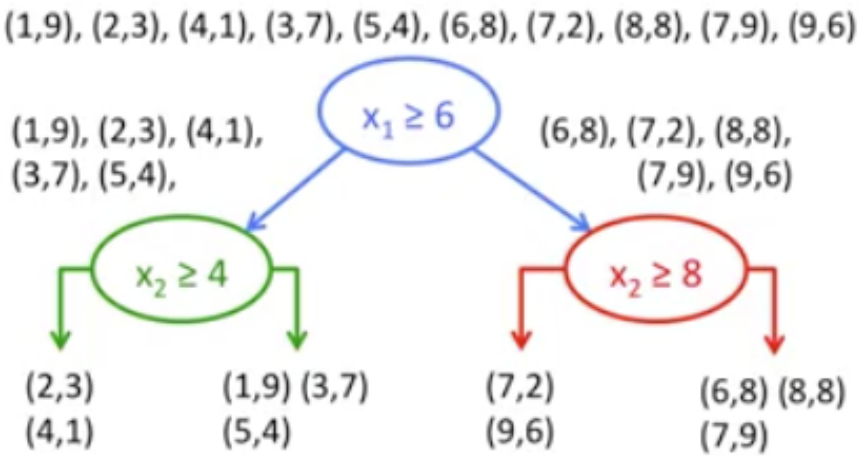
- Na sequencia, escolhemos um outro atributo. Nesse exemplo, temos x_2
- Repetimos o processo considerando os dados que sobraram para o lado esquerdo ...

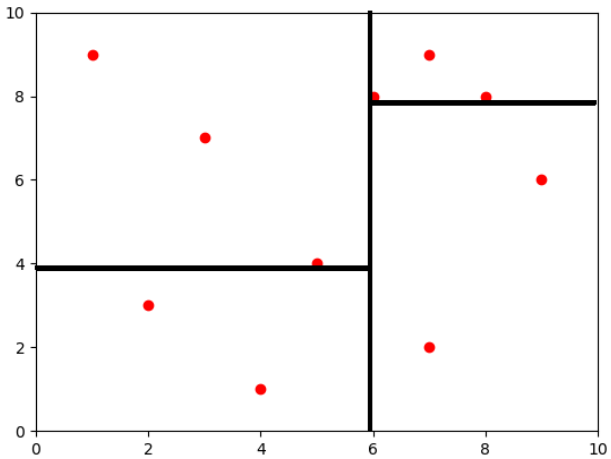




kNN - Space Partition Tree

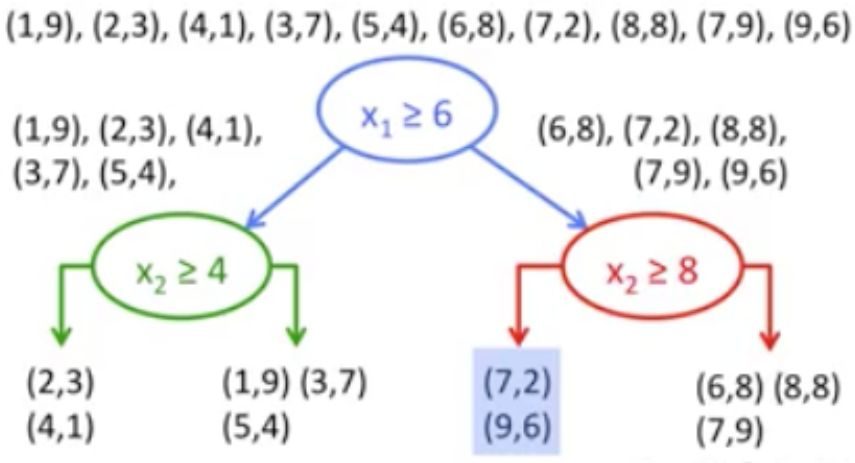
- ... e lado direito da árvore.
- O processo é repetido até que o processo de parada seja satisfeito (e.g., 3 exemplos na folha)
- Ao final temos o espaço particionado. Note que a base é dividida em 2 a cada nível. Desta forma, a complexidade nesse caso é de $\log_2(n)$, em que n é o total de instâncias na base de treinamento.

