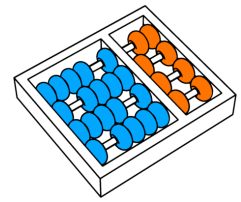


Introdução ao Processamento Digital de Imagem
MC920 / MO443

Prof. Hélio Pedrini

Instituto de Computação

UNICAMP

http://www.ic.unicamp.br/~helio

1º Semestre de 2019

Roteiro

- O *reconhecimento de padrões* visa determinar um mapeamento para relacionar *características* ou *propriedades* extraídas de amostras com um conjunto de rótulos.
- Amostras com características semelhantes devem ser mapeadas ao mesmo rótulo.
- Quando se atribui um mesmo rótulo a amostras distintas, diz-se que tais elementos pertencem a uma mesma *classe*, esta caracterizada por compreender elementos que compartilham propriedades em comum.
- Cada classe recebe um dentre os rótulos $C_1, C_2, \ldots, C_m$, em que m denota o número de classes de interesse em um dado problema.

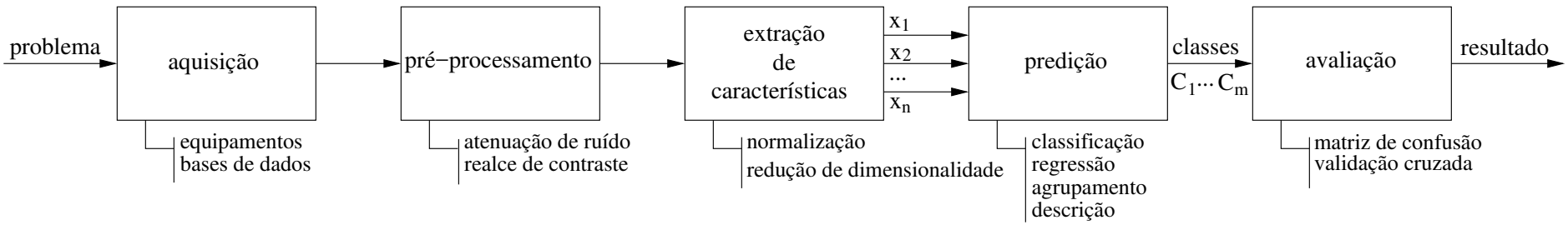
Reconhecimento de Padrões

Áreas de aplicação

- A tarefa de classificação ocorre em uma variedade de atividades humanas:
  - Medicina
  - Sensoriamento Remoto
  - Biometria (reconhecimento de faces, íris, impressões digitais, voz)
  - Automação Industrial
  - Biologia
  - Astronomia
  - Meteorologia

Reconhecimento de Padrões

Componentes



- Aquisição
  - Provavelmente o componente mais custoso
  - Quantas amostras são suficientes?
- Pré-processamento
  - Preparação dos dados
- Extração de características
  - Crítico para resolução do problema
  - Requer conhecimento prévio
- Predição
  - Adaptação do modelo para explicar os dados
  - Treinamento
  - Escolha do modelo: sintático, estatístico, estrutural, rede neural
- Avaliação
  - Acurácia
  - Sobreajuste $\times$ generalização

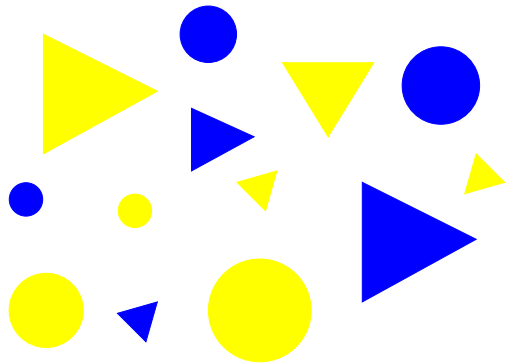
- Uma característica é uma propriedade mensurável de uma amostra.
- Características devem ser idealmente discriminativas e independentes.
- O conjunto de características normalmente é agrupado em um *vetor* ou *descritor de características*, representado como

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

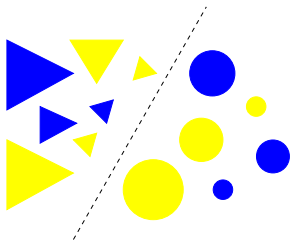
em que x_i denota o i-ésimo descritor e n é o número total desses descritores.

- O espaço n-dimensional definido pelo vetor de características é chamado de espaço de características.

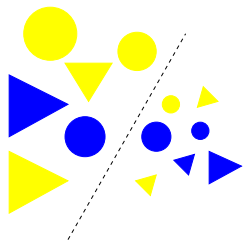
- possíveis características: forma, tamanho, cor, textura.



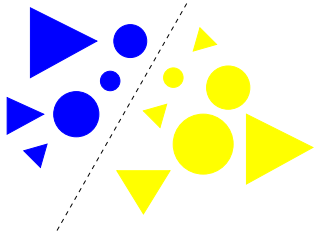
- separação em diferentes classes



forma



tamanho



cor

- dependência do domínio / informação a priori
- custo de extração
- preferencialmente baixa dimensionalidade
- características discriminativas
  - valores similares para padrões similares (baixa variabilidade intraclasse)
  - valores diferentes para padrões diferentes (alta variabilidade interclasse)
- características invariantes com respeito à rotação, escala, translação, oclusão, deformações
- robustez em relação a ruído
- não correlação entre características
- normalização

- Padrão é uma composição de características de amostras.
- Em tarefas de reconhecimento, um padrão é um par de variáveis $\{\mathbf{x}, c\}$, tal que
  - $\mathbf{x}$ é uma coleção de observações ou características (vetor de características)
  - c é o conceito associado à observação (rótulo)

- Classificação
  - Atribuição de um objeto a uma classe.
  - O sistema retorna um rótulo inteiro.
    - exemplo: classificar um produto como "bom" ou "ruim" em um teste de controle de qualidade.
- Regressão
  - Generalização de uma tarefa de classificação
  - O sistema retorna um número real.
    - exemplo: predizer o valor das ações de uma empresa com base no desempenho passado e indicadores do mercado de bolsas.
- Agrupamento
  - Organização de objetos em grupos representativos.
  - O sistema retorna um grupo de objetos.
    - exemplo: organizar formas de vida em uma taxonomia de espécies.
- Descrição
  - Representação de um objeto em termos de uma série de primitivas.
  - O sistema produz um descrição estrutural ou linguística.
    - exemplo: rotular um sinal de eletrocardiograma em termos de seus componentes (P, QRS, T, U)

- Classificação automática de objetos.

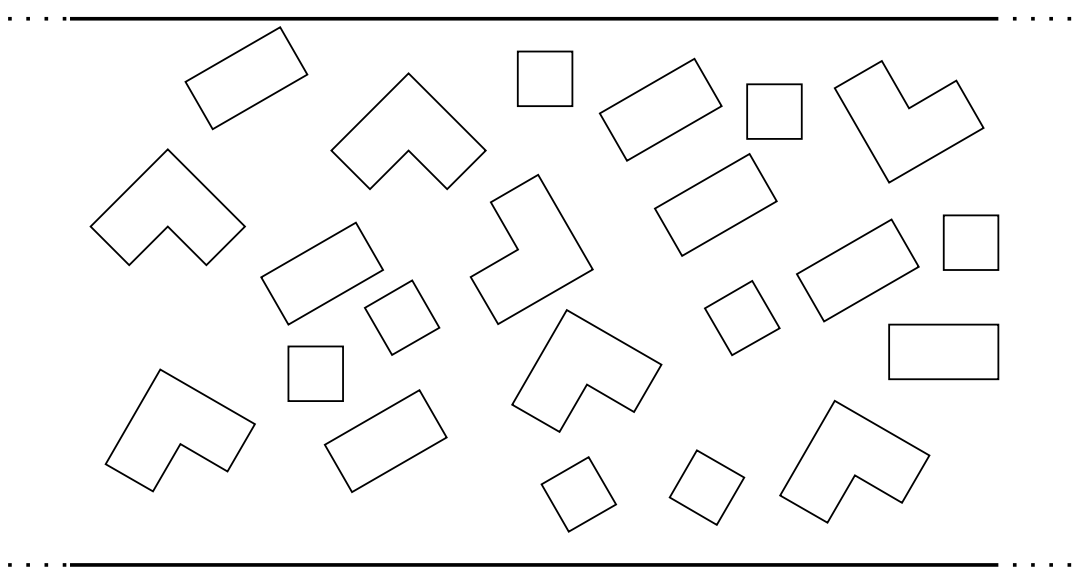


Figura: Visão superior de uma esteira de rolagem. Os objetos devem ser classificados de acordo com a forma apresentada.

- Neste exemplo, o objetivo do reconhecimento de padrões é encontrar um mapeamento, a partir da forma de cada objeto, para o conjunto $Y = \{C_1, C_2, C_3\}$.

