Aprendizado de Máquina II

Introdução aos Métodos Supervisionados

Profª. Renata De Paris

Especialização em Ciência de Dados

UCS
UNIVERSIDADE
DE CAXIAS DO SUL

Roteiro da Aula

- Introdução
  - Definição
  - Exemplo de aplicações
  - Como funciona os algoritmos supervisionados
- Algoritmo Knn
- Avaliação de desempenho
  - Matriz de confusão
  - Validação do conjunto de teste
- Atividade

Classificação: Definição

- Dado um conjunto de registros (*conjunto de treino*)
  - Cada registro contém um conjunto de *atributos*, e um dos atributos é o atributo *classe*.
- Encontrar um *modelo* para o atributo classe como uma função dos valores dos outros atributos.
- Objetivo: aos registros previamente desconhecidos devem ser atribuídas classes, com a maior precisão possível.
  - O conjunto de registros inicial é dividido em conjuntos de treino e teste.
  - O conjunto de treino é usado para construir o modelo e, o de teste, para validá-lo.
  - Um *conjunto de teste* é usado para determinar a precisão do modelo.

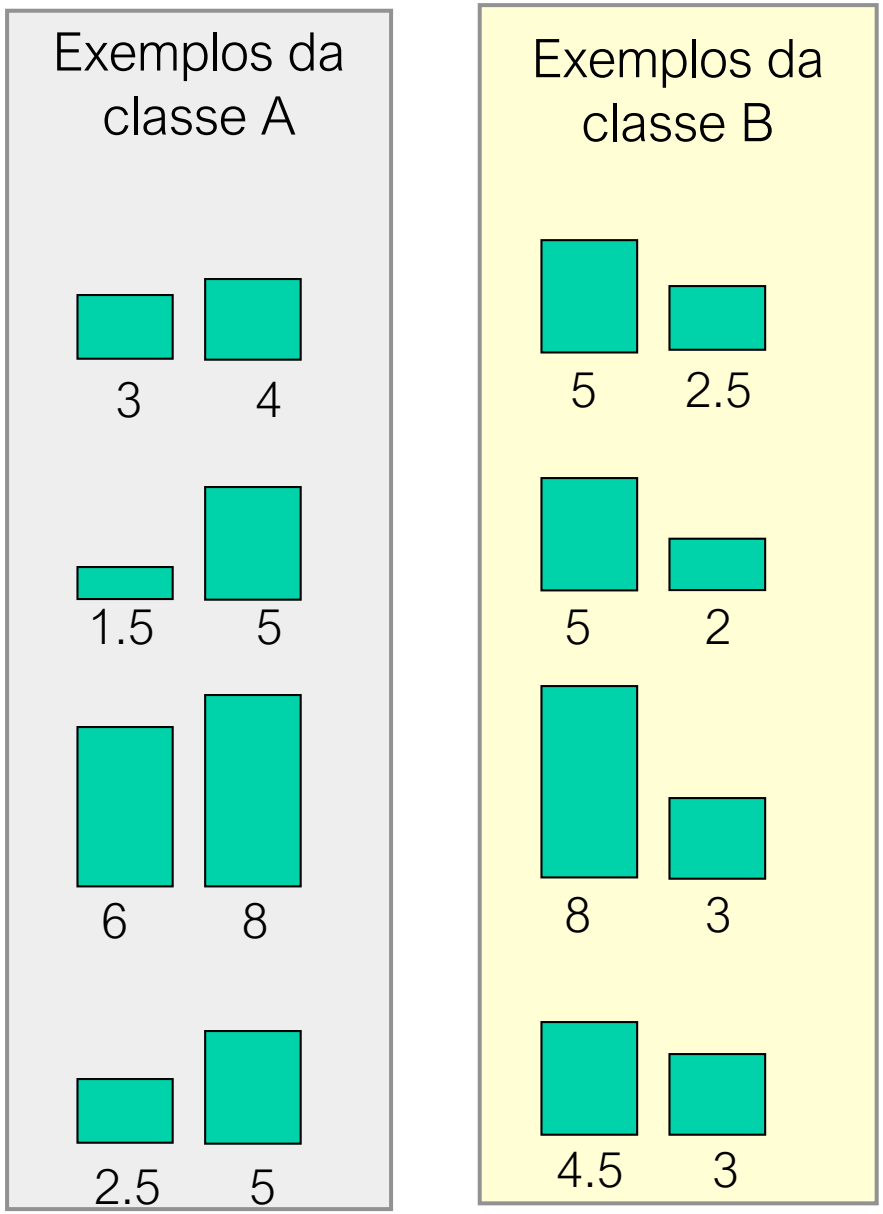
Classificação: Aplicação

AlphaGo - Lee Sedol

Máquina ganhou a partida por 4 X 1.

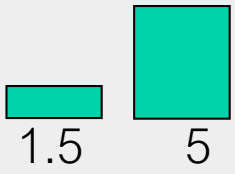
Classificação: Aplicação

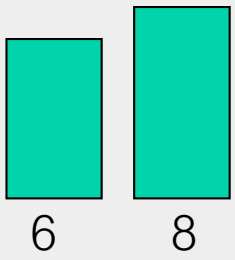
Classificação: Problema do Pombo 1

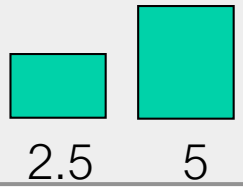


Classificação: Problema do Pombo 1

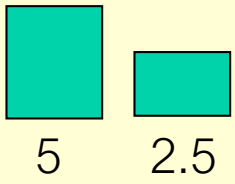
Exemplos da classe A

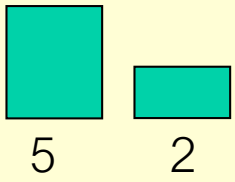


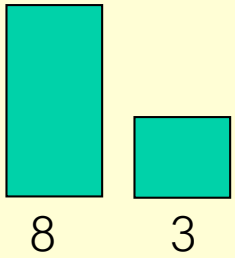


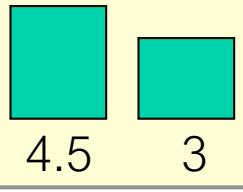


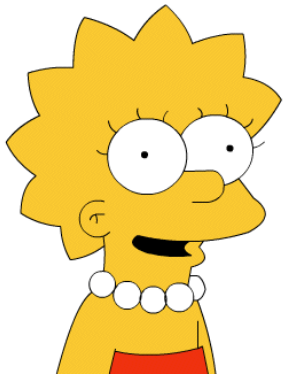
Exemplos da classe B



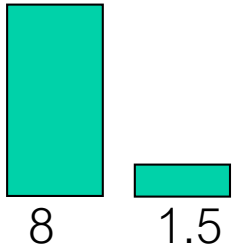




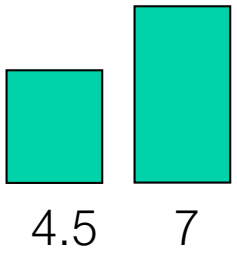




De qual classe é este objeto?

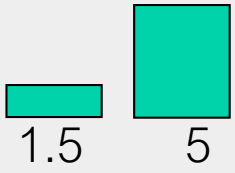


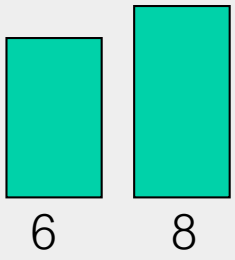
Que tal este, A ou B?

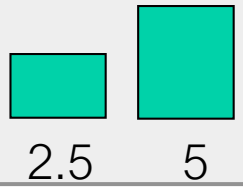


Classificação: Problema do Pombo 1

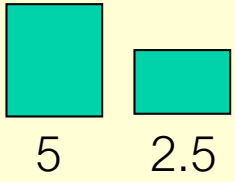
Exemplos da classe A

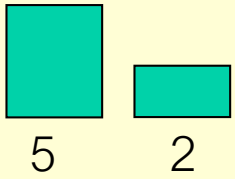


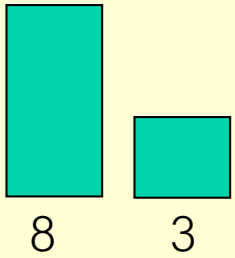


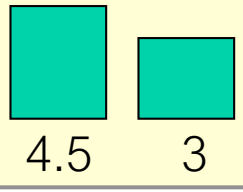


Exemplos da classe B

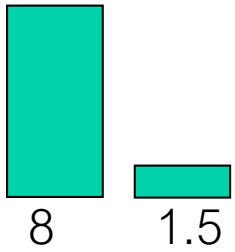








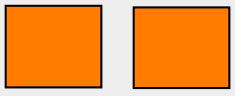
Este é um B!

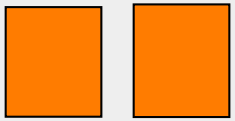


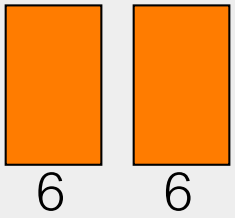
Regra: se a barra esquerda é menor que a direita, é um A, caso contrário é um B.

Classificação: Problema do Pombo 2

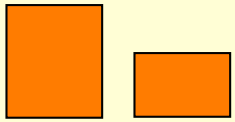
Exemplos da classe A

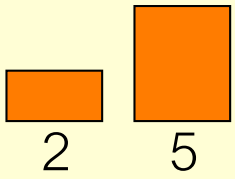


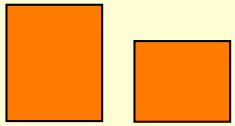


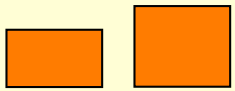


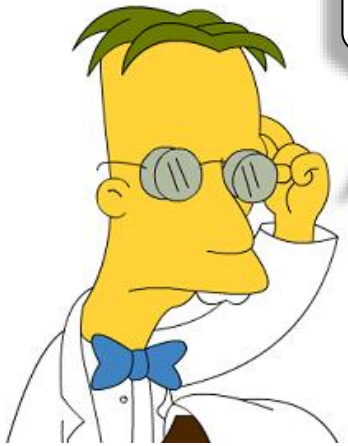
Exemplos da classe B



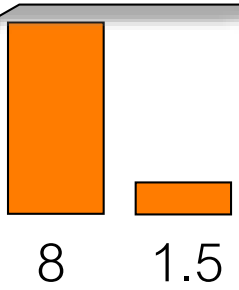




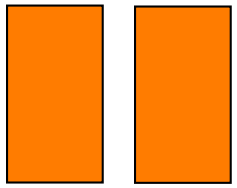




Oh! Este aqui é difícil!

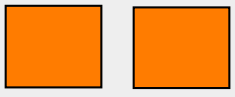


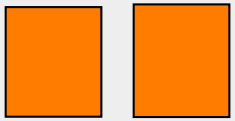
Até eu sei este!

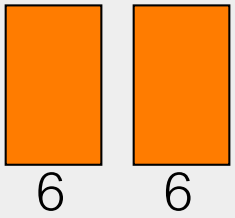


Classificação: Problema do Pombo 2

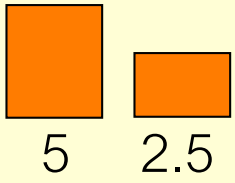
Exemplos da classe A

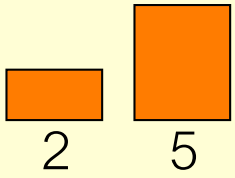


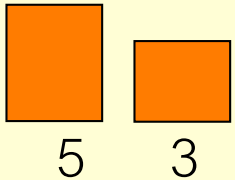


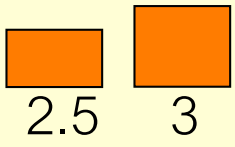


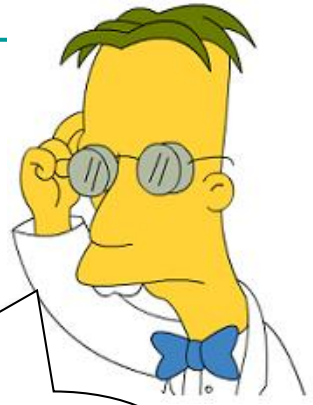
Exemplos da classe B





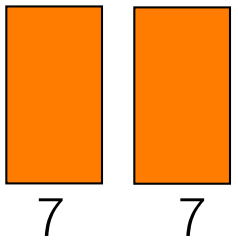






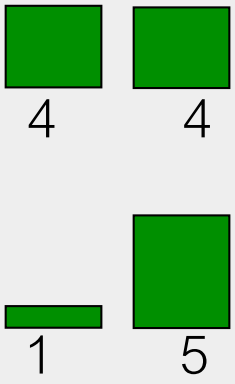
A regra é: se duas barras são iguais em tamanho é um A. Caso contrário é um B.

Então este é um A.

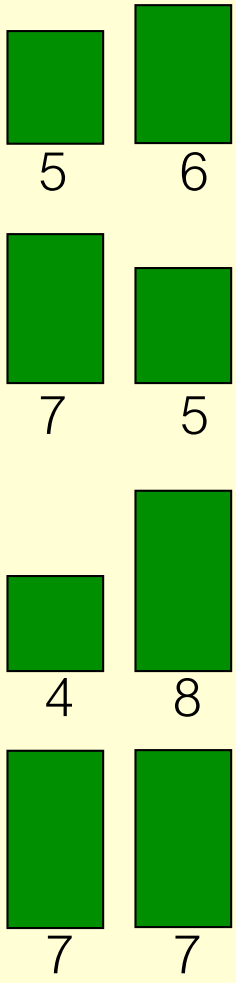


Classificação: Problema do Pombo 3

Exemplos da classe A

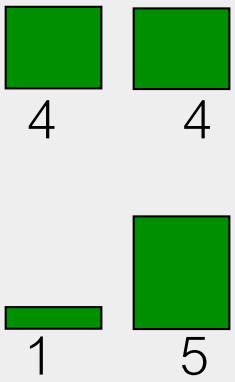


Exemplos da classe B

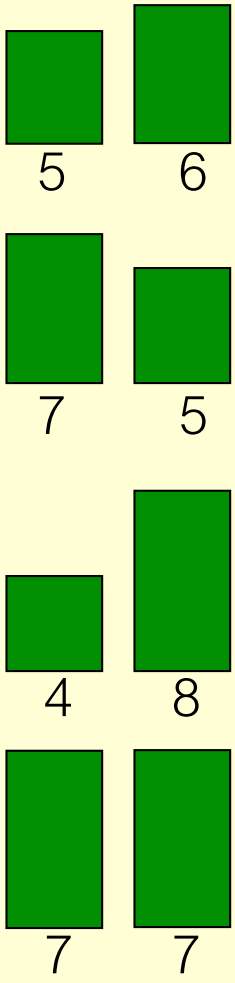


Classificação: Problema do Pombo 3

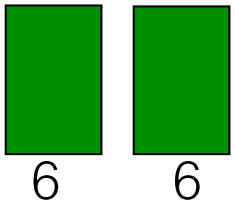
Exemplos da classe A



Exemplos da classe B



É um B!

