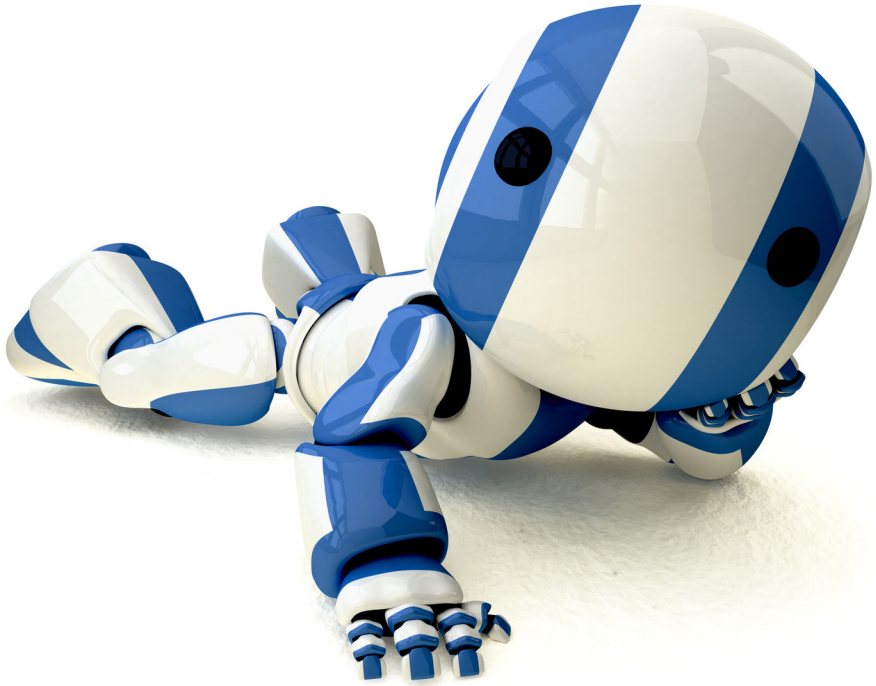
PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL
ESCOLA POLITÉCNICA

Aprendizado de Máquina



Introdução ao Aprendizado Supervisionado
Paradigma baseado em Instâncias (e Distâncias)

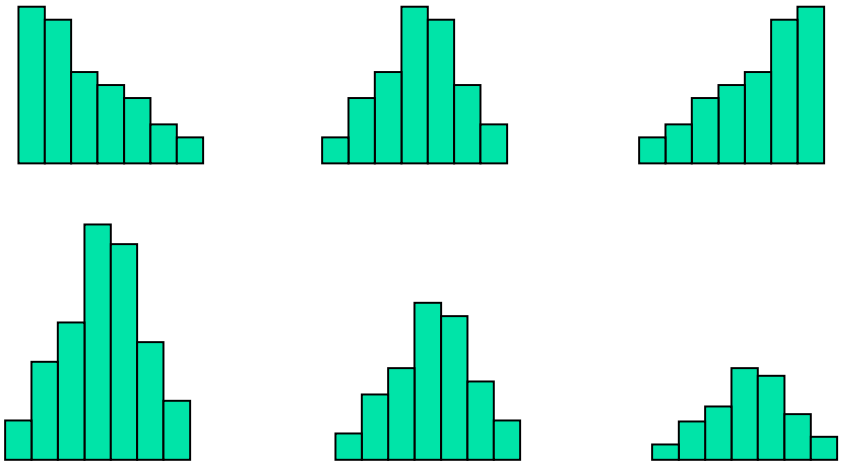
Prof. Me. Otávio Parraga

MALTA

Machine Learning Theory and Applications Lab

Aula Passada

- Dados
  - Análise
  - Pré-Processamento



	$\mathbf{x}_1$	$\mathbf{x}_2$		$\mathbf{x}_m$
$\mathbf{x}^{(1)}$	$x_1^{(1)}$	$x_2^{(1)}$	...	$x_m^{(1)}$
$\mathbf{x}^{(2)}$	$x_1^{(2)}$	$x_2^{(2)}$	...	$x_m^{(2)}$
			...	
$\mathbf{x}^{(N)}$	$x_1^{(N)}$	$x_2^{(N)}$	...	$x_m^{(N)}$

Aula de Hoje

- Introdução ao Aprendizado Supervisionado
- Aprendizado Baseado em Instâncias (e Distâncias)
- Algoritmos Baseados em Instâncias

Aula de Hoje

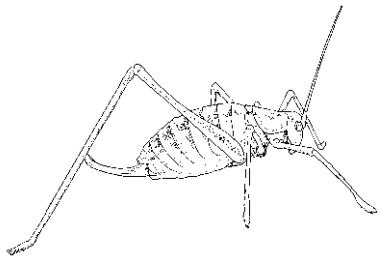
- Introdução ao Aprendizado Supervisionado
  - Classificação
  - Regressão
- Aprendizado Baseado em Instâncias (e Distâncias)
- Algoritmos Baseados em Instâncias

O problema de classificação

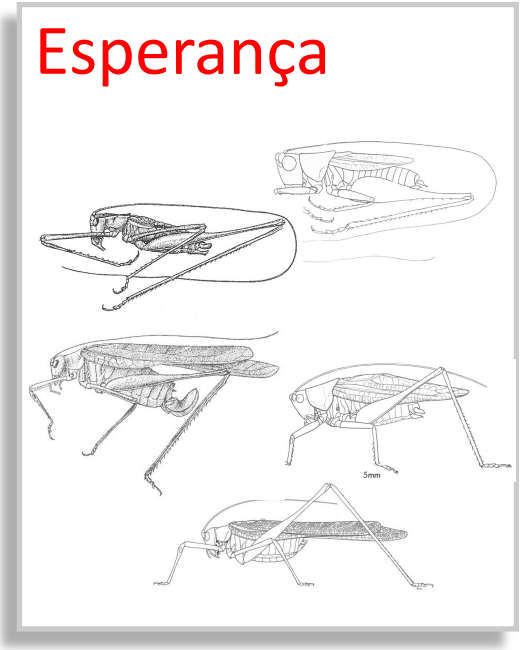
(definição informal)

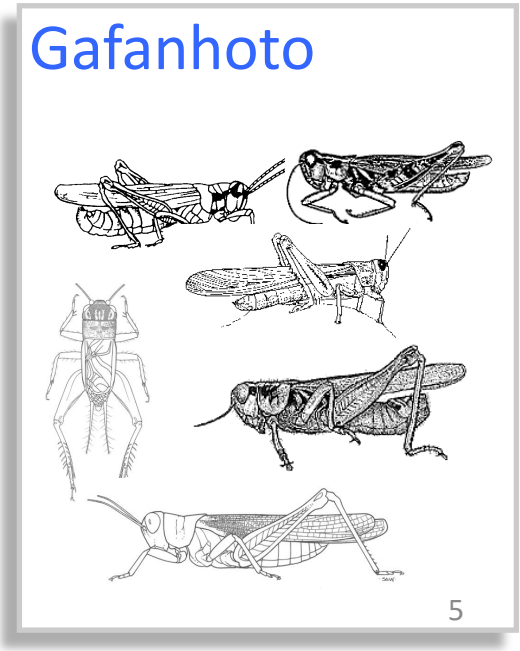
- Dada uma coleção de dados detalhados (neste caso 5 exemplos de **Esperança** e 5 de **Gafanhoto**), decida a qual tipo de inseto o exemplo não rotulado abaixo pertence:

Obs: **Esperança** = tipo de gafanhoto verde



Esperança ou Gafanhoto?

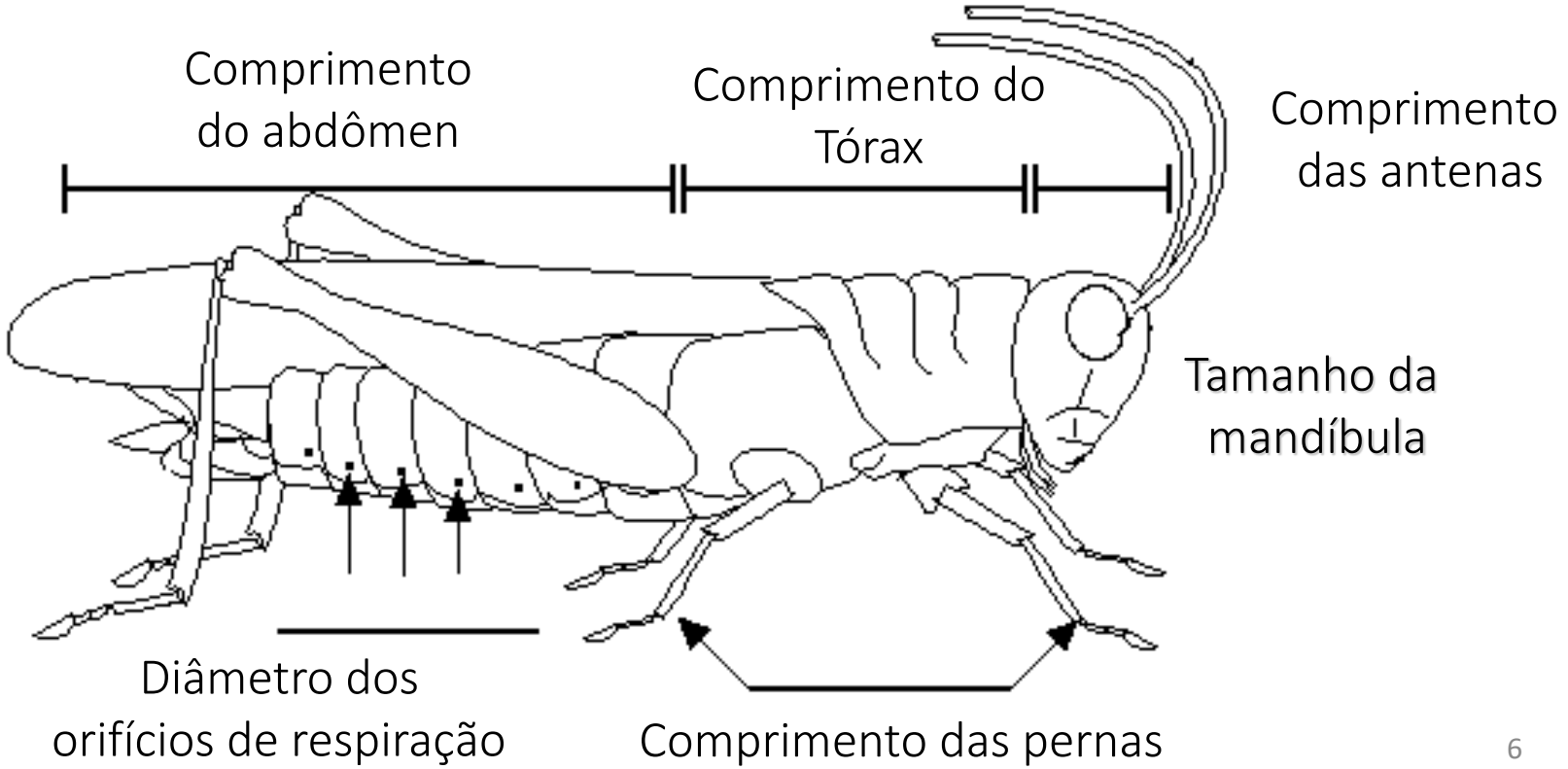




Para qualquer domínio de interesse podemos medir características

Cor: {Verde, Marrom, Cinza, Outra}

Tem asas?



Podemos armazenar as características em *datasets*

O problema de classificação agora pode ser expresso da seguinte forma:

- Dada uma base de treino (Base), preveja o rótulo da classe dos exemplos ainda não vistos

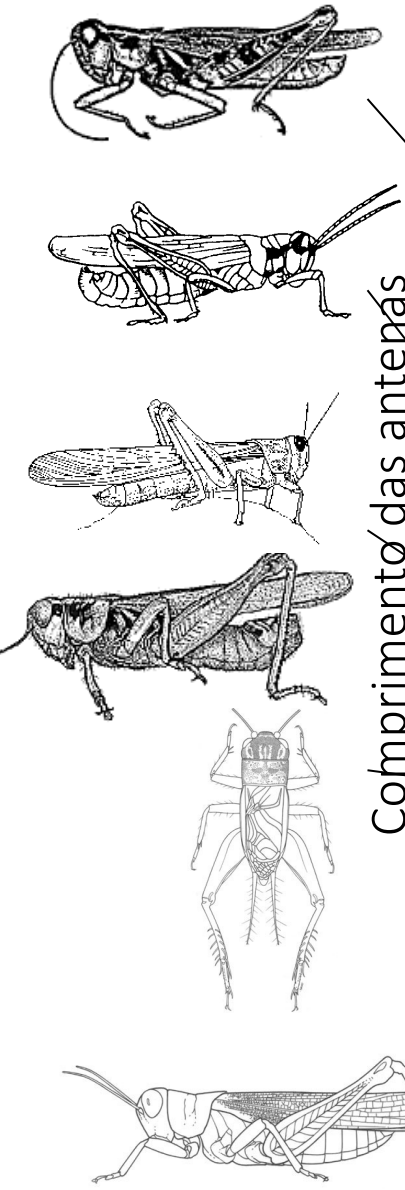
Base de Treinamento

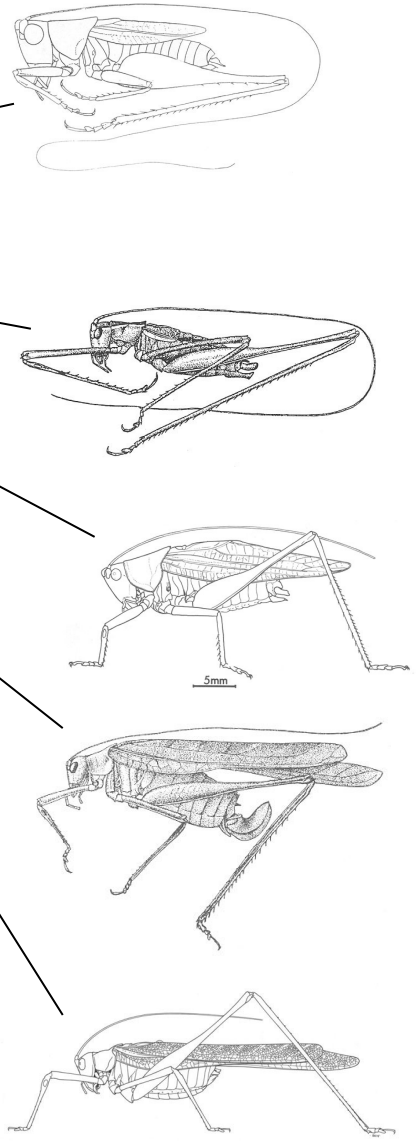
ID do inseto	Comp. do abdômen	Comp. das antenas	Classe do inseto
1	2.7	5.5	Gafanhoto
2	8.0	9.1	Esperança
3	0.9	4.7	Gafanhoto
4	1.1	3.1	Gafanhoto
5	5.4	8.5	Esperança
6	2.9	1.9	Gafanhoto
7	6.1	6.6	Esperança
8	0.5	1.0	Gafanhoto
9	8.3	6.6	Esperança
10	8.1	4.7	Esperança