- Os algoritmos que visam estabelecer o mapeamento entre as propriedades das amostras e o conjunto de rótulos são denominados *classificadores*.
- A determinação da rotulação para o conjunto de amostras pode ser realizada por um processo de classificação:
  - supervisionado
  - não supervisionado

- Quando o processo de classificação considera classes previamente definidas, este é denotado como *classificação supervisionada*.
- Uma etapa denominada *treinamento* deve ser executada anteriormente à aplicação do algoritmo de classificação para obtenção dos parâmetros que caracterizam cada classe.
- O conjunto formado por amostras previamente identificadas (rotuladas) chama-se *conjunto de treinamento*, no qual cada elemento apresenta dois componentes, o primeiro composto de medidas responsáveis pela descrição de suas propriedades e o segundo representando a classe a qual ele pertence.

- Ilustração da classificação que utiliza classes definidas previamente.

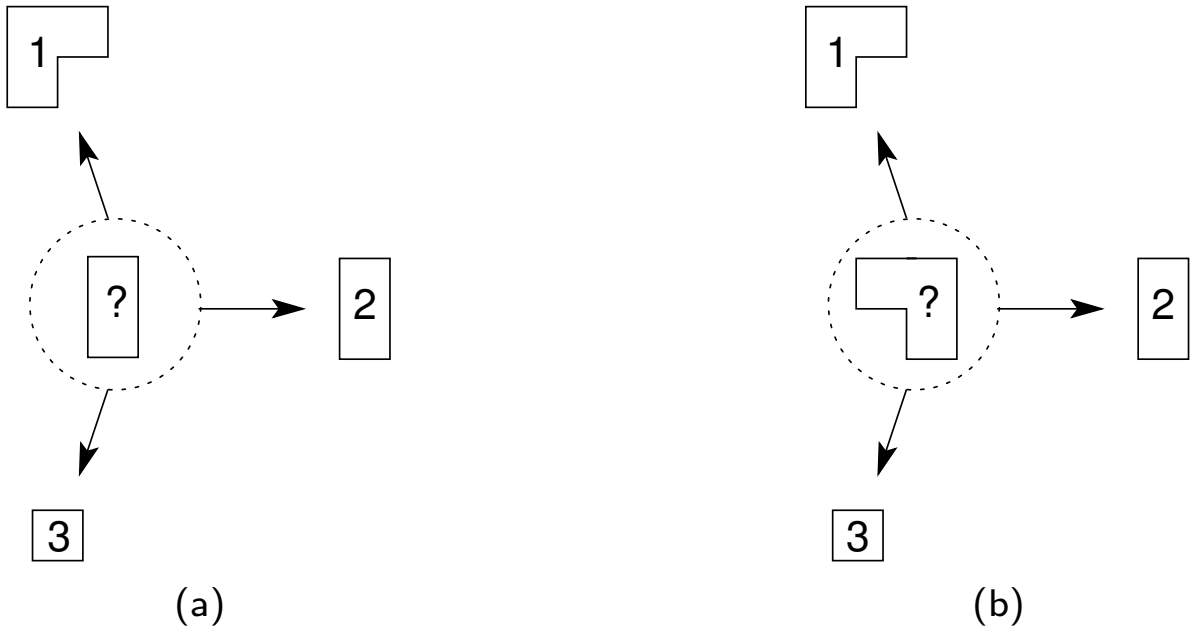


Figura: Classificação supervisionada. Atribuição de amostras desconhecidas (indicadas pelas linhas tracejadas) a classes previamente definidas.

- As classes são definidas por meio de descritores capazes de resumir as propriedades das amostras que as compõem.
- No exemplo da esteira rolante, o algoritmo de classificação deve atribuir cada objeto à classe que apresentar características mais próximas.
- No caso da figura (a), a atribuição dá-se de maneira direta, em razão da forma apresentada pela amostra em questão ser idêntica aquelas representadas pela classe C_2.
- Entretanto, na classificação mostrada na figura (b), deve-se utilizar uma medida de similaridade que seja robusta à transformação de rotação para que a amostra seja classificada como pertencente à classe C_1.

- Quando não se dispõe de parâmetros ou informações coletadas previamente à aplicação do algoritmo de classificação, o processo é denotado como *não supervisionado*.
- Todas as informações de interesse devem ser obtidas a partir das próprias amostras a serem rotuladas.
- Assim como na classificação supervisionada, amostras que compartilham propriedades semelhantes devem receber o mesmo rótulo na classificação não supervisionada.
- No entanto, diferentemente da classificação supervisionada, as classes não apresentam um significado previamente conhecido, associado aos rótulos.

- No exemplo da classificação supervisionada, a classe C_2 é composta de objetos que apresentam formato retangular, entretanto, quando um conjunto de treinamento não se encontra disponível, pode-se afirmar apenas que os elementos que compõem a classe C_2 possuem propriedades semelhantes.

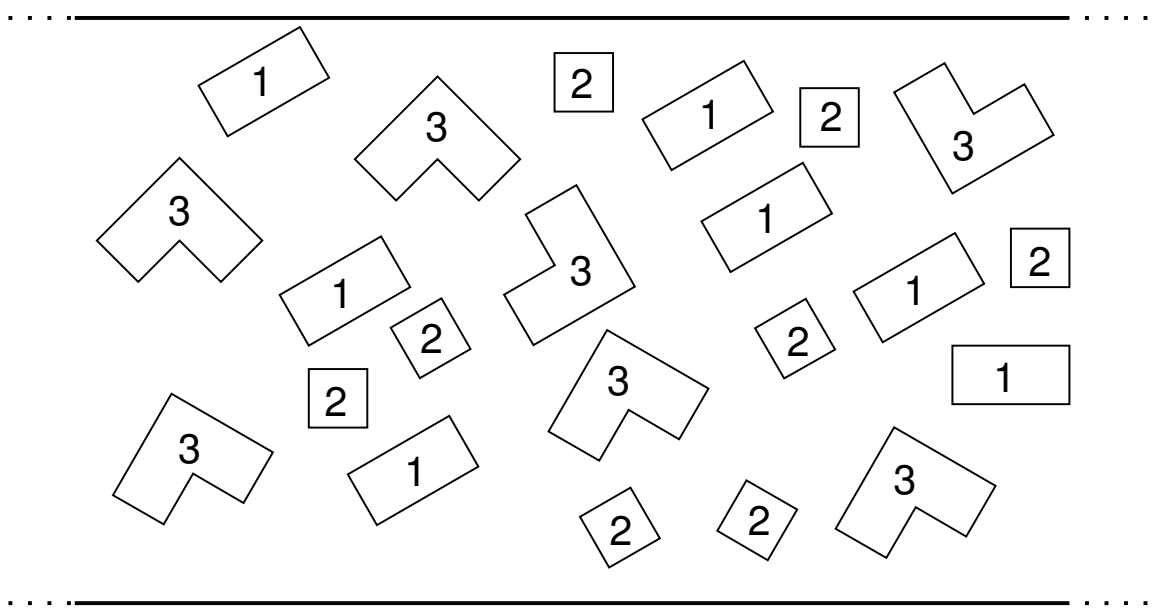
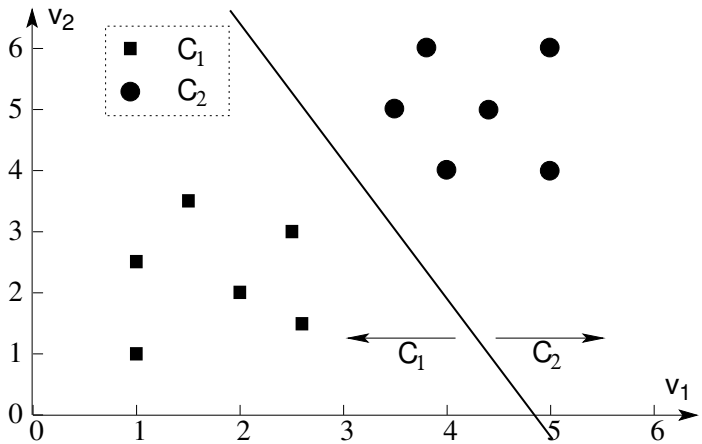
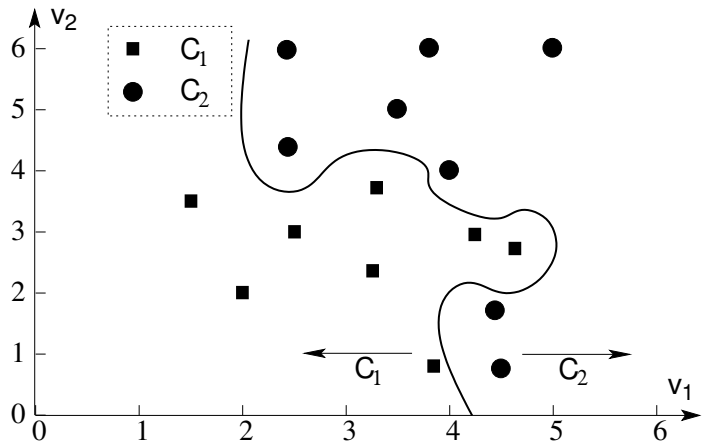


Figura: Classificação não supervisionada. Agrupamento de amostras que possuem propriedades semelhantes.

