CC-226
Introdução à Análise de Padrões

Prof. Carlos Henrique Q. Forster

Visão Geral do Curso e
Introdução a Classificadores

Padrões

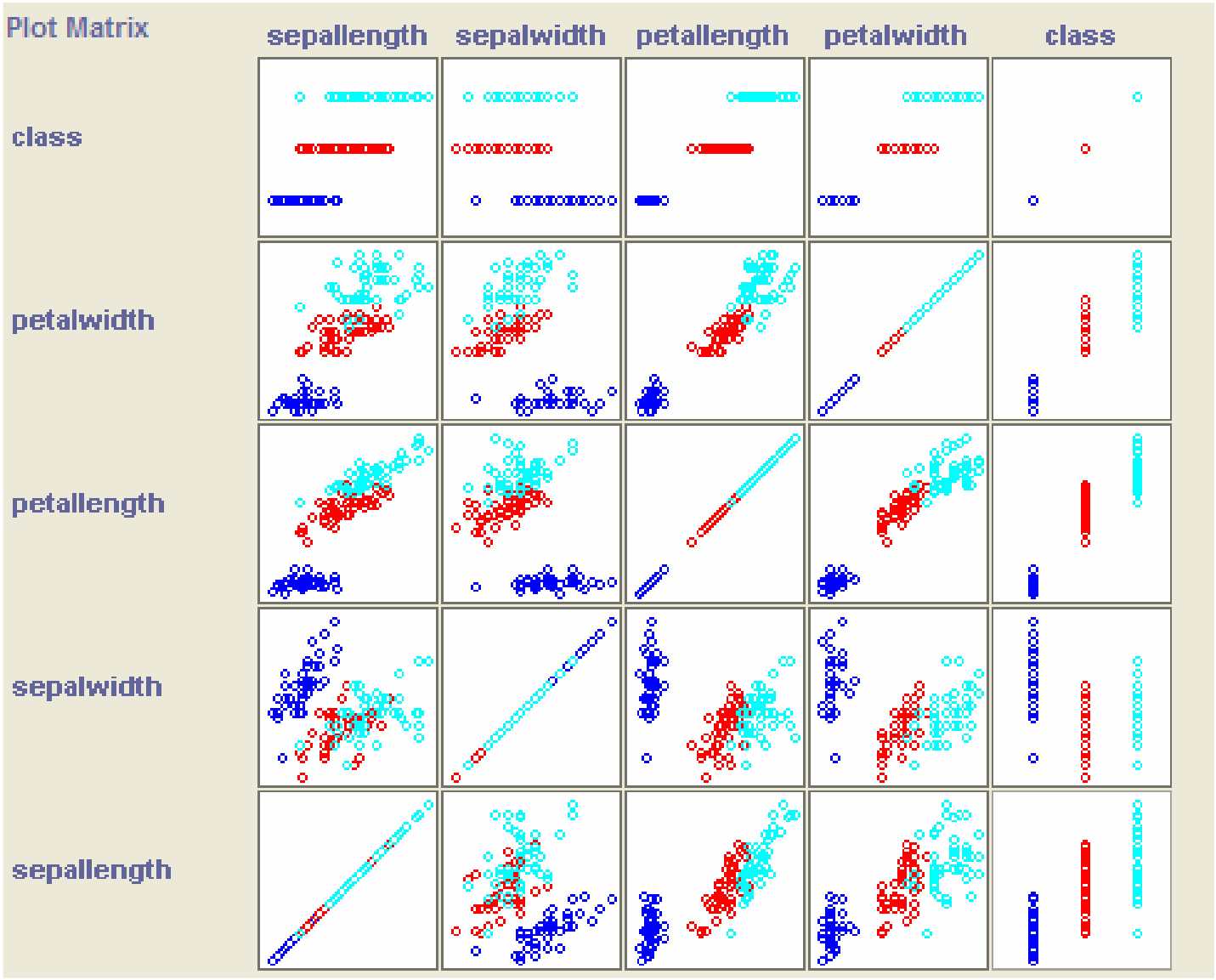
- São apresentados como tuplas de variáveis aleatórias
- O conjunto amostra é utilizado como entrada para se estimar um modelo que explique a distribuição dos dados e a relação entre as variáveis
- Cada tupla da amostra é chamada instância e foi sorteada de forma independente de qualquer outro elemento do conjunto (instâncias são independentes e identicamente distribuídas)

Exemplo – base Iris

- Identificar a espécie da planta pelas medidas realizadas nas flores
  - 4 atributos medidos: largura e comprimento de pétala e sépala. 3 espécies. 150 instâncias (50 cada).

No.	sepallength Numeric	sepalwidth Numeric	petallength Numeric	petalwidth Numeric	**class** Nominal
61	5.0	2.0	3.5	1.0	Iris-versicolor
63	6.0	2.2	4.0	1.0	Iris-versicolor
69	6.2	2.2	4.5	1.5	Iris-versicolor
120	6.0	2.2	5.0	1.5	Iris-virginica
42	4.5	2.3	1.3	0.3	Iris-setosa
54	5.5	2.3	4.0	1.3	Iris-versicolor

Exemplo – base Iris – Scatter plot



virgínica

versicolor

setosa

Padrões - classificação

- Os padrões ou instâncias podem ser rotulados ou não.
- O rótulo pode ser uma das variáveis da tupla e se deseja predizê-lo através da classificação
- Neste caso fala-se em aprendizado supervisionado: o classificador é construído a partir de dados rotulados

Padrões - regressão

- Tipos de variáveis (level of measurement)
  - Nominal (discreto sem ordenação)
  - Ordinal (discreto com ordem total)
  - Intervalo (contínuo, comprimentos, sem origem)
  - Razão (contínuo, proporção entre comprimentos)
- Regressão vai predizer uma variável contínua
- É comum a conversão entre tipos, definindo limiares, agrupando em “bins”

Padrões – agrupamento e indução de regras

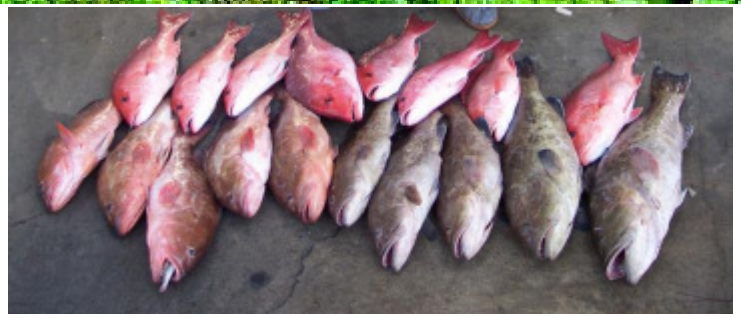
- Não há definição de rótulo
- Não se sabe a priori qual variável deve-se estimar a partir das outras
- Formação de agrupamentos define um rótulo que corresponde à identificação do agrupamento
- Análise da dependência estabelece relação entre as variáveis
- Aprendizado não-supervisionado

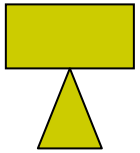
Avaliação de um classificador

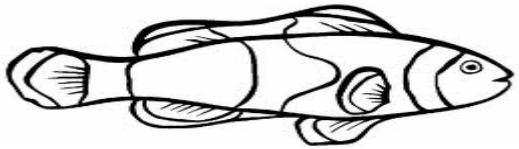
- Fases da classificação:
  - Aprendizado ou treinamento: a partir de dados rotulados é construído um modelo de decisão.
  - Predição ou aplicação: a partir de uma tupla de entrada, é estimado o seu rótulo.
- Objetivo:
  - Predizer com mínimo de erros.
  - Predizer com mínimo risco ou custo.

Exemplo de classificador

- Imagine que queremos construir uma máquina para separar automaticamente o pescado em duas categorias de peixe (robalo e salmão).
- Os peixes chegam um a um por uma esteira e uma câmera captura sua imagem.
- Aplicamos algoritmos de processamento de imagem para medir algumas características dos peixes, por exemplo: brilho e comprimento
- Dadas as medidas de um peixe, a máquina estima sua categoria e o desvia para a esteira apropriada.





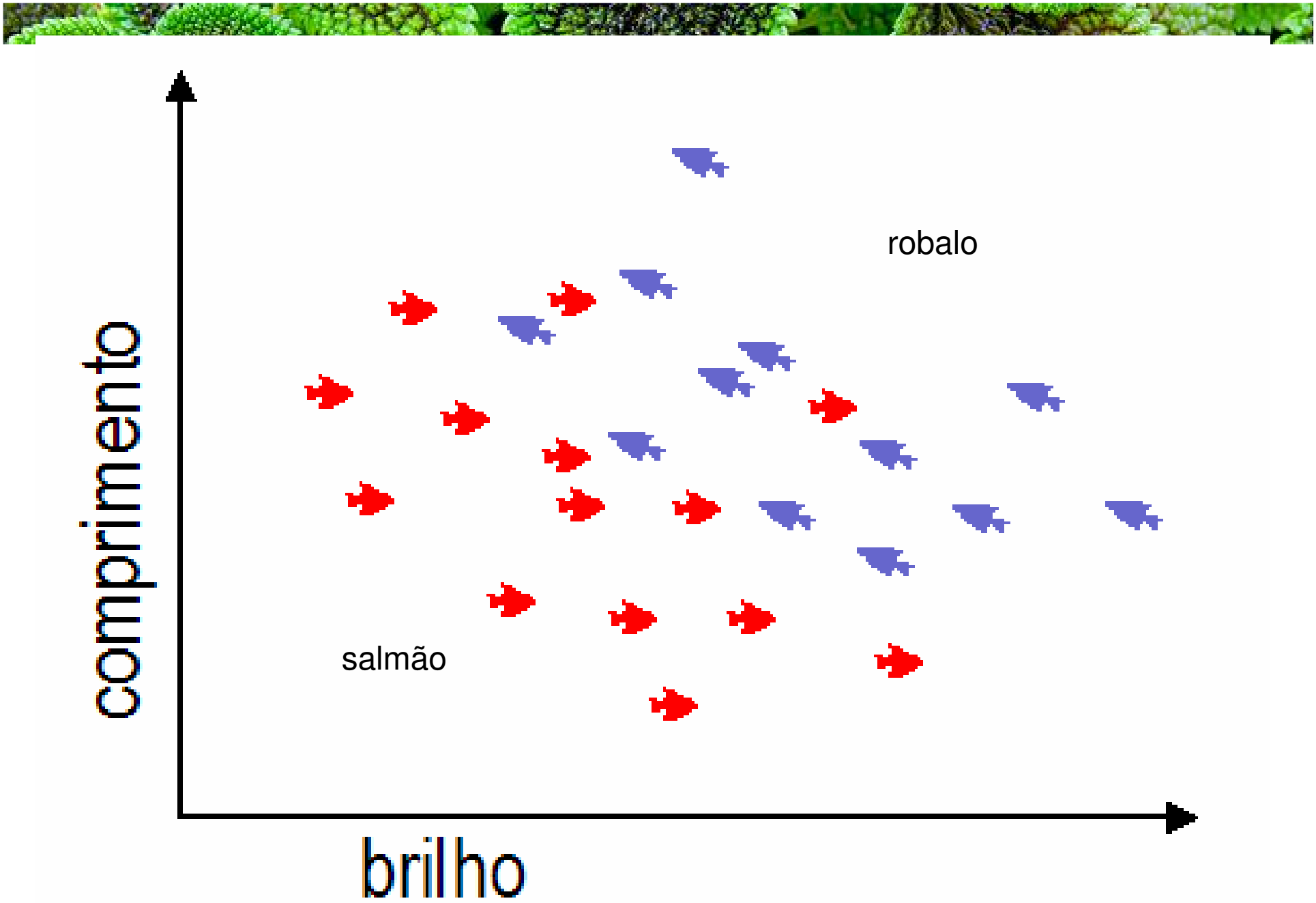


salmão

robalo

Construção do dataset

- Pegamos uma amostra de peixes e a rotulamos manualmente.
- Para cada peixe da amostra tabelamos brilho, comprimento e rótulo da classe correspondente. (maior comprimento variando a direção, brilho médio na imagem)
- Cada peixe pode ser representado como um ponto num gráfico de dispersão com eixos: comprimento e brilho


robalo
salmão
comprimento
brilho

Construção do classificador

- Vamos estabelecer visualmente uma fronteira entre as duas categorias de peixe, no caso uma linha reta
- Observe que não foi possível separá-los com totalidade de acerto
- A linha construída é, no caso geral, chamada superfície de separação
- Neste caso, o classificador seria representado pelos parâmetros da reta (orientada)