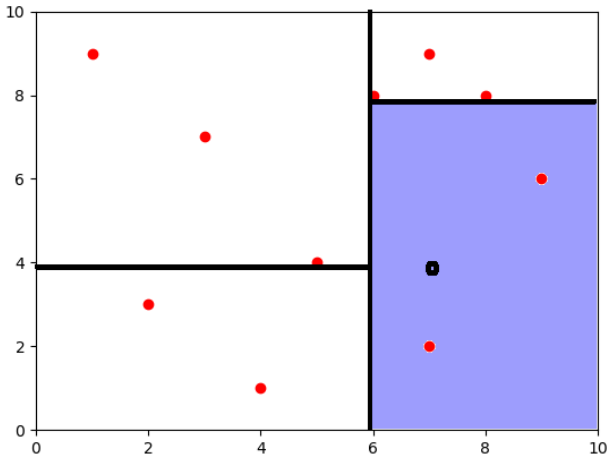




kNN - Space Partition Tree

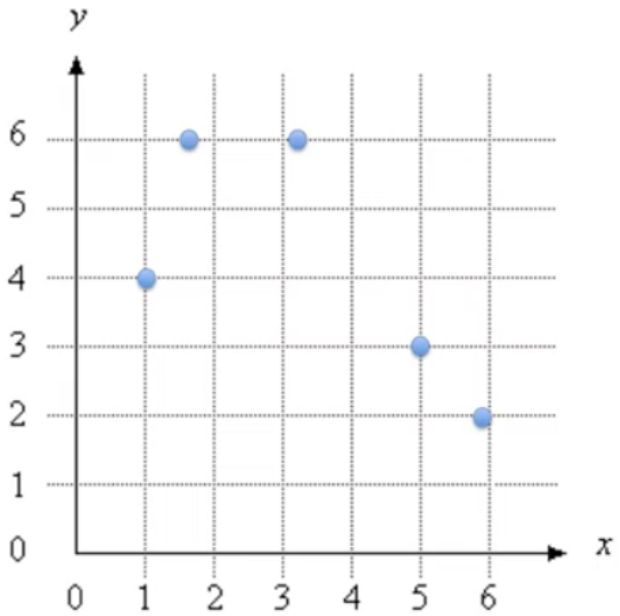
- Para classificar um exemplo X, basta encontrar a região do espaço que contém X e comparar com os vizinhos daquela região.
- Por exemplo, $X = (7, 4)$
- Note que com esse particionamento podemos perder alguns vizinhos mais próximos.

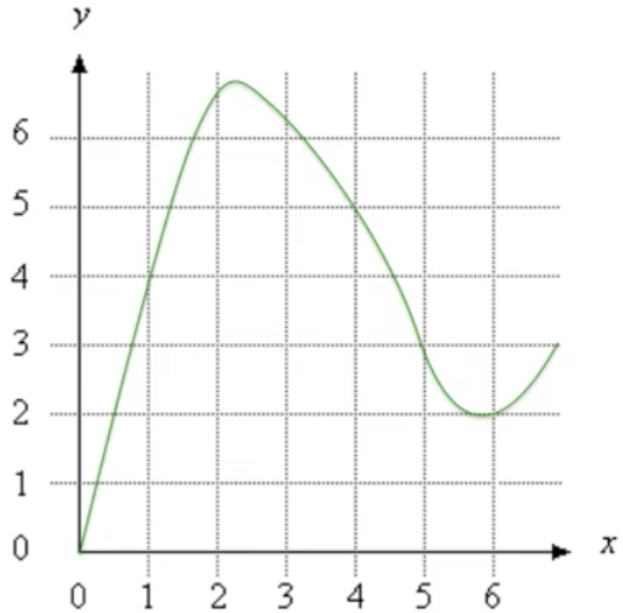




KNN para regressão

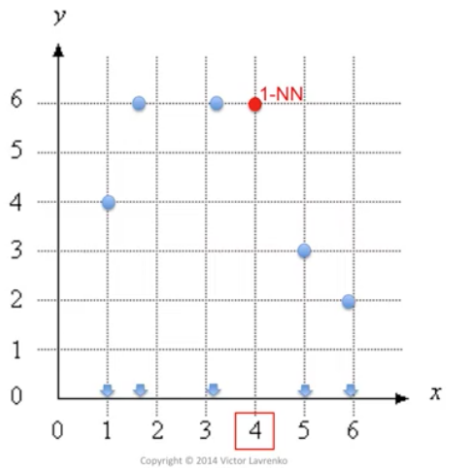
- O KNN também pode ser utilizado para regressão.
- Considere os dados de treinamento abaixo e uma função de descreve esses dados.