- Alguns modelos de classificação utilizam medidas de similaridade para determinar a qual classe uma amostra deve ser atribuída pela geração de uma superfície de decisão.
- As superfícies de decisão podem ser obtidas a partir das amostras contidas no conjunto de treinamento.
- Utiliza-se o vetor de características de cada amostra para determinar a posição em relação a essa superfície e, consequentemente, sua classificação.



(a) separabilidade linear



(b) separabilidade não-linear

- As abordagens para a classificação de padrões dividem-se normalmente em duas categorias:
  - abordagem sintática ou estrutural: baseia-se na relação existente entre primitivas que compõem as amostras.
  - abordagem estatística: considera que as amostras são obtidas de maneira independente a partir de uma distribuição de probabilidades fixa, porém, desconhecida.

Abordagem Sintática

- Na abordagem sintática, as amostras são representadas por meio de sentenças decompostas em primitivas, ou seja, símbolos contidos em uma gramática.
- Cada classe é representada por uma gramática, a qual gera linguagens que descrevem suas amostras.
- Dessa maneira, uma amostra é classificada conforme a pertinência de sua sentença às linguagens geradas pelas gramáticas que representam cada classe.

Abordagem Sintática

Exemplo:

Considerando a abordagem sintática para a classificação de padrões, este exemplo descreve como amostras que apresentam a forma retangular poderiam ser classificadas. Sejam as primitivas mostradas na figura (a). Um retângulo pode ser descrito como $\{a^m b^n a^m b^n | m, n \geq 1\}$, conforme ilustrado na figura (b).

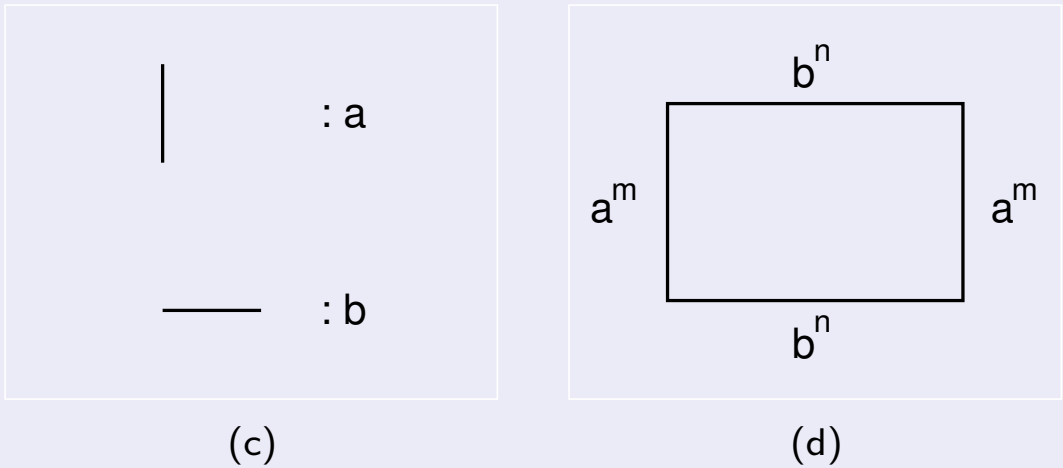


Figura: Abordagem sintática para a classificação de padrões. (a) primitivas que compõem um retângulo; (b) definição de um retângulo.

Abordagem Sintática

A gramática que gera a linguagem que representa um retângulo é dada por $G = (V_n, V_t, P, S)$ definidos como

$$
\begin{aligned}
V_n &= \{S, H, V, A, B\} \\
V_t &= \{a, b\} \\
P &= \{S \to aH, S \to VH, V \to aVA, H \to BHb, AB \to BA, \\
&\qquad aV \to aa, Hb \to ab, aB \to ab, bB \to bb, Aa \to aa\}
\end{aligned}
$$

em que V_n e V_t denotam os conjuntos de símbolos não-terminais e terminais, respectivamente, P o conjunto de regras de produção e S o símbolo inicial.
Com a definição da gramática, se uma forma geométrica for decomposta nas primitivas a e b e a sentença pertencer à linguagem gerada pela gramática G, essa forma geométrica será classificada como um retângulo.

Abordagem Estatística

- Decisões tomadas com base em distribuições de probabilidade em conjunto com os dados observados.
- As distribuições de probabilidade podem ser previamente conhecidas ou estimadas.
- Teoria de decisão Bayesiana.

Classificador Bayesiano

- Seja um conjunto de amostras rotuladas como pertencentes a uma dentre m classes distintas, $C_1, C_2, \ldots, C_m$.
- Uma possível atribuição dos rótulos é dada de modo que cada amostra pertença à classe que maximize a probabilidade $P(C_i|\boldsymbol{x})$, para $i = 1, 2, \ldots, m$, ou seja, dado o vetor de características $\boldsymbol{x}$, atribui-se a amostra à classe C_i que apresenta a maior probabilidade condicional.
- Conforme a descrição sobre a atribuição dos rótulos, a classificação de uma amostra específica segue a regra de decisão mostrada na equação 1.

$$P(C_i|\boldsymbol{x}) > P(C_j|\boldsymbol{x}) \qquad j = 1, \ldots, m;\ i \neq j \tag{1}$$

- Nessa regra, $\boldsymbol{x}$ é atribuída à classe C_i caso $P(C_i|\boldsymbol{x})$, denominada probabilidade *a posteriori*, seja maior que qualquer outra $P(C_j|\boldsymbol{x})$.

Classificador Bayesiano

- Ambos os lados da regra de decisão da equação 1 podem ser avaliados por meio do teorema de Bayes, mostrado na equação 2, em que $P(\boldsymbol{x}|C_i)$ denota a probabilidade de ocorrer $\boldsymbol{x}$ dado que a amostra pertence à classe C_i e $P(C_i)$ é denominada probabilidade *a priori*, ou seja, a probabilidade incondicional de ocorrência de uma determinada classe.

$$P(C_i|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|C_i)P(C_i)}{P(\boldsymbol{x})} = \frac{P(\boldsymbol{x}|C_i)P(C_i)}{\sum_{j=1}^{n} P(\boldsymbol{x}|C_j)P(C_j)} \tag{2}$$

- A substituição da equação 2 em 1 resulta na equação mostrada em 3, conhecida como *regra de Bayes para taxa mínima de erro*.
- Essa regra particiona o espaço de características nas regiões R_1, R_2, ..., R_m, tal que, se $\boldsymbol{x} \in R_i$, $\boldsymbol{x}$ deve ser atribuído à classe C_i.

Classificador Bayesiano

- Dado que a probabilidade $P(\boldsymbol{x})$ do denominador da equação 2 é termo comum para todas as classes, esse é removido da equação 3.

$$P(\boldsymbol{x}|C_i)P(C_i) > P(\boldsymbol{x}|C_j)P(C_j) \qquad j = 1, \ldots, m;\ i \neq j \tag{3}$$

- Considerando a existência de apenas duas classes, pode-se transformar a equação 3 na razão mostrada na equação 4, denominada *razão de verossimilhança*.

$$\Lambda(\boldsymbol{x}) = \frac{P(\boldsymbol{x}|C_1)}{P(\boldsymbol{x}|C_2)} > \frac{P(C_2)}{P(C_1)} \qquad \text{implica que } \boldsymbol{x} \in C_1 \tag{4}$$

- Utiliza-se essa equação como função discriminante entre as duas classes, ou seja, a atribuição de $\boldsymbol{x}$ passa a depender de sua localização espacial no espaço de características.

Classificador Bayesiano

Exemplo:

Ilustração da distribuição de probabilidade para as classes C_1 e C_2, em que $P(x|C_1) = N(0, 0.5)$ e $P(x|C_2) = 0.4N(1, 0.5) + 0.6N(-1, 1)$.

Na figura (a), as duas classes são equiprováveis, enquanto em (b), $P(C_1) = 0.70$ e $P(C_2) = 0.30$. Nesta última classe, segundo a regra de decisão mostrada na equação 3, um vetor $\boldsymbol{x} = [x_1]$ deve pertencer à classe C_1 quando $-1 < x_1 < 1$, aproximadamente.