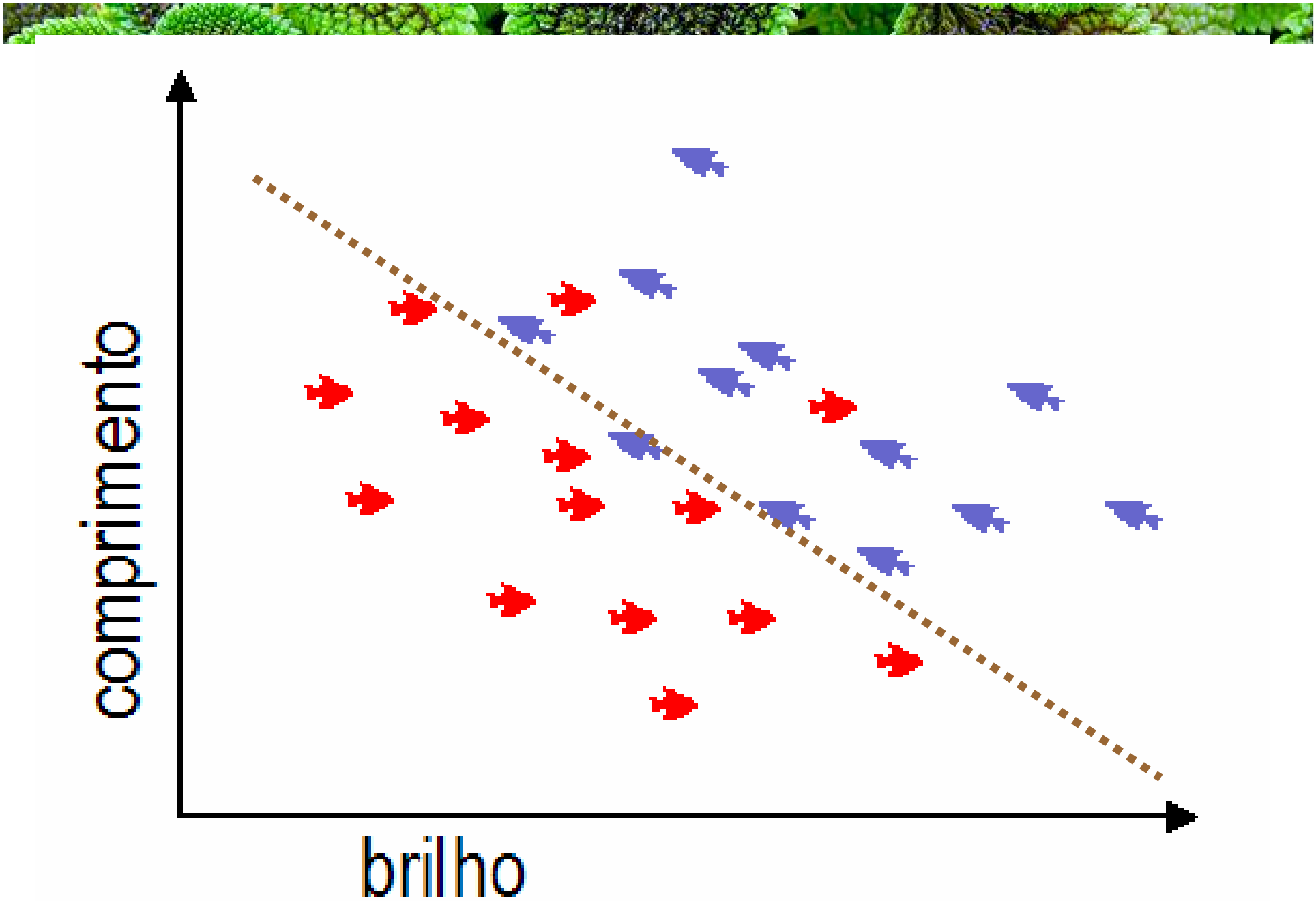
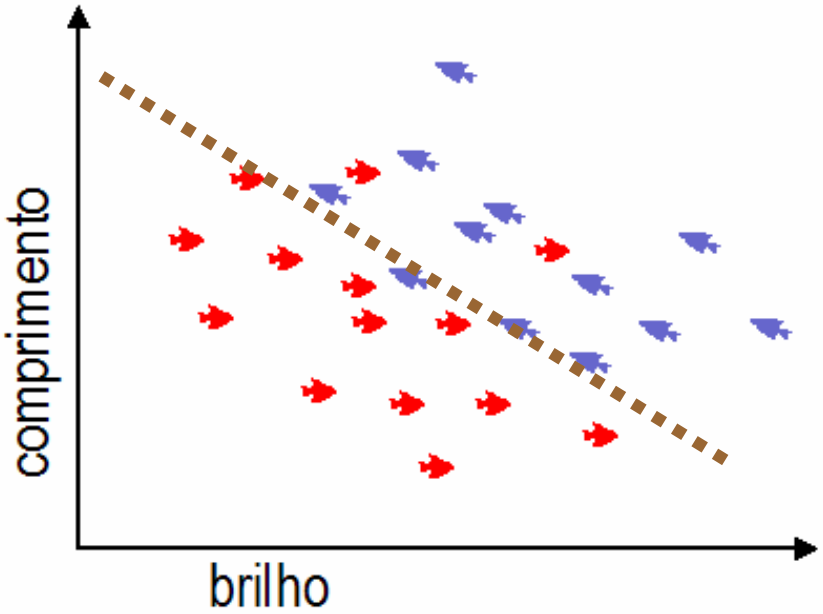

comprimento
brilho

Tabela de contingência

- Também é chamada matriz de confusão e estabelece os tipos de erro cometidos pelo classificador
- Suponha que queremos identificar salmão
  - Verdadeiro positivo é cada instância que representa um salmão e é classificada como salmão
  - Verdadeiro negativo é cada instância que não representa salmão e não é classificada como salmão
  - Falso positivo (tipo I) é cada instância classificada erroneamente como salmão, por se tratar de outro peixe
  - Falso negativo (tipo II) é cada instância que representa um salmão, mas é classificada erroneamente como outro peixe
- Taxa de acerto ou acurácia:
  - Número das classificações corretas dividido pelo total



	Classificado como salmão	Classificado como robalo
Na verdade é salmão	12 (verdadeiros positivos)	2 (falsos negativos)
Na verdade é robalo	1 (falso positivo)	11 (verdadeiros negativos)

Mais métricas

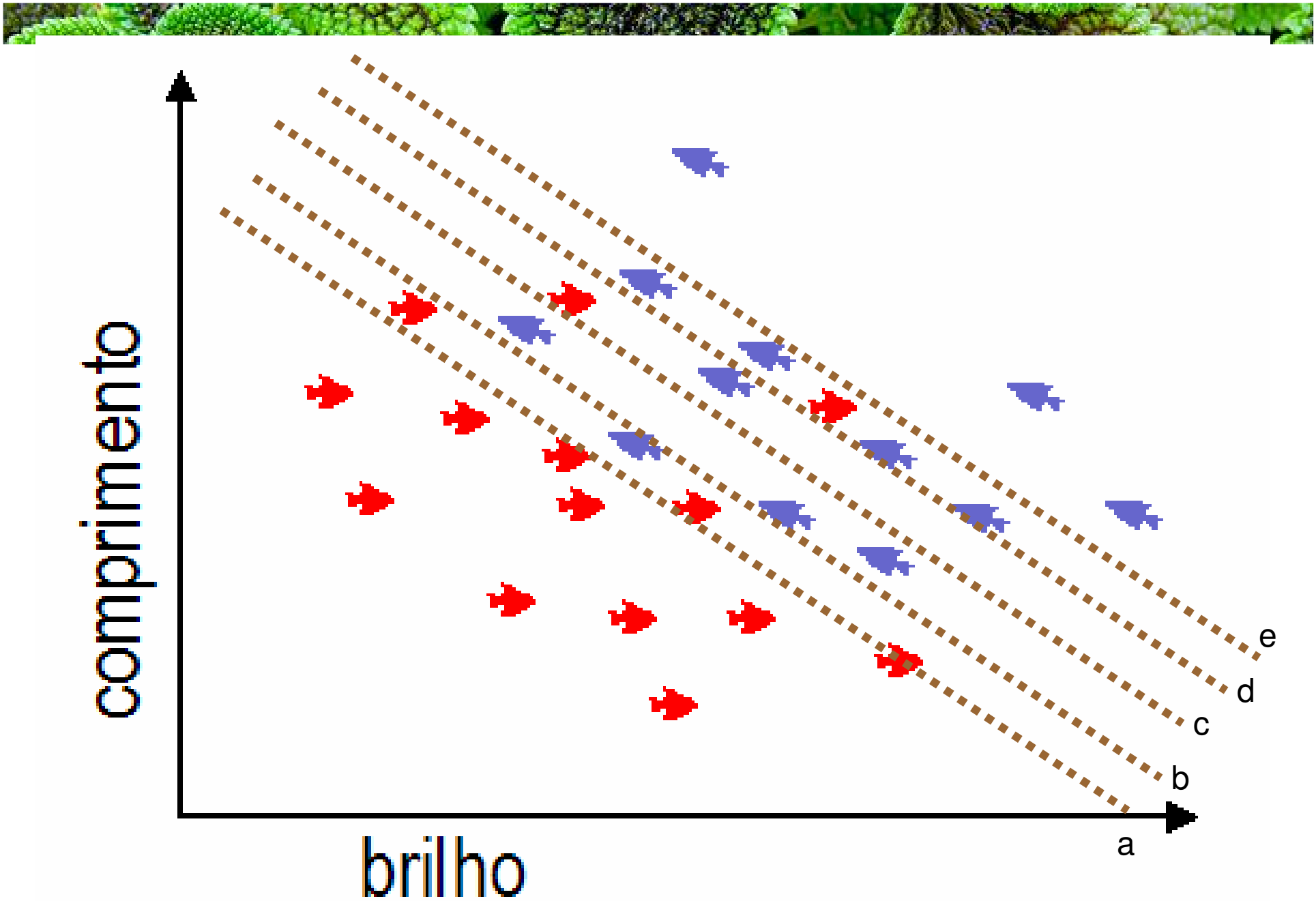
- Recall (taxa de verdadeiros positivos ou sensitividade) é a fração dos positivos que é classificada corretamente
  recall=VP/(VP+FN)
  representa o quanto dos objetos procurados conseguem ser recuperados por uma busca
- No exemplo: do conjunto de salmões, qual porcentagem é corretamente identificada
- O que acontece com o recall se classificarmos todos os peixes como salmão?

Mais métricas

- Taxa de falsos positivos (falsos alarmes)
  FPR=FP/(FP+VN)
  dos negativos, quantos foram selecionados como positivos
- Precisão
  prec=VP/(VP+FP)
  dos identificados como positivos, quantos são de fato positivos
- O que acontece com a taxa de falsos positivos e a precisão se classificamos todos como positivos?

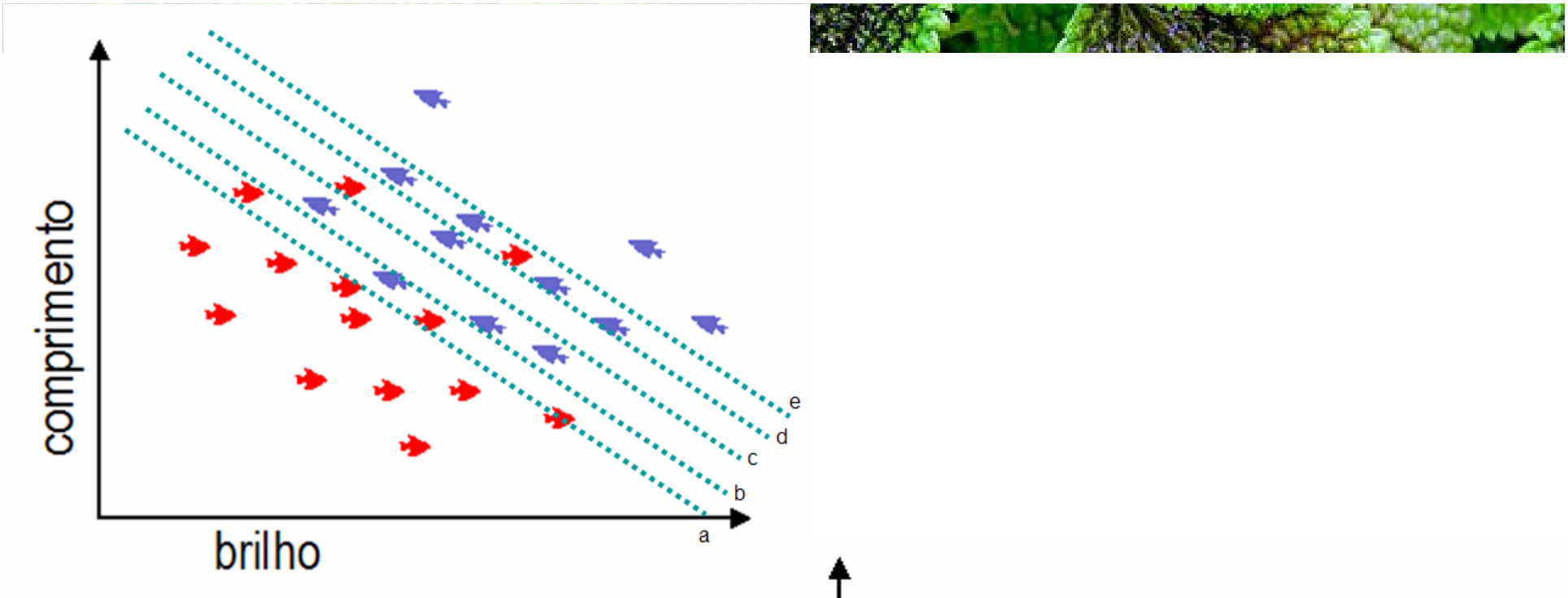
Limiar da classificação

- Vamos considerar um parâmetro para mover a superfície de separação criada
- Ao invés de utilizar apenas o lado da fronteira, utilizamos a “distância algébrica” de cada ponto à fronteira, obtendo um valor numérico do quanto cada ponto deve pertencer a cada categoria
- Alterando esse limiar, movemos a fronteira e obtemos diferentes resultados de classificação