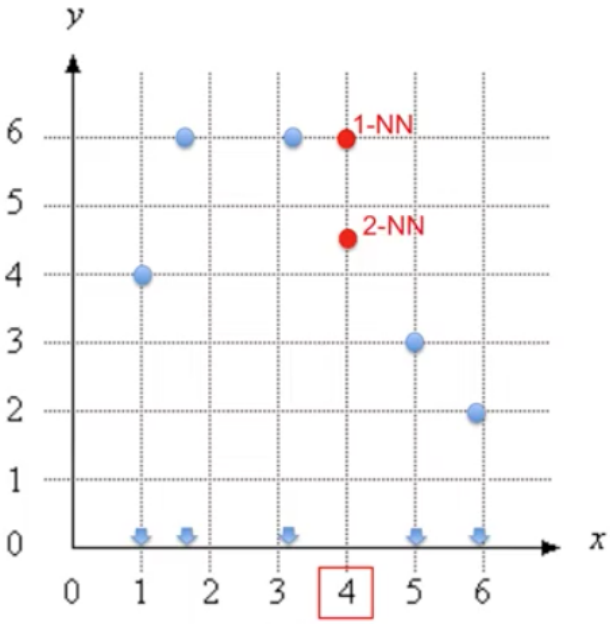
KNN para regressão

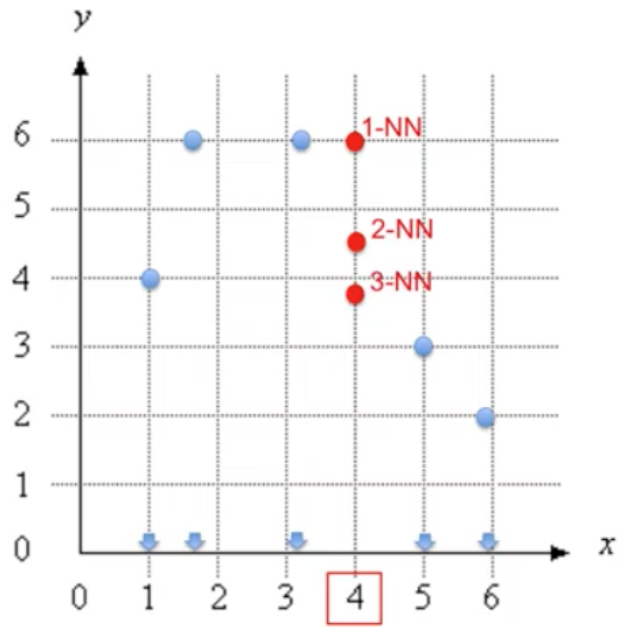
- kNN para regressão consiste em fazer a média do "Target" dos k vizinhos mais próximos.
- Por exemplo, vamos estimar o valor para y com base em x.
- Para $x = 4$, qual seria o valor de y?
- Encontrar os k vizinhos mais próximos de $x = 4$, nesse caso, 3.2 é o vizinho mais próximo.



KNN para regressão

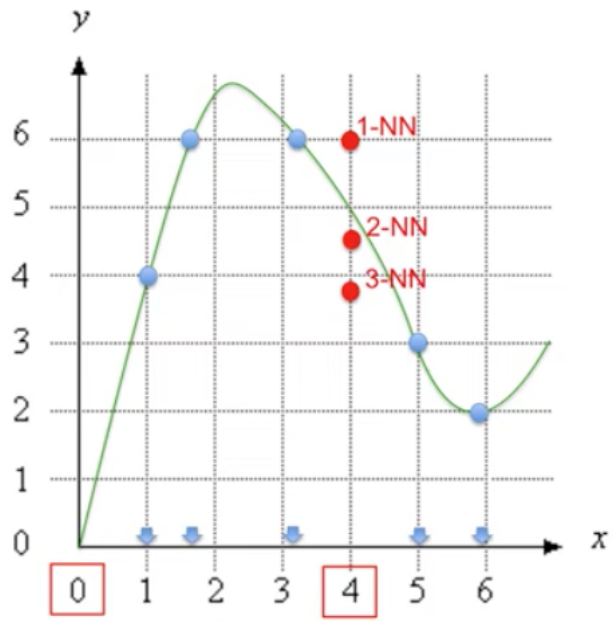
- Para 2 e 3 vizinho teríamos os seguintes valores para y





KNN para regressão

- Note que aumentar o número de vizinhos não significa uma predição melhor.
- Se k=tamanho da base, você terá um valor médio.
- Predição ruim nas extremidades. Não existem dados para fazer a interpolação.



Exercício

- Considere os dados abaixo. Para alguém com 48 anos de idade e empréstimos de 142k, calcule o HPI (House Price Index) para $k = 1$ e $k = 3$

X1 (Idade)	X2 (Emprestimo)	Label (HPI)
25	40	135
35	60	256
45	80	231
20	20	267
35	120	139
52	18	150
23	85	127
40	62	216
60	100	139
48	220	250
33	150	264