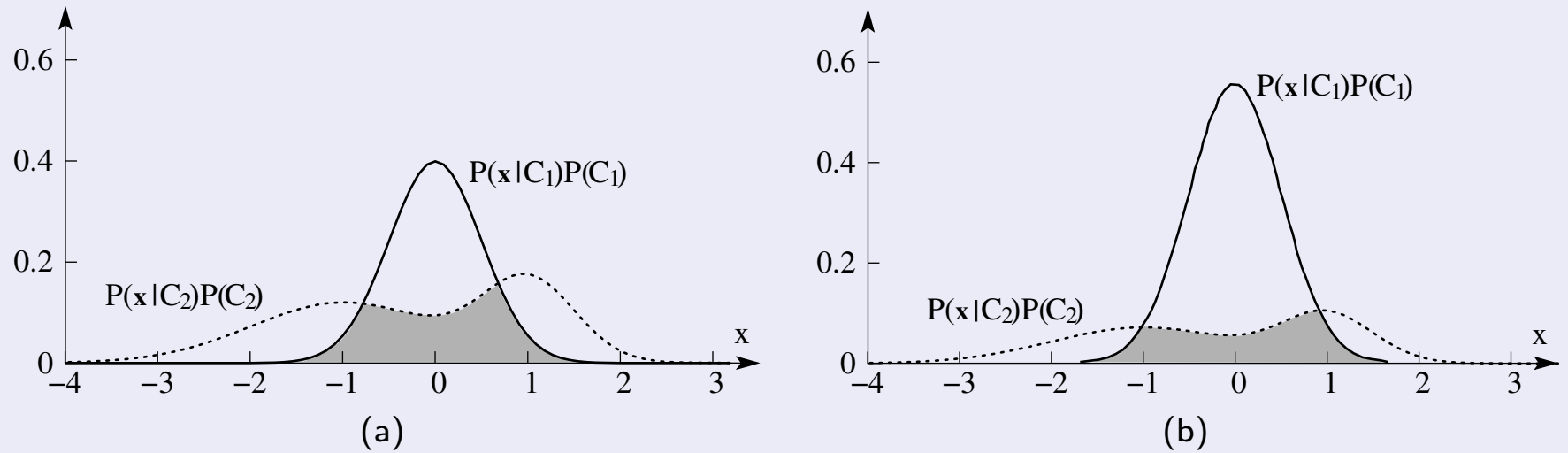


Figura: Aproximação da probabilidade *a posteriori* para as classes C_1 e C_2. (a) $P(C_1) = P(C_2) = 0.50$; (b) $P(C_1) = 0.70$ e $P(C_2) = 0.30$.

Classificador Bayesiano

Exemplo:

Para facilitar a visualização da regra de decisão expressa pela equação 3, apresenta-se a razão de verossimilhança definida na equação 4. O vetor $\boldsymbol{x}$ deve ser atribuído à classe C_1 quando $\Lambda(\boldsymbol{x}) > P(C_2)/P(C_1)$.

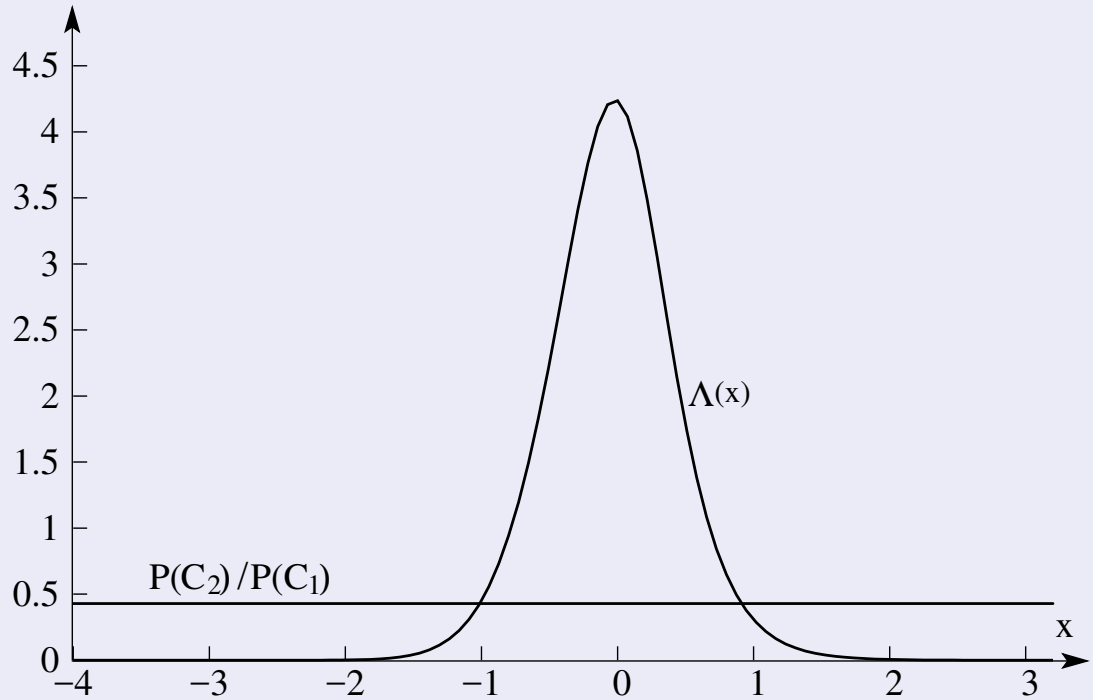


Figura: Razão de verossimilhança entre as distribuições de probabilidade mostradas na figura anterior (b).

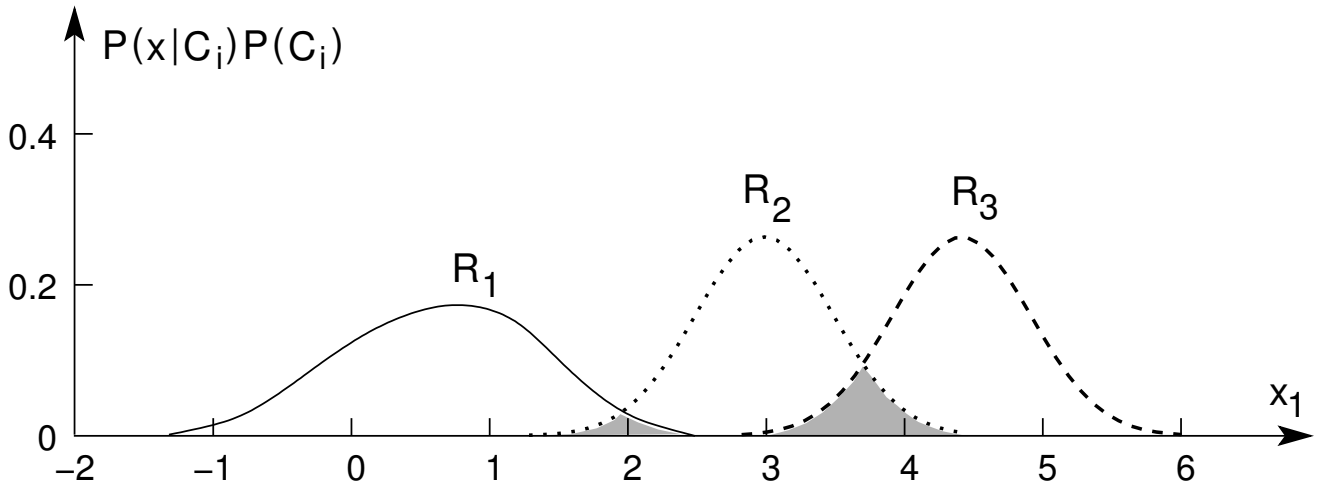
Classificador Bayesiano

- Conforme descrito, a regra de decisão particiona o espaço de características nas regiões R_1, R_2, ..., R_m, denominadas *regiões de decisão*, dentro das quais uma determinada classe apresenta maior probabilidade de ocorrência, enquanto em suas fronteiras, denominadas *fronteiras de decisão*, a ocorrência de classes distintas é equiprovável.
- Pode-se definir as fronteiras de decisão entre duas classes igualando-se as probabilidades *a posteriori*, conforme mostra a equação 5.

$$P(C_1)P(\boldsymbol{x}|C_1) = P(C_2)P(\boldsymbol{x}|C_2) \tag{5}$$

Classificador Bayesiano

- Ilustração de um espaço de características unidimensional dividido em três regiões de decisão, R_1, R_2 e R_3, representando as classes C_1, C_2 e C_3, respectivamente.



- As fronteiras de decisão ocorrem em pontos nos quais duas distribuições *a posteriori* são equiprováveis.
- Dessa maneira, para efetuar a classificação de uma amostra $\boldsymbol{x}$, deve-se determinar a qual região de decisão esta pertence.
- Por exemplo, uma amostra representada pelo vetor de características $\boldsymbol{x} = [3]$ deve ser atribuída à classe C_2, desde que $\boldsymbol{x}$ localize-se dentro da região de decisão R_2.