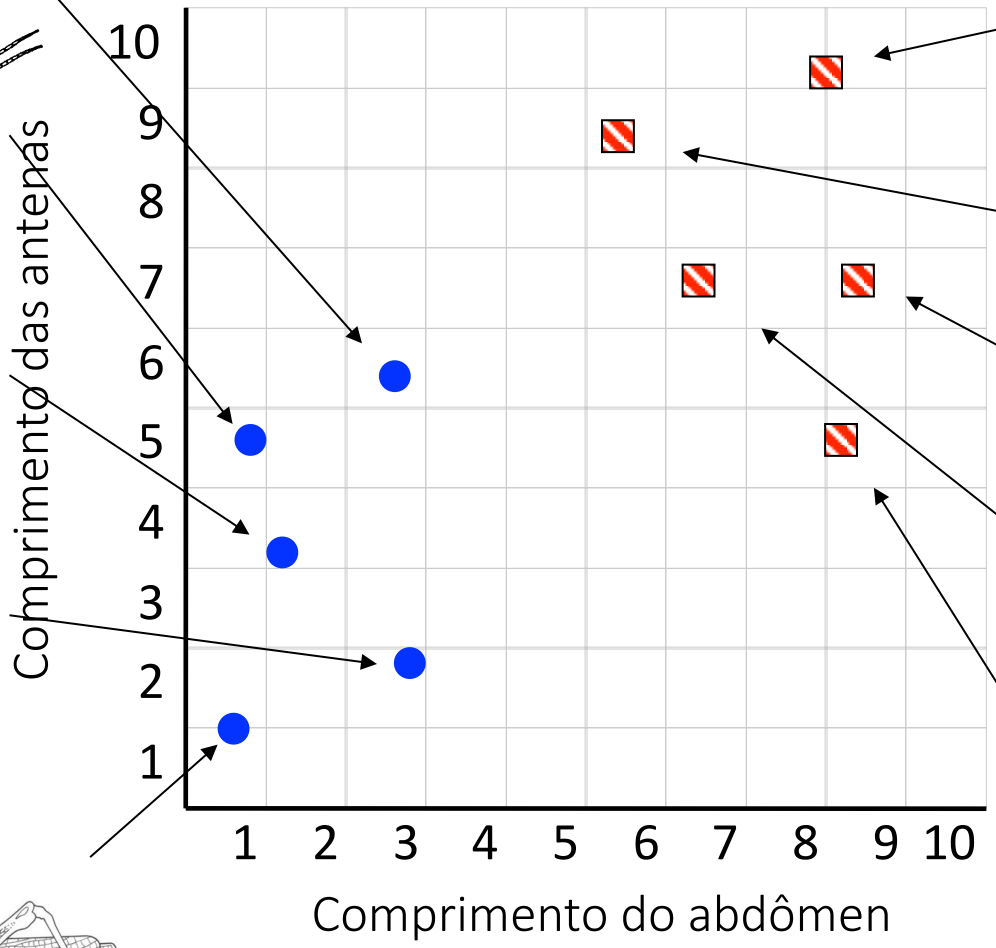
Exemplo não visto =	11	5.1	7.0	??????????

Gafanhoto

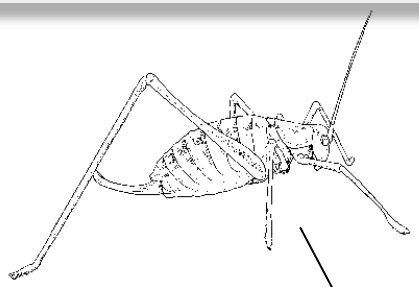
Esperança

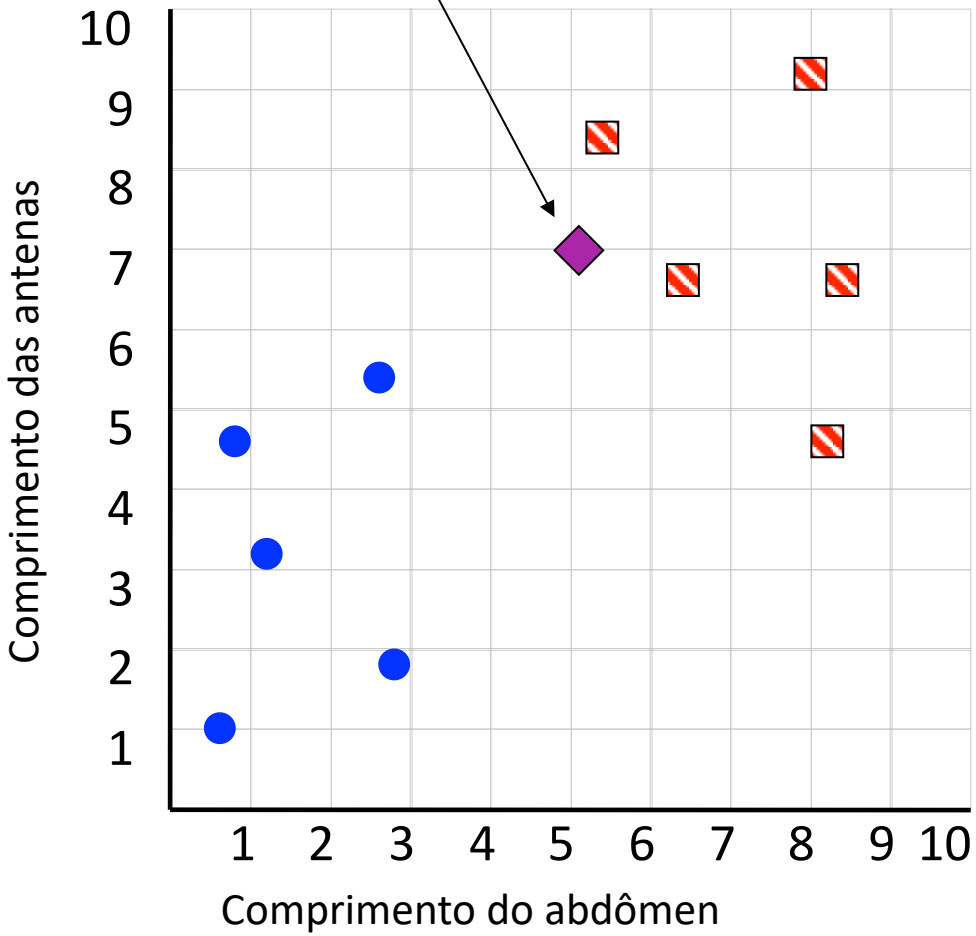






Exemplo não visto antes =	11	5.1	7.0	???????



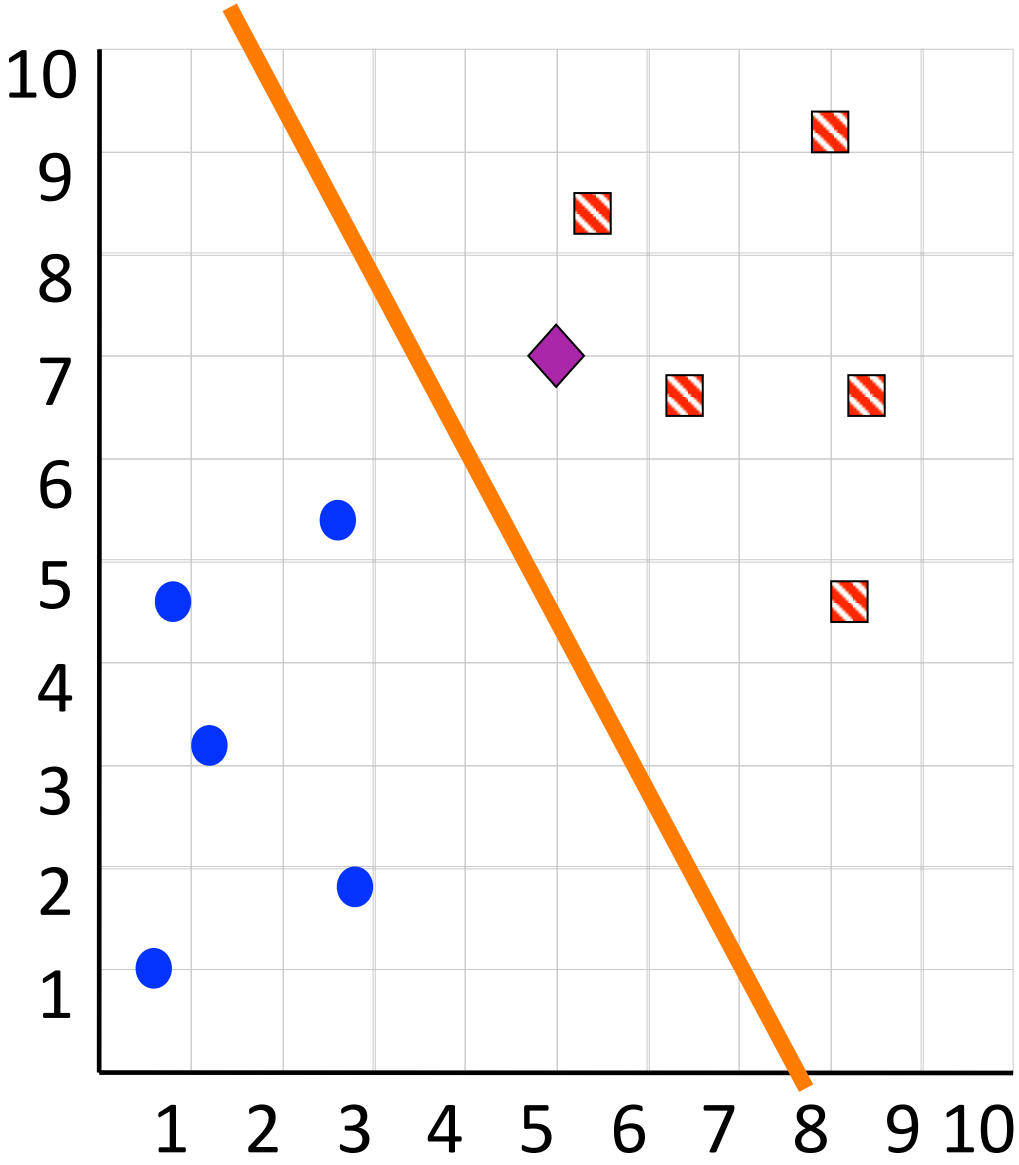


- Podemos **projetar** o exemplo não visto antes dentro do mesmo espaço que os dados de treino

- Acabamos de abstrair os detalhes do nosso problema particular. Será muito mais fácil falar de pontos no espaço