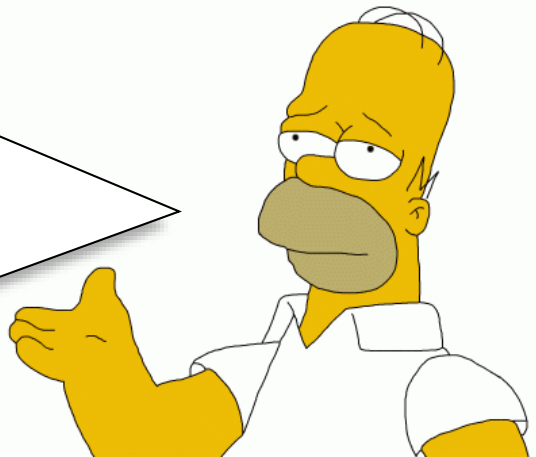


A regra é a seguinte: se o quadrado da soma das duas barras é menor ou igual a 100, é um A. Caso contrário é um B.

Classificação

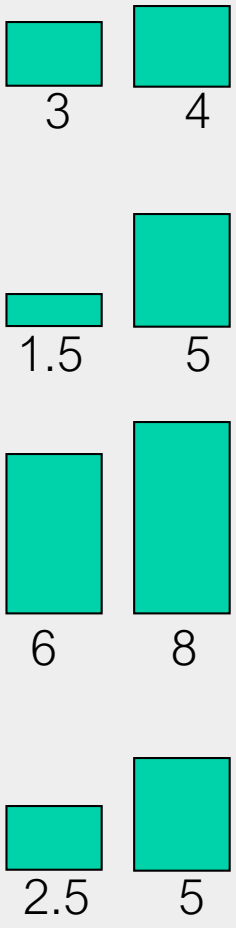
Por que gastamos tanto tempo com este joguinho?

Porque quero mostrar que quase todos os problemas de classificação tem uma interpretação geométrica. Confira os próximos 3 slides…

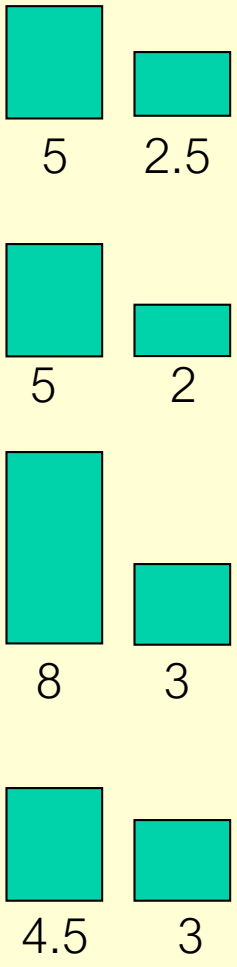


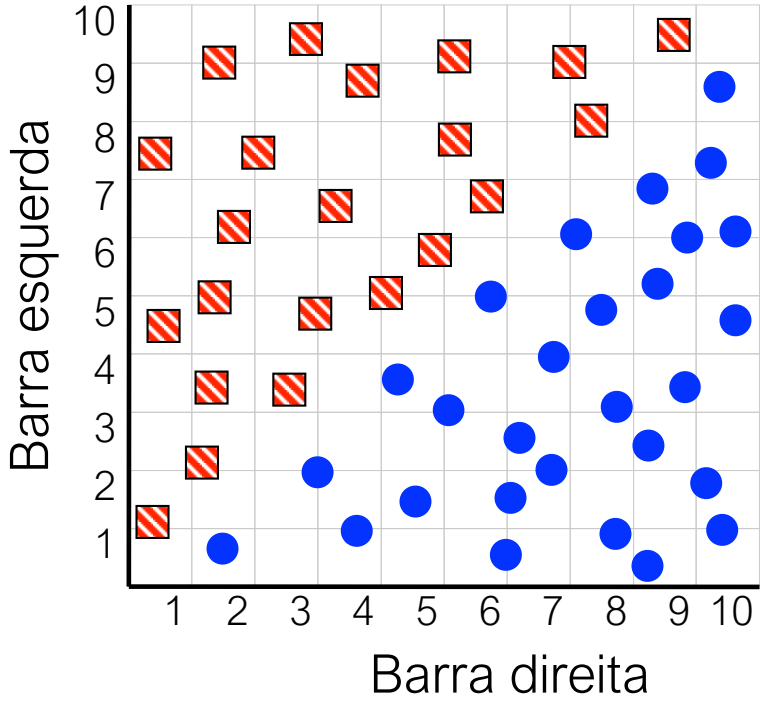
Classificação: Problema do Pombo 1

Exemplos da classe A



Exemplos da classe B

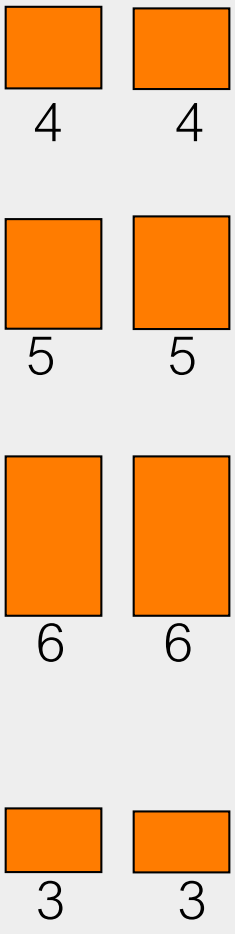




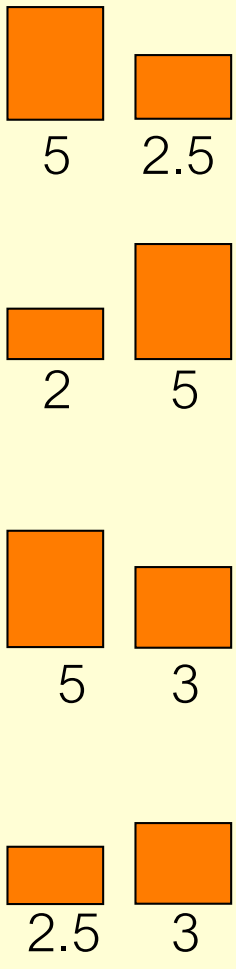
Eis a regra novamente. Se a barra esquerda é menor que a direita, é um A, caso contrário é um B.

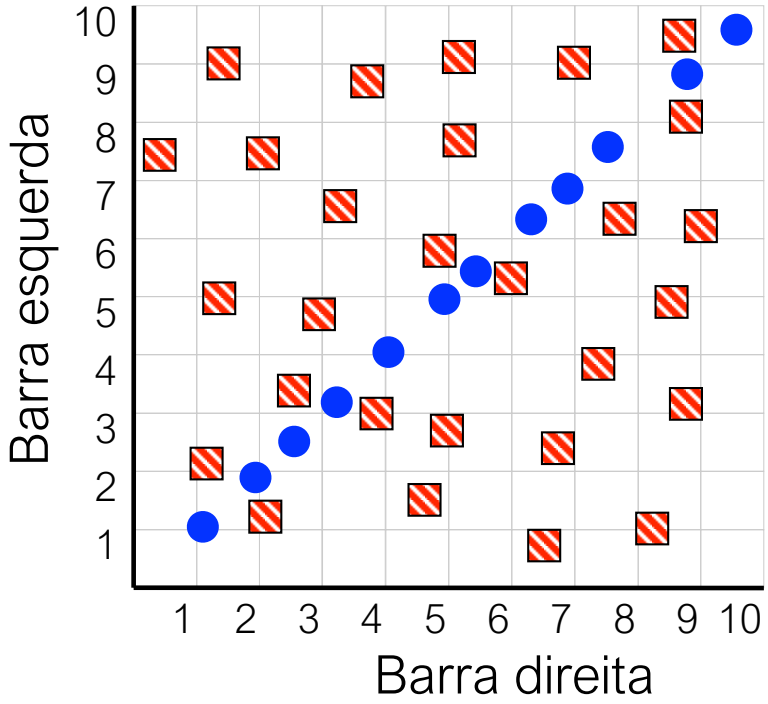
Classificação: Problema do Pombo 2

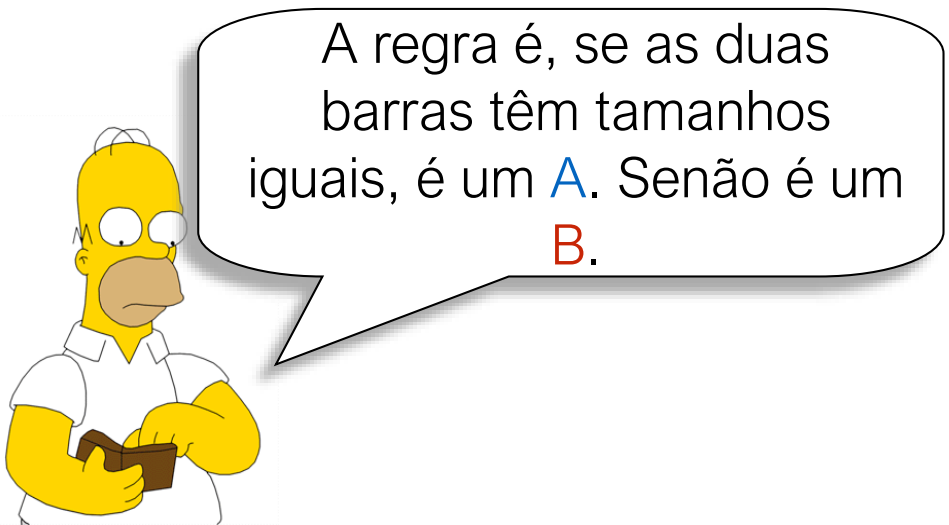
Exemplos da classe A



Exemplos da classe B

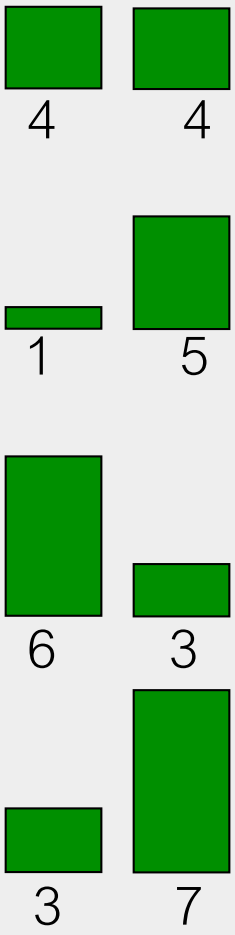




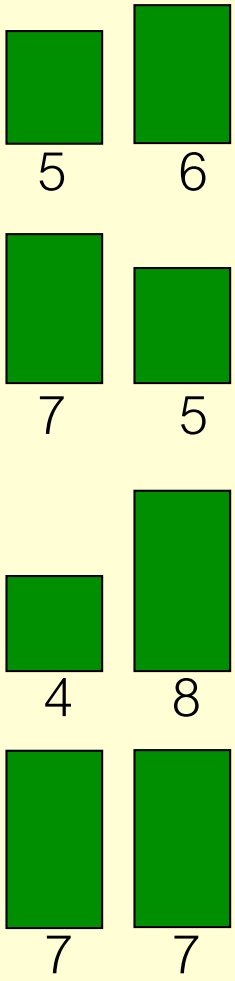


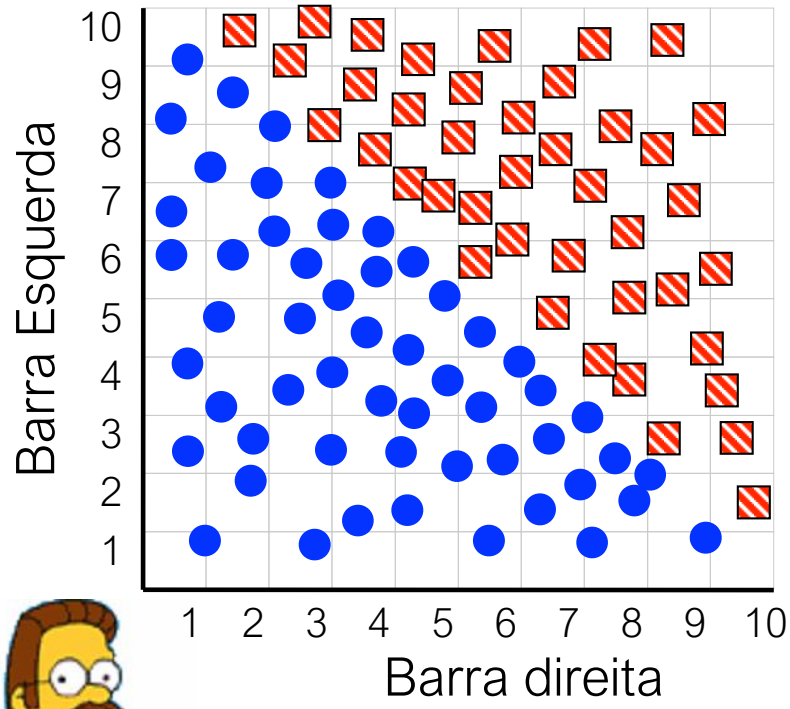
Classificação: Problema do Pombo 3

Exemplos da classe A



Exemplos da classe B





A regra novamente:
Se o quadrado da soma das duas barras é menor ou igual a 100, é um A. Senão é um B.