- O risco Bayesiano pode ser formalizado como uma função λ_{ij} que representa o custo de se escolher a classe C_i quando C_j é a classe correta.

Classificador Bayesiano

Variações da razão de verossimilhança

- Critério de Bayes
  - minimiza o risco Bayesiano

$$\Lambda_{\text{Bayes}}(\boldsymbol{x}) = \frac{P(\boldsymbol{x}|C_1)}{P(\boldsymbol{x}|C_2)} \underset{C_2}{\overset{C_1}{\gtrless}} \frac{(\lambda_{12} - \lambda_{22})P(C_2)}{(\lambda_{21} - \lambda_{11})P(C_1)}$$

- Critério de Máximo a Posteriori (MAP)
  - caso especial de $\Lambda_{\text{Bayes}}(\boldsymbol{x})$ que usa função de custo 0/1, $\lambda_{ij} = \begin{cases} 0, & i = j \\ 1, & i \neq j \end{cases}$
  - procura maximizar $P(C_i|\boldsymbol{x})$

$$\Lambda_{\text{MAP}}(\boldsymbol{x}) = \frac{P(\boldsymbol{x}|C_1)}{P(\boldsymbol{x}|C_2)} \underset{C_2}{\overset{C_1}{\gtrless}} \frac{P(C_2)}{P(C_1)} \Rightarrow \frac{P(C_1|\boldsymbol{x})}{P(C_2|\boldsymbol{x})} \underset{C_2}{\overset{C_1}{\gtrless}} 1$$

- Critério da Máxima Verossimilhança (MV)
  - para probabilidade a priori $P(C_1) = 1/2$ e função de custo 0/1, o critério MV procura maximizar $P(\boldsymbol{x}|C_i)$

$$\Lambda_{\text{MV}}(\boldsymbol{x}) = \frac{P(\boldsymbol{x}|C_1)}{P(\boldsymbol{x}|C_2)} \underset{C_2}{\overset{C_1}{\gtrless}} 1$$

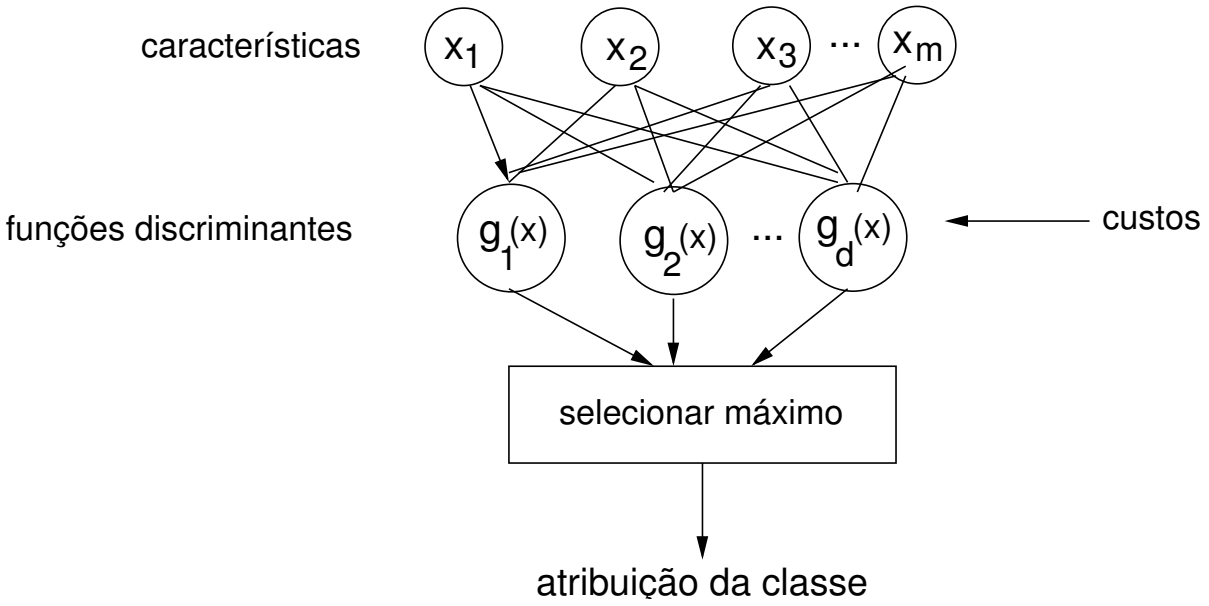
Classificador Bayesiano

Funções discriminantes

- Critérios anteriores possuem a mesma estrutura:
  - Em cada ponto $\boldsymbol{x}$ do espaço de características, escolher a classe C_i que maximiza (ou minimiza) alguma medida $g_i(x)$.
  - esta estrutura pode ser formalizada como um conjunto de funções discriminantes $g_i(\boldsymbol{x})$, $i = 1, 2, \ldots d$ e a regra de decisão:

    atribuir $\boldsymbol{x}$ *à classe* C_i *se* $g_i(\boldsymbol{x}) > g_j(\boldsymbol{x}) \quad \forall i \neq j$

  - assim, pode-se visualizar a regra de decisão como uma rede que computa d funções discriminantes e seleciona a classe com maior valor discriminante.



Modelo de Análise Discriminante Gaussiana

- Até o momento, definiu-se a regra de Bayes para taxa de erro mínimo, no entanto, nada foi especificado sobre a forma da distribuição de probabilidade $P(\boldsymbol{x}|C_i)$.
- Agora, considera-se que essa probabilidade segue a distribuição Gaussiana, esquema conhecido como modelo de análise discriminante Gaussiana e aplicado para classificação.
- Utilizando a distribuição Gaussiana multivariada, a distribuição de probabilidade $P(\boldsymbol{x}|C_i)$ é

$$P(\boldsymbol{x}|C_i) = \frac{1}{(2\pi)^{m/2}|\boldsymbol{\Sigma}_i|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)^T\right)$$

  em que m denota o número de elementos contidos no vetor de características e $|\boldsymbol{\Sigma}_i|$ o determinante da matriz $\boldsymbol{\Sigma_i}$.
- Representa-se cada classe C_i por meio da matriz de covariância $\boldsymbol{\Sigma}_i$ e pelo vetor de médias $\boldsymbol{\mu}_i$, ambos obtidos a partir de um conjunto de treinamento previamente definido.

Modelo de Análise Discriminante Gaussiana

- Como se utiliza a regra de Bayes para erro mínimo, uma amostra descrita pelo vetor $\boldsymbol{x}$ é atribuída à classe C_i que maximiza a probabilidade *a posteriori*.
- Dada a monotonicidade da função logarítmica, torna-se conveniente reescrever a aproximação da probabilidade $P(C_i|\boldsymbol{x})$, desconsiderando $P(\boldsymbol{x})$, termo em comum não utilizado pela regra de decisão da equação 3.

$$\begin{aligned}\ln(P(C_i|\boldsymbol{x})) &\approx \ln(P(\boldsymbol{x}|C_i)) + \ln(P(C_i)) \\ &\approx -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)^T - \frac{1}{2}\ln(|\boldsymbol{\Sigma}_i|) - \frac{l}{2}\ln(2\pi) + \ln(P(C_i)) \\ &\approx -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)^T - \frac{1}{2}\ln(|\boldsymbol{\Sigma}_i|) + \ln(P(C_i)) \qquad (6)\end{aligned}$$

- Aplicando-se a regra de Bayes, uma amostra descrita pelo vetor de características $\boldsymbol{x}$ será atribuída à classe C_i que maximiza a equação a seguir, com $i = 1, 2, \ldots, m$.

$$Y = \arg\max_{C_i} \{\ln(P(C_i|\boldsymbol{x}))\} \qquad (7)$$

Modelo de Análise Discriminante Gaussiana

Algoritmo para classificação baseada no modelo de análise discriminante Gaussiano

Seja o conjunto de treinamento $T = \{< \boldsymbol{x}_1, y_1 >, < \boldsymbol{x}_2, y_2 >, \ldots, < \boldsymbol{x}_n, y_n >\}$, em que a amostra representada pelo vetor de características $\boldsymbol{x}_i$ pertença à classe $y_i = \{C_1, C_2, \ldots, C_m\}$. Após a etapa de treinamento, cada amostra $\boldsymbol{x}$ apresentada ao classificador é atribuída à classe C_i que maximiza $P(C_i|\boldsymbol{x})$.