Esperança

Gafanhoto



Definindo classificação formalmente:

$$\mathbf{x}^{(i)} = \left[x_j^{(i)}\right]_{j=1}^{m} \in X^m$$

$$Y = \{y_1, \ldots, y_k\}$$

$$D = \left\{\mathbf{x}^{(i)}, f\left(\mathbf{x}^{(i)}\right)\right\}_{i=1}^{N}$$

$$\hat{f} = X^m \rightarrow Y$$

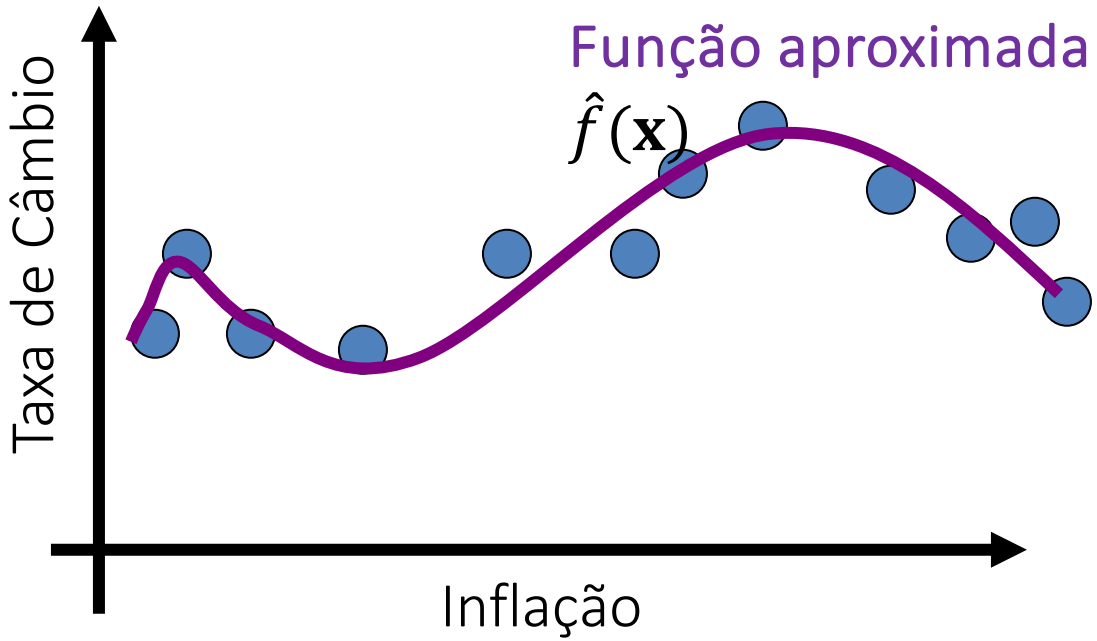
Esperança

Gafanhoto

Problema de Regressão

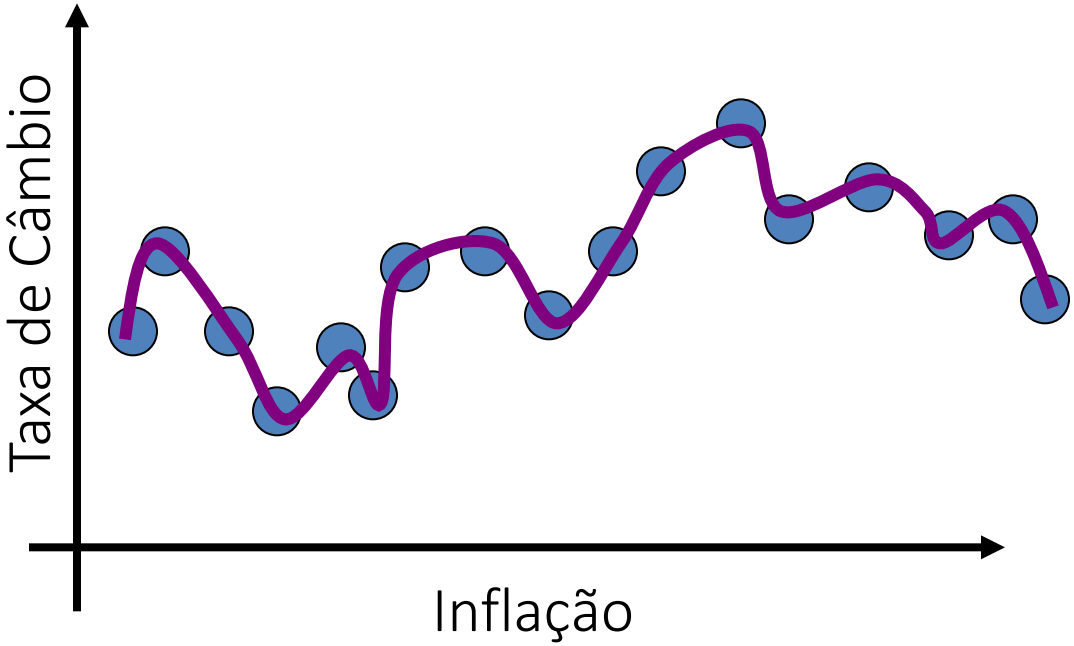
- Funciona exatamente como a classificação, mas atributo meta é contínuo em vez de discreto
- Também pode ser visto sob a ótica de aproximação de funções
  - Descobrir a função que mapeia os atributos preditivos em um valor real
  - Em geral, busca-se minimizar uma função de custo
    - Erro quadrático médio, etc.

Problema de Regressão



Problema de Regressão

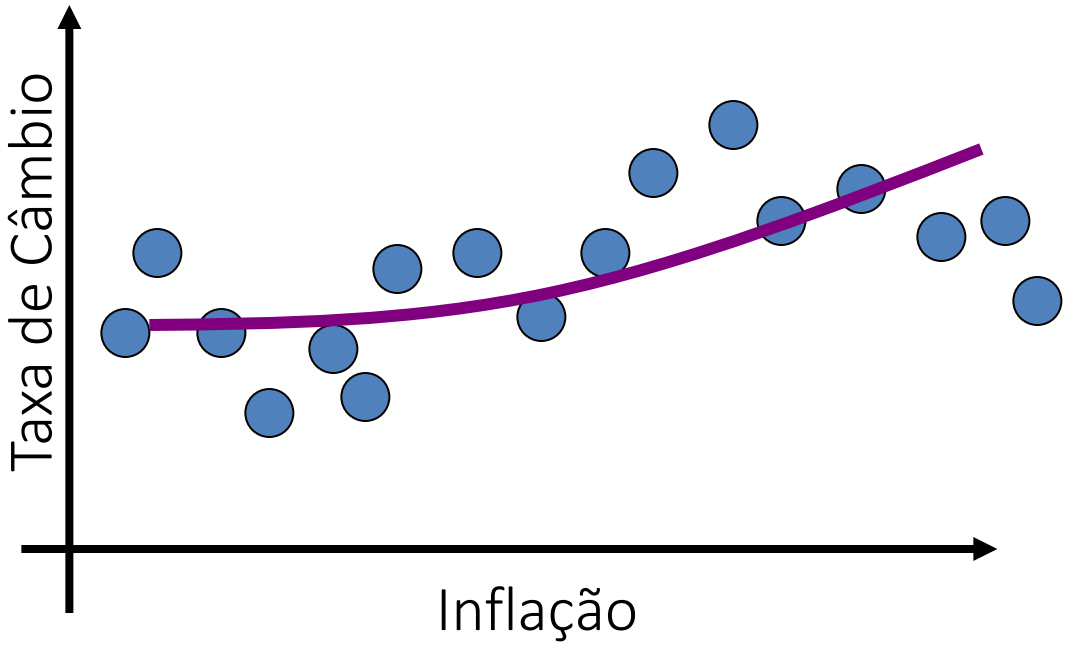
Overfitting



Também presente
em classificação

Problema de Regressão

Underfitting



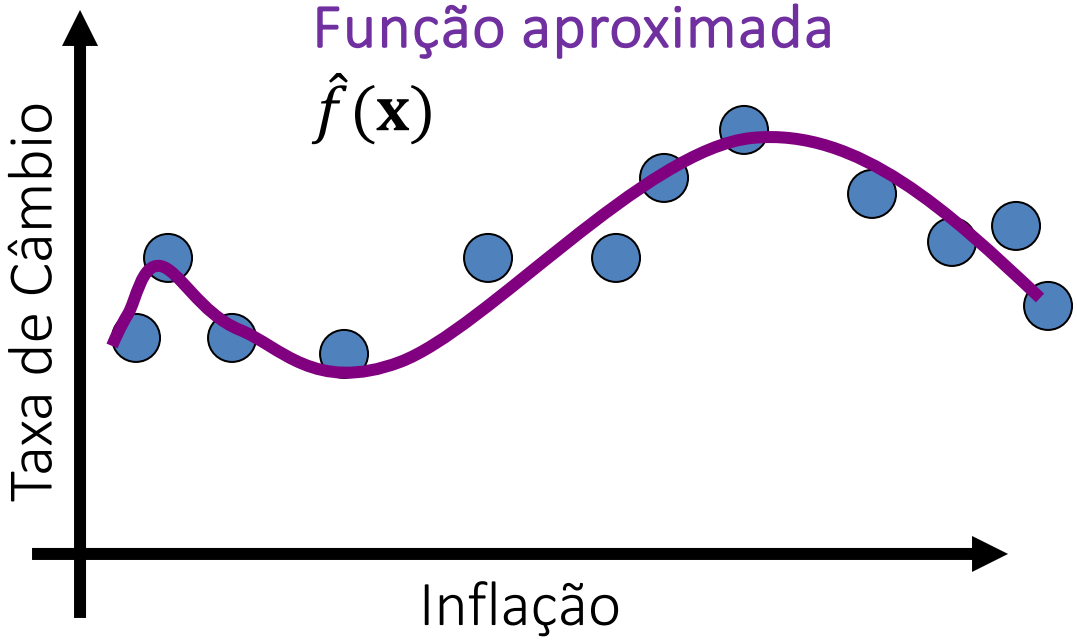
Também presente
em classificação

Formalizando Regressão

$$\mathbf{x}^{(i)} = \left[x_j^{(i)}\right]_{j=1}^{m} \in X^m$$

$$f\left(\mathbf{x}^{(i)}\right) \in \mathbb{R}$$

$$D = \left\{\mathbf{x}^{(i)}, f\left(\mathbf{x}^{(i)}\right)\right\}_{i=1}^{N}$$



Descobrir $\hat{f}$ que aproxima f
minimizando uma função de erro e (ou $\mathcal{L}$)

Aula de Hoje

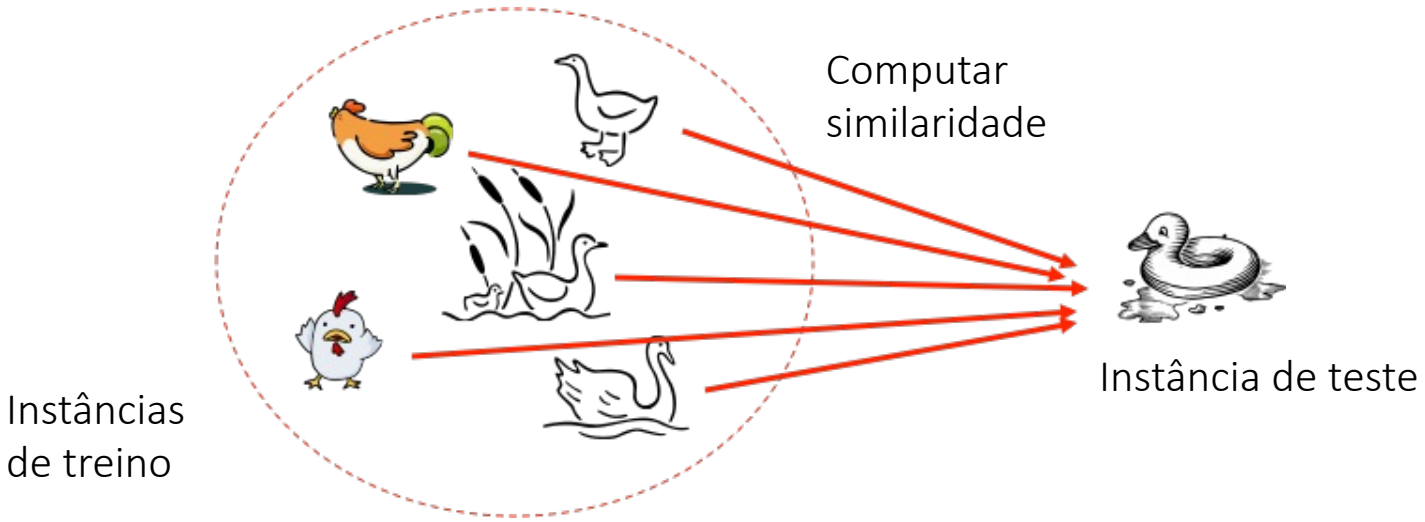
- Introdução ao Aprendizado Supervisionado
- Aprendizado Baseado em Instâncias (e Distâncias)
  - Conceitos
  - Medindo Proximidade
    - Medidas de Similaridade
    - Medidas de Dissimilaridade
- Algoritmos Baseados em Instâncias

Instance-based Learning

- Paradigma baseado em instâncias
  - Ou em "memória" (*memory-based learning*)
- Não constrói um modelo preditivo
  - Aprendizado preguiçoso (lazy)
    - Só olha dados de treino quando precisa classificar um objeto novo
  - Tem como premissa:
    - Instâncias similares pertencem à mesma classe! (classificação)
    - Instâncias similares têm valores (contínuos) semelhantes de atributo alvo (regressão)

Instance-based Learning

- **Ideia básica:** se caminha como um pato, faz *quack* como um pato e parece um pato, então provavelmente **é um pato!**



O que é Similaridade?

O que é Similaridade?

O que é Similaridade?

É difícil definir similaridade mas… "sabemos quando vemos"

Como descobrir o real sentido de similaridade é uma questão filosófica, vamos partir para uma abordagem mais pragmática!