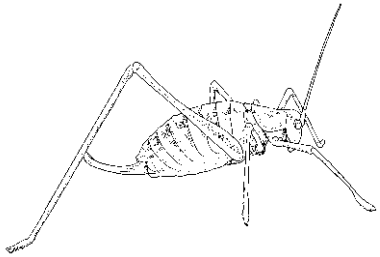
Métodos Preditivos

- **Métodos Baseados em Distância**
  - Algoritmo *k*-NN
- **Métodos Baseados em Procura**
  - Árvores de Decisão e Regressão
  - Regras de Decisão
- **Métodos Probabilísticos**
  - Classificador *Naive* Bayes
  - Redes Bayesianas para Classificação
- **Métodos Baseados em Otimização**
  - Redes Neurais
  - SVM: máquinas de vetores de suporte
- **Outros Métodos**
  - Algoritmos Genéticos
  - Conjuntos Fuzzy

Classificação: O Problema (definição informal)

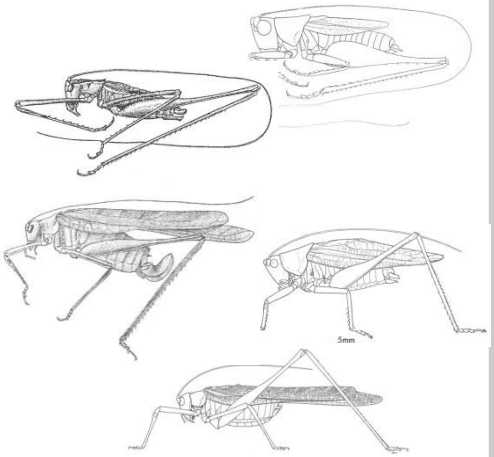
- Dada uma coleção de dados detalhados (neste caso 5 exemplos de Esperança e 5 do Gafanhoto), decida a qual tipo de inseto o exemplo não rotulado pertence.

Obs: Esperança = tipo de gafanhoto verde

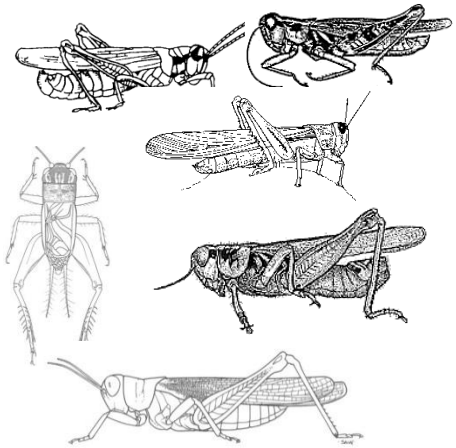


Esperança ou Gafanhoto?

Esperança



Gafanhoto

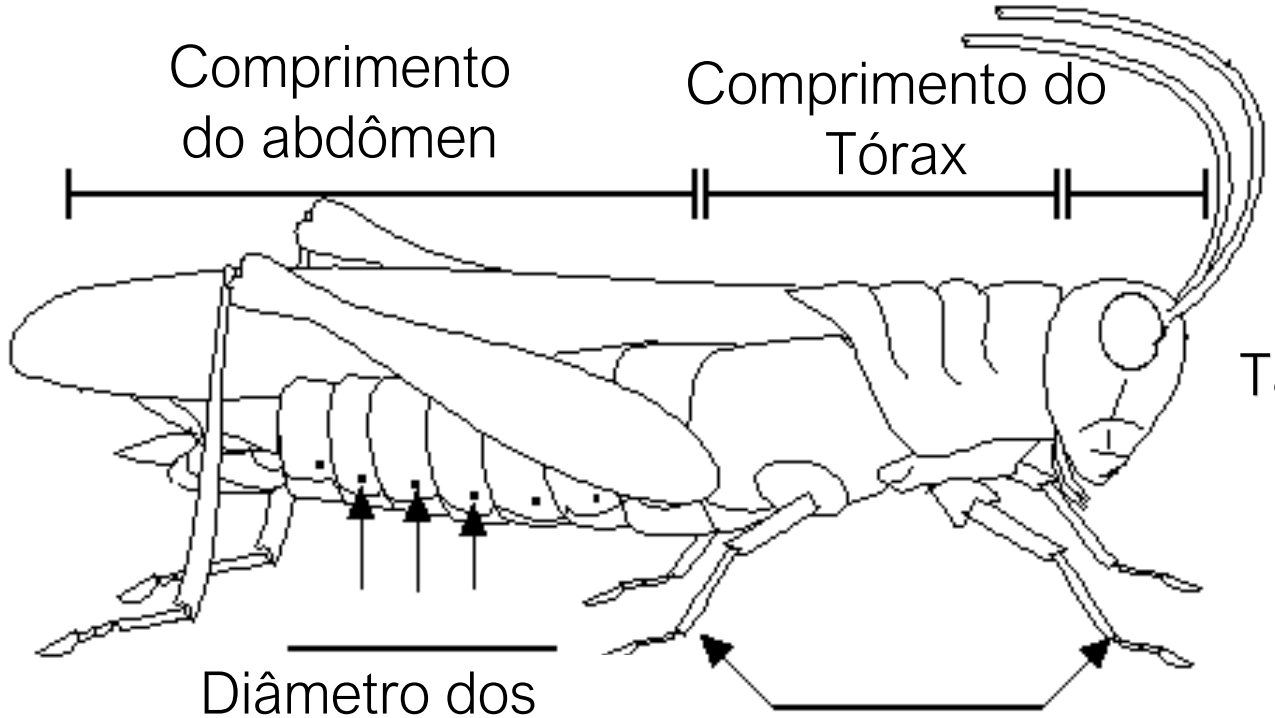


Classificação: Domínio de Interesse

Medir Características

Cor: {Verde, Marrom, Cinza, Outra}

Tem asas?



Classificação: Armazenamento de Características em Datasets

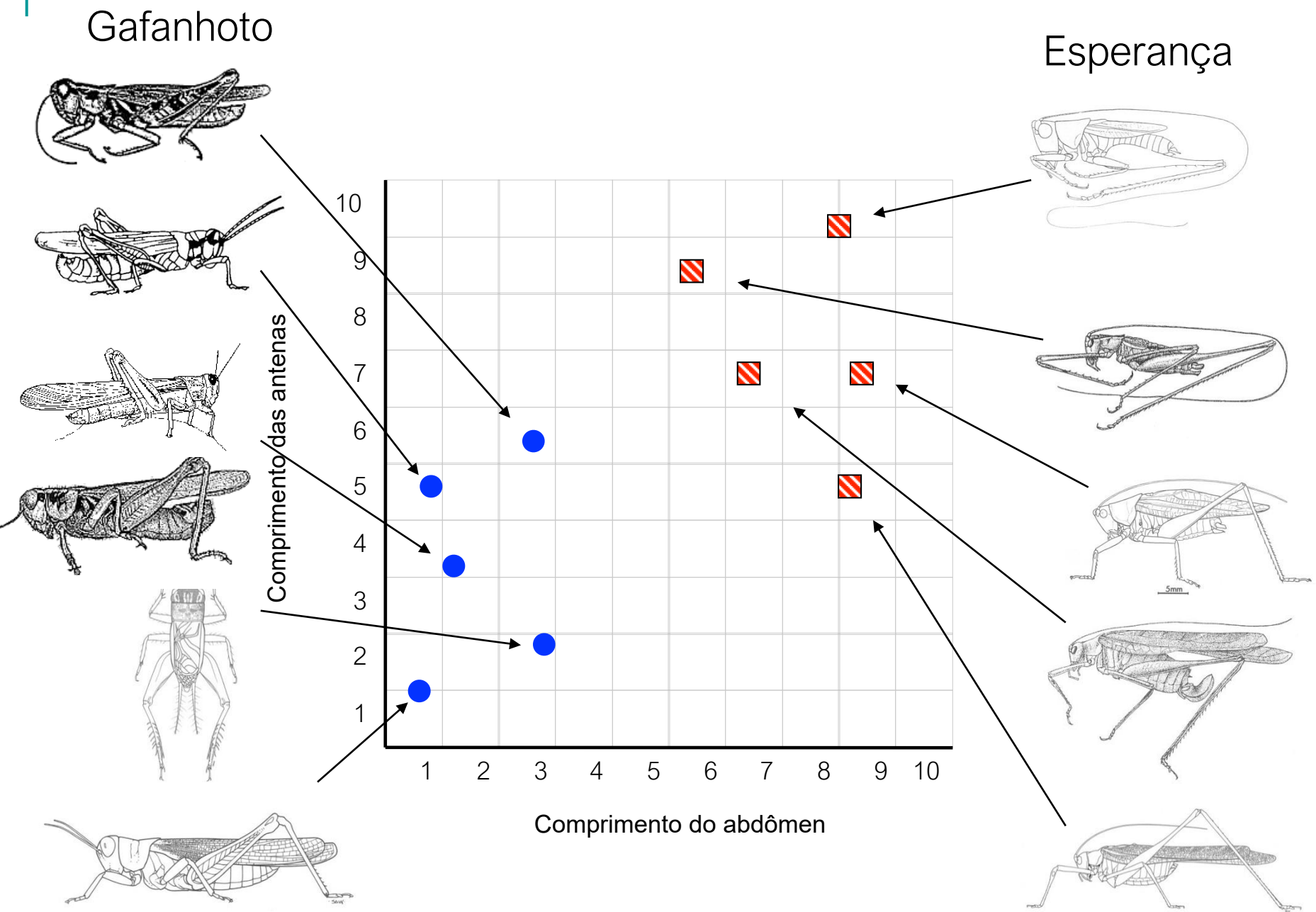
O problema de classificação agora pode ser expresso da seguinte forma:

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

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**

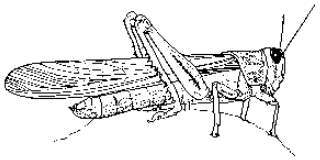
Exemplo não visto =	*11*	5.1	7.0	????????