1: // etapa de treinamento
2: para cada classe C_i faça
3: selecionar as amostras $< \boldsymbol{x}_i, C_i >$
4: calcular o vetor de médias $\boldsymbol{\mu}_i$
5: calcular a matriz de covariância $\boldsymbol{\Sigma}_i$
6:
7: // etapa de classificação
8: para cada amostra $\boldsymbol{x}$ a ser classificada faça
9: calcular a aproximação para $\ln(P(C_i|\boldsymbol{x}))$ conforme a equação 6
10:
11: atribuir $\boldsymbol{x}$ à classe C_i resultante da equação 7

Modelo de Análise Discriminante Gaussiana

- Ilustração da distribuição de probabilidade condicional referente a duas classes representadas por vetores de características unidimensionais que seguem a distribuição Gaussiana. A área hachurada contém as regiões em que ocorrem erros na classificação segundo a regra de Bayes para taxa mínima de erro, em outras palavras, quando uma amostra pertencente à classe C_i é atribuída à classe C_j, com $i \neq j$.

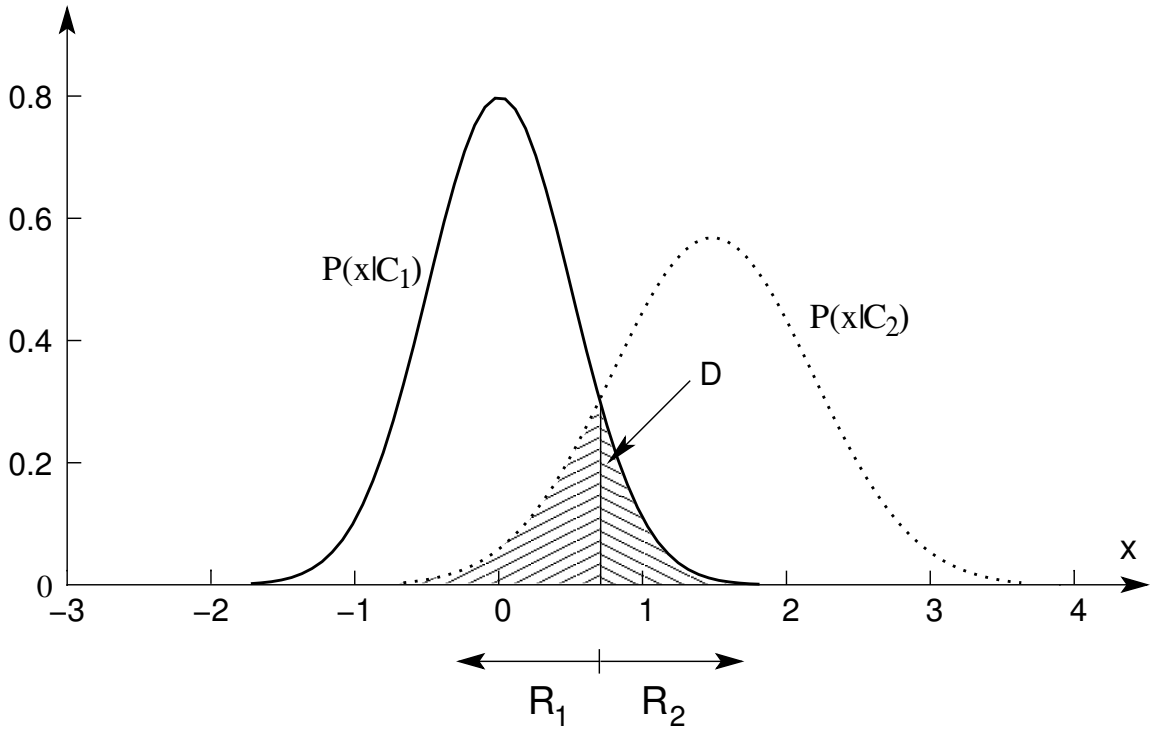


Figura: Probabilidade de ocorrência de erros na classificação, em que $P(\boldsymbol{x}|C_1) = N(0, 0.5)$ e $P(\boldsymbol{x}|C_2) = N(1.5, 0.7)$.

- O Naïve Bayes é um classificador estatístico que assume indepedência condicional entre as características, dada a classe alvo.
- Assim, o valor de uma dada característica não está relacionada à presença ou ausência de qualquer outra característica, dada a classe.
- Pelo critério de máximo a posteriori, a classificação é realizada de forma que a amostra $\boldsymbol{x}$ é atribuída à classe C_i que maximiza $P(\boldsymbol{x}|C_i)P(C_i)$, ou seja

$$Y = \arg\max_{C_i} P(C_i)P(\boldsymbol{x}|C_i) \qquad i = 1, \ldots, m \tag{8}$$

- O primeiro termo à direita da equação é a probabilidade *a priori* para a classe C_i, reescrito como $P(Y = C_i)$, que pode ser computado conforme a equação 9, em que $\#\{Y = C_i\}$ denota o número de amostras do conjunto de treinamento pertencentes à classe C_i.

$$\hat{P}(Y = C_i) = \frac{\#\{Y = C_i\}}{\sum_{j=1}^{m} \#\{Y = C_j\}} \tag{9}$$

- Considerando que o vetor de características apresenta n elementos, o segundo termo da equação 8 é decomposto conforme a equação 10.

$$P(\boldsymbol{x}|C_i) = P(X_1|X_2, \ldots, X_n, C_i)P(X_2|X_3, \ldots, X_n, C_i) \ldots P(X_n|C_i) \tag{10}$$

- Com a utilização da equação 10, o número de parâmetros a serem estimados cresce exponencialmente conforme a dimensionalidade do espaço de características.

- Para possibilitar que a regra de classificação apresentada na equação 8 seja utilizada sem a necessidade de se estimar um número exponencial de parâmetros, assume-se a independência condicional entre os elementos X_i, denominada Naïve Bayes.
- Assim, a equação 10 pode ser reescrita como mostrado na equação 11, tal que o número de parâmetros a serem estimados é significativamente reduzido.

$$P(\boldsymbol{x}|C_j) = \prod_{i=1}^{n} P(X_i = x_i|Y = C_j) \tag{11}$$

- Após considerar a independência condicional, a estimação de cada um dos parâmetros $P(X_i = x_i|C_j)$ pode ser efetuada conforme a equação 12, em que $\#\{X_i = x_i \ \wedge \ Y = C_j\}$ denota o número de amostras no conjunto de treinamento pertencentes à classe C_j apresentando valor x_i para o i-ésimo elemento do vetor de características.

$$\hat{P}(X_i = x_i|Y = C_j) = \frac{\#\{X_i = x_i \ \wedge \ Y = C_j\}}{\#\{Y = C_j\}} \tag{12}$$

- O algoritmo a seguir descreve os passos necessários para efetuar a classificação utilizando Naïve Bayes.
- Nesse algoritmo, o logaritmo natural é utilizado para determinar a estimação da probabilidade *a posteriori* $\hat{P}(C_i|\boldsymbol{x}) \approx \hat{P}(C_i)\hat{P}(\boldsymbol{x}|C_i)$, assim, o produtório mostrado na equação 11 é convertido em um somatório.
- Embora a hipótese de que as características sejam independentes dada a classe, o classificador Naïve Bayes apresenta bons resultados em vários problemas, com eficácia comparável às redes neurais e árvores de decisão.

Classificador Naïve Bayes

Algoritmo para Naïve Bayes

Seja o conjunto de treinamento $T = \{< \boldsymbol{x}_1, y_1 >, < \boldsymbol{x}_2, y_2 >, \ldots, < \boldsymbol{x}_n, y_n >\}$, composto de amostras classificadas entre as m classes de interesse do experimento. Cada vetor de características é composto de q elementos, representados individualmente pelas variáveis discretas X_i.

// etapa de treinamento (estimação dos parâmetros)
para cada classe C_i faça
    $\hat{P}(Y = C_i) = \dfrac{\#\{Y = C_i\}}{n}$
para cada classe C_j faça
    para cada variável X_i faça
        para cada valor x_i faça
            $\hat{P}(X_i = x_i | Y = C_j) = \dfrac{\#\{X_i = x_i \ \wedge \ Y = C_j\}}{\#\{Y = C_j\}}$
// etapa de classificação
para cada amostra $\boldsymbol{x}$ a ser classificada faça
    para cada classe C_j faça
        $\ln(\hat{P}(C_j|\boldsymbol{x})) = \ln(\hat{P}(C_j)) + \sum_{i=1}^{q} \ln(\hat{P}(X_i = x_i | Y = C_j))$
atribuir $\boldsymbol{x}$ à classe C_j que maximiza $\ln(\hat{P}(C_j|\boldsymbol{x}))$

- Uma rede neural artificial é um modelo inspirado na aprendizagem de sistemas biológicos formados por redes complexas de neurônios interconectados.
- O modelo é um grafo orientado em que os nós representam neurônios artificiais e as arestas denotam as conexões entre as entradas e as saídas dos neurônios.
- Redes neurais podem ser vistas como máquinas massivamente paralelas com muitos processadores simples e diversas interconexões.
- A popularidade das redes neurais cresceu devido ao fato de que, aparentemente, elas possuem uma baixa dependência a um domínio específico, de forma que a mesma rede pode ser utilizada em problemas distintos.
- As redes neurais são capazes de aprender relações entre entradas e saídas complexas, bem como generalizar a informação aprendida.