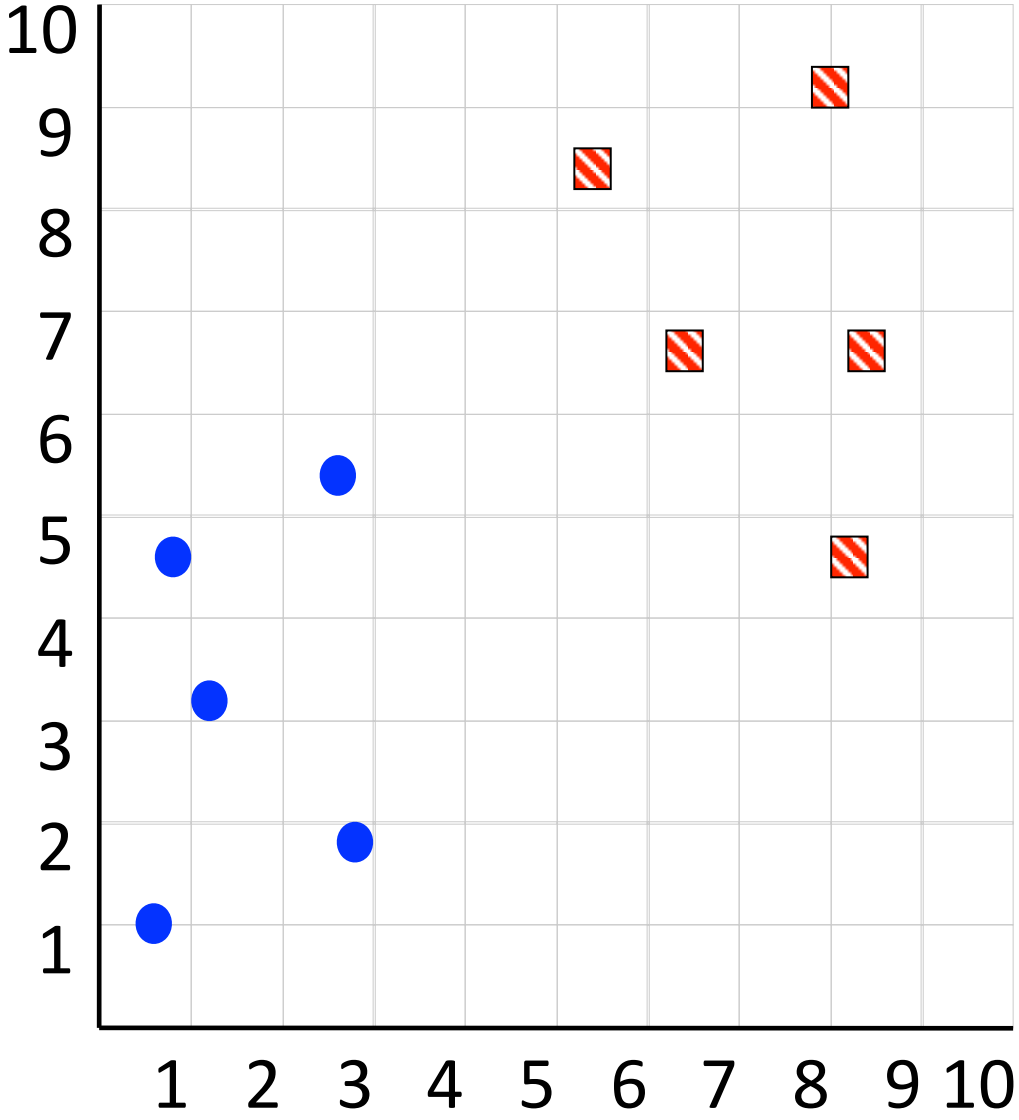
Similaridade vs Dissimilaridade

- Similaridade
  - Medida que indica nível de semelhança entre dois objetos
  - Quanto mais semelhantes, maior o seu valor
  - Geralmente valor $\in [0, 1]$

- Dissimilaridade
  - Medida que indica o quanto dois objetos são diferentes
  - Quanto mais diferentes, maior o seu valor
  - Geralmente valor $\in [0, \mathrm{dmax}]$ ou $[0, +\infty]$

- Medidas de similaridade e dissimilaridade são chamadas genericamente de medidas de proximidade

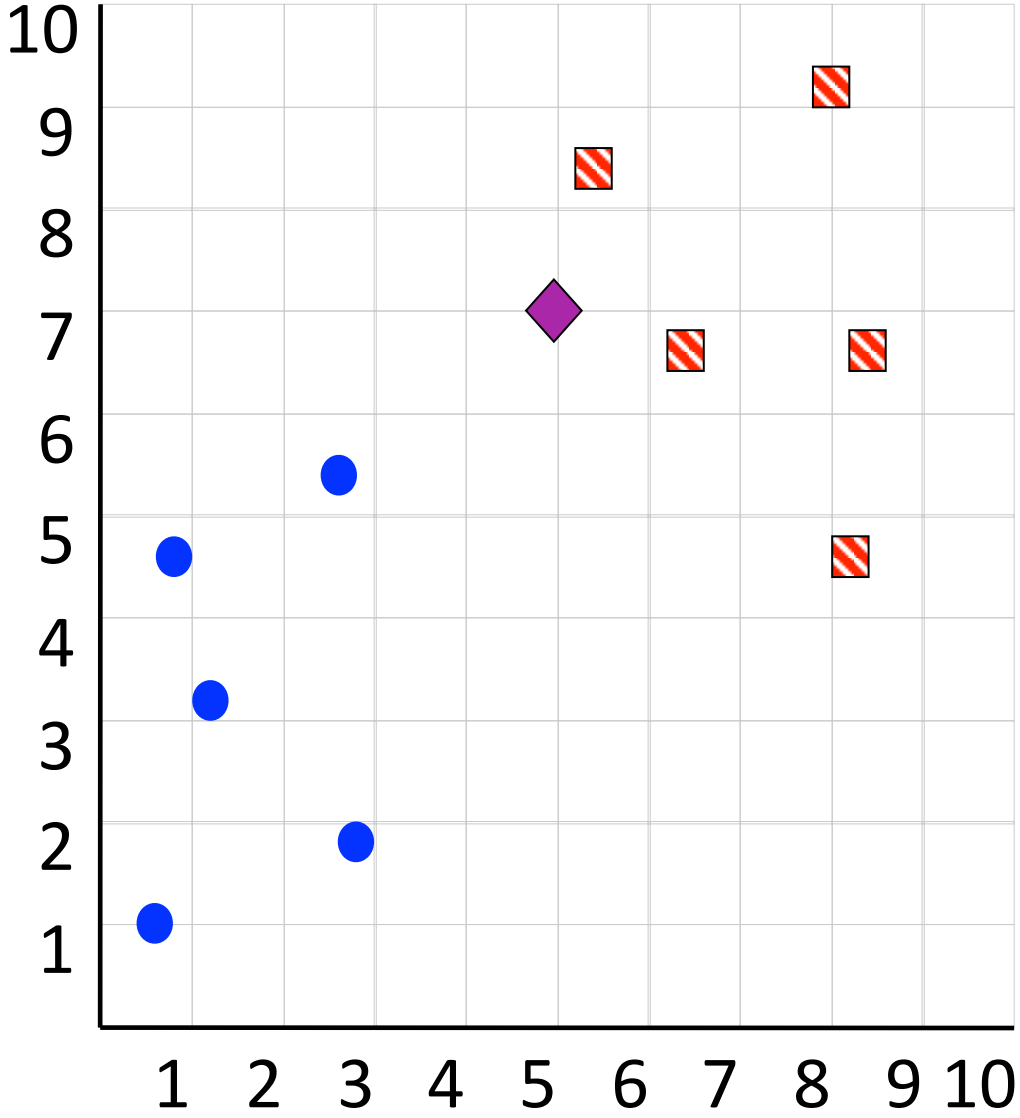
Proximidade



Podemos analisar o nível de (dis)similaridade entre instâncias conforme a proximidade delas no espaço de instâncias

Para tanto, precisamos definir uma medida de distância!

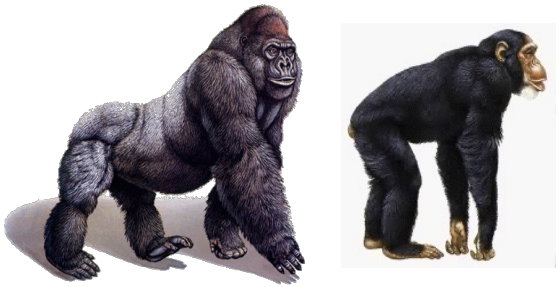
O Algoritmo que Veremos



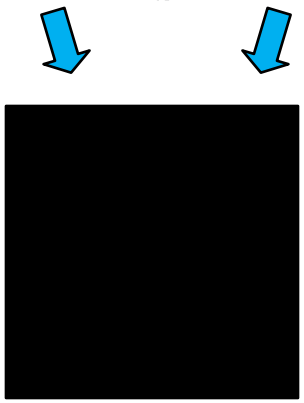
- Necessita de 3 coisas:
  - Base de treinamento
  - Medida de distância
  - Quantidade de objetos para comparar
- Para classificar uma instância não-vista:
  - Calcule a distância para todas as instâncias de treino
  - Obtenha as mais próximas
  - Classifique a instância não vista na classe da maioria

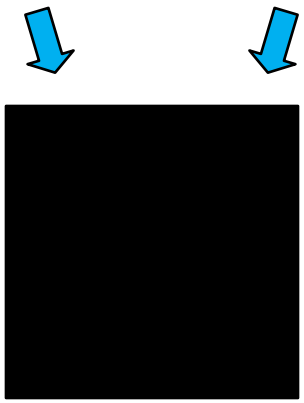
Definindo Medidas de Distância

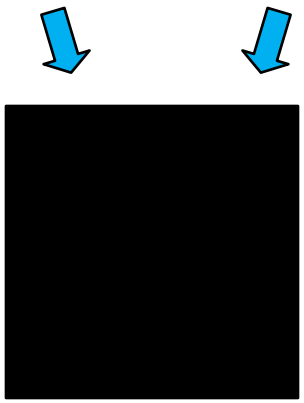
Definição: sejam $\mathbf{x}^{(1)}$ e $\mathbf{x}^{(2)}$ dois objetos do universo de possíveis objetos. A distância (dissimilaridade) entre $\mathbf{x}^{(1)}$ e $\mathbf{x}^{(2)}$ é um número real denotado por $d(\mathbf{x}^{(1)}, \mathbf{x}^{(2)})$



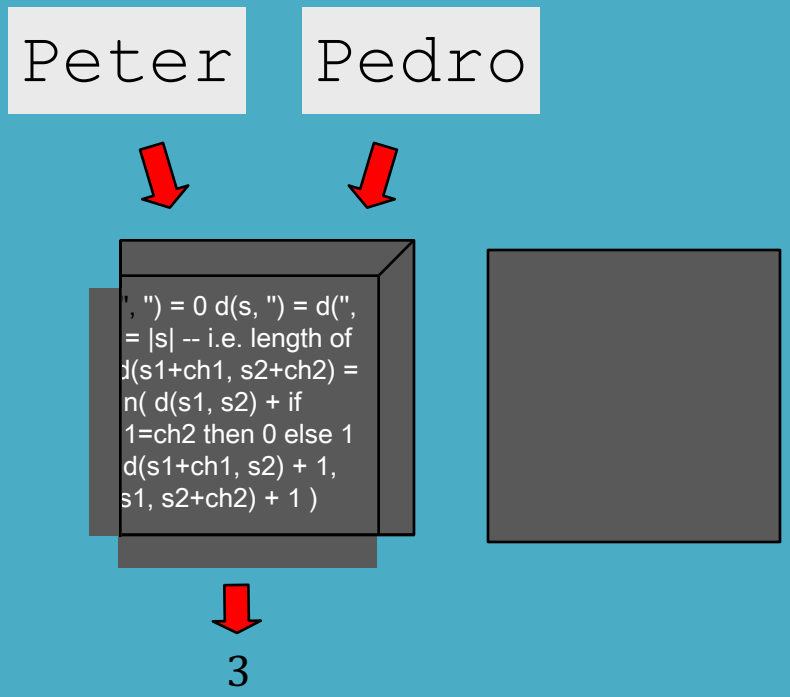
Peter Pedro





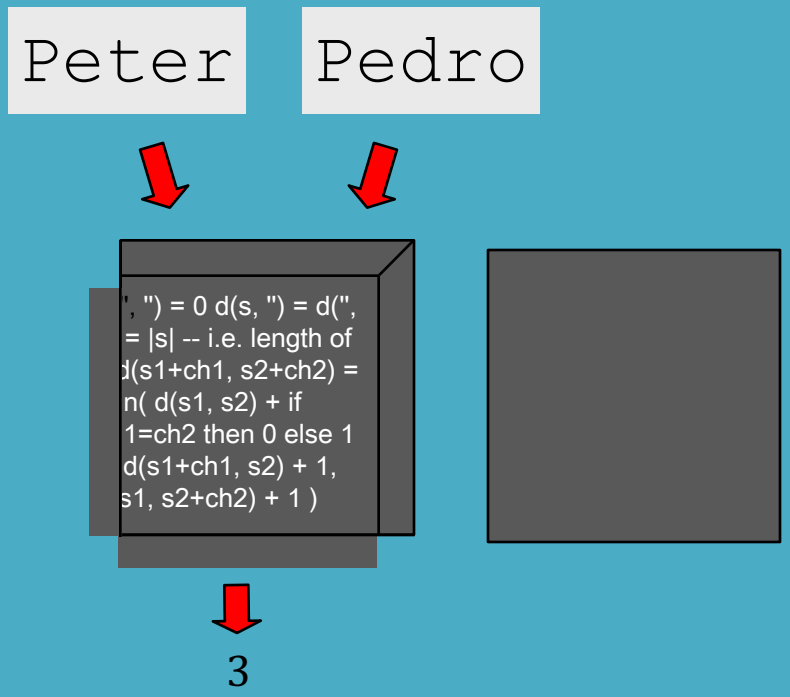


0.23 3 342.7



Quando olhamos dentro de uma destas caixas pretas, observamos uma função aplicável a duas variáveis. Tais funções podem ser muito simples ou muito complexas

Em qualquer caso, é natural perguntarmos: que propriedades tais funções têm?