Classificação: Representação das Características



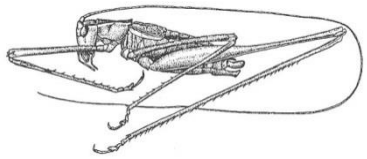
Classificação: Mais exemplos na Base

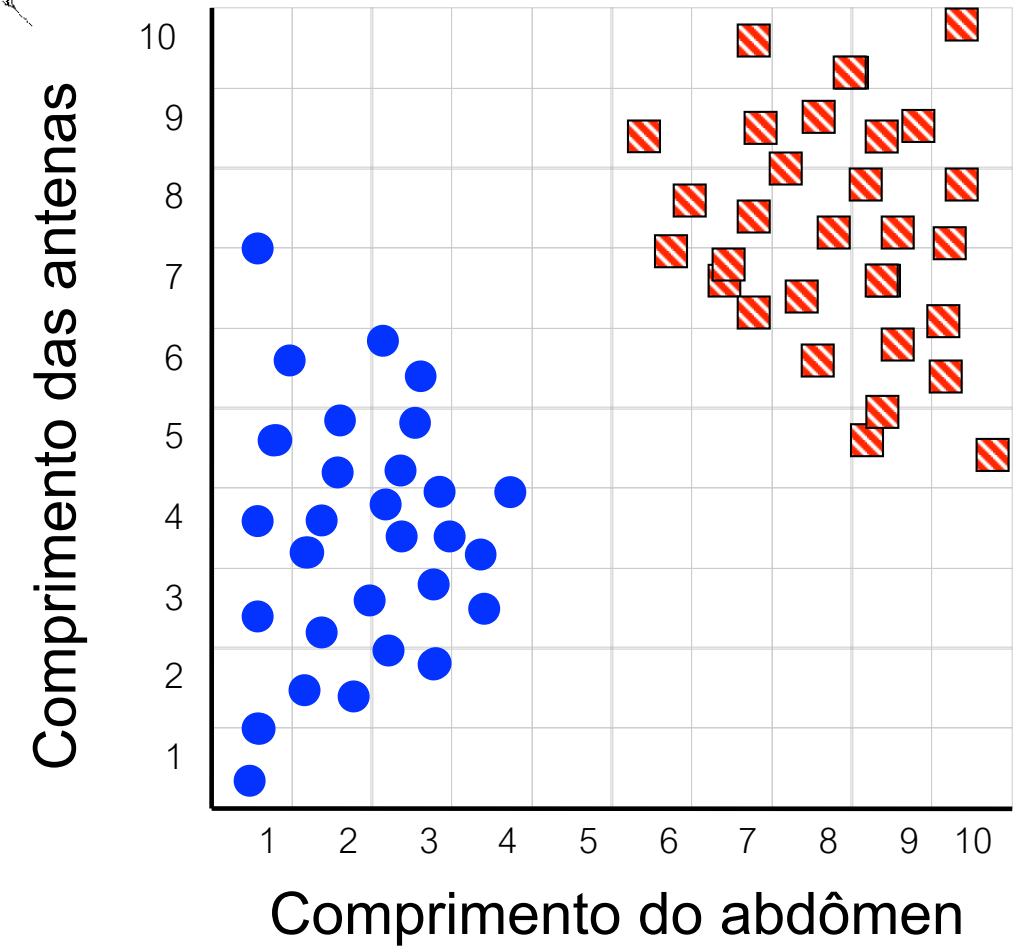
Gafanhoto



Suponha que a base tenha mais exemplos, como apresentado abaixo

Esperança

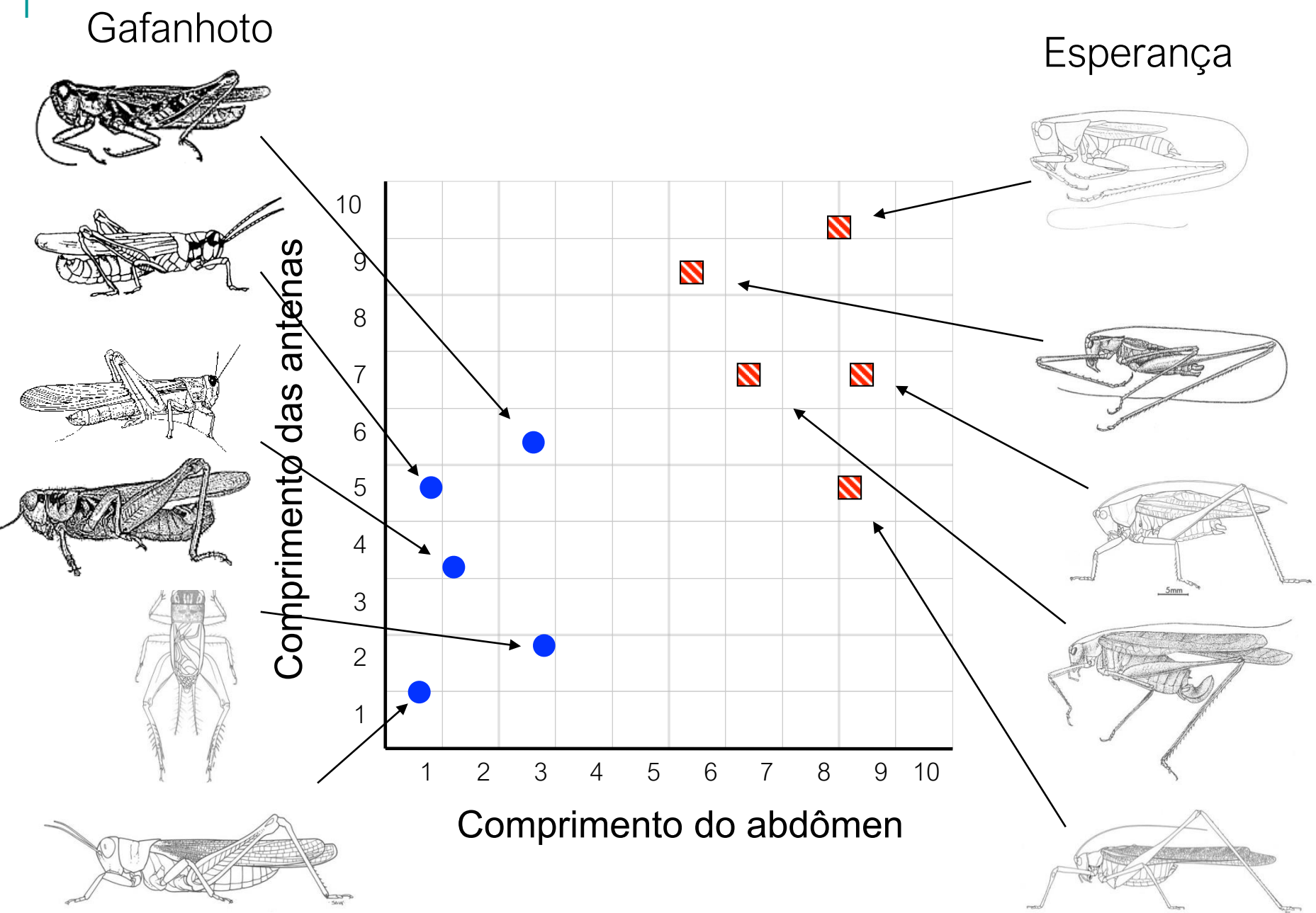




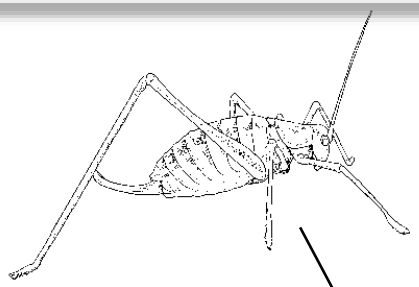
Cada um destes objetos de dados é chamado de:

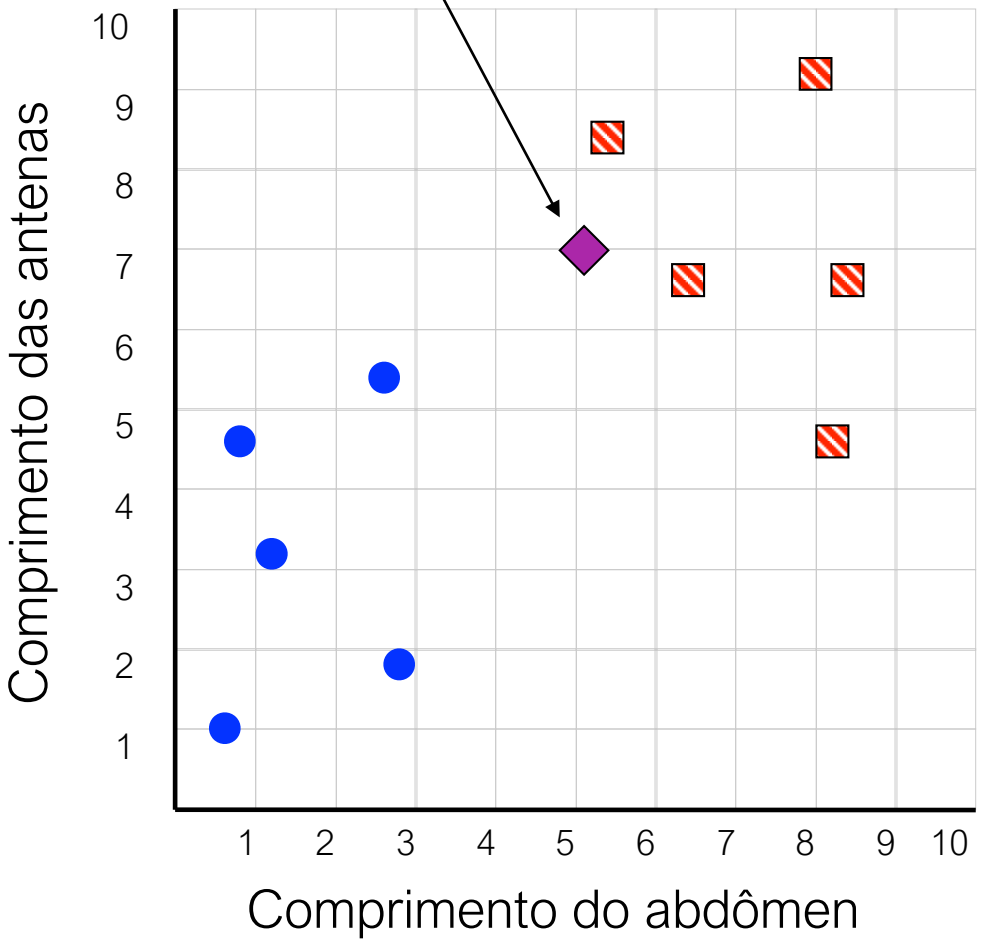
- exemplo (de treino)
- instância
- linha
- tupla
- exemplar
- objeto

Classificação: Representação das Características



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





- Podemos "projetar" o exemplo não visto antes dentro do mesmo espaço que a base de dados.

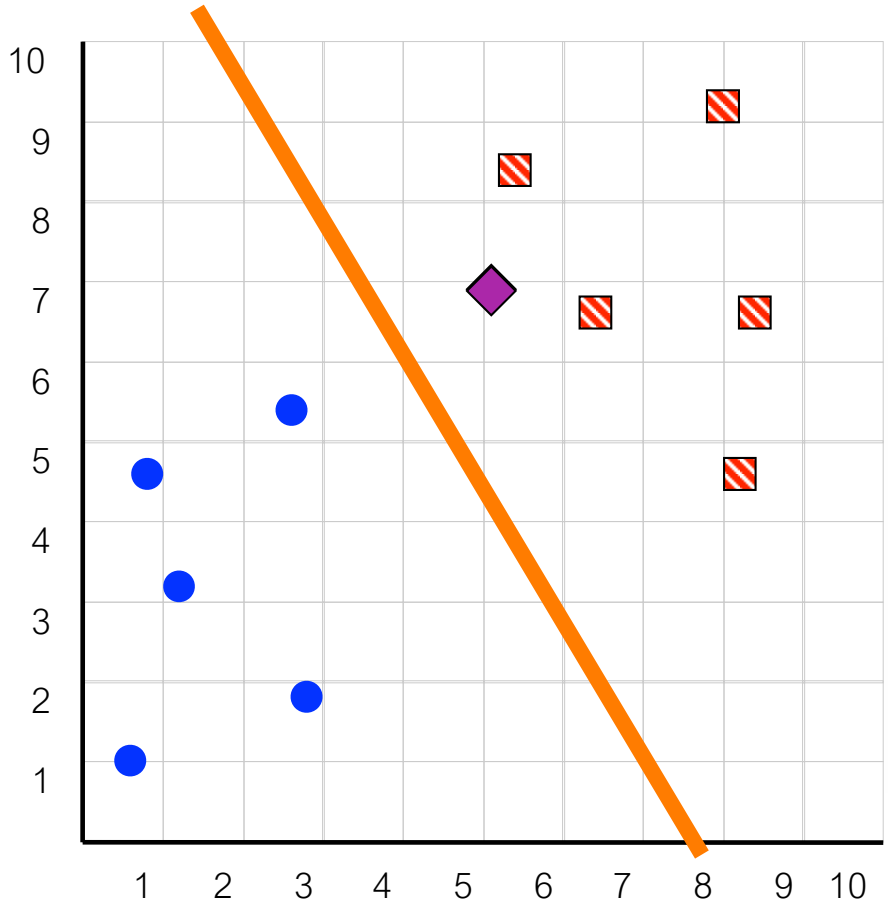
- Acabamos de abstrair os detalhes do nosso problema particular. Será muito mais fácil falar de pontos no espaço.

Esperança
Gafanhoto

Classificador Linear Simples

Ronald.A. Fisher
1890-1962

1936

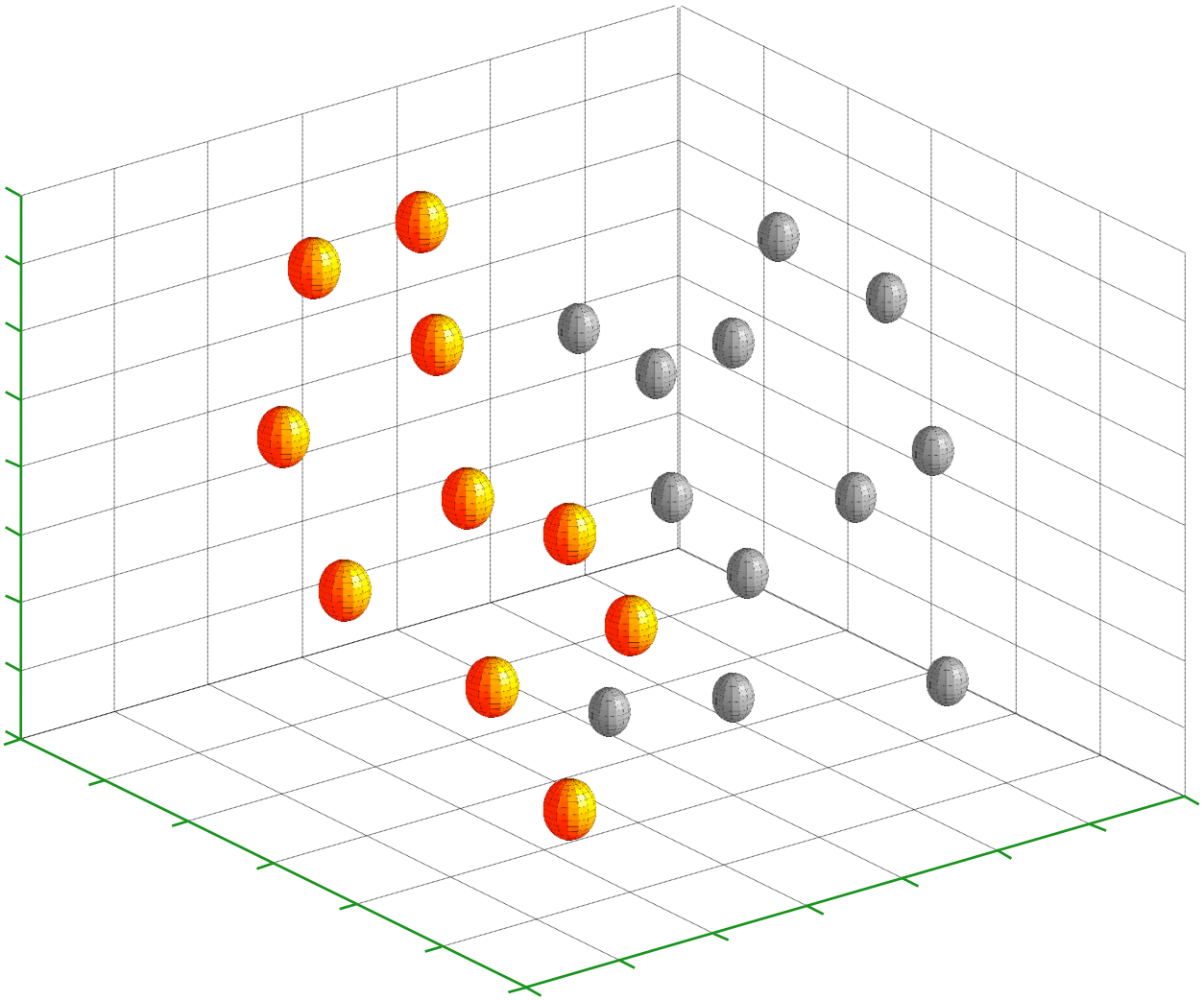


se exemplo desconhecido está acima da linha
então
    classe é Esperança
senão
    classe é Gafanhoto