- O modelo mais simples de rede neural é composto de apenas uma unidade denominada *perceptron*.
- Essa rede mapeia múltiplas entradas, compostas de valores reais, para uma saída representada por um valor binário com os estados normalmente denotados como ativado ou não-ativado.
- Obtém-se o valor da saída pela aplicação da função de ativação mostrada na equação 13, em que $\boldsymbol{x} = [x_1\ x_2\ \ldots\ x_n]$ é o vetor de características e os elementos de $\boldsymbol{w} = [w_0\ w_1\ \ldots\ w_n]$ representam os pesos utilizados para o treinamento da rede neural.

$$s(\boldsymbol{x}) = \begin{cases} 1 & \text{se } w_0 + w_1 x_1 + \ldots + w_n x_n > 0 \\ -1 & \text{caso contrário} \end{cases} \tag{13}$$

- Ilustração da unidade perceptron: o coeficiente $-w_0$ representa o valor do limiar que deve ser ultrapassado pela soma ponderada das entradas x_i, de modo que a saída $s(\boldsymbol{x})$ se apresente no estado ativado, ou seja, $s(\boldsymbol{x}) = 1$.

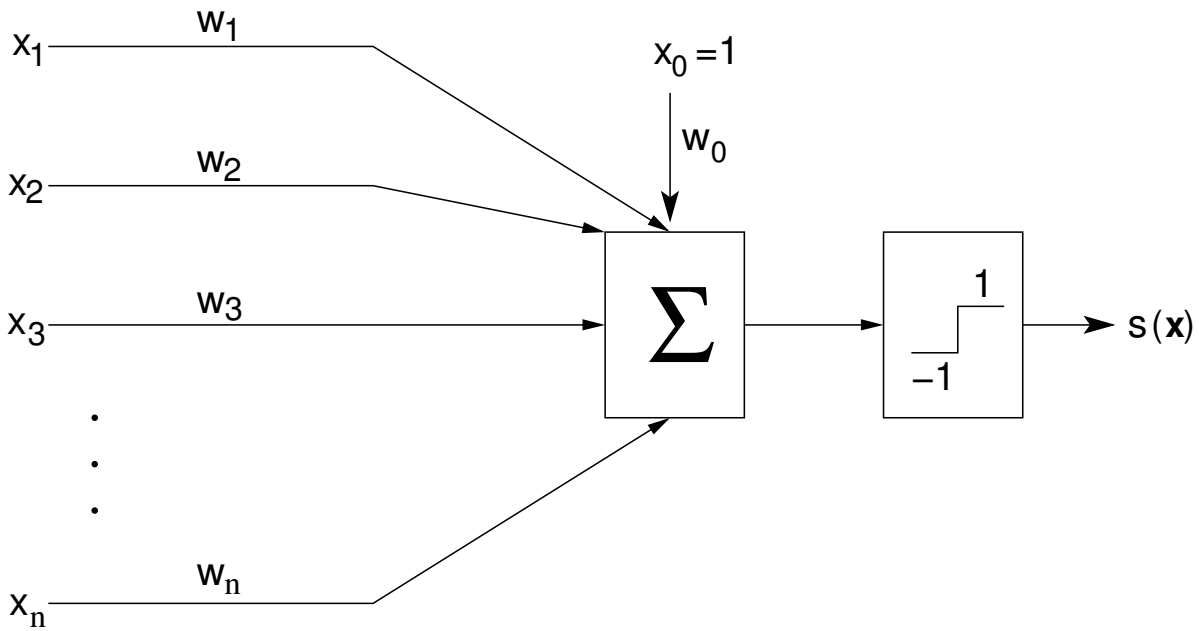


Figura: Modelo de rede neural composto da unidade perceptron.

- O treinamento de uma rede neural dá-se por meio da atualização iterativa dos pesos, ou seja, as variáveis $w_0, w_1, \ldots, w_n$, com objetivo de encontrar uma função que permita classificar corretamente todas as amostras contidas no conjunto de treinamento.
- Duas regras são normalmente aplicadas para o treinamento da rede neural com camada única:
  - *regra perceptron*: encontra um hiperplano em um espaço n-dimensional de amostras que sejam linearmente separáveis.
  - *regra delta*: permite boa adaptação aos dados que compõem o conjunto de treinamento, mesmo que não sejam linearmente separáveis.

- A regra perceptron atualiza os pesos w_i conforme a equação 14, em que y denota a classe a qual pertence a amostra sendo apresentada à rede neural e $s(\boldsymbol{x})$ o valor obtido em sua saída.

$$w_i \leftarrow w_i + \alpha(y - s(\boldsymbol{x}))x_i \tag{14}$$

- A constante α é chamada de taxa de aprendizagem, que determina quão rápido os pesos w_i devem ser alterados (normalmente, utiliza-se $\alpha = 0.1$).
- Com essa regra, tem-se que o peso w_i sofrerá alterações quando a amostra em questão for classificada incorretamente, pois o valor de y e $s(\boldsymbol{x})$ serão distintos; por outro lado, w_i não sofrerá alterações quando ocorrer a classificação correta, pois a segunda parte da equação 14 resulta no valor zero.

- O algoritmo a seguir descreve os passos utilizados para efetuar o treinamento e a classificação com rede neural pela aplicação da regra perceptron.
- Após a convergência dos valores dos pesos w_i, a rotulação de amostras que apresentam classificação desconhecida pode ser efetuada.
- Uma amostra descrita pelo vetor de características $\boldsymbol{x}$, composto de n elementos, é apresentada na camada de entrada da rede e sua classificação é dada conforme o valor da saída $s(\boldsymbol{x})$.
- Como $s(\boldsymbol{x})$ assume os valores -1 ou 1, a classificação com rede neural constituída de apenas uma unidade perceptron é binária.

Dado o conjunto de treinamento $T = \{< \boldsymbol{x}_1, y_1 >, < \boldsymbol{x}_2, y_2 >, \ldots, < \boldsymbol{x}_n, y_n >\}$, em que $y_i = \{-1, 1\}$. O vetor de pesos $\boldsymbol{w} = [w_0, w_1, \ldots, w_n]$ é inicializado aleatoriamente com valores próximos a zero. $x_{i,j}$ representa a j-ésima medida do vetor de caracterísstica da i-ésima amostra.

1: // etapa de treinamento
2: inicializar os pesos $\boldsymbol{w}$
3: enquanto houver alterações nos pesos $\boldsymbol{w}$ faça
4: para cada amostra $< \boldsymbol{x}_i, y_i > \in T$ faça
5: apresentar os valores de $\boldsymbol{x}_i$ na entrada da rede neural
6: $w_0 = w_0 + \alpha(y_i - s(\boldsymbol{x}))$
7: para j de 1 até n faça
8: $w_j = w_j + \alpha(y_i - s(\boldsymbol{x}))x_{i,j}$
9: // etapa de classificação
10: para cada amostra $\boldsymbol{x}$ a ser classificada faça
11: apresentar os valores de $\boldsymbol{x}$ na entrada da rede neural
12: atribuir $\boldsymbol{x}$ à classe obtida pela saída $s(\boldsymbol{x})$

- A regra delta é utilizada quando as amostras não são linearmente separáveis.
- A regra aplica o método de otimização do gradiente descendente[1] para encontrar o hiperplano que melhor se adapta aos dados.
- Efetuando a minimização da função de erro mostrada na equação 15, utiliza-se a regra de treinamento conhecida como regra delta

$$E(w_0, w_1, \ldots, w_n) \equiv \frac{1}{2} \sum_{<\boldsymbol{x}, y> \in T} (y - s(\boldsymbol{x}))^2 \tag{15}$$

em que T denota o conjunto de treinamento e $s(\boldsymbol{x})$ representa a saída da rede quando se apresenta o vetor de características $\boldsymbol{x}$ na sua entrada.

[1] O gradiente descendente é um algoritmo que aproxima o mínimo local de uma função tomando-se passos proporcionais ao negativo do gradiente da função como o ponto corrente.

- A atualização iterativa dos pesos $\boldsymbol{w}$ pela regra delta é efetuada por meio da equação 16.