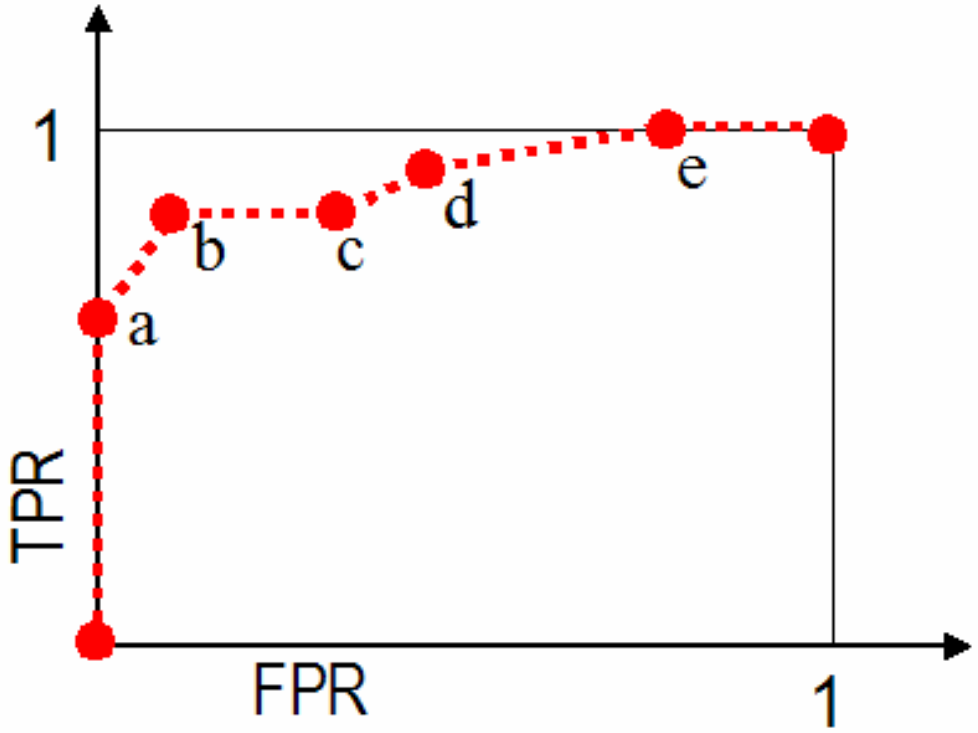

comprimento
brilho
a
b
c
d
e

Análise ROC

- Considerando os vários valores de limiar, plotamos a taxa de verdadeiros positivos (ou recall) em função da taxa de falsos positivos (taxa de alarmes falsos)
- Essa curva é chamada ROC (receiver operating characteristic curve) e permite analisar a classificação em diversos pontos de operação
- A reta diagonal de (0,0) a (1,1) corresponde a um classificador aleatório
- A poligonal (0,0)-(0,1)-(1,1) é o classificador ideal
- A área sob a curva ROC chamada AUC é uma medida para comparação de classificadores



limiar	FRP	TPR
a	0	9/14
b	1/12	12/14
c	4/12	12/14
d	5/12	13/14
e	9/12	14/14



Exemplo – Iris classificada

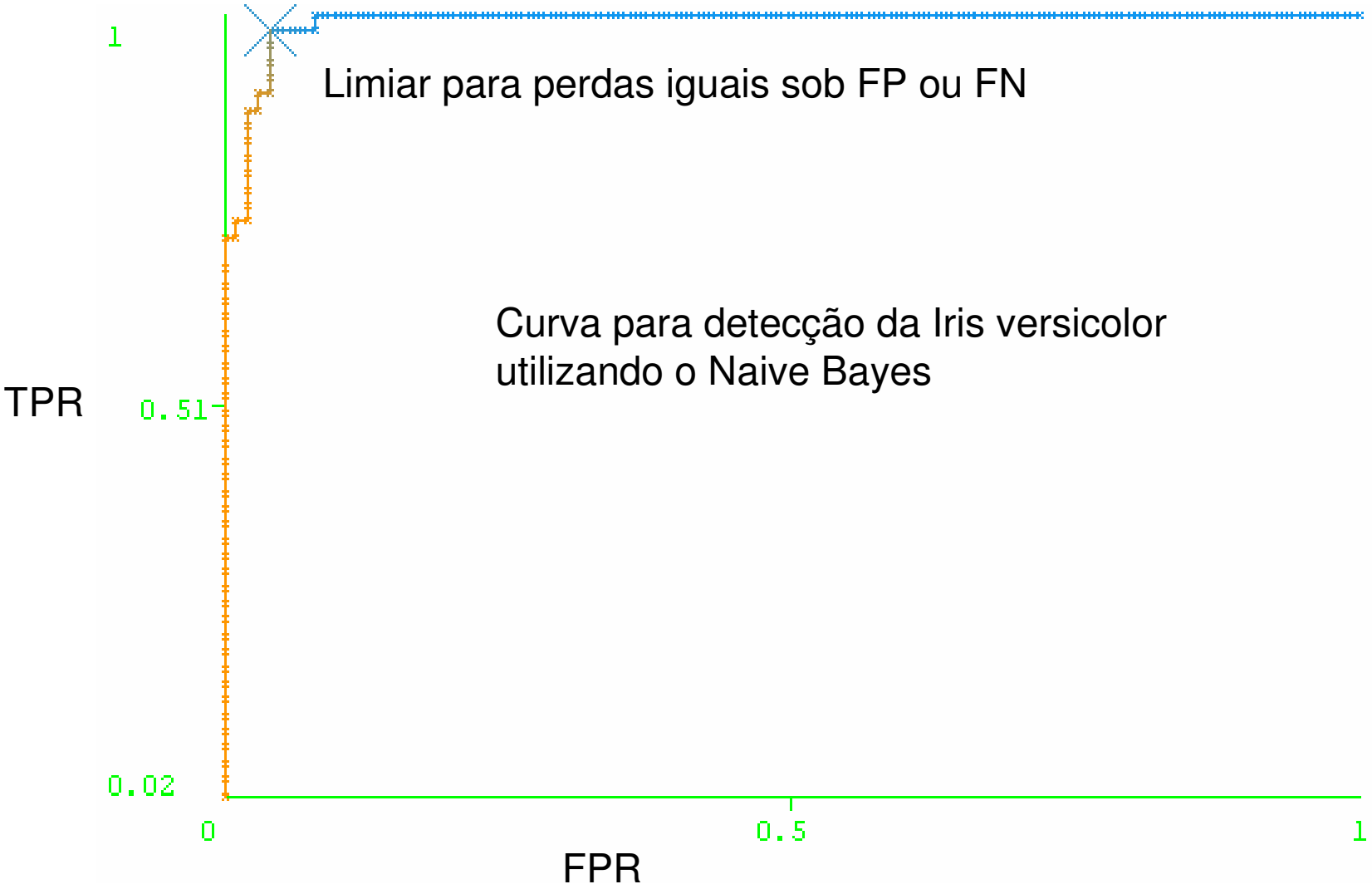
- Utilizando o Naive Bayes
- Validação Cruzada de 10 dobras

```
=== Confusion Matrix ===

  a  b  c   <-- classified as
 50  0  0 |  a = Iris-setosa
  0 48  2 |  b = Iris-versicolor
  0  4 46 |  c = Iris-virginica
```

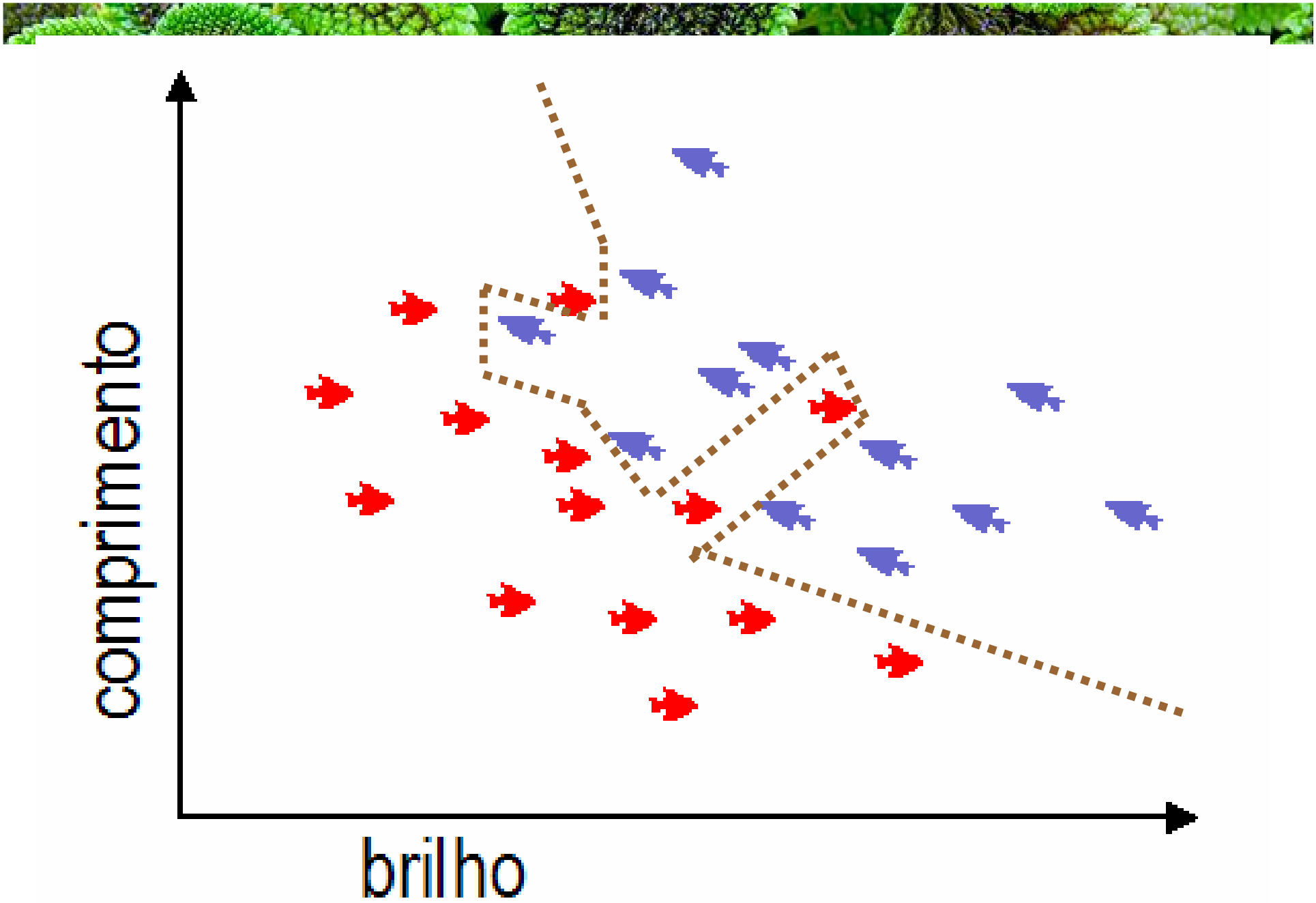
TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
1	0	1	1	1	1	Iris-setosa
0.96	0.04	0.923	0.96	0.941	0.992	Iris-versicolor
0.92	0.02	0.958	0.92	0.939	0.992	Iris-virginica

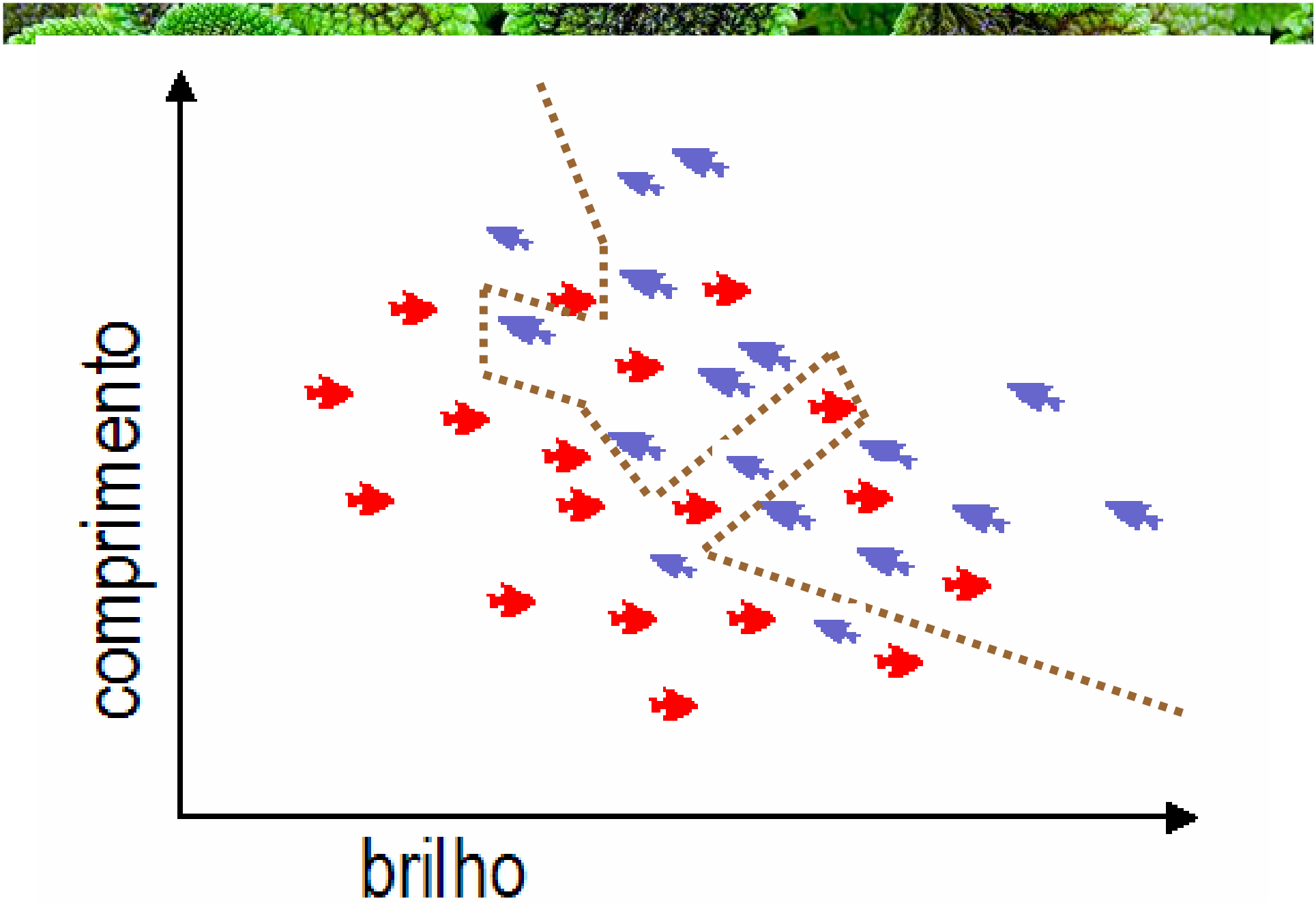
Curva ROC – Iris



Generalização

- Imagine que ao invés de utilizarmos uma reta podemos utilizar um polinômio de grau mais alto, podendo contornar os pontos da amostra para obter 100% de acerto.
- O quão melhor será esse classificador?
- Observe o que acontece se acrescentamos exemplos que não estão na amostra


comprimento
brilho


comprimento
brilho

Generalização

- A capacidade de generalização de um classificador está relacionada à simplicidade do modelo adotado.
- O efeito adverso observado chama-se overfitting, o modelo é excessivamente influenciado pelo caso particular da amostra.
- Se o grau de liberdade do modelo é alto, a técnica de regularização pode ser utilizada para promover uma restrição de suavidade ao modelo.

Medindo a generalização

- Para avaliar um classificador considerando seu poder de generalização existem várias estratégias
  - Separação de conjunto de treinamento e conjunto de teste
  - Leave-one-out: treina-se removendo um elemento da amostra e testa-se a classificação para o elemento removido, faz-se isso para cada elemento
  - N-fold cross validation: divide-se a amostra em N subconjuntos, remove-se um subconjunto, treina-se e avalia-se o classificador, repete para cada subconjunto

Seleção de modelo

- A avaliação de um classificador com generalização permite a comparação entre classificadores diferentes e a escolha de parâmetros que otimizem os critérios desejados do classificador.
- Se um conjunto de treinamento é utilizado para construir o classificador e um conjunto de teste é utilizado para escolher o melhor modelo, é necessário um terceiro conjunto de dados para poder avaliar o classificador obtido.