Esperança
Gafanhoto

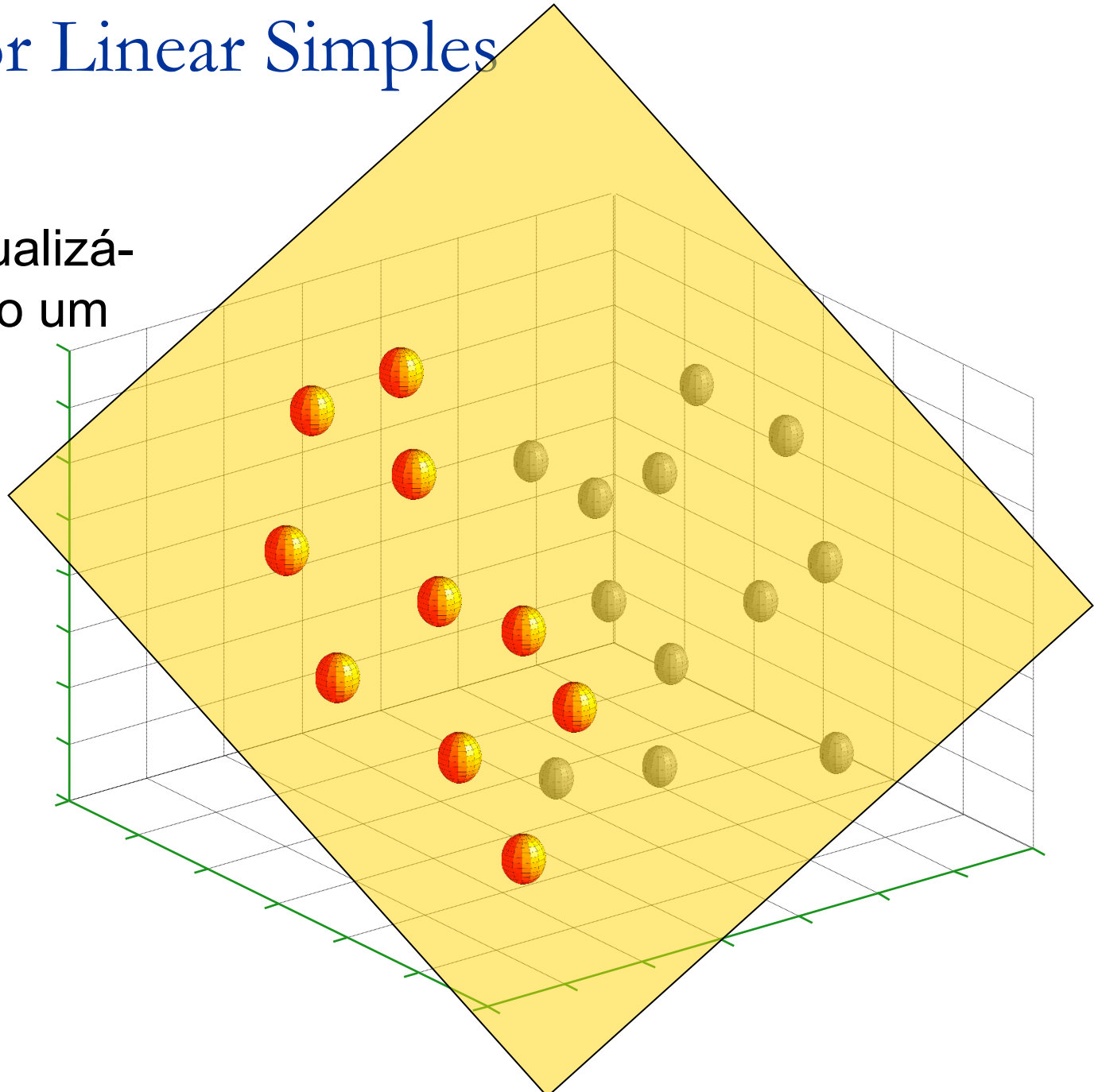
Classificador Linear Simples

- Está definido para espaços dimensionais maiores



Classificador Linear Simples

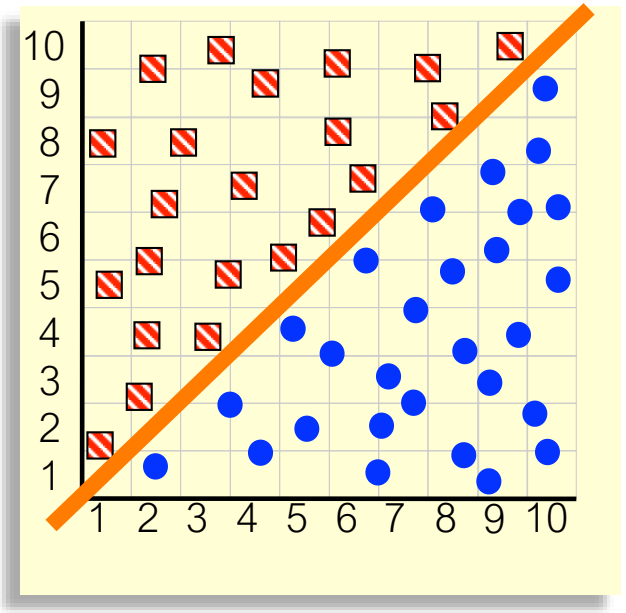
- Podemos visualizá-lo como sendo um hiperplano n-dimensional

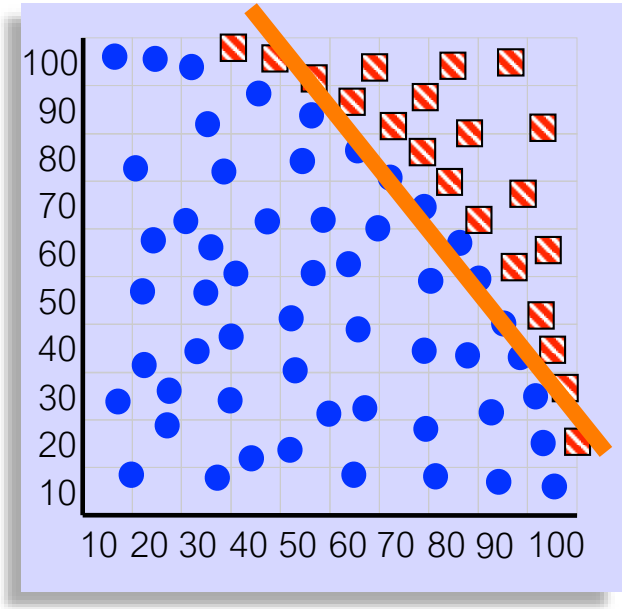


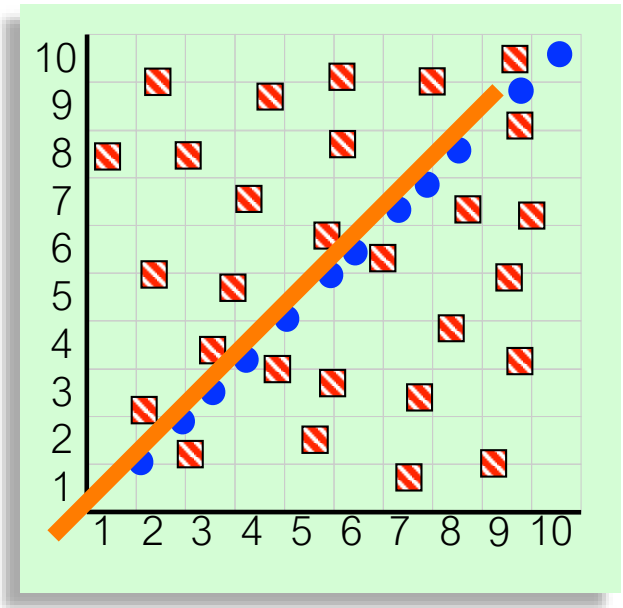
Classificação: Problema do Pombo

- Quais dos “Problemas do Pombo” podem ser resolvidos pelo Classificador Linear Simples?

1) Perfeito
2) Inútil
3) Muito bom

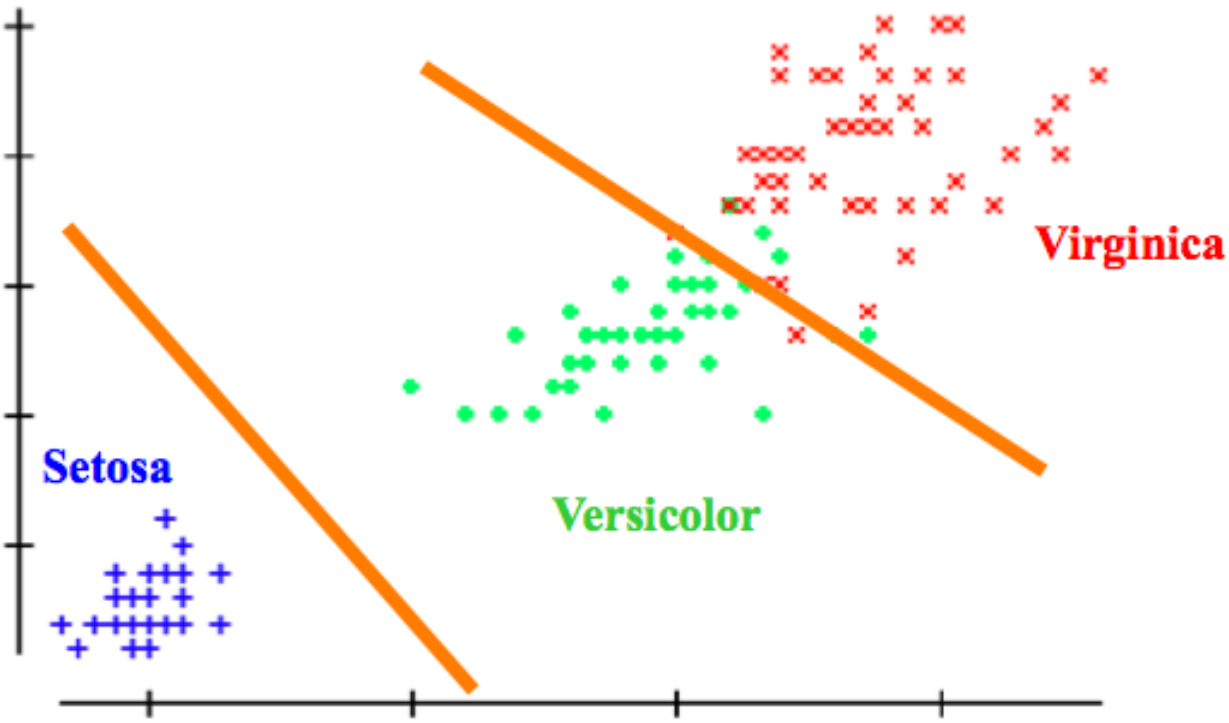






Problemas que podem ser resolvidos por um classificador linear são chamados de linearmente separáveis.

Classificação: Um Problema Famoso



Iris Setosa

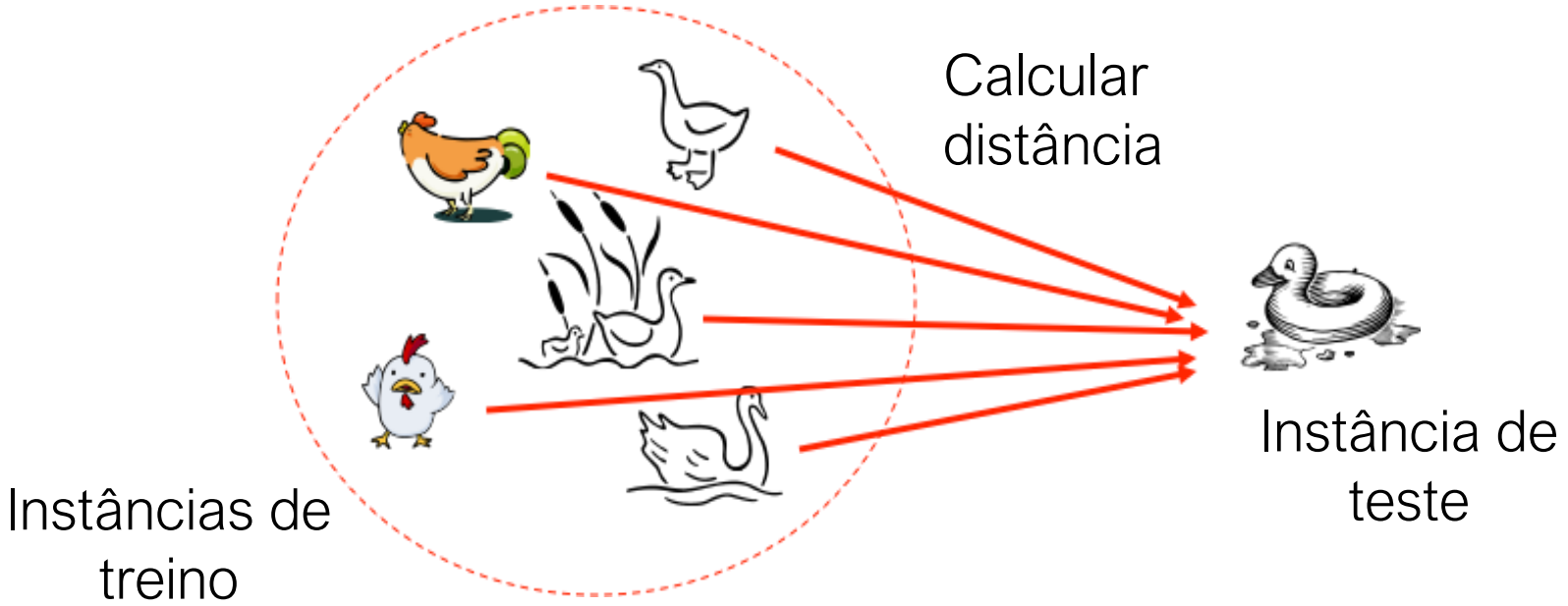
Iris Versicolor

Iris Virginica

se largura_petala > 3.72 - (0.325 * comprimento_petala)
    então classe = Virginica
senão
    se largura_petala…

Classificação: Algoritmo K-NN

- K-NN = *k nearest neighbors*
  - O classificador dos *k* vizinhos mais próximos.
  - Idéia básica: se caminha como um pato, faz “quack" como um pato, então provavelmente é um pato!