Quando olhamos dentro de uma destas caixas pretas, observamos uma função aplicável a duas variáveis. Tais funções podem ser muito simples ou muito complexas

Em qualquer caso, é natural perguntarmos: que propriedades tais funções têm?

Que propriedades são desejáveis a uma medida de distância?

- $d(a,b) = d(b,a)$ Simetria
- $d(a,\mathrm{a}) = 0$ Constância da auto-similaridade
- $d(a,b) = 0 \Leftrightarrow a = b$ Positividade (separação)
- $d(a,c) \leq d(a,b) + d(b,c)$ Desigualdade triangular

Escolhendo uma medida de (dis)similaridade

"A escolha da medida de (dis)similaridade é importante para aplicações, e a melhor escolha é frequentemente obtida via uma combinação de experiência, habilidade, conhecimento e sorte."

Gan, G., Ma, C., Wu, J**. Data Clustering: Theory, Algorithms, and Applications**. SIAM Series on Statistics and Applied Probability, 2007.

Medidas de (Dis)similaridade:

- Espaço de Atributos Contínuo
- Espaço de Atributos Discreto
- Espaço de Atributos Misto

- Nosso foco será nas medidas mais amplamente utilizadas na prática
  - Literatura sobre o assunto é vasta! Pesquise!

Medidas de (Dis)similaridade:

a) Atributos contínuos

a.1) Distância Euclidiana

$$d^E\left(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}\right) = \left\|\mathbf{x}^{(i)} - \mathbf{x}^{(j)}\right\|_2 = \sqrt{\sum_{k=1}^{m}\left(x_k^{(i)} - x_k^{(j)}\right)^2}$$

- Métrica (satisfaz as 4 propriedades vistas anteriormente)
- Visualização geométrica é uma hiper-esfera
- Implementações computacionais eficientes não computam raiz (operação monotônica)
- Atributos com maiores valores e variâncias tendem a dominar os demais…

Generalizando a Distância Euclidiana

a.2) Distância de Minkowski

$$d^p\left(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}\right) = \left\|\mathbf{x}^{(i)} - \mathbf{x}^{(j)}\right\|_p = \left(\sum_{k=1}^{m} \left|x_k^{(i)} - x_k^{(j)}\right|^p\right)^{1/p}$$

- Para $p = 2$: Distância Euclidiana
- Para $p = 1$: Distância de Manhattan (**city block**)
- Para $p \to \infty$: Distância Suprema

$$d^\infty\left(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}\right) = \left\|\mathbf{x}^{(i)} - \mathbf{x}^{(j)}\right\|_\infty = \max_{1 \le k \le m} \left|x_k^{(i)} - x_k^{(j)}\right|$$

a.3) Distância de Mahalanobis

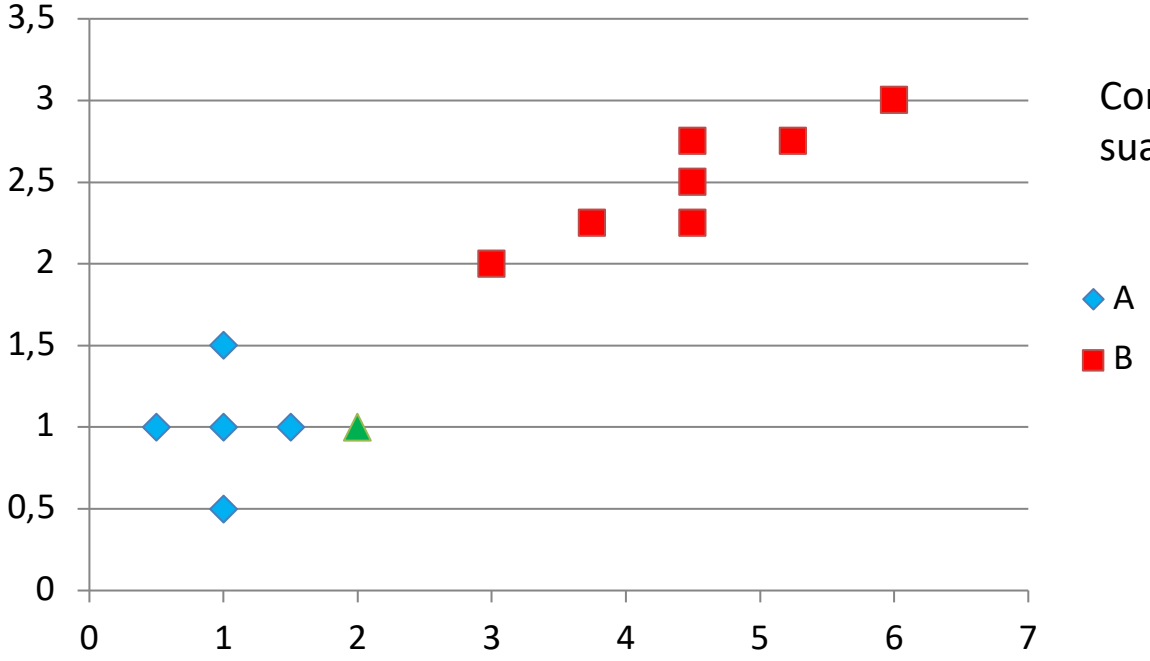
$$d^m\left(\mathbf{x}^{(i)}, \boldsymbol{\mu}_j\right) = \sqrt{\left(\mathbf{x}^{(i)} - \boldsymbol{\mu}_j\right)^T \Sigma_j^{-1} \left(\mathbf{x}^{(i)} - \boldsymbol{\mu}_j\right)}$$

- É a distância entre uma instância ($\mathbf{x}^{(i)}$) ao centro de um grupo de instâncias ($\boldsymbol{\mu}_j$)
- Ou seja:

$$\boldsymbol{\mu}_j = \frac{1}{N_j} \sum_{l=1}^{N_j} \mathbf{x}^{(l)}$$

$$\Sigma_j = \begin{pmatrix} var_1 & cov_{12} & \cdots & cov_{1m} \\ cov_{12} & var_2 & \ldots & cov_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ cov_{m1} & cov_{m2} & \ldots & var_m \end{pmatrix} = \frac{1}{N_j} \sum_{l=1}^{N_j} (\mathbf{x}^{(l)} - \boldsymbol{\mu}_j)\,(\mathbf{x}^{(l)} - \boldsymbol{\mu}_j)^T$$

Exemplo Pedagógico



Considere a instância: $\mathbf{x}^{(i)} = [2\ 1]^T$ e suas distâncias aos dois grupos de dados:

$$d^m\left(\mathbf{x}^{(i)}, \mu_A\right) = 10$$

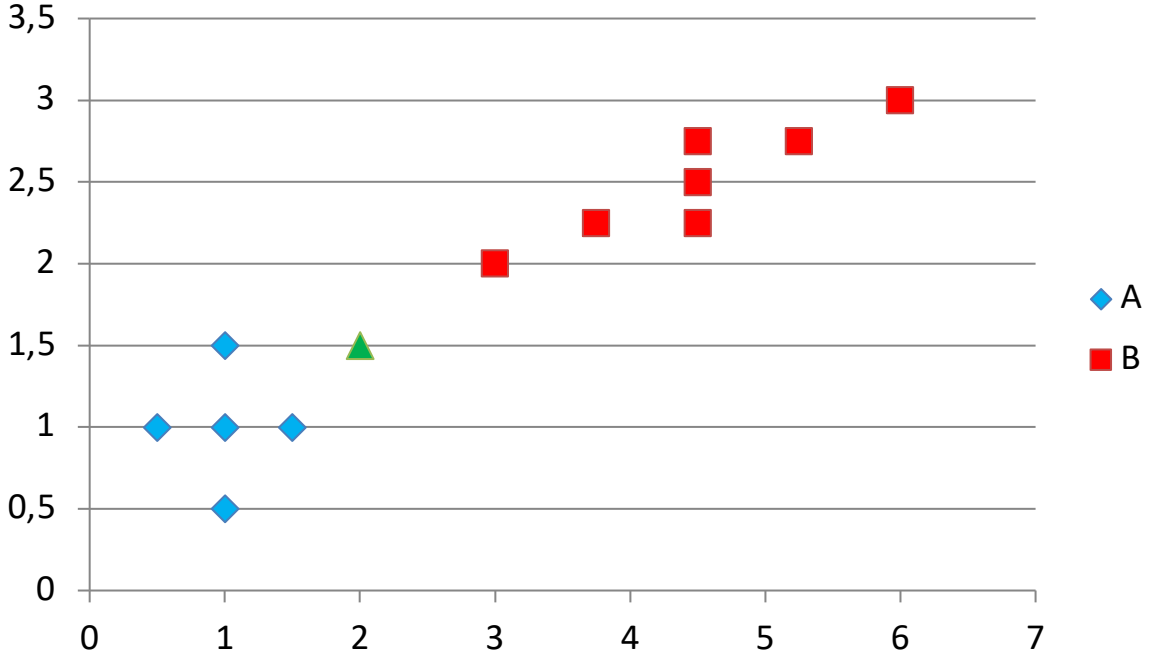
$$d^m\left(\mathbf{x}^{(i)}, \mu_B\right) = 29$$

Consideremos agora que esse ponto se mova para cima...

$$\Sigma_A = \begin{bmatrix} 0.1 & 0 \\ 0 & 0.1 \end{bmatrix} \qquad \Sigma_B = \begin{bmatrix} 0.80 & 0.27 \\ 0.27 & 0.11 \end{bmatrix}$$

$$\Sigma_A^{-1} = \begin{bmatrix} 10 & 0 \\ 0 & 10 \end{bmatrix} \qquad \Sigma_B^{-1} = \begin{bmatrix} 7 & -17 \\ -17 & 50 \end{bmatrix}$$

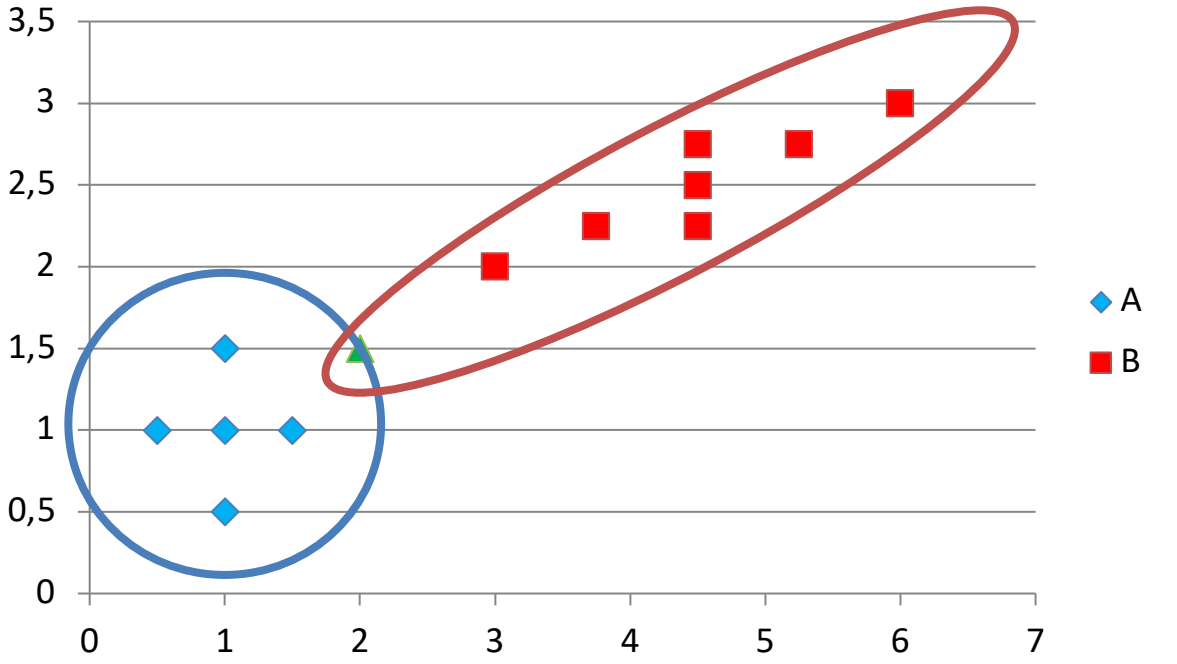
Exemplo Pedagógico



Qual o grupo mais próximo?