Teoria da decisão

- Vamos estabelecer um custo para cada tipo de erro.
  - Qual o prejuízo em se colocar um salmão num engradado de robalos?
  - Qual o prejuízo em se colocar um robalo no engradado de salmão?

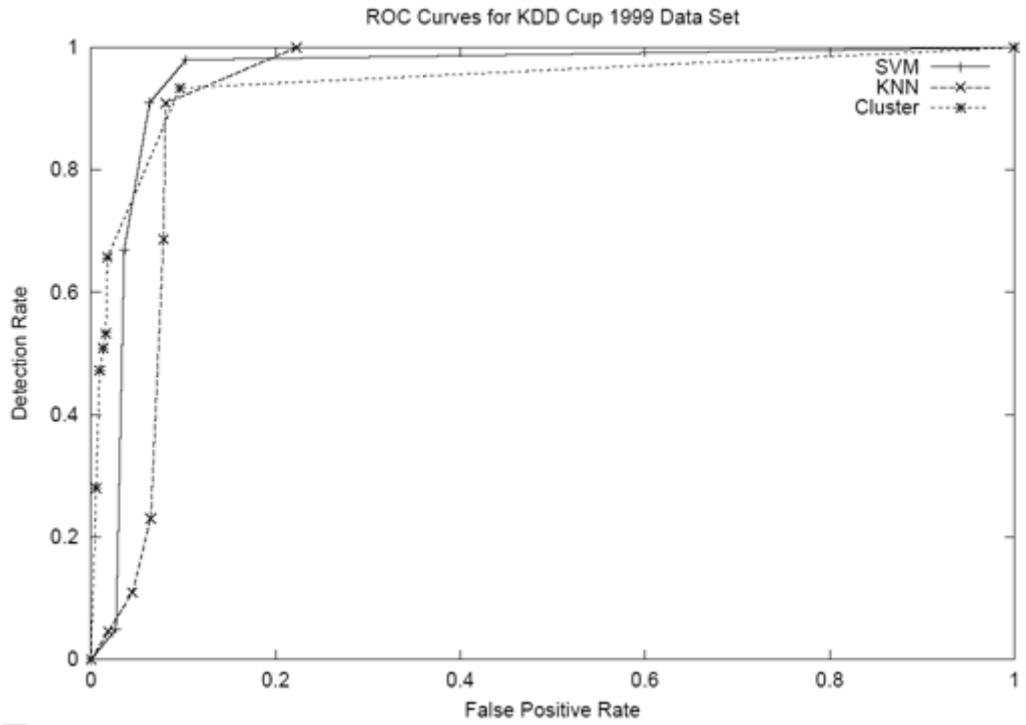
  - Qual o prejuízo em deixar de detectar um intruso?
  - Qual o prejuízo em cada falso alarme que é gerado?

Risco

- O risco compreende a somatória da chance de cometer cada tipo de erro vezes o custo desse erro.
- O limiar do classificador pode ser obtido através da minimização desse risco.

Exemplo – Detecção de intrusão

- O custo dos falsos positivos é muito alto, pois o evento da intrusão é muito raro.

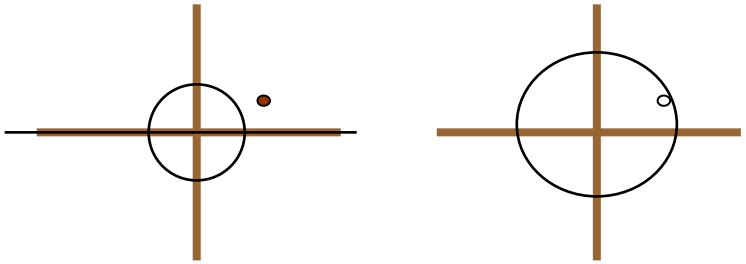


http://www.music.columbia.edu/~mike/research/ids/

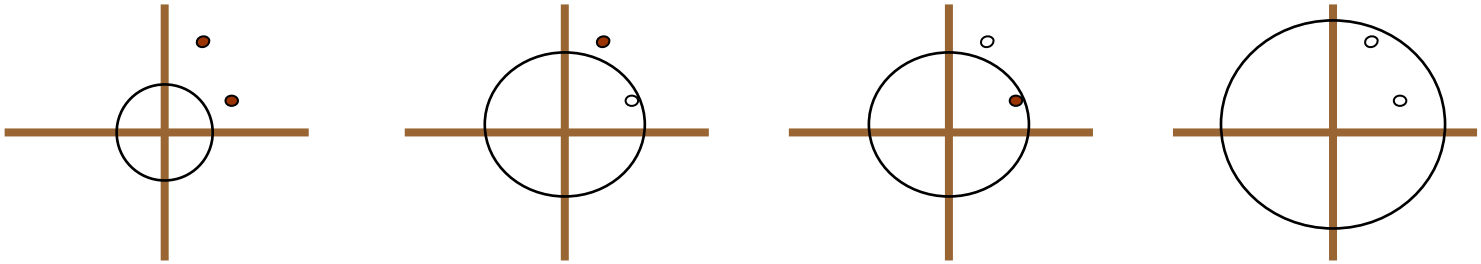
VC-dimension

- Considere N pontos rotulados +/- das 2^N formas possíveis
- Um modelo de classificador divide N pontos se existe uma instância desse classificador que classifique corretamente para cada uma das possíveis 2^N configurações
- A dimensão VC (*Vapnik–Chervonenkis*) de um modelo de classificação é o maior N para que o modelo consegue dividir

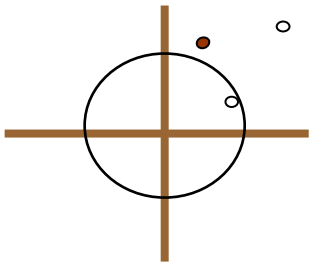
f(x,b) = sign(x.x-b)



f(x,q,b) = sign(qx.x-b)



Mas...



Algumas representações

- Classificador bayesiano
- Naive Bayes
- K-Vizinhos
- Árvore de Decisão
- Rede neural
- Máquina de vetor de suporte
- Composição de classificadores

Regra de Bayes

$C_1 \ldots C_n$ são eventos mutuamente exclusivos e exaustivos com probabilidades $P(C_i)$ conhecidas.

B é um evento para o qual se conhece $P(B|C_i)$.

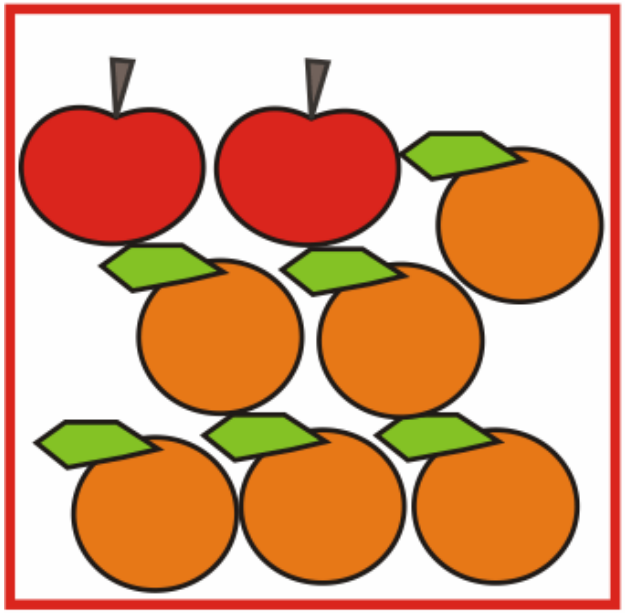
Interpretar C_i como possíveis causas para o evento B.

Computar $P(C_i|B)$

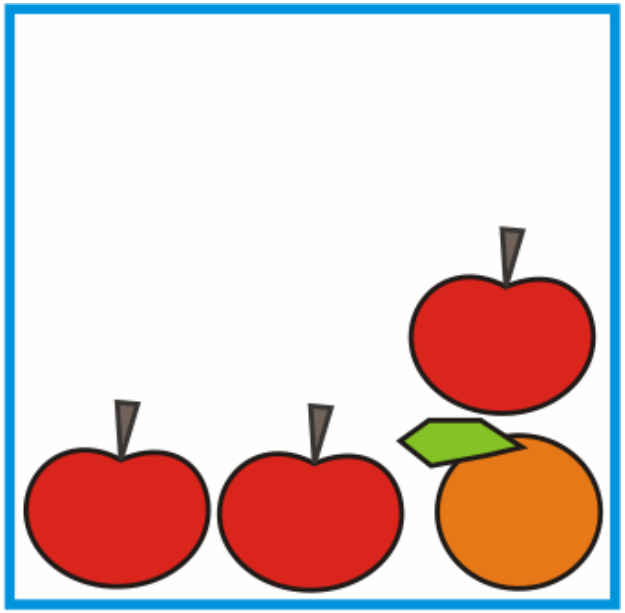
$$P(C_i|B) = \frac{P(BC_i)}{P(B)} = \frac{P(B|C_i)P(C_i)}{\sum_{j=1}^{n} P(B|C_j)P(C_j)}$$

Inferência bayesiana

- Exemplo de aplicação da regra de bayes
  Eu pego 40% das vezes uma fruta da caixa vermelha e 60% da caixa azul. Sabendo que peguei uma laranja, qual a probabilidade de tê-la pego da caixa azul?



40%



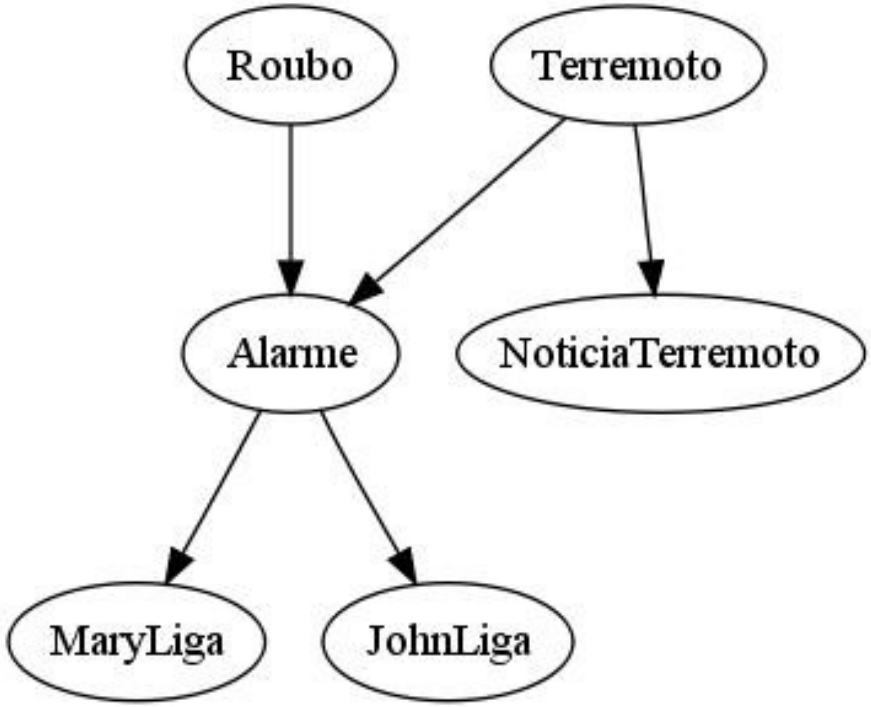
60%

Classificador bayesiano

- O classificador bayesiano assume que cada classe é distribuída na forma de uma gaussiana multidimensional
- Da amostra correspondente a cada classe, obtém-se o vetor média e a matriz de covariância, suficientes para representar a gaussiana
- Um ponto é classificado de acordo com a classe com maior valor de probabilidade (verossimilhança)

Rede Bayesiana

- Grafo direcionado acíclic
  que representa
  dependência entre
  variáveis
- Forma resumida de
  representação da
  distribuição conjunta



Para cada nó, é armazenada, ou a probabilidade a priori de cada possível valor, ou as probabilidades condicionais para cada possível valor dado cada possível tupla dos nós pais.

Naive Bayes

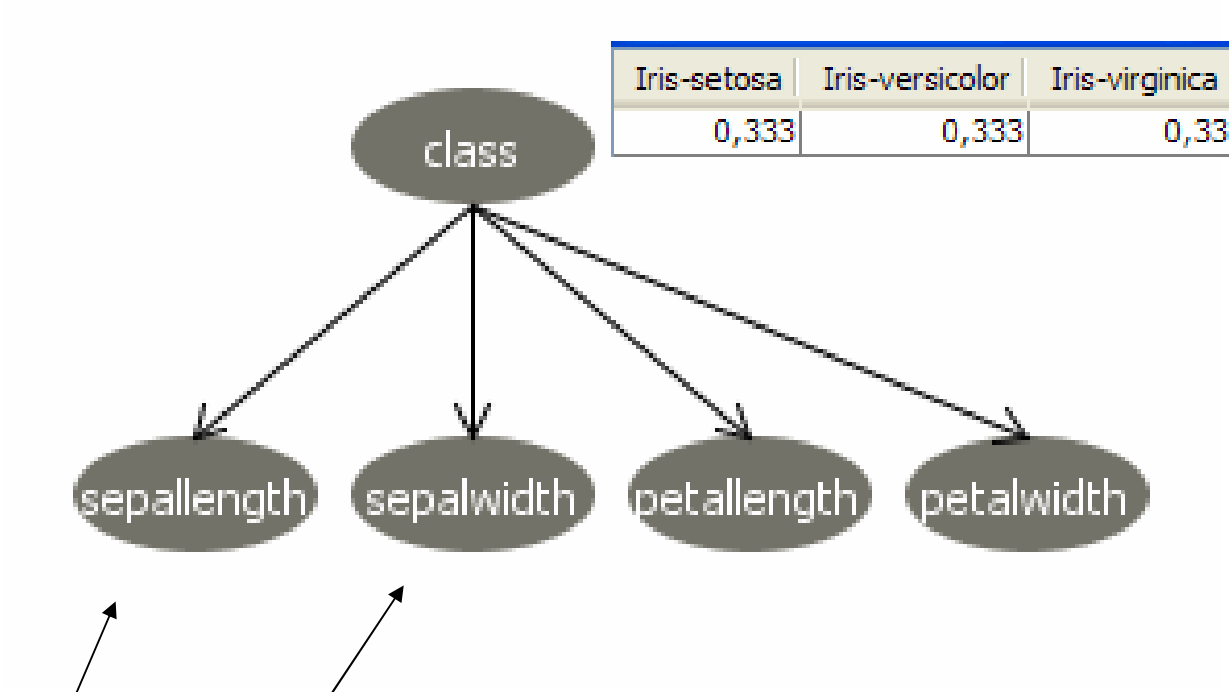
Supomos os atributos x_j estatisticamente independentes.

$$p(\mathbf{x}|C_k) = \prod_{j=1}^{M} p(x_j|C_k), k = 1, \ldots, L$$

Associar $\mathbf{x}$ à classe C_m que:

$$C_m = arg\max_{C_k} \prod_{j=1}^{M} p(x_j|C_k), k = 1, \ldots, L$$

O classificador Naive Bayes é um caso particular de rede bayesiana.