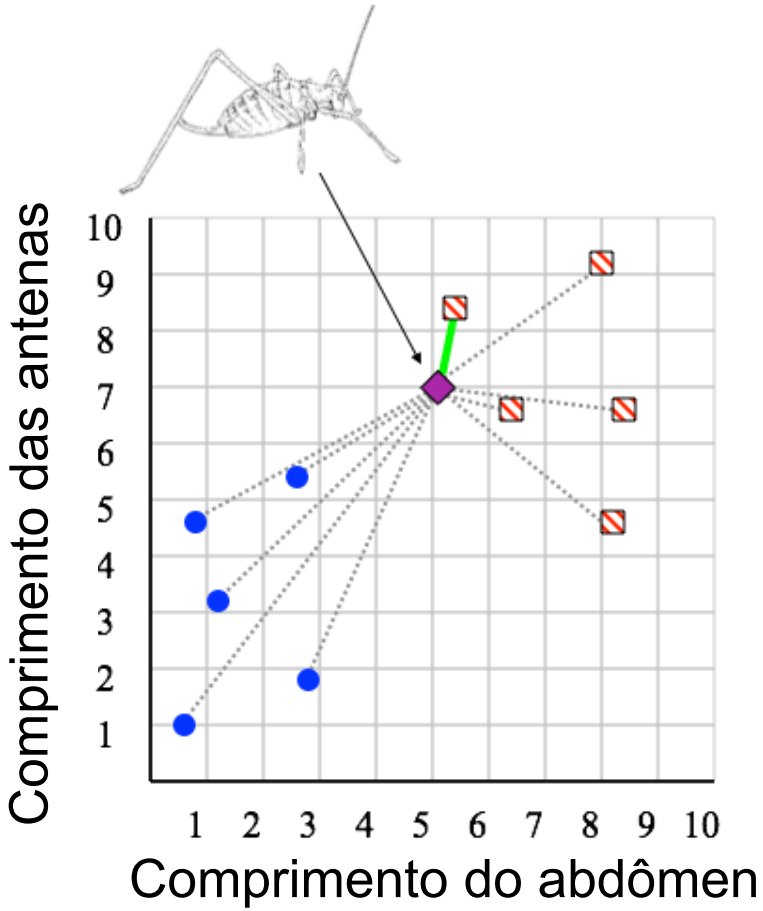


Classificação: Algoritmo K-NN com k=1

Evelyn Fix
1904-1965

Joseph Hodges
1922-2000

Criado em 1951

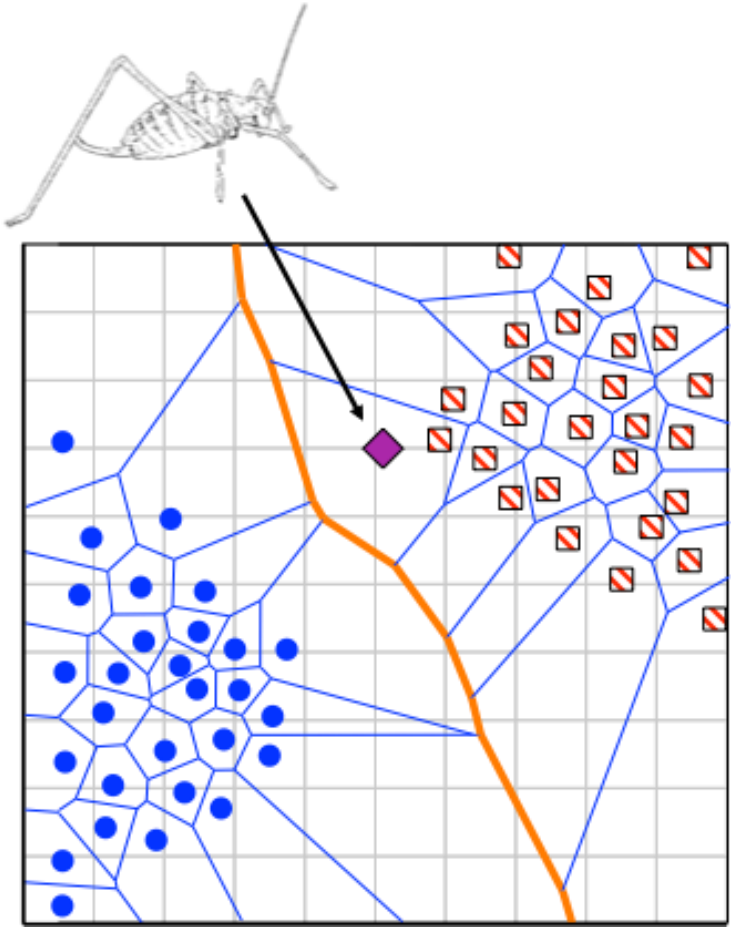


se o exemplo mais próximo ao exemplo desconhecido é da classe Esperança
então
classe é Esperança
senão
classe é Gafanhoto

Esperança
Gafanhoto

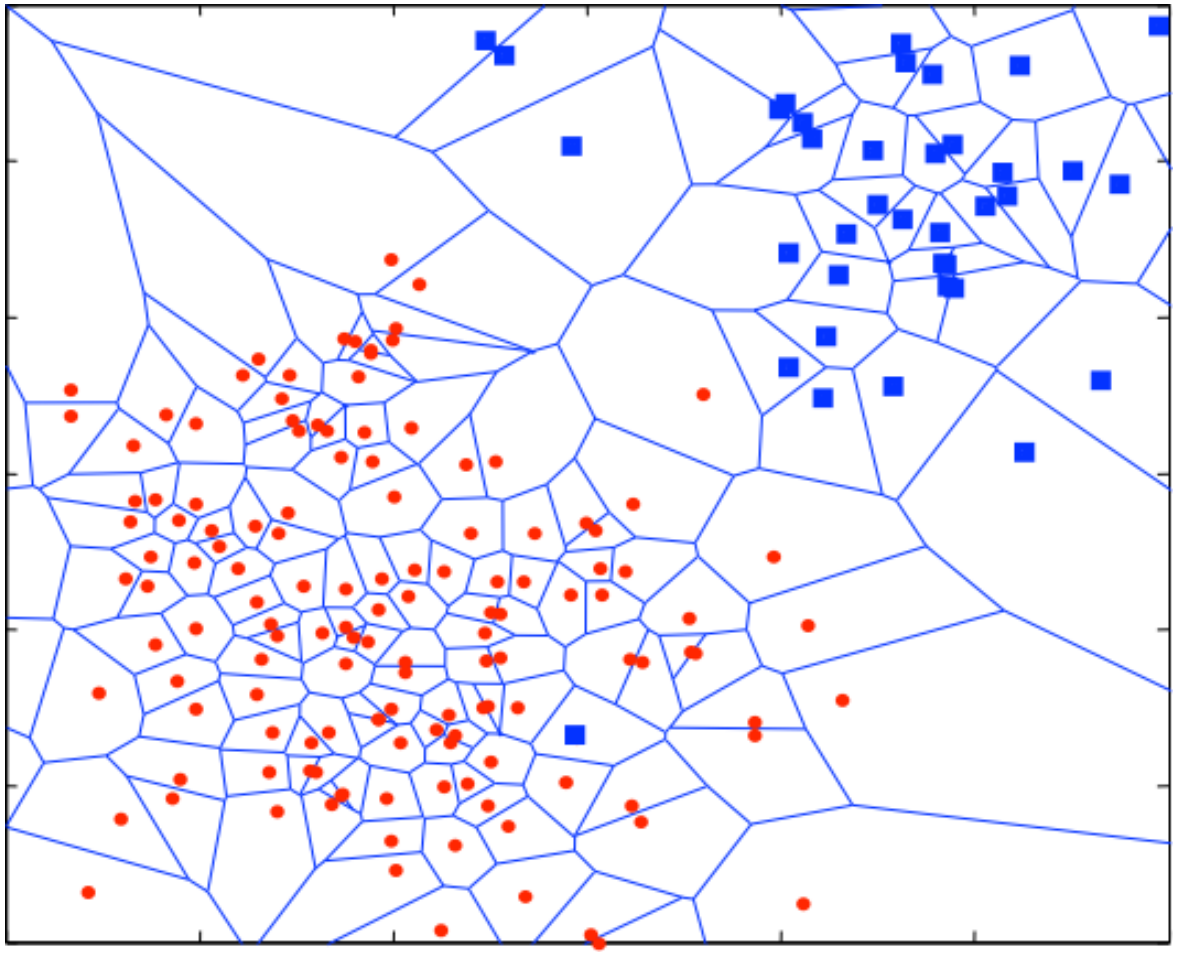
Classificação: Algoritmo K-NN com k=1

- É possível visualizar o K-NN (com k=1) em termos de uma fronteira de decisão!
- Fronteiras implícitas que delimitam as zonas pertencentes a cada exemplo de treino
- Esse tipo de divisão é chamada de **Diagrama de Voronoi** compostas por poliedros convexos.



Classificação: Algoritmo K-NN com k=1

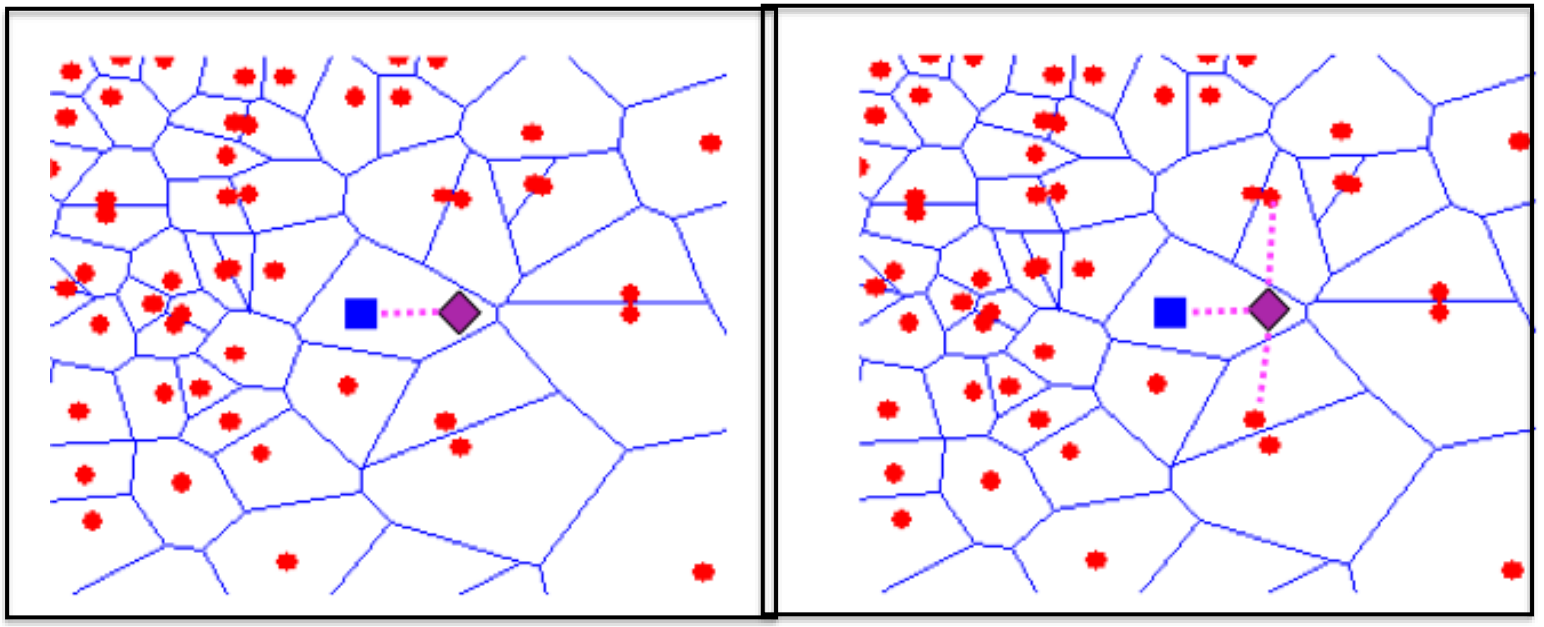
É sensível a "*outliers*"!



A solução para isso é….

Classificação: Algoritmo K-NN

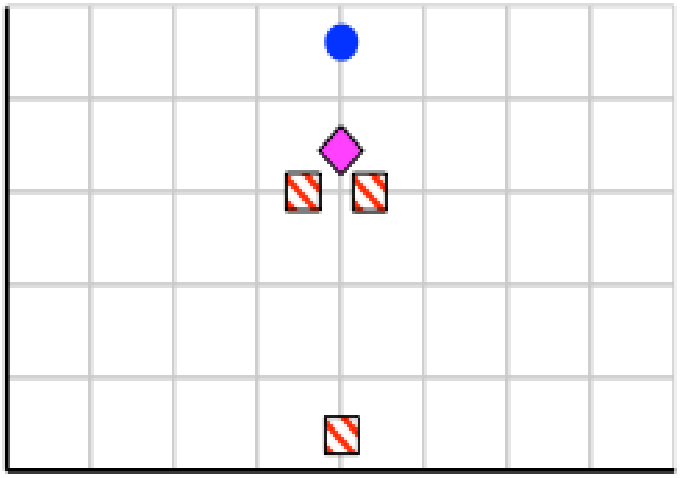
- Aumentar o valor de K!
- Mede-se a distância para os *k* exemplos mais próximos.
- Computa o voto da maioria para definição da classe do exemplo desconhecido.
- O objeto de teste é classificado na classe mais votada.



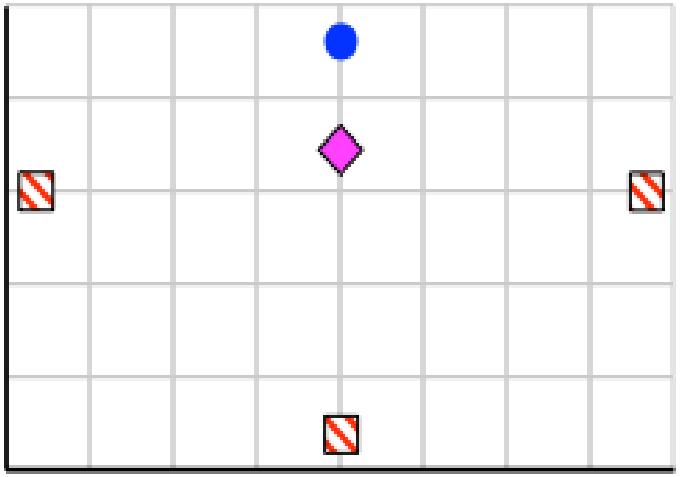
k=1 k=3

Classificação: Algoritmo K-NN

K-NN é sensível às unidades de medidas utilizadas.



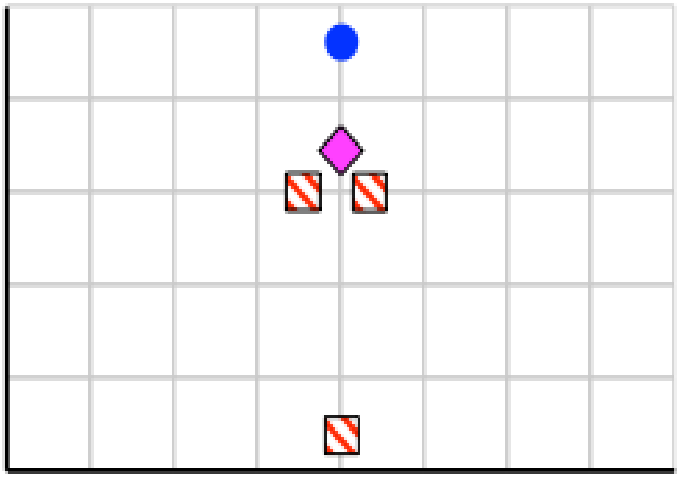
Eixo x em centímetros.
Eixo y em reais.
Exemplo mais próximo do rosa desconhecido é vermelho.