$$\mathbf{x}^{(i)} = [2 \quad 1.5]^T$$

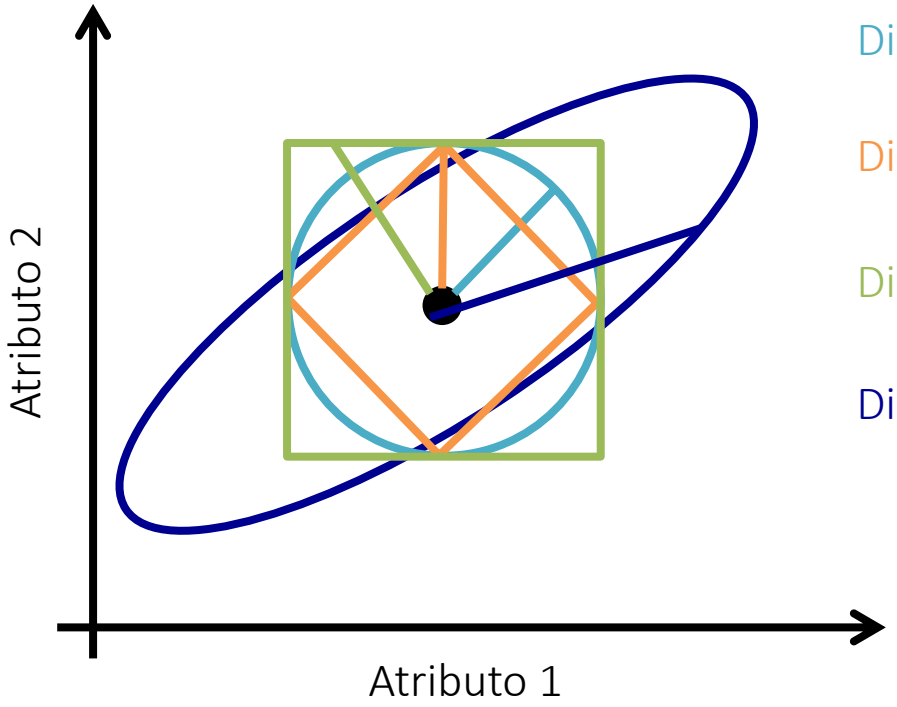
Exemplo Pedagógico



$$d^m\left(\mathbf{x}^{(i)}, \boldsymbol{\mu}_A\right) = 12.5$$
$$d^m\left(\mathbf{x}^{(i)}, \boldsymbol{\mu}_B\right) = 8.80$$

- Problema da distância de Mahalanobis?
  - Custo computacional (inverter matriz de covariância)

Visualização Geométrica



Distância Euclidiana

Distância de Manhattan

Distância Suprema

Distância de Mahalanobis

a.4) Correlação Linear de Pearson

$$corr\left(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}\right) == \frac{cov\left(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}\right)}{\sigma_{\mathbf{x}^{(i)}} \sigma_{\mathbf{x}^{(j)}}}$$

- Varia em [-1,+1]
- Para se transformar em medida de similaridade:

$$s(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = \left|corr\left(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}\right)\right|$$

  - Nova faixa de valores = $[0, 1]$

a.5) Cosseno

- Correlação de Pearson enxerga as instâncias como sequências de valores e captura a tendência linear destas sequências
  - Não trata os valores como assimétricos
  - Valores nulos interferem no cálculo
- Similaridade Cosseno , embora matematicamente similar, possui características diferentes:

$$\cos\left(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}\right) = \frac{{\mathbf{x}^{(i)}}^T \mathbf{x}^{(j)}}{\left\|\mathbf{x}^{(i)}\right\| \left\|\mathbf{x}^{(j)}\right\|}$$

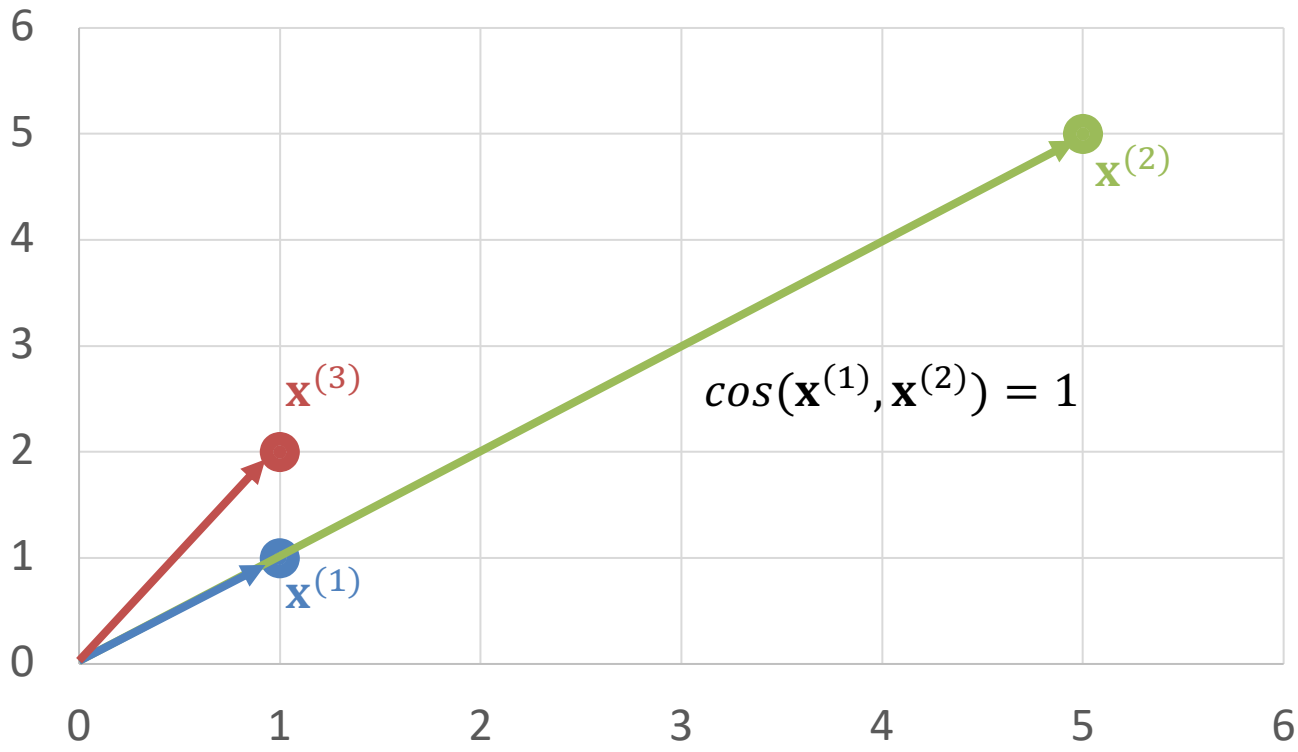
Lembrando que ${\mathbf{x}^{(i)}}^T \mathbf{x}^{(j)} = \mathbf{x}^{(i)} \circ \mathbf{x}^{(j)}$

Cosseno

- Apropriado para **atributos assimétricos**
  - Muito utilizada em NLP
    - Alta dimensionalidade e esparsidade
      - Muitos atributos, poucos não-nulos
  - Mede o cosseno do ângulo entre os respectivos versores!
    - Assim como a correlação, o cosseno mede similaridade através de:

$$s(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = \left| cos\left(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}\right) \right|$$

Exemplo (Gráfico):



$$cos(\mathbf{x}^{(1)}, \mathbf{x}^{(3)}) = cos(\mathbf{x}^{(2)}, \mathbf{x}^{(3)}) = 0.95$$

(ângulo ≈18 graus)

Exemplo (Numérico)

- Considere as instâncias $\mathbf{x}^{(1)}$ e $\mathbf{x}^{(2)}$ abaixo:

$$\mathbf{x}^{(1)} = [3 \quad 2 \quad 0 \quad 5 \quad 0 \quad 0 \quad 0 \quad 2 \quad 0 \quad 0]^T$$

$$\mathbf{x}^{(2)} = [1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 1 \quad 0 \quad 2]^T$$

$$\mathbf{x}^{(1)^T}\mathbf{x}^{(2)} = (3\times1) + (2\times0) + (0\times0) + (5\times0) + (0\times0) + (0\times0)+(0\times0) + (2\times1) + (0\times0) + (0\times2) = 5$$

$$\left\|\mathbf{x}^{(1)}\right\| = \sqrt{3^2 + 2^2 + 0^2 + 5^2 + 0^2 + 0^2 + 0^2 + 2^2 + 0^2 + 0^2} = \sqrt{42} = 6.48$$

$$\left\|\mathbf{x}^{(2)}\right\| = \sqrt{1^2 + 0^2 + 0^2 + 0^2 + 0^2 + 0^2 + 0^2 + 1^2 + 0^2 + 2^2} = \sqrt{6} = 2.45$$

$$cos\left(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}\right) = \frac{5}{6.48\times2.45} = 0.315$$

b) Atributos Discretos

Motivação

	Sexo	País	Estado Civil	Comprar?
$\mathbf{x}^{(1)}$	M	França	Solteiro	Sim
$\mathbf{x}^{(2)}$	M	China	Divorciado	Sim
$\mathbf{x}^{(3)}$	F	EUA	Divorciado	Não
$\mathbf{x}^{(4)}$	F	Inglaterra	Viúvo	Sim
$\mathbf{x}^{(5)}$	F	França	Solteiro	Sim
$\mathbf{x}^{(6)}$	M	Alemanha	Casado	Sim
$\mathbf{x}^{(7)}$	M	Brasil	Casado	Não
$\mathbf{x}^{(8)}$	F	Brasil	Casado	Não
$\mathbf{x}^{(9)}$	M	Inglaterra	Divorciado	Sim
$\mathbf{x}^{(10)}$	M	EUA	Solteiro	Não

$$d(\mathbf{x}^{(1)}, \mathbf{x}^{(6)}) = ?$$
$$d(\mathbf{x}^{(1)}, \mathbf{x}^{(7)}) = ?$$

b.1) Atributos Binários:

- Calcular a distância entre: $\mathbf{x}^{(1)} = [1\ 0\ 0\ 1\ 1\ 0\ 0\ 1\ 0\ 0]^T$

$$\mathbf{x}^{(2)} = [0\ 0\ 0\ 1\ 0\ 1\ 1\ 0\ 0\ 0]^T$$

- Usando uma tabela de contingência, temos:

Objeto $\mathbf{x}^{(j)}$				
Objeto $\mathbf{x}^{(i)}$		1	0	Total
	1	m_{11}	m_{10}	$m_{11} + m_{10}$
	0	m_{01}	m_{00}	$m_{01} + m_{00}$
	Total	$m_{11} + m_{01}$	$m_{10} + m_{00}$	m

$$s^{SM}\left(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}\right) = \frac{m_{11} + m_{00}}{m_{11} + m_{10} + m_{01} + m_{00}} = \frac{m_{11} + m_{00}}{m}$$

Simple-Matching Coefficient
(Zubin, 1938)

$$1 - s^{SM}\left(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}\right) = \frac{m_{10} + m_{01}}{m} = \frac{d^{Hamming}(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})}{m}$$

- Entretanto, como vimos na última aula, os atributos binários podem ser simétricos ou assimétricos!

- Exemplo: considere 3 instâncias que apresentam (1) ou não (0) dez sintomas para uma determinada doença:

$$\mathbf{x}^{(1)} = [1\ 0\ 0\ 1\ 1\ 0\ 0\ 1\ 0\ 1]$$

$$\mathbf{x}^{(2)} = [1\ 0\ 1\ 1\ 0\ 1\ 1\ 1\ 0\ 0]$$

$$\mathbf{x}^{(3)} = [0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0]$$

$$s^{SM}\left(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}\right) = 0.5$$

$$s^{SM}\left(\mathbf{x}^{(1)}, \mathbf{x}^{(3)}\right) = 0.5$$

Perceberam o problema?

- Para atributos binários assimétricos, pode-se usar, por exemplo, o Coeficiente de Jaccard (1908):

$$s^{Jaccard}\left(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}\right) = \frac{m_{11}}{m_{11} + m_{10} + m_{01}}$$

  - Foca apenas nos casamentos $1 - 1$
  - Despreza casamentos $0 - 0$

- Existem várias medidas similares, mas SM e Jaccard são as mais utilizadas

Outro Exemplo

$$\mathbf{x}^{(1)} = [1\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0]$$

$$\mathbf{x}^{(2)} = [0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 1\ \ 0\ \ 0\ \ 1]$$

$$m_{01} = 2$$
$$m_{10} = 1$$
$$m_{00} = 7$$
$$m_{11} = 0$$

$$s^{SM}\left(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}\right) = \frac{m_{11} + m_{00}}{m} = \frac{(0+7)}{10} = 0.7$$

$$s^{Jaccard}\left(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}\right) = \frac{m_{11}}{m_{11} + m_{01} + m_{10}} = \frac{0}{(0+2+1)} = 0$$

c) Atributos Mistos (Contínuos e Discretos)

- Método de Gower (1971):

$$s^{Gower}\left(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}\right) = \frac{1}{m}\sum_{k=1}^{m} S_{ijk}$$

  - Para atributos nominais/binários:

$$\begin{cases} x_k^{(i)} = x_k^{(j)} \rightarrow S_{ijk} = 1 \\ x_k^{(i)} \neq x_k^{(j)} \rightarrow S_{ijk} = 0 \end{cases}$$

  - Para atributos ordinais ou contínuos:

$$S_{ijk} = 1 - \left[\frac{\left|x_k^{(i)} - x_k^{(j)}\right|}{\max(k) - \min(k)}\right]$$