Iris-setosa	Iris-versicolor	Iris-virginica
0,333	0,333	0,333

class	'(-inf-2.95]'	'(2.95-3.35]'	'(3.35-inf)'
Iris-setosa	0,049	0,359	0,592
Iris-versicolor	0,67	0,301	0,029
Iris-virginica	0,417	0,476	0,107

class	'(-inf-5.55]'	'(5.55-6.15]'	'(6.15-inf)'
Iris-setosa	0,922	0,068	0,01
Iris-versicolor	0,223	0,456	0,32
Iris-virginica	0,029	0,204	0,767

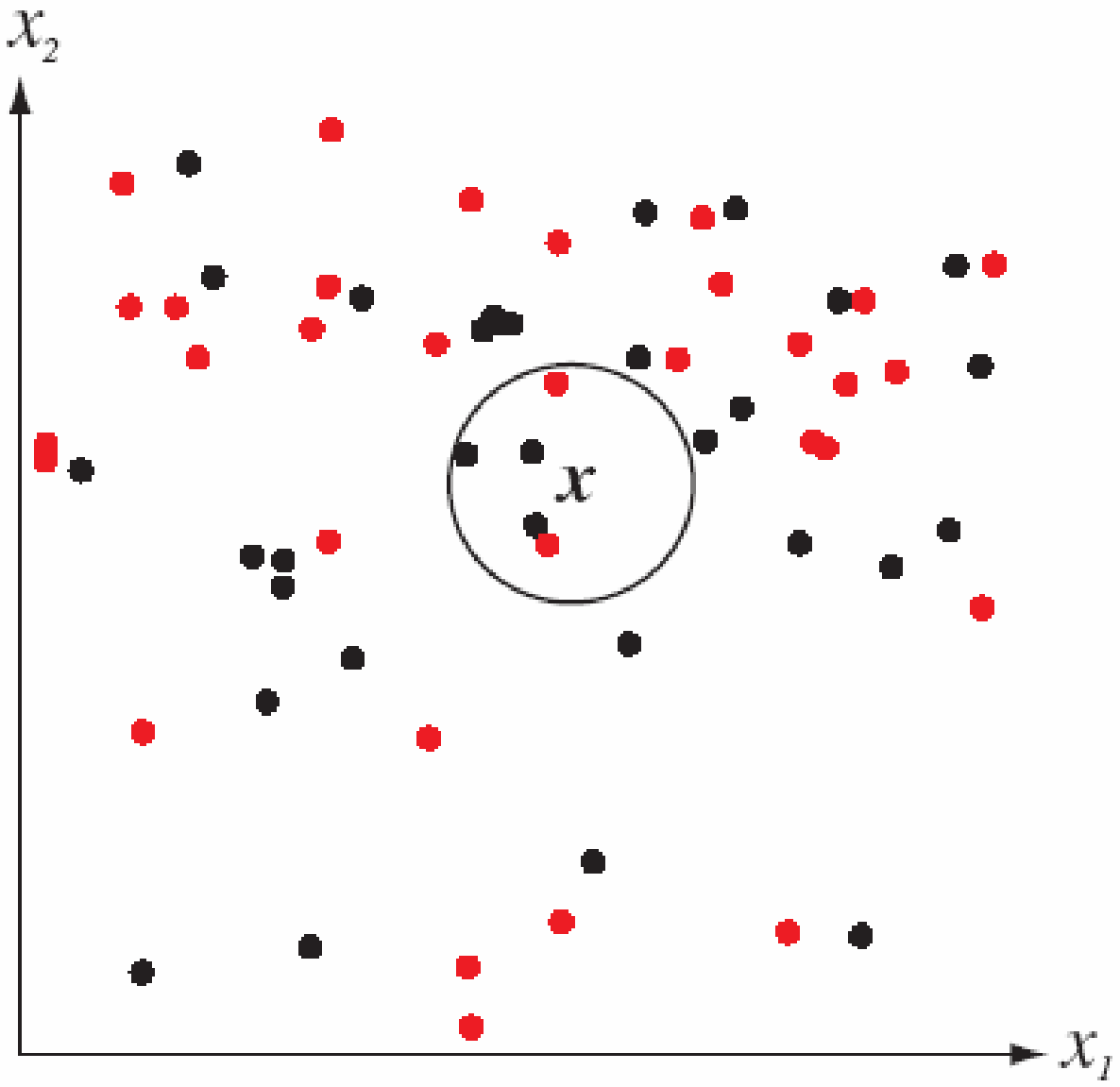
Exemplo - Iris

Naive Bayes Classifier

Attribute	Class Iris-setosa (0.33)	Iris-versicolor (0.33)	Iris-virginica (0.33)
sepallength			
mean	4.9913	5.9379	6.5795
std. dev.	0.355	0.5042	0.6353
weight sum	50	50	50
precision	0.1059	0.1059	0.1059
sepalwidth			
mean	3.4015	2.7687	2.9629
std. dev.	0.3925	0.3038	0.3088
weight sum	50	50	50
precision	0.1091	0.1091	0.1091
petallength			
mean	1.4694	4.2452	5.5516
std. dev.	0.1782	0.4712	0.5529
weight sum	50	50	50
precision	0.1405	0.1405	0.1405
petalwidth			
mean	0.2743	1.3097	2.0343
std. dev.	0.1096	0.1915	0.2646
weight sum	50	50	50
precision	0.1143	0.1143	0.1143

K-Vizinhos

- Para classificar um ponto, procuram-se os K pontos da amostra mais próximos (de menor distância ao ponto dado)
- Dos vizinhos, verifica-se qual a classe mais frequente
- Atribui-se esta classe ao ponto

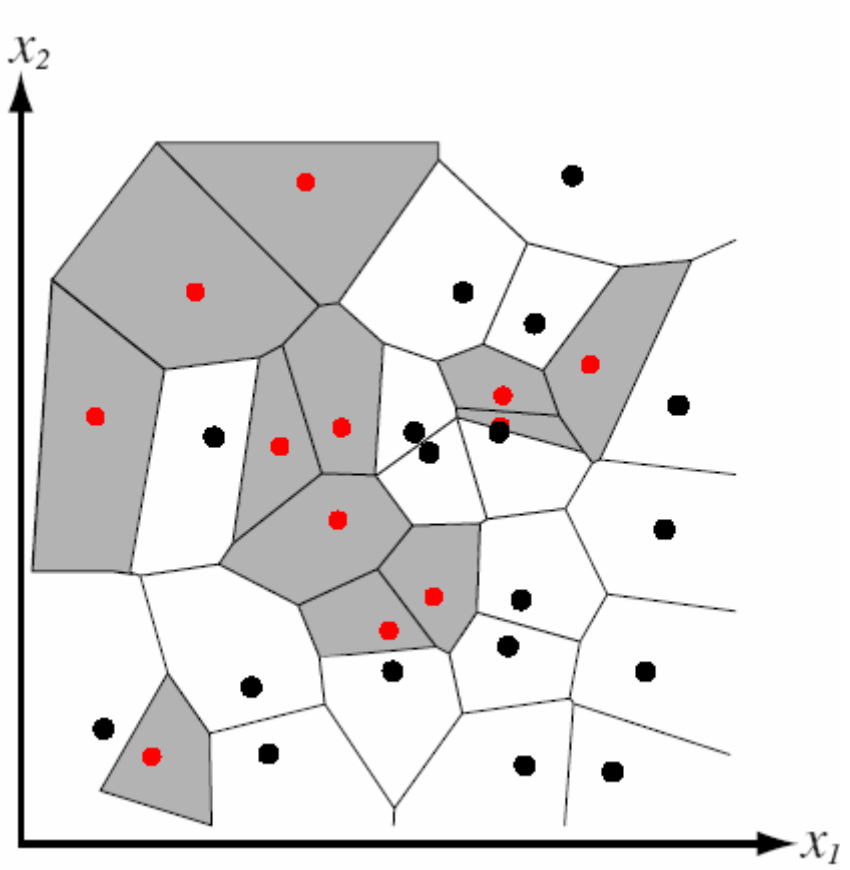

x_2
x
x_1

Método dos K-Vizinhos mais próximos

- Não há necessidade de treinamento
- O custo é maior na classificação
- Uma boa estrutura de dados para facilitar a busca dos vizinhos mais próximos acelera o algoritmo
- O modelo de classificador se adapta melhor a distribuições pouco convencionais

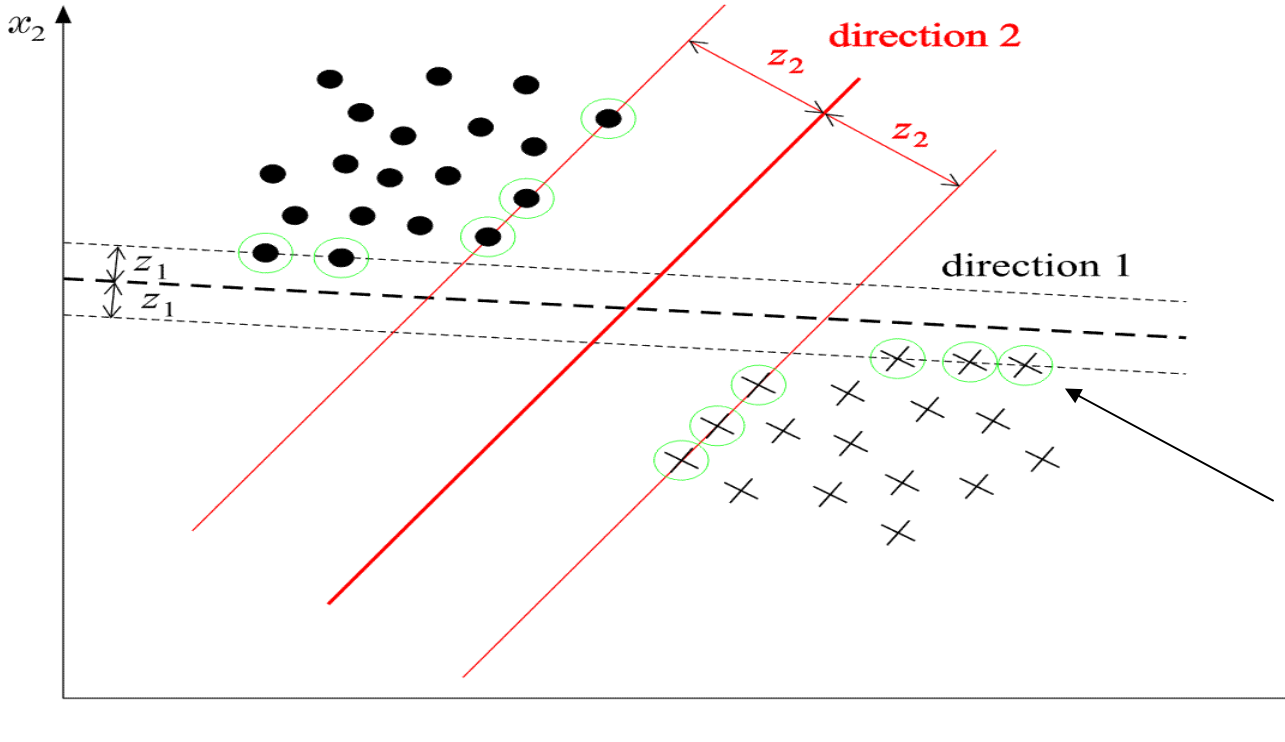
Método dos K-Vizinhos mais próximos

- Para K=1 é obtido um mosaico de Voronoi
- Para K maiores há maior suavização da superfície de separação
- Podem-se utilizar ponderações da importância de acordo com a distância, configurando um método de kernel
- Essa é uma forma de estimação de densidade



SVM (máquina de vetor de suporte)

- Classificador da forma $g(\underline{x}) = \underline{w}^T \underline{x} + w_0 = 0$ que mantém a máxima margem entre conjuntos linearmente separáveis