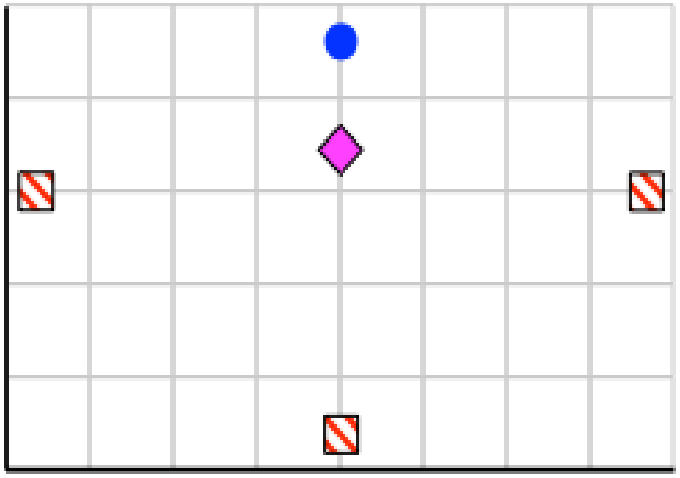
Eixo x em milímetros.
Eixo y em reais.
Exemplo mais próximo do rosa desconhecido é azul.

Classificação: Algoritmo K-NN

K-NN é sensível às unidades de medidas utilizadas.



Eixo x em centímetros.
Eixo y em reais.
Exemplo mais próximo do rosa desconhecido é vermelho.



Eixo x em milímetros.
Eixo y em reais.
Exemplo mais próximo do rosa desconhecido é azul.

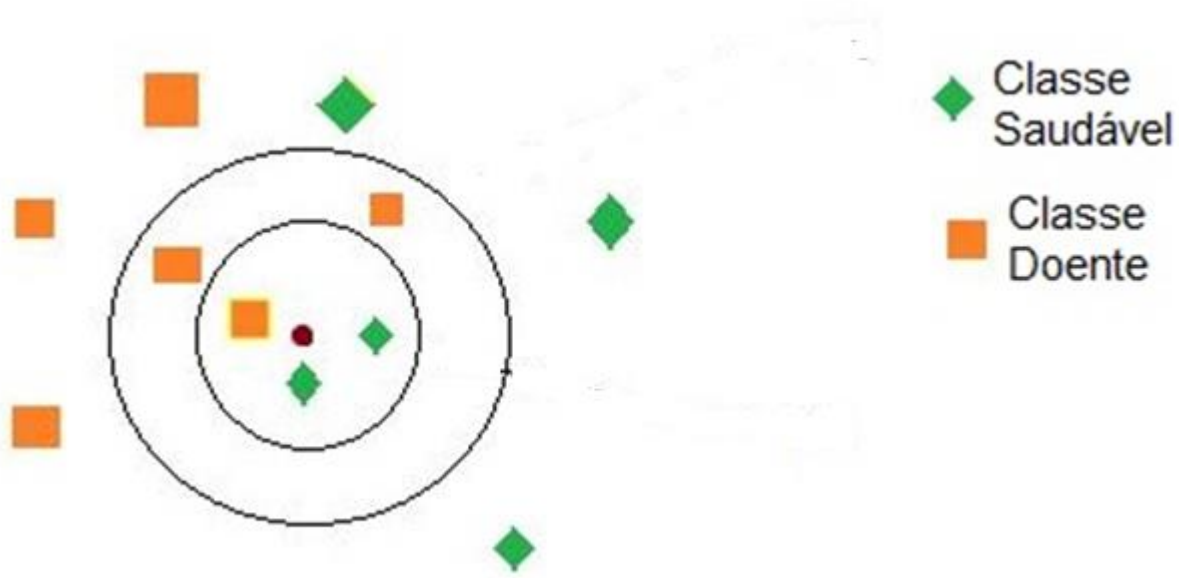
Solução? Normalizar os dados!

Classificação: Algoritmo K-NN

- A escolha do valor de k mais apropriado para um problema de decisão específico pode não ser trivial.

- Qual seria a classe do objeto desconhecido abaixo se o K fosse igual a 3?

- E se K fosse igual a 5?

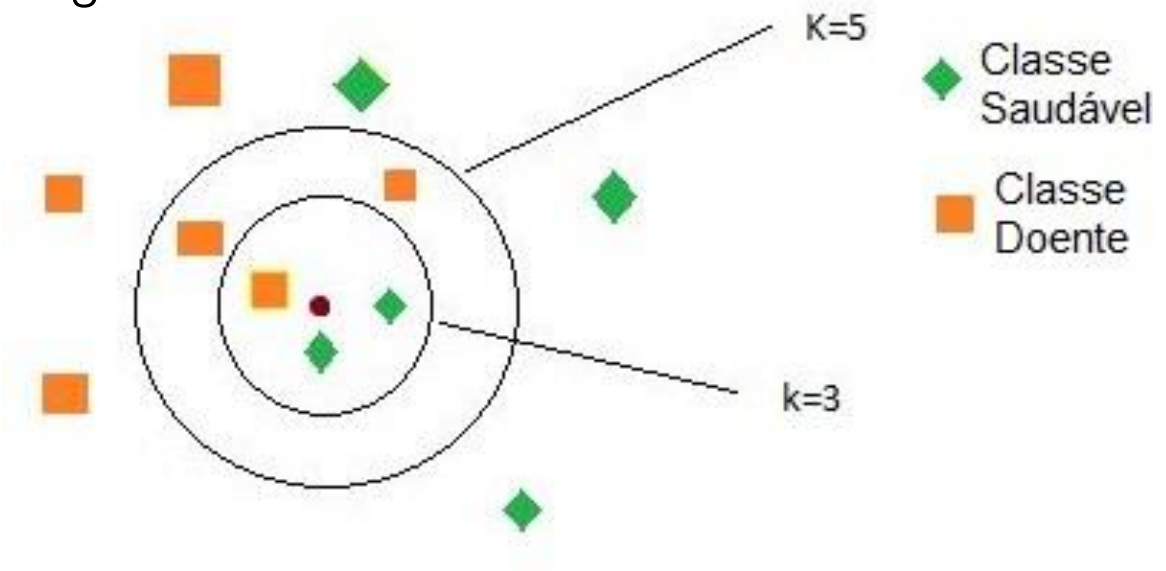


Classificação: Algoritmo K-NN

- A escolha do valor de k mais apropriado para um problema de decisão específico pode não ser trivial.

- Qual seria a classe do objeto desconhecido abaixo se o K fosse igual a 3?

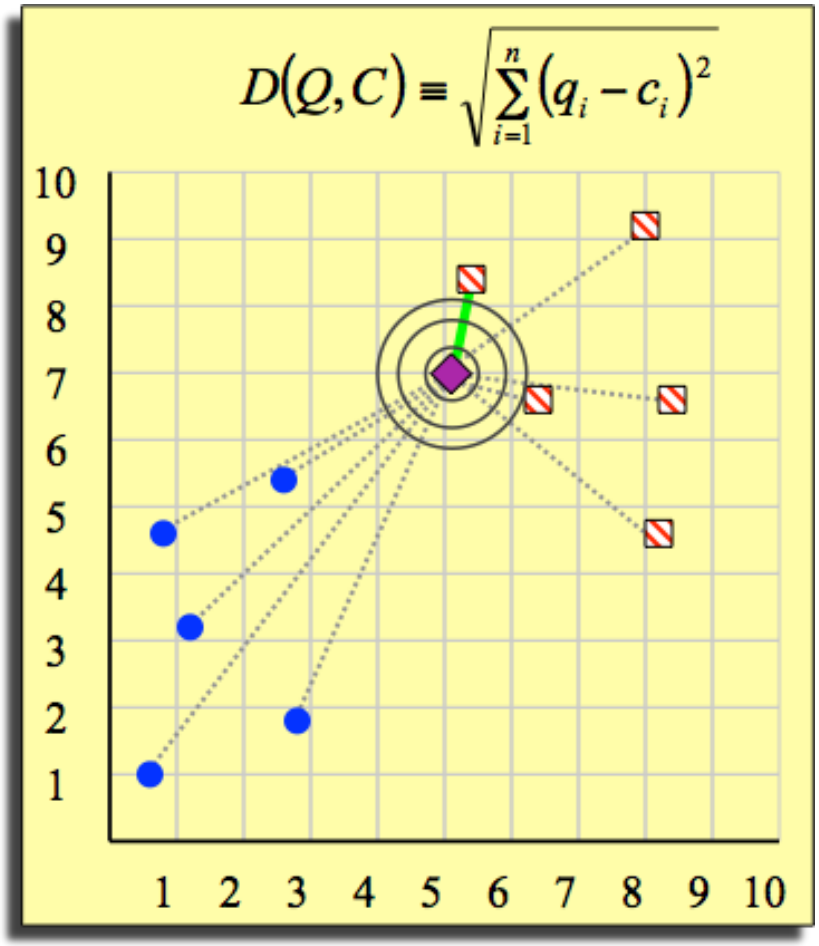
- E se K fosse igual a 5?

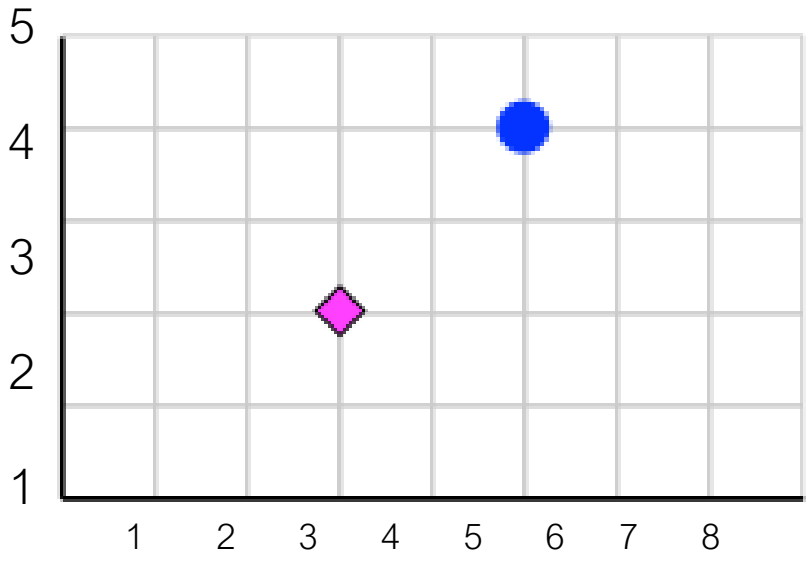


- Normalmente o valor de k é pequeno e ímpar (k = 3,5,7, ...)
- Para evitar empates em problemas de classificação, não é usual utilizar k=2 ou valores pares

Classificação: Algoritmo K-NN

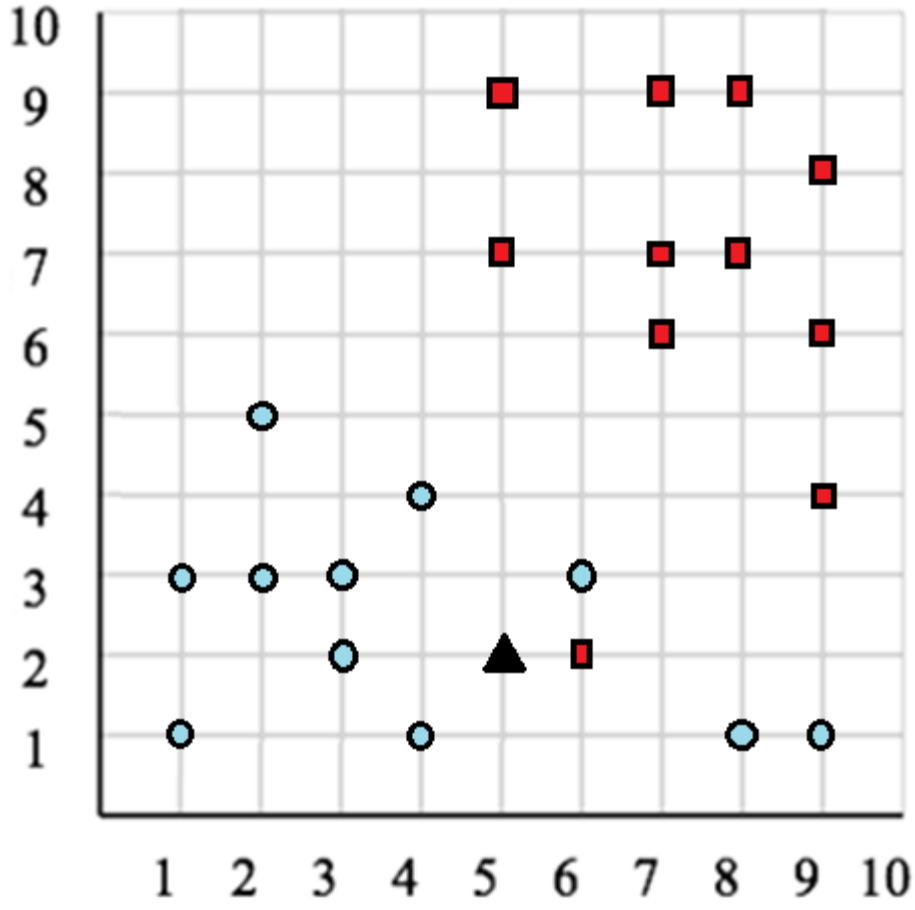
Calculando a distância Euclidiana.





Classificação: K-NN - Exercício

Descubra a classe do exemplo desconhecido com k=1 e com k=3



Objeto	X	y	Classe
1	5	9	□
2	7	9	□
3	8	9	□
4	9	8	□
5	5	7	□
6	7	7	□
7	8	7	□
8	7	6	□
9	9	6	□
10	9	4	□
11	6	2	□
12	2	5	○
13	4	4	○
14	1	3	○
15	2	3	○
16	3	3	○
17	6	3	○
18	3	2	○
19	1	1	○
20	4	1	○
21	8	1	○
22	9	1	○
23	5	2	?

Classificação: K-NN Exercício

$$D(1) = \sqrt{(5-5)^2+(2-9)^2} = \sqrt{49} = 7$$

D(2) ≈ 7
D(3) ≈ 7
D(4) ≈ 7
D(5) = 7
D(6) ≈ 5
D(7) ≈ 5
D(8) ≈ 4
D(9) ≈ 5
D(10) ≈ 4
D(11) = 1
D(12) ≈ 4
D(13) ≈ 2
D(14) ≈ 4
D(15) ≈ 3
D(16) ≈ 2
D(17) = 1
D(18) = 2
D(19) ≈ 4
D(20) = 1
D(21) ≈ 3
D(22) ≈ 4
D(23) ????