Aula de Hoje

- Introdução ao Aprendizado Supervisionado
- Aprendizado Baseado em Instâncias (e Distâncias)
- Algoritmos Baseados em Instâncias
  - k-NN

k-NN

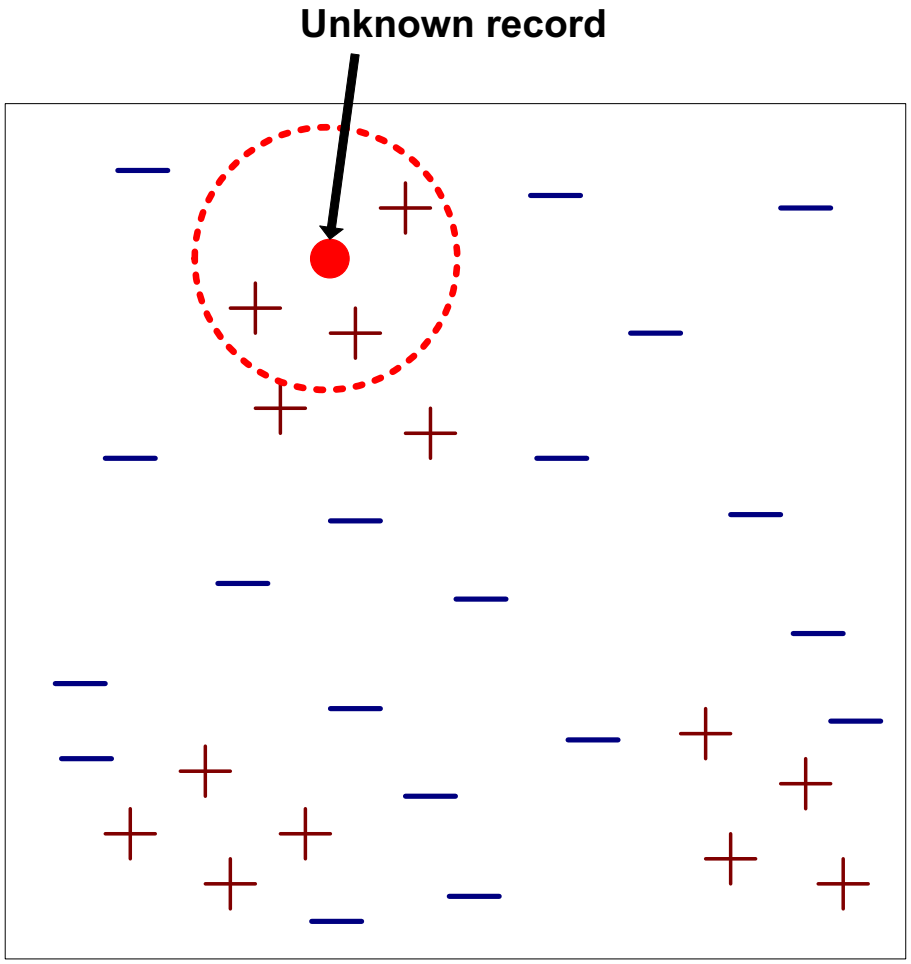
- k-Nearest Neighbors
  - k-vizinhos mais próximos
  - Utiliza as k instâncias mais próximas (similares) para prever o atributo meta de uma instância ainda não vista

k-NN

Algorithm 6.2 The k-nearest neighbor classifier.

1: Let k be the number of nearest neighbors and D be the set of training examples.
2: **for** each test instance $z = (\mathbf{x}', y')$ **do**
3: Compute $d(\mathbf{x}', \mathbf{x})$, the distance between z and every example, $(\mathbf{x}, y) \in D$.
4: Select $D_z \subseteq D$, the set of k closest training examples to z.
5: $y' = \underset{v}{\operatorname{argmax}} \sum_{(\mathbf{x}_i, y_i) \in D_z} I(v = y_i)$
6: **end for**

k-NN



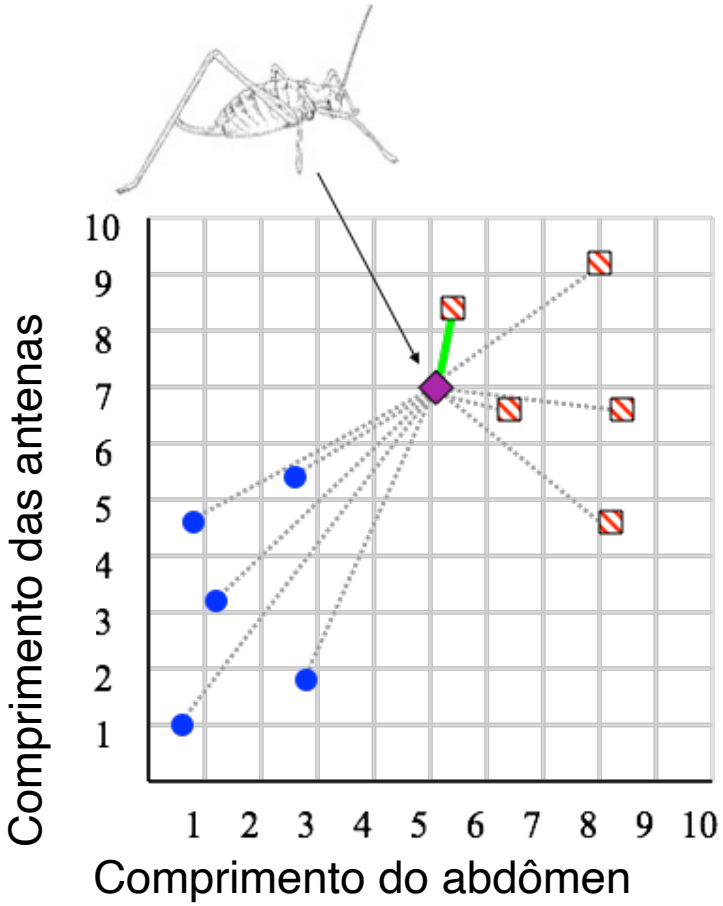
- Necessita de 3 coisas:
  - Base de treinamento
  - Medida de (dis)similaridade
  - Valor de k (número de vizinhos)

- Para classificar uma instância não-vista:
  - Calcule a (dis)similaridade para todas as instâncias de treino
  - Obtenha as k instâncias de treino mais similares (mais próximas)
  - Classifique a instância não vista na classe da maioria dos k vizinhos

k-NN com $k = 1$

Evelyn Fix
1904-1965

Joe Hodges
1922-2000

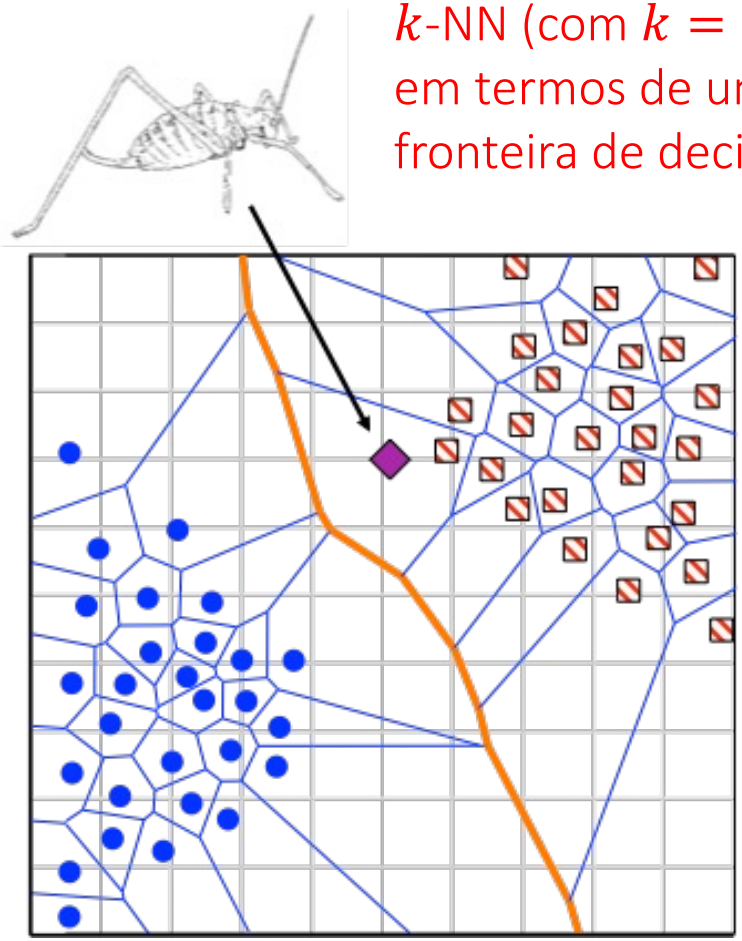


se o exemplo mais próximo ao **exemplo desconhecido** é da classe **Esperança**
então
classe é **Esperança**
senão
classe é **Gafanhoto**

Esperança
Gafanhoto

k-NN com $k = 1$

É possível visualizar o k-NN (com $k = 1$) em termos de uma fronteira de decisão!

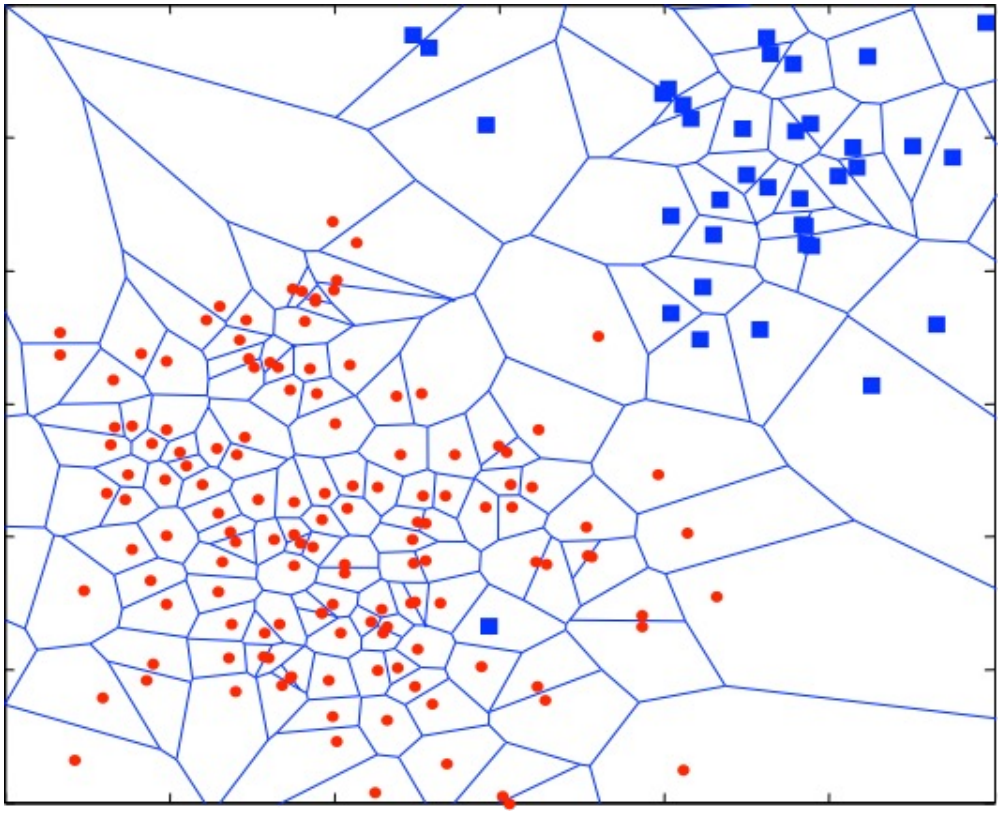


Note que não precisamos construir essas divisões. São apenas fronteiras implícitas que delimitam as zonas *pertencentes* a cada exemplo de treino

Esse tipo de divisão é chamada de Diagrama de Voronoi (ou Regiões de Theissen, ou Mosaico de Dirichlet)

k-NN com $k = 1$

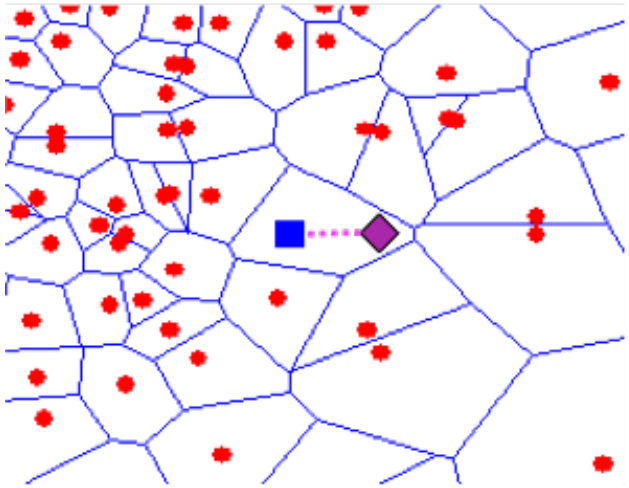
É sensível a *outliers*!



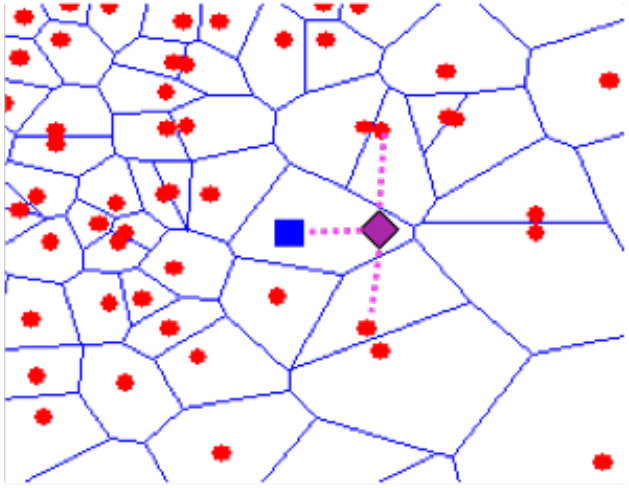
A solução para isso é....

Aumentar o valor de k!