A solução é um problema de programação quadrática com restrições lineares

A solução equivale a encontrar um conjunto de vetores de suporte

Boosting

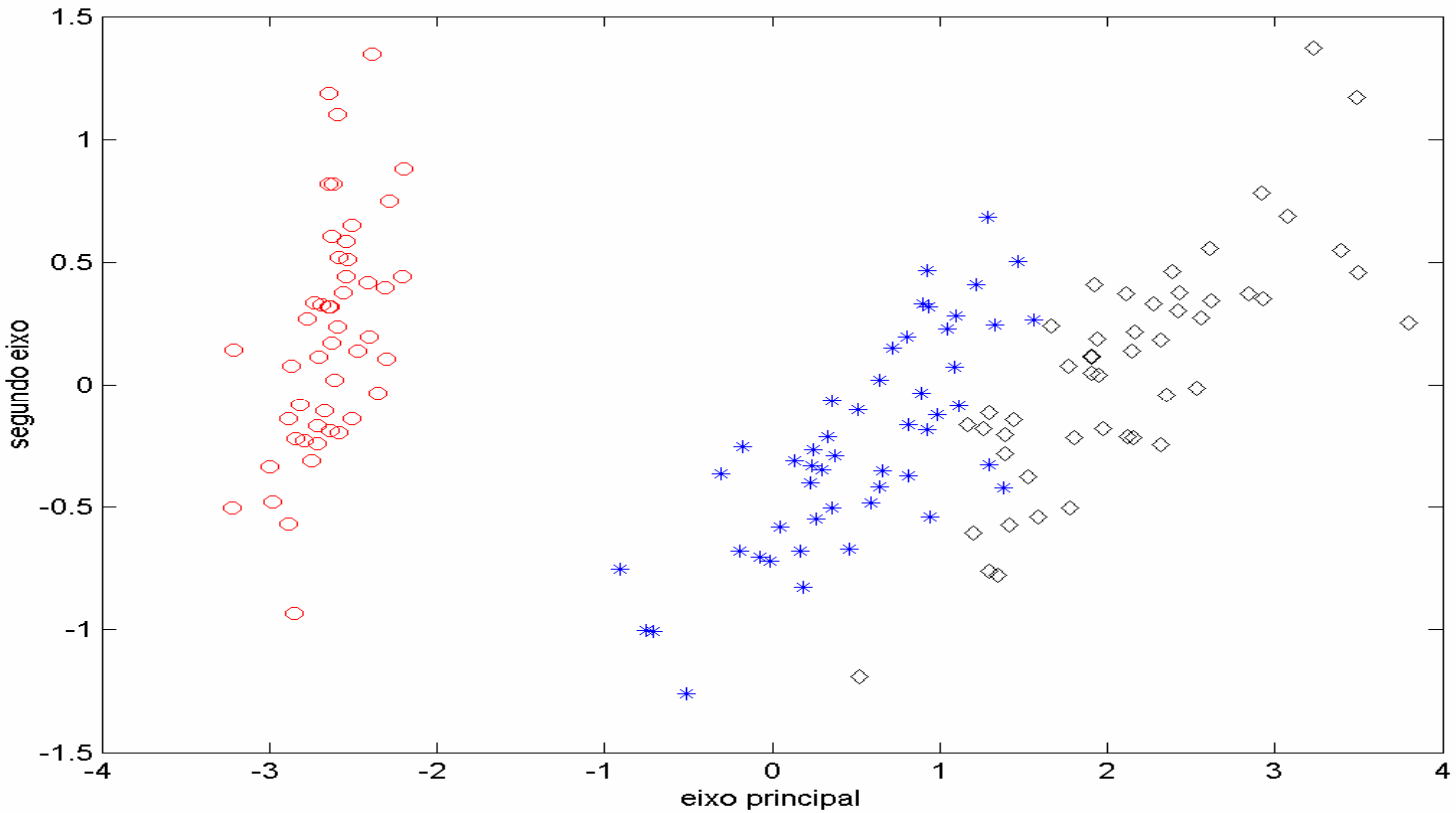
- Partindo-se de um classificador fraco (acerta mais que 50% que corresponde ao classificador aleatório)
- Atribuem-se pesos às instâncias
- Aumentam-se os pesos das instâncias classificadas erroneamente pelo classificador i, e trina-se o classificador i+1
- Combinam-se os resultados dos classificadores utilizando uma média ponderada

Seleção de feições

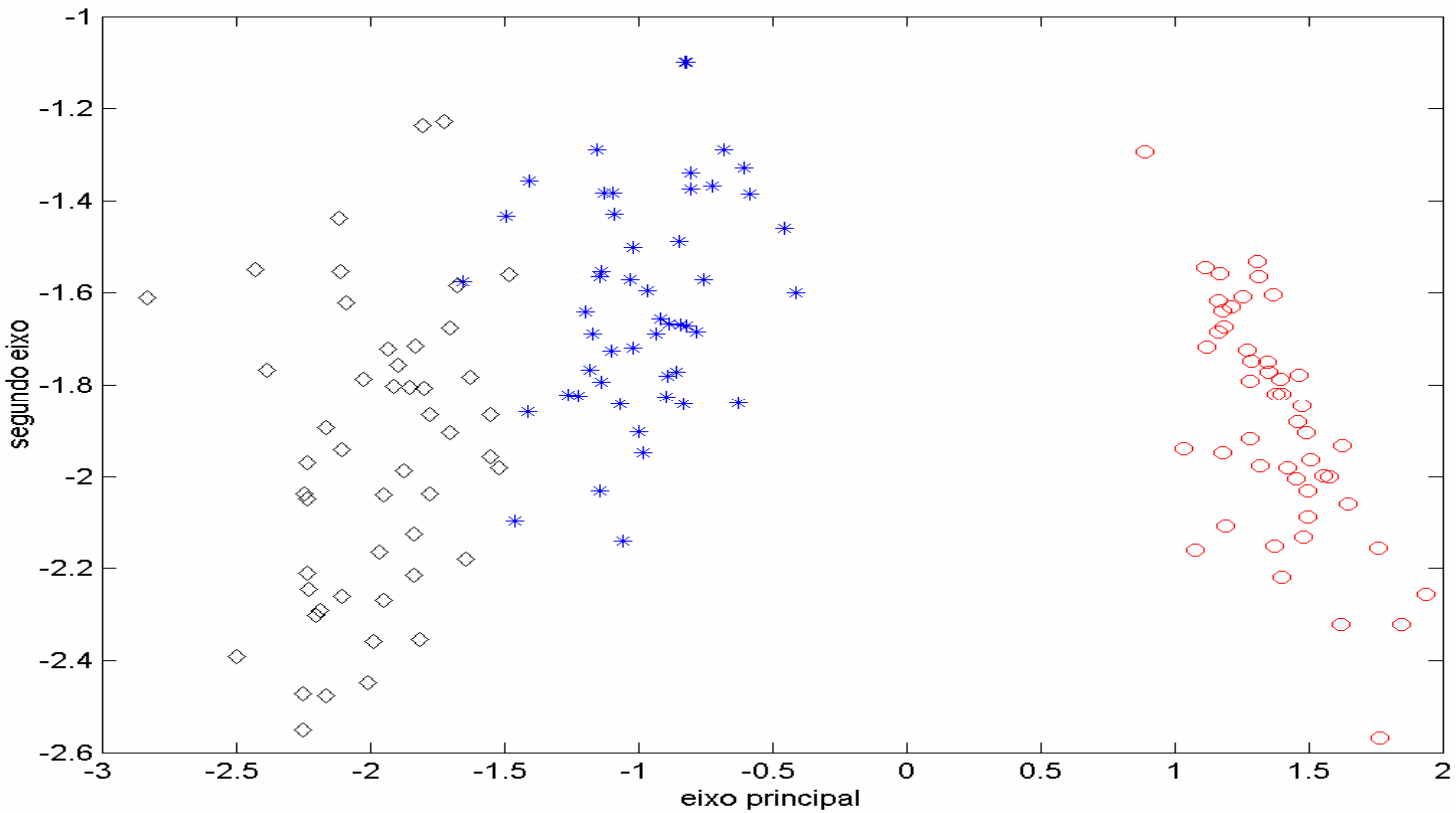
- Escolha de subconjunto de atributos
- Busca de projeção
  - PCA (Análise de componentes principais)
  - LDA (Linear Discriminant Analysis)
  - ICA (Independent components analysis)

  - PCA é uma base ortogonal fundada nos autovetores da matriz de covariância, LDA se baseia em matrizes de espalhamento inter-classe e intra-classe (supervisionado), ICA se baseia em métricas de não-gaussianidade (curtose, negentropia)

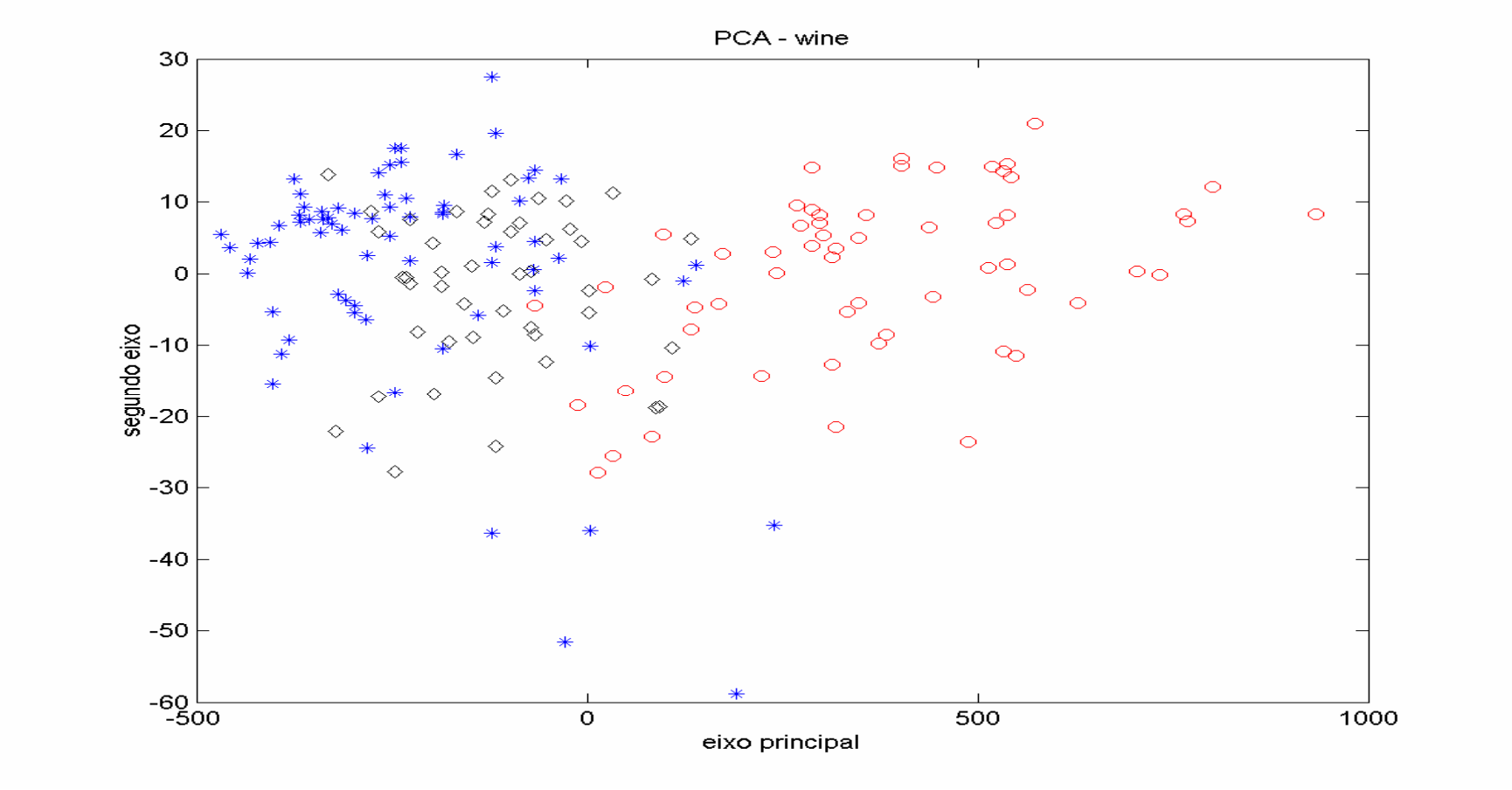
PCA - Iris



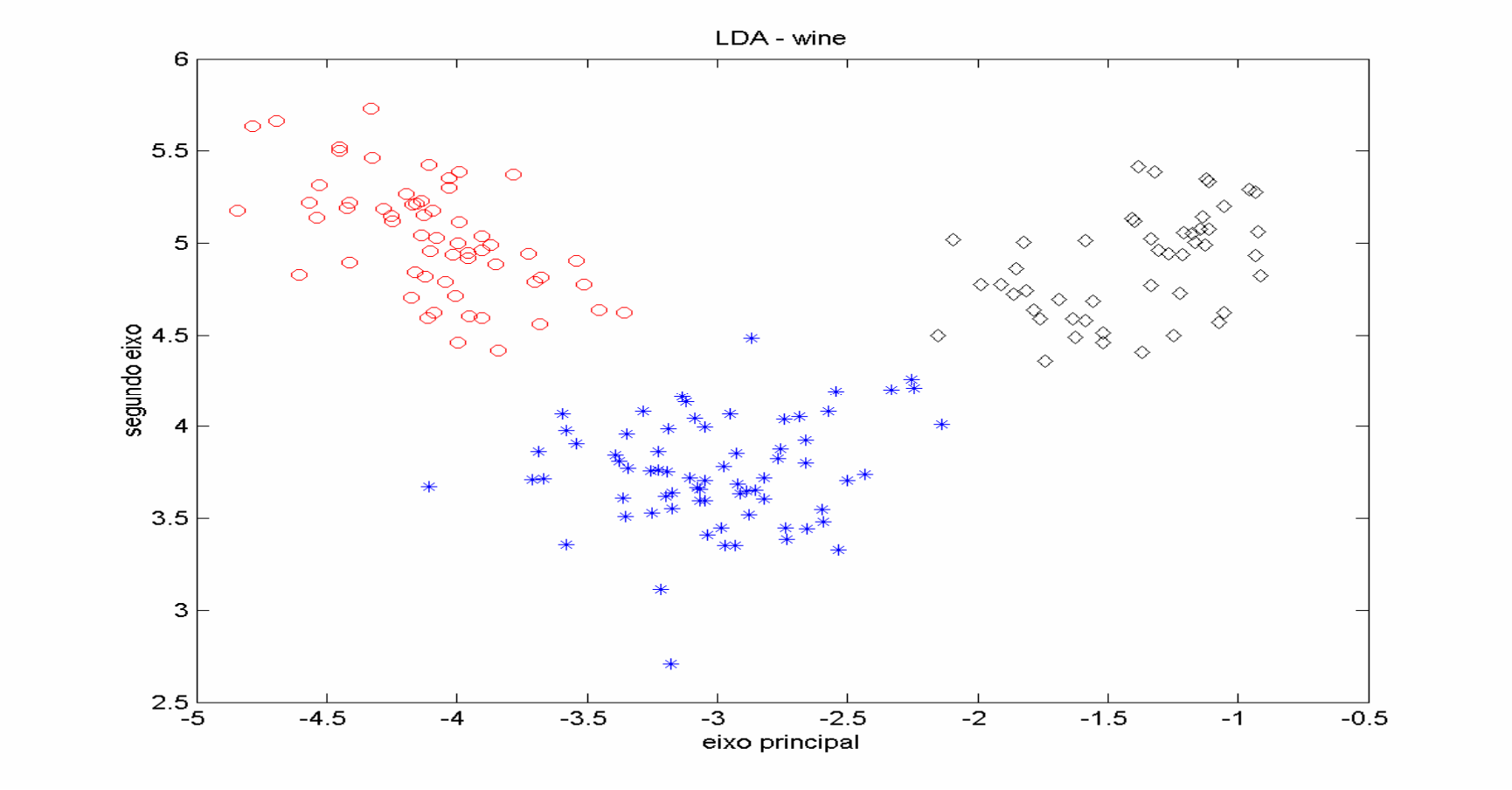
LDA - Iris


segundo eixo
eixo principal
-1
-1.2
-1.4
-1.6
-1.8
-2
-2.2
-2.4
-2.6
-3
-2.5
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2