Objeto	X	y	Classe
1	5	9	□
2	7	9	□
3	8	9	□
4	9	8	□
5	5	7	□
6	7	7	□
7	8	7	□
8	7	6	□
9	9	6	□
10	9	4	□
11	6	2	□
12	2	5	○
13	4	4	○
14	1	3	○
15	2	3	○
16	3	3	○
17	6	3	○
18	3	2	○
19	1	1	○
20	4	1	○
21	8	1	○
22	9	1	○
23	5	2	?

Classificação: Algoritmo K-NN

- **Aspectos Positivos**
  - Algoritmo de treinamento é simples.
  - Não gera um modelo.
  - Algoritmo é incremental: quando novos exemplos de treinamento estão disponíveis, basta armazená-los na memória.
  - Existem apenas dois parâmetros necessários para a implementação, o valor de K e a função de distância
- **Aspectos Negativos**
  - Algoritmo *lazy*.
  - Predição custosa: classificar um objeto de teste requer calcular a distância desse objeto a todos os objetos de treinamento.
  - Afetado pela presença de atributos redundantes e de atributos irrelevantes.
  - O número de atributos define o número de dimensões do espaço, o qual cresce exponencialmente.
  - Não aceita atributos categóricos

Avaliação de Desempenho

- Matriz de Confusão:

		CLASSE PREVISTA	
CLASSE REAL		Classe=SIM	Classe=NAO
	Classe=SIM	a (TP)	b (FN)
	Classe=NAO	c (FP)	d (TN)

a: **TP** (true positive)
verdadeiro positivo

b: **FN** (false negative)
falso negativo

c: **FP** (false positive)
falso positivo

d: **TN** (true negative)
verdadeiro negativo

Métricas para Avaliação de Desempenho

		CLASSE PREVISTA	
		Classe=SIM	Classe=NAO
CLASSE REAL	Classe=SIM	a (TP)	b (FN)
	Classe=NAO	c (FP)	d (TN)

- Métrica mais usada: Acurácia (Accuracy)

$$\text{Acurácia} = \frac{a+d}{a+b+c+d} = \frac{TP+TN}{TP+TN+FP+FN}$$

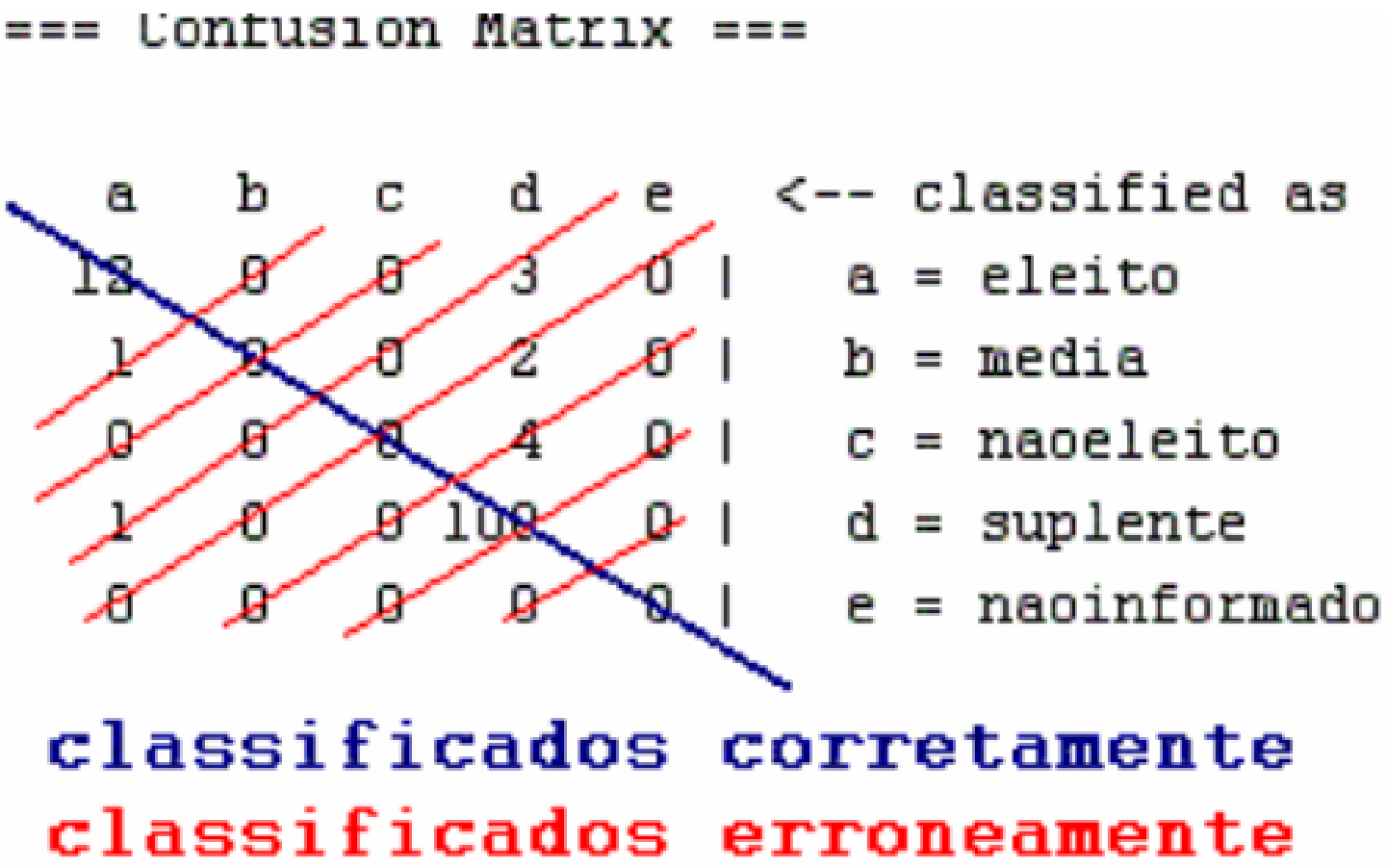
$$\text{Sensibilidade (}\textit{Recall}\text{)} = \frac{TP}{TP+FN} \qquad \text{Especificidade (}\textit{Precision}\text{)} = \frac{TN}{TN+FP}$$

Matriz de Confusão

- Sensibilidade
  - Proporção de verdadeiros positivos.
  - Capacidade do sistema em predizer corretamente a condição para casos que realmente a têm.

- Especificidade
  - Proporção de verdadeiros negativos.
  - Capacidade do sistema em predizer corretamente a ausência da condição para casos que realmente não a têm.

Métricas para Avaliação de Desempenho



Limitações de Precisão

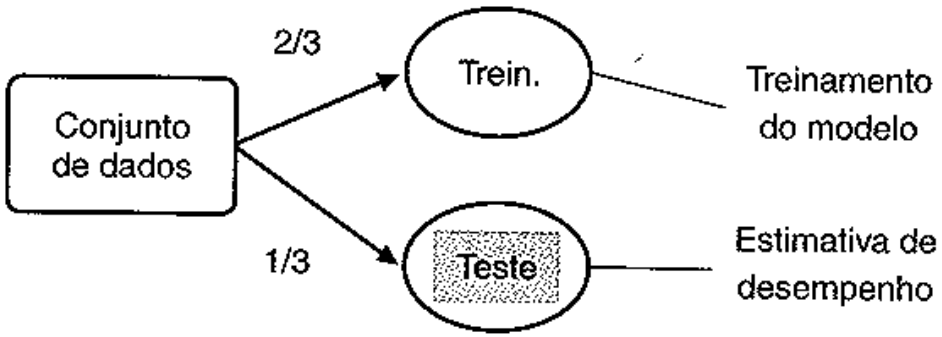
- Considere um problema de 2 classes
  - Número de exemplos da Classe 0 = 9990
  - Número de exemplos da Classe 1 = 10

- Se o modelo prevê que todos os objetos são da classe 0, a acurácia é 9990/10000 = 99.9 %
  - Acurácia é enganosa porque o modelo não detectou nenhum exemplo da classe 1.

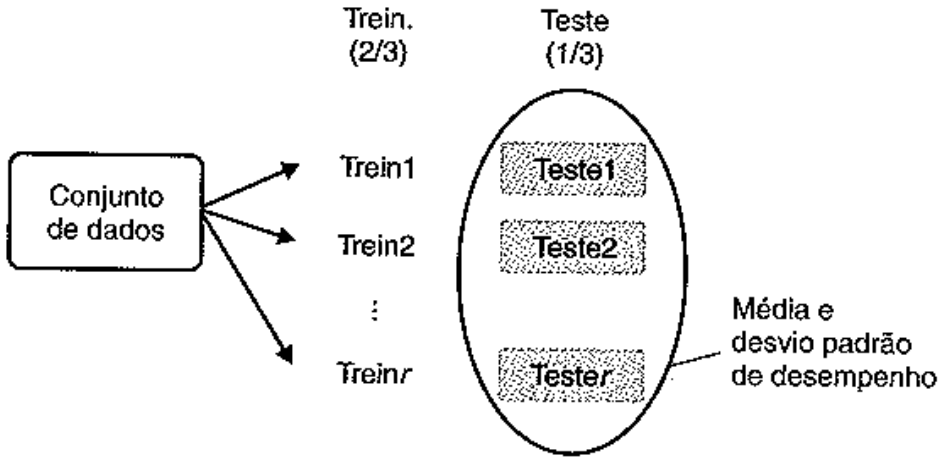
Métodos de Amostragem

- Calcular o erro estimado do modelo durante o treino.
- O cálculo do erro estimado auxilia o algoritmo de aprendizagem a fazer uma seleção do modelo.
- Objetivo: encontrar o modelo com uma complexidade precisa que não é suscetível a *overfitting*.
- Métodos comumente utilizados:
  - Holdout
  - Validação Cruzada (Cross-Validation)
  - Bootstrap

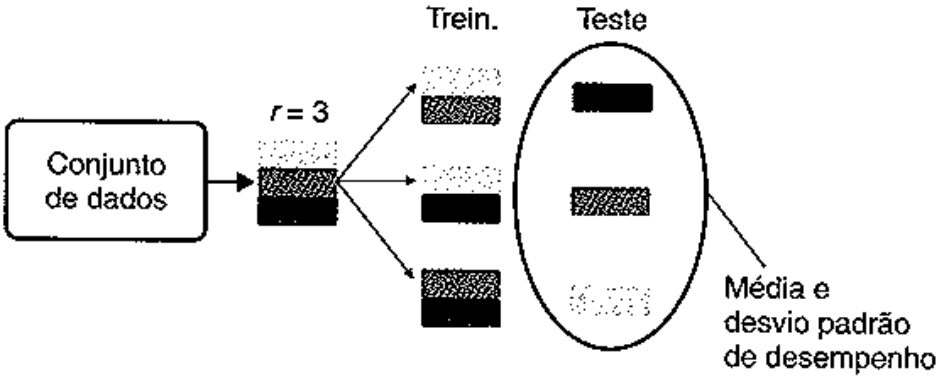
Métodos de Amostragem



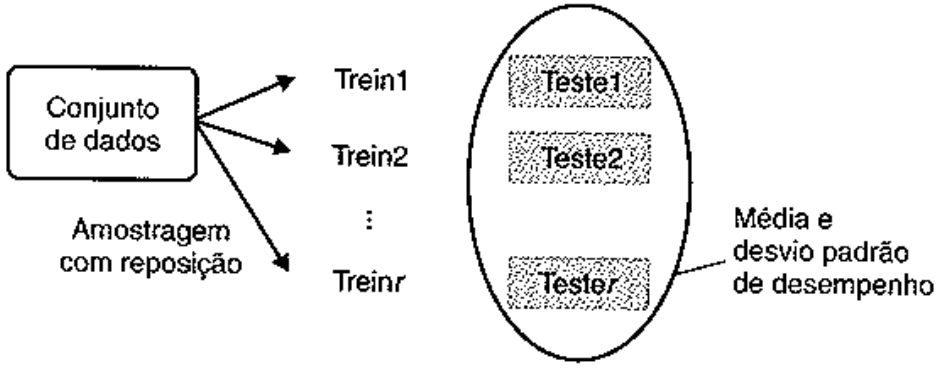
(a) *Holdout*



(b) Amostragem aleatória



(c) Validação cruzada



(d) *Bootstrap*

Atividade

- A partir do dataset escolhido para trabalhar com os métodos supervisionados e a biblioteca *scikitlearn* do Python, realize as seguintes tarefas:

1. Identifique o atributo alvo.
2. Execute o algoritmo K-NN aplicando *cross-validation* (validação cruzada) para 10 *KFolds*.
3. Execute o algoritmo K-NN aplicando *Holdout*, dividindo o dataset em 30% para teste.
4. Altere o número de k para obter uma melhor acurácia na validação.
5. Analise e compare os resultados obtidos, utilizando a matriz de confusão e computando a acurácia.

Créditos

- Adaptação dos slides de Pang-Ning Tan
  - Michigan State University
  - http://www.cse.msu.edu/~ptan/
  - ptan@cse.msu.edu
- Adaptação dos slides de Eamon Keogh
  - University of California at Riverside
  - http://www.cs.ucr.edu/~eamonn/
  - eamonn@cs.ucr.edu
- Adaptação dos slides de Ricardo Campello e Eduardo Hruschka
  - Universidade de São Paulo (ICMC)
- Adaptação dos slides de Rodrigo Barros
  - Pontifícia Universidade Católica do Rio Grande do Sul (PPGCC)

Referências

- Breiman, L., Freidman, J., Olshen, R. e Stone, C. (1984). *Classification and Regression Trees.* Wadsworth International Group., USA.
- Faceli, K.; Lorena, A.C.; Gama, J.; de Carvalho, A.C.P.L.F. *Inteligência Artificial: Uma abordagem de aprendizado de máquina.* LTC, Rio de Janeiro, 2011.
- Quilan, R. (1979). *Discovering rules by induction from large collections of examples.* In: Michie, D. (Ed.) Expert Systems in the Microelectronic Age, p. 168-201. Edinburgh University Press.
- Quilan, J.R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Mateo, CA, USA.
- TAN, P-N; STEINBACH, M.; KUMAR,V. *Introduction to Data Mining.* Pearson, 2006.