Mede-se a distância para os k exemplos mais próximos e depois basta computar o voto da maioria para definição da classe do exemplo desconhecido

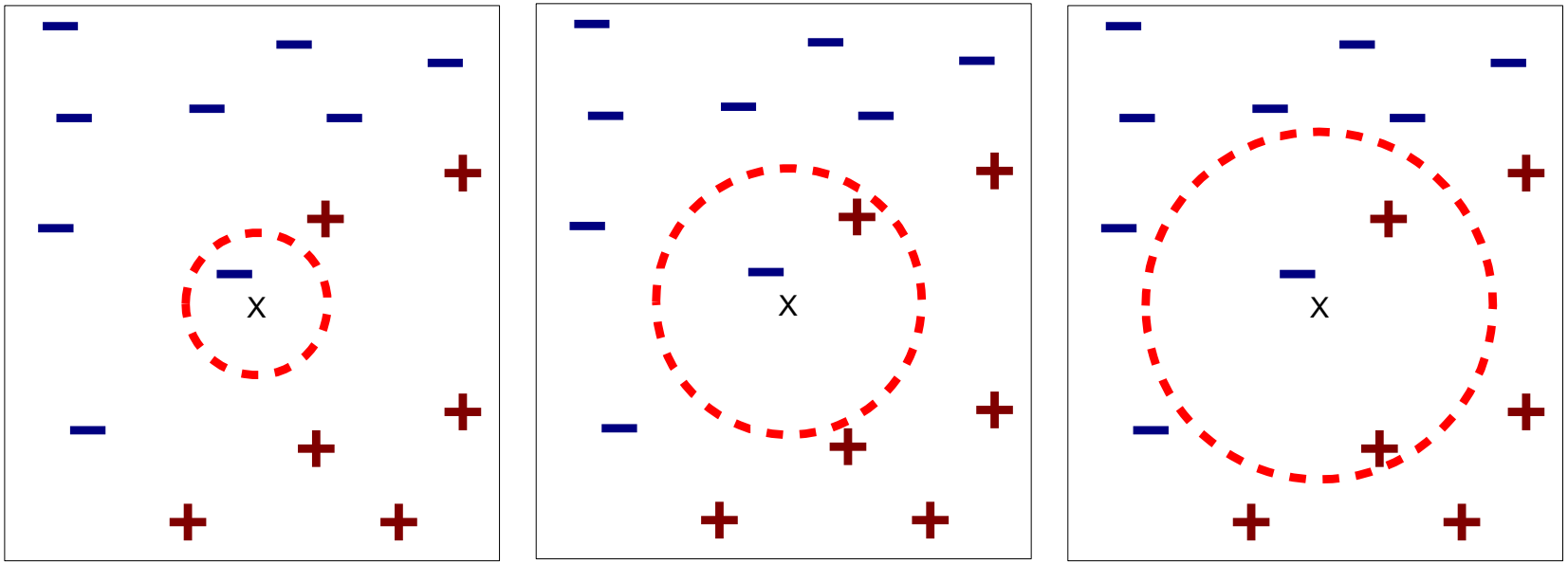


$k = 1$



$k = 3$

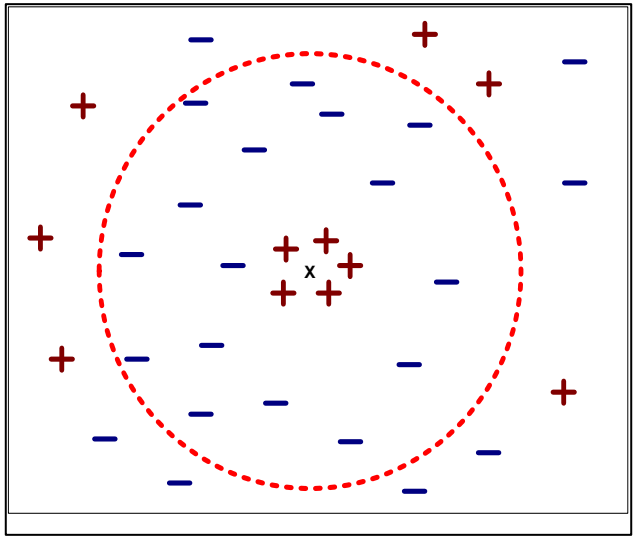
k-NN



(a) 1-nearest neighbor (b) 2-nearest neighbor (c) 3-nearest neighbor

k-NN: visão geométrica para 2 atributos contínuos e dissimilaridade por distância Euclidiana

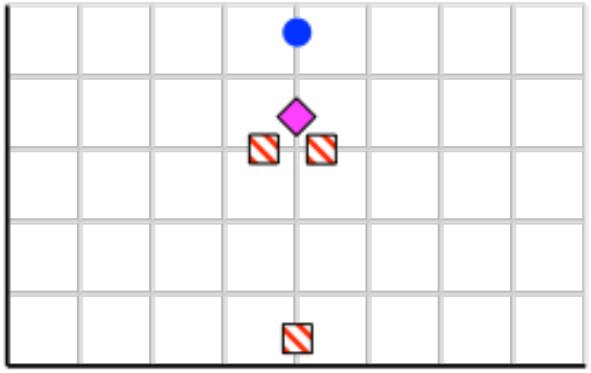
k-NN



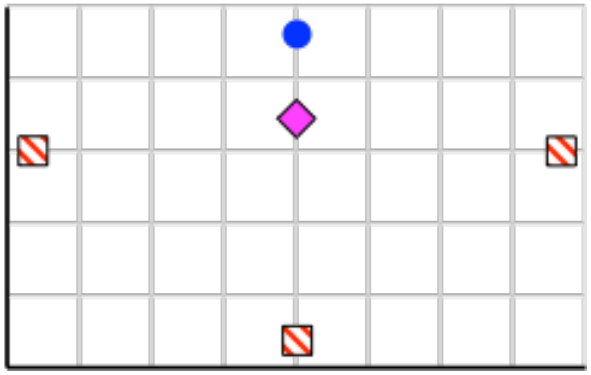
- Sobre a escolha de k:
  - Valor muito pequeno
    - Função de discriminação muito flexível
    - Porém, sensível a ruído
      - Classificação pode ser instável!
    - *Overfitting*!!!!!
  - Valor muito grande
    - Robusto a ruído
    - Porém, vizinhança tende a ser heterogênea
      - Privilegia classe majoritária
      - Reduz flexibilidade da função de discriminação
    - *Underfitting*!!!

Atenção 1!

k-NN é sensível às unidades de medidas utilizadas



Atributo $\mathbf{x}_1$ em centímetros.
Atributo $\mathbf{x}_2$ em reais.
Objeto mais próximo do rosa desconhecido é vermelho



Atributo $\mathbf{x}_1$ em milímetros.
Atributo $\mathbf{x}_2$ em reais.
Objeto mais próximo do rosa desconhecido é azul

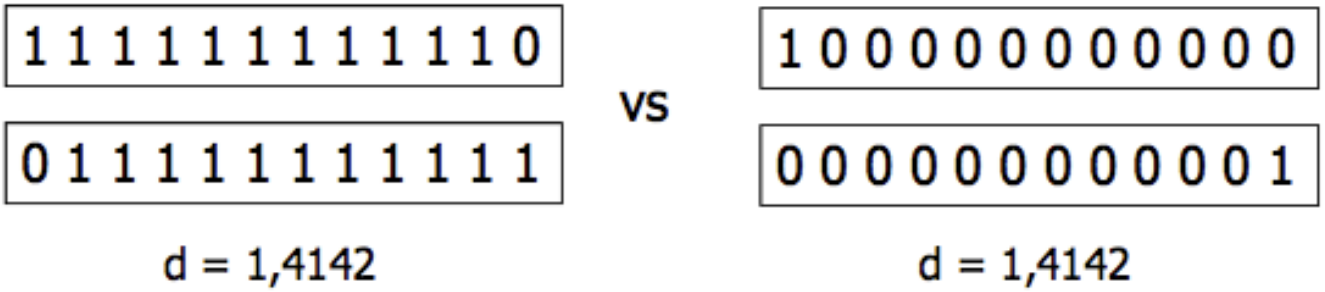
Solução? Normalizar os dados! Possibilidades:

$z_i = (x_i - xmin)/(x_{max} - x_{min})$ → dados entre 0 e 1 (re-escalar)
$z_i = (x_i - \mu_x)/\sigma(x)$ → dados com média zero e desvio 1 (padronizar)

Atenção 2!

- Já vimos que a escolha da medida de (dis)similaridade mais apropriada depende:
  - do(s) tipo(s) dos atributos
  - do domínio de aplicação!
    - Conheça seus dados!!

- Exemplo de escolha inapropriada:
  - Euclidiana para atributos binários assimétricos



Atenção 3!

- Na versão básica do algoritmo, a indicação da classe de cada vizinho possui o mesmo peso
  - 1 voto por vizinho mais próximo
- Isso torna o algoritmo sensível à escolha de k
- Uma alternativa para reduzir esta sensibilidade e permitir, assim, o aumento de k (aumentando a robustez a ruído) é ponderar cada voto pela respectiva distância

$$\hat{f}\left(\mathbf{x}^{(t)}\right) = \underset{y_j}{arg\max} \sum_{\left(\mathbf{x}^{(i)}, f\left(\mathbf{x}^{(i)}\right)\right) \in NN} w_i \times I\left(y_j = f\left(\mathbf{x}^{(i)}\right)\right)$$

$$w_i = \frac{1}{d\left(\mathbf{x}^{(t)}, \mathbf{x}^{(i)}\right)^2}$$

$$I\left(y_j = f\left(\mathbf{x}^{(i)}\right)\right) = \begin{cases} 1 \text{ se } y_j = f\left(\mathbf{x}^{(i)}\right) \\ 0 \text{ se } y_j \neq f\left(\mathbf{x}^{(i)}\right) \end{cases}$$

Atenção 4!

- k-NN é um classificador **lazy**
  - Não constrói modelo e atrasa a discriminação até a chegada dos dados não vistos
  - Isso torna a classificação de novos objetos custosa computacionalmente!
    - Precisa calcular as distâncias de cada objeto a ser classificado para todos os objetos de treino
    - Possível solução?
      - Estruturas de dados eficientes!
        - KD-Tree

k-NN para Regressão

- Adaptação é trivial:

$$\hat{f}\left(\mathbf{x}^{(t)}\right) = \frac{\sum_{\left(\mathbf{x}^{(i)}, f\left(\mathbf{x}^{(i)}\right)\right) \in NN} w_i \times f\left(\mathbf{x}^{(i)}\right)}{\sum_i w_i}$$

$$w_i = \frac{1}{d\left(\mathbf{x}^{(t)}, \mathbf{x}^{(i)}\right)^2}$$

k-NN: Sumário

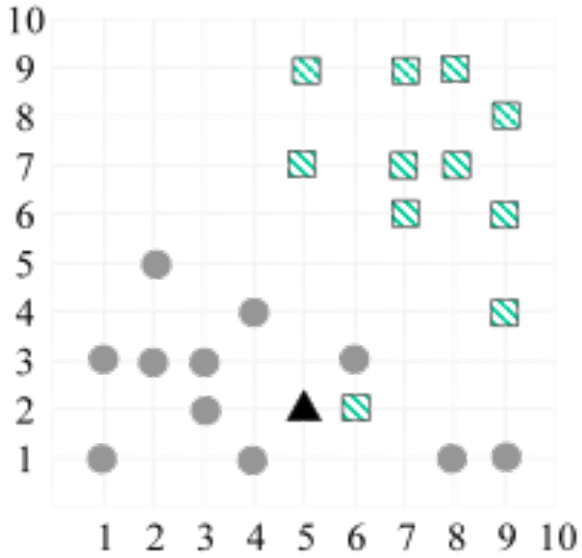
- Características sensíveis ao projeto:
  - Escolha de k
  - Escolha da medida de (dis)similaridade
- Pode ter poder de classificação elevado
  - Função de discriminação muito flexível para k pequeno
- Incrivelmente simples de implementar!
  - Tarefa para casa: implemente o $\mathbf{k}$-NN!

k-NN: Sumário

- Sensível a ruído
  - Principalmente para k pequeno (*overfitting*!)
  - Robusto com k grande, porém as custas de flexibilidade (*underfitting*!)
- Sensível a atributos irrelevantes
  - Distorcem o cálculo das distâncias
  - Maldição da dimensionalidade
    - Seleção de atributos, análise exploratória dos dados

k-NN: exercício

- Descubra a classe do exemplo desconhecido com $k = 1$ e com $k = 3$
- Utilize a distância Euclidiana e outras duas medidas de sua preferência, comparando os resultados



$\mathbf{x}_1$	$\mathbf{x}_2$	Classe
5	9	□
7	9	□
8	9	□
9	8	□
5	7	□
7	7	□
8	7	□
7	6	□
9	6	□
9	4	□
6	2	□
2	5	○
4	4	○
1	3	○
2	3	○
3	3	○
6	3	○
3	2	○
1	1	○
4	1	○
8	1	○
9	1	○
5	2	?

Sugestão de Leituras

- Seção 5.2 (Tan et al., 2006)
- Capítulo 4 (Faceli et al., 2011)

Créditos e Referências

Slides adaptados dos originais gentilmente cedidos por:

- Rodrigo Coelho Barros (PUCRS)
- André Carvalho, Eduardo Hruschka, Ricardo Campello (ICMC-USP)
- Pang-Ning Tan (Michigan State University)
- Eamon Keogh (University of California at Riverside)
  - http://www.cs.ucr.edu/~eamonn/
  - eamonn@cs.ucr.edu

- TAN, P. N. STEINBACH, M. KUMAR, V. **Introduction to Data Mining.** Addison-Wesley, 2005. 769 p.

- Faceli et al. **Inteligência Artificial: Uma Abordagem de Aprendizado de Máquina**. LTC, 2011. 378 p.