PCA - Wine


PCA - wine
segundo eixo
eixo principal
30
20
10
0
-10
-20
-30
-40
-50
-60
-500
0
500
1000

LDA - Wine


LDA - wine
segundo eixo
eixo principal
6
5.5
5
4.5
4
3.5
3
2.5
-5
-4.5
-4
-3.5
-3
-2.5
-2
-1.5
-1
-0.5

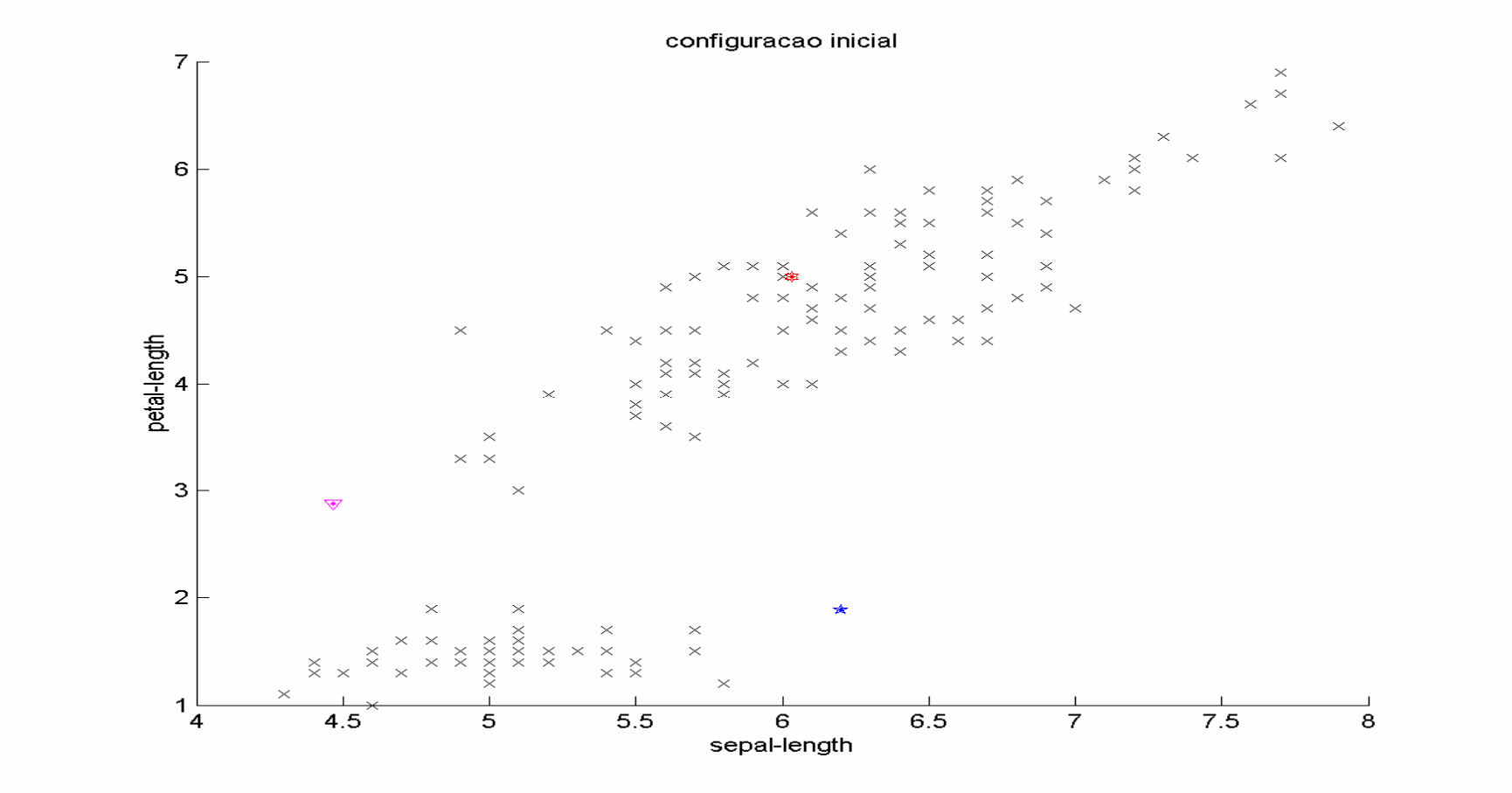
Clustering

- Método K-Means
  - Iniciam-se K centros aleatórios
  - Repetir:
    - Associar cada ponto ao centro mais próximo
    - Recalcular a nova posição dos centros
  - Parar quando não houver mudança de associação (rótulo)

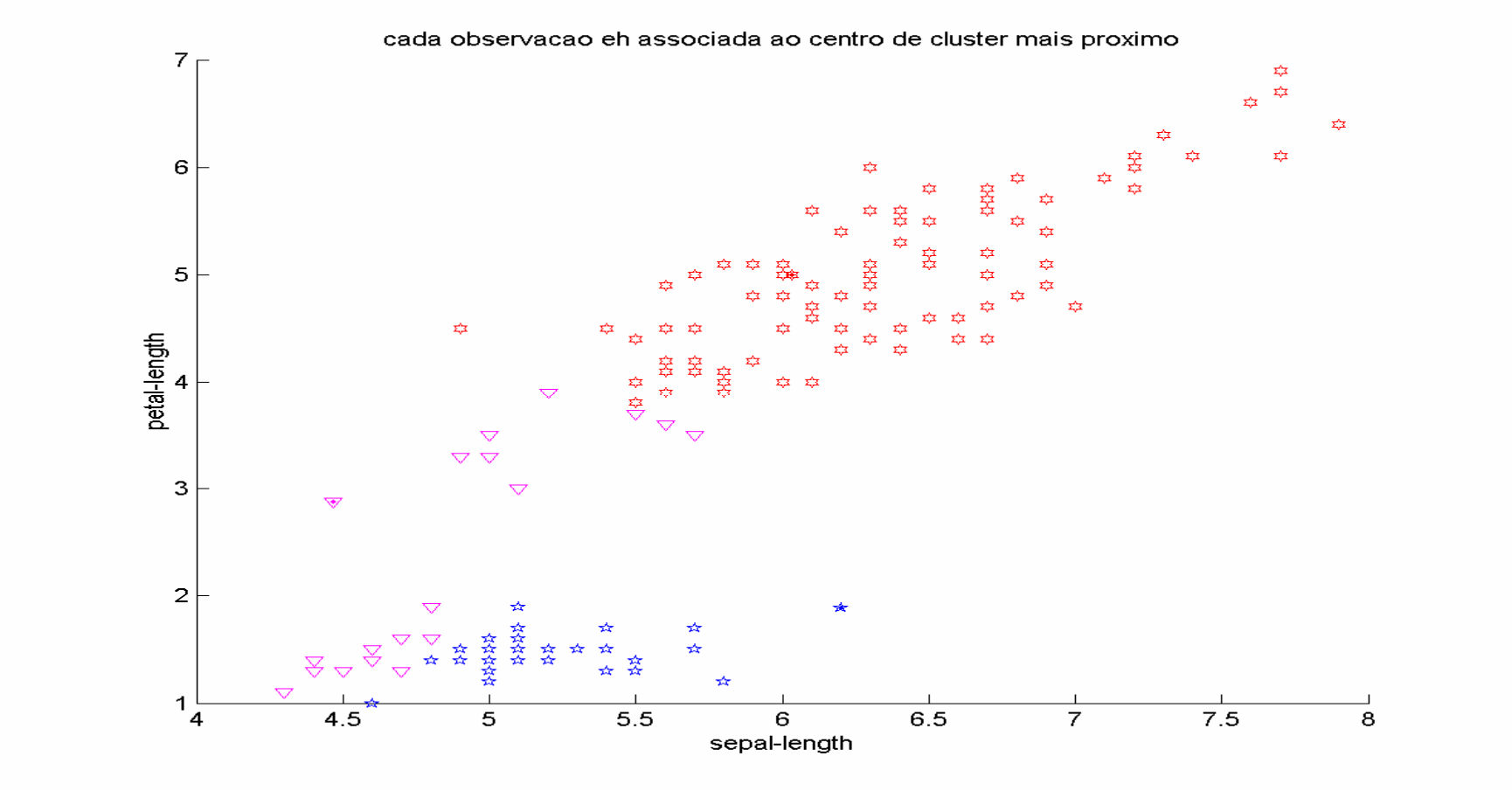
K-Means

- Sempre converge, mas nem sempre para o máximo global e nem sempre tão rápido quanto deveria
- Depende muito da escolha inicial dos centros
- Há esquemas para se reiniciar o método e unir os resultados obtidos
- Alguns centros podem não ter nenhum ponto associado no final

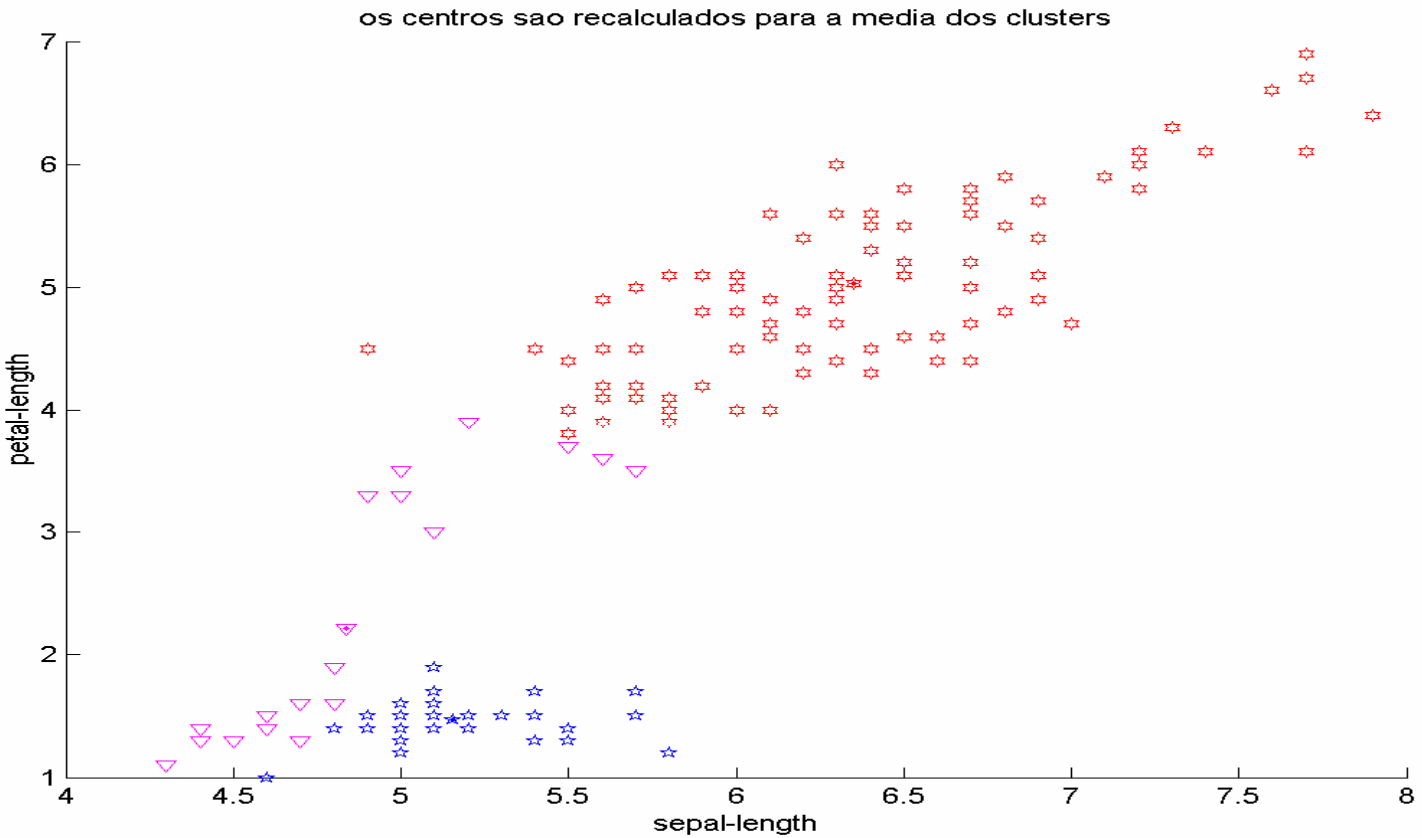
(início)


configuracao inicial
petal-length
sepal-length
1
2
3
4
5
6
7
4
4.5
5
5.5
6
6.5
7
7.5
8