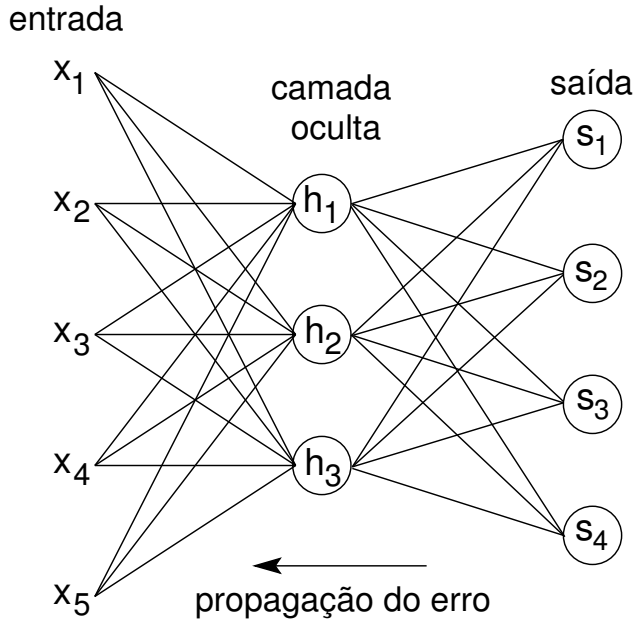
$$w_i \leftarrow w_i + \alpha \sum_{j=1}^{n} (y_j - s_j(\boldsymbol{x}))x_{j,i} \tag{16}$$

- Ao contrário da regra perceptron, atualiza-se os pesos w_i apenas após considerar o resultado da classificação para cada uma das amostras que compõem o conjunto de treinamento.
- Quando se utiliza a regra delta, obtém-se a saída sem a aplicação da limiarização como na regra perceptron, resultando, portanto, em uma função linear, dependente da soma ponderada da multiplicação dos pesos $\boldsymbol{w}$ e das entradas $\boldsymbol{x}$, como mostrado na equação 17.

$$s(\boldsymbol{x}) = w_0 + \sum_{i=1}^{n} x_i w_i \tag{17}$$

- Diferentemente da regra perceptron, em que o valor de saída $s(\boldsymbol{x})$ é binário, a saída na regra delta é um número real no intervalo $[0, 1]$.

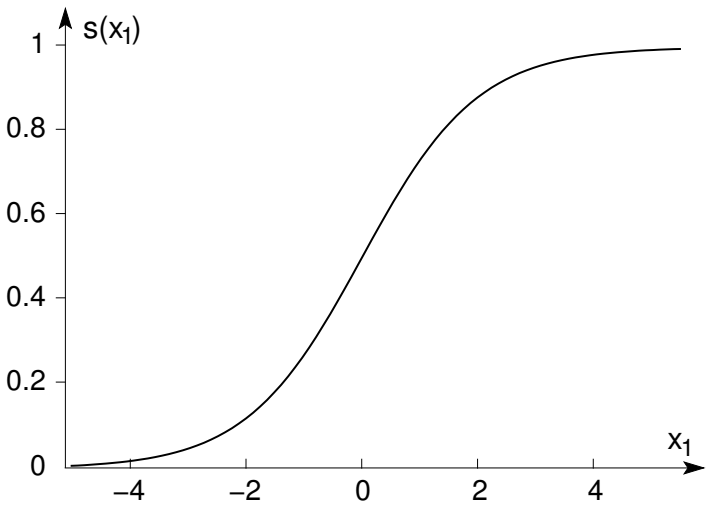
- Redes multicamadas são capazes de expressar uma variedade de superfícies de decisão não-lineares para a classificação de amostras.



- Nessa arquitetura, a saída de uma camada é utilizada como entrada para a camada subsequente e a função de ativação aplicada é não-linear, permitindo assim que funções complexas sejam obtidas para o mapeamento entre a entrada e a saída.

- Utiliza-se a função de ativação sigmóide, mostrada na equação 18, que apresenta a característica de ser não-linear e diferenciável.
- Diferentemente da saída da unidade perceptron, que retorna valores binários, essa função retorna valores reais entre 0 e 1.

$$s(\boldsymbol{x}) = \frac{1}{1 + \exp\left(-w_0 - \sum_{i=1}^{n} x_i w_i\right)} \tag{18}$$

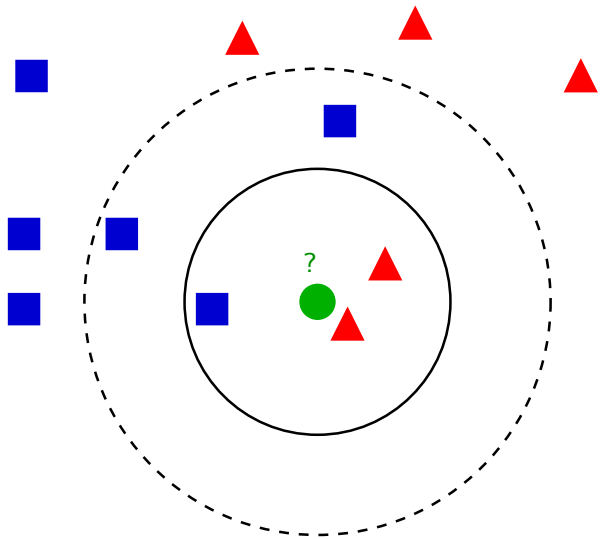


- O treinamento é realizado por meio da retropropagação do erro obtido pela classificação.
- Esse algoritmo aplica a regra do gradiente descendente para minimizar uma função de erro que depende do resultado obtido na saída e do valor correto para a classificação, contido no conjunto de treinamento.
- Uma vez que apenas o resultado da última camada é conhecido por meio do conjunto de treinamento, o erro das camadas intermediárias é minimizado com a retropropagação do erro obtido nas camadas posteriores.
- Durante a etapa de classificação, atribui-se cada amostra à classe relacionada ao nodo da camada de saída que apresenta o valor máximo.

Classificador K-Vizinhos Mais Próximos

Fundamentos

- A regra dos vizinhos mais próximos classifica uma amostra $\boldsymbol{x}$ atribuindo a ela o rótulo que mais frequentemente ocorre dentre os k exemplos de treinamento mais próximos.
- O classificador requer:
  - um valor inteiro k
  - um conjunto de amostras rotuladas (dados de treinamento)
  - uma métrica de distância



- O voto de cada vizinho pode ser ponderado de acordo com o quadrado do inverso de sua distância à amostra $\boldsymbol{x}$.

Classificador K-Vizinhos Mais Próximos

Vantagens e desvantagens

- Algumas vantagens
  - implementação simples
  - usa informação local, o que pode resultar em comportamento adaptativo
  - pode ser naturalmente paralelizado
  - permite a criação de superfície de decisão complexa
- Algumas desvantagens
  - sensível a ruído nas características
  - sensível à alta dimensionalidade dos dados
  - função de distância deve ser calculada entre a amostra e todos os exemplos no conjunto de treinamento, cujo processo torna-se custoso para conjuntos de treinamento compostos de um elevado número de amostras.
- Complexidade para algoritmo básico (todas as n amostras de dimensão d em memória)
  - $O(d)$: complexidade para computar distância para uma amostra
  - $O(nd)$: complexidade para encontrar um vizinho mais próximo
  - $O(knd)$: complexidade para encontrar k vizinhos mais próximos
  - complexidade pode ser reduzida com diagrama de Voronoi ou *k-d tree*

Classificador K-Vizinhos Mais Próximos

Algoritmo para k-vizinhos mais próximos

Dado o conjunto de treinamento T, o algoritmo visa encontrar um subconjunto $S \subset T$ composto de k amostras mais próximas de $\boldsymbol{x}$ no espaço de características. O rótulo atribuído a $\boldsymbol{x}$ é aquele que apresenta maior frequência dentre as amostras em S.

Seja o conjunto de treinamento $T = \{< \boldsymbol{x}_1, y_1 >, < \boldsymbol{x}_2, y_2 >, \ldots, < \boldsymbol{x}_n, y_n >\}$, em que $y_i = \{C_1, C_2, \ldots, C_m\}$. $\delta(x) = 1$ quando $x = 0$ e $\delta(x) = 0$, caso contrário.

1: // treinamento
2: $T =$ amostras no conjunto de treinamento
3: // classificação
4: determinar as k amostras em T mais próximas de $\boldsymbol{x}$
5: para i de 1 até m faça
6: $$c_i = \sum_{j=1}^{k} \delta(y_j - C_i)$$
7: $$Y = \arg\max_i \{c_i\}$$

- Os algoritmos anteriores necessitam de um conjunto de treinamento rotulado previamente à classificação.
- A aquisição prévia de informações sobre as classes presentes em um experimento pode-se caracterizar como uma tarefa árdua ou mesmo inviável para alguns domínios de aplicação.
- Torna-se de interesse a utilização de classificadores que não necessitem de um conjunto de treinamento, denotados de maneira geral como métodos de agrupamento de dados.
- A partir de amostras não-identificadas, os métodos de agrupamento de dados aplicam um processo de decisão de modo a agrupar amostras que apresentem características semelhantes e separar aquelas com características