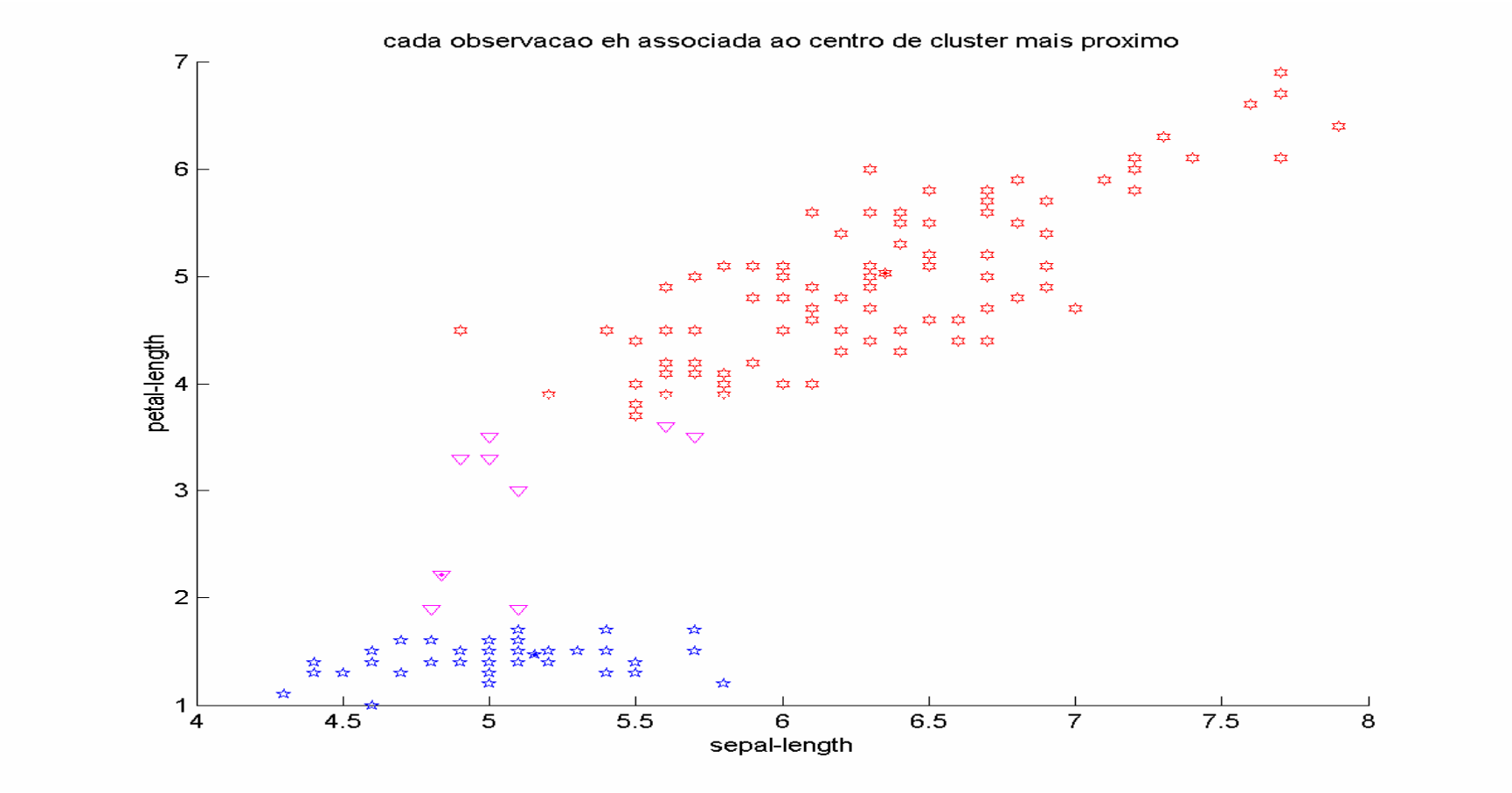
(associação)


cada observacao eh associada ao centro de cluster mais proximo
petal-length
sepal-length
1
2
3
4
5
6
7
4
4.5
5
5.5
6
6.5
7
7.5
8

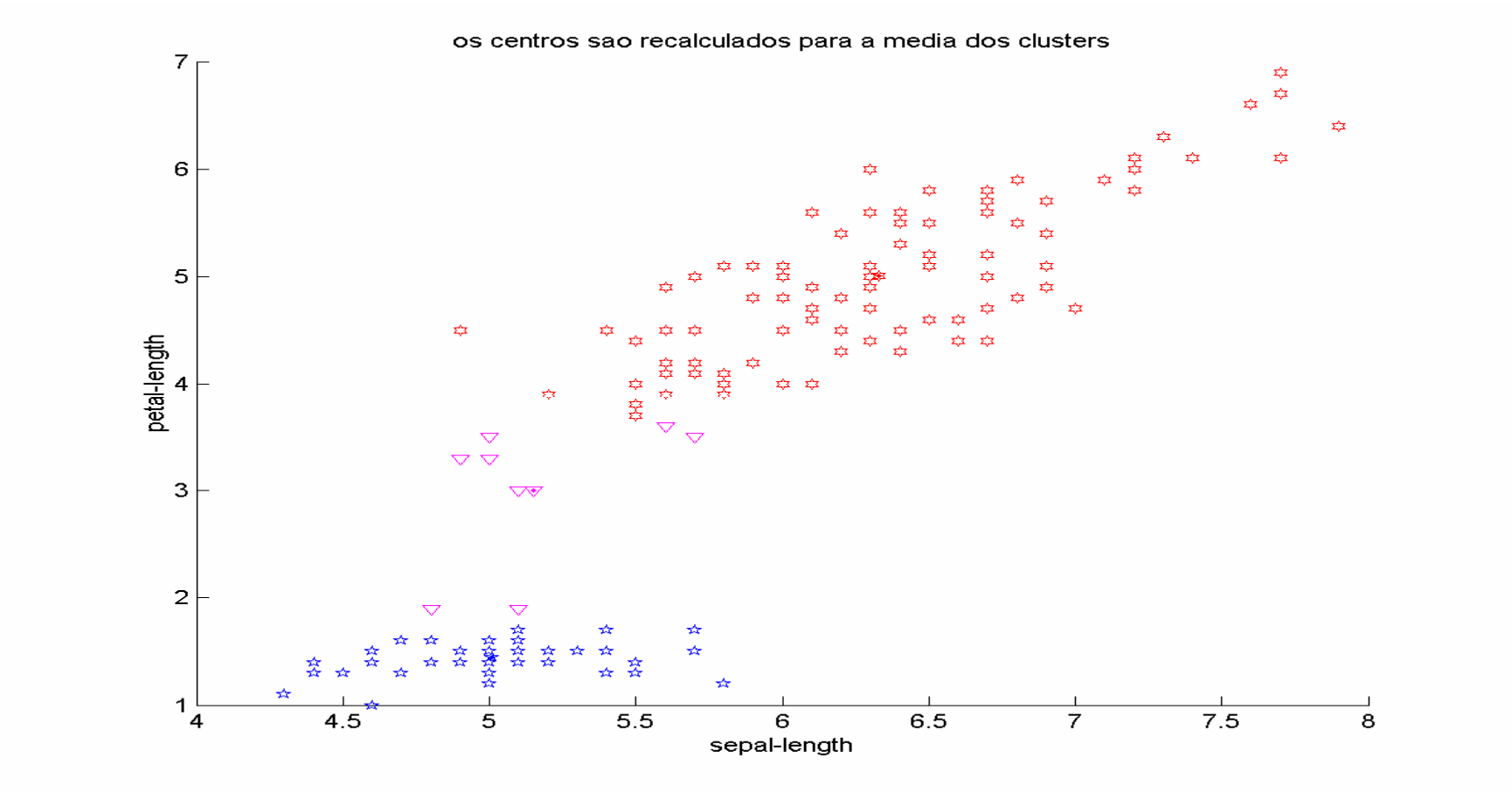
(reposicionamento)



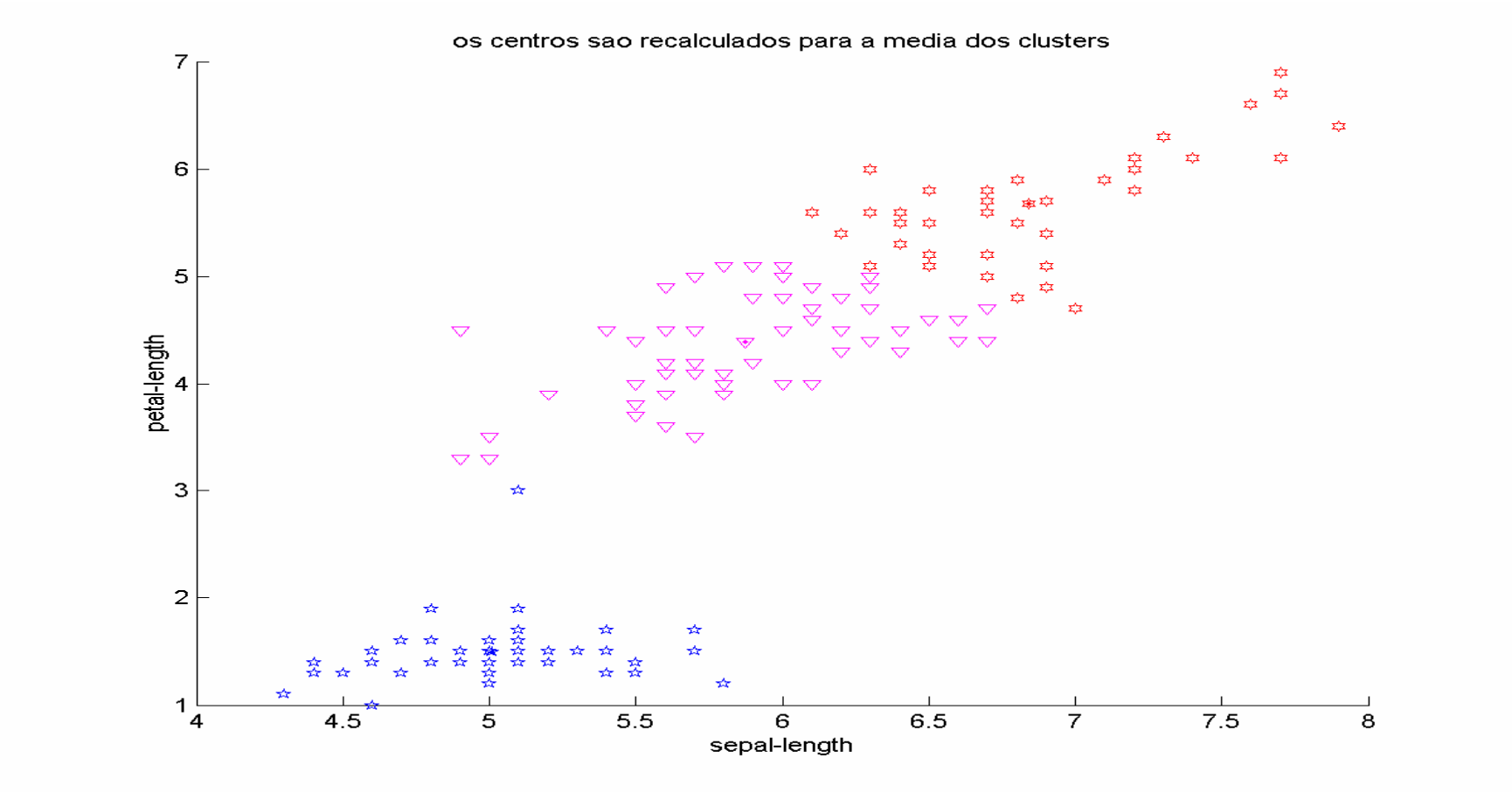
(associação)


cada observacao eh associada ao centro de cluster mais proximo
petal-length
sepal-length
1
2
3
4
5
6
7
4
4.5
5
5.5
6
6.5
7
7.5
8

(reposicionamento)



(final)


os centros sao recalculados para a media dos clusters
petal-length
sepal-length
1
2
3
4
5
6
7
4
4.5
5
5.5
6
6.5
7
7.5
8

Clustering

- Minimum Spanning Tree
  - Considerando os pontos a serem agrupados e uma matriz de dissimilaridade representada num grafo completo
  - A árvore é um grafo sem ciclos
  - A árvore que contém todos os nós do grafo e as arestas cuja soma dos custos é mínima é a árvore geradora mínima
  - Se removermos qualquer aresta da árvore, dividimos o grafo em 2 subgrafos conexos
  - As arestas exageradamente longas costumam caracterizar a existência de agrupamentos

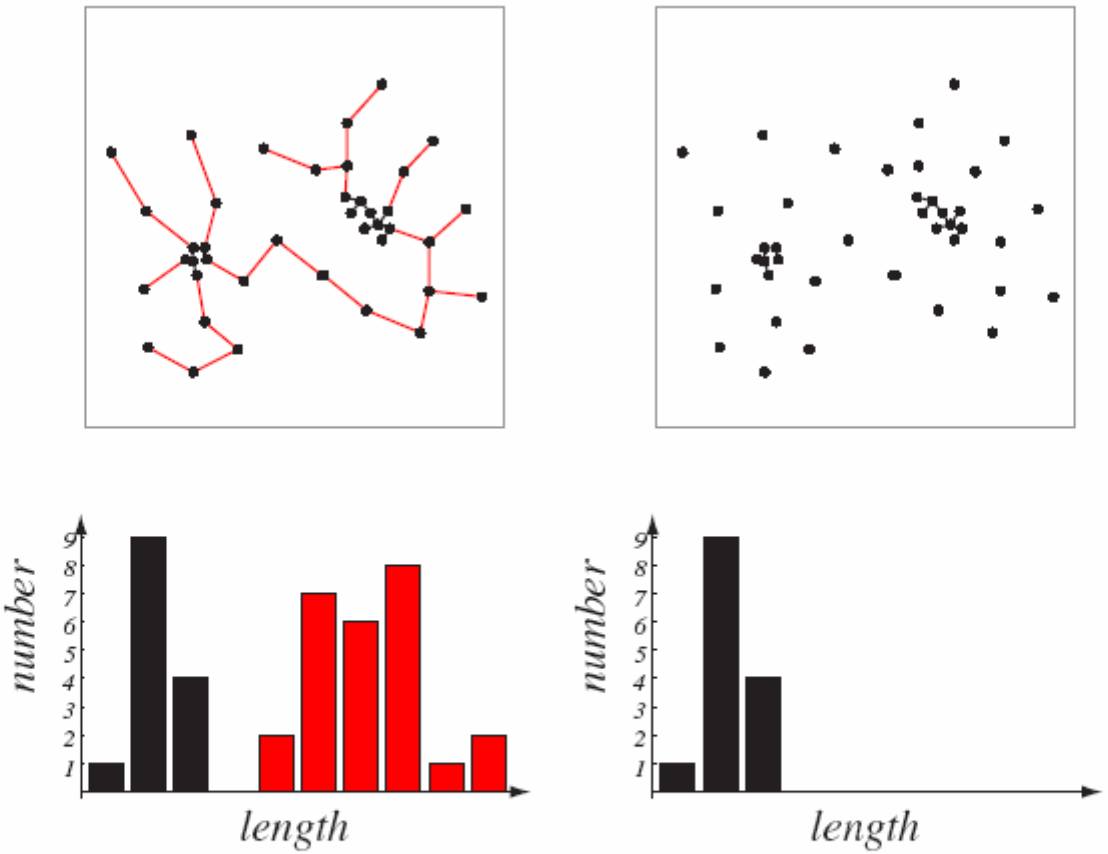


FIGURE 10.21. A minimal spanning tree is shown at the left; its bimodal edge length distribution is evident in the histogram below. If all links of intermediate or high length are removed (red), the two natural clusters are revealed (right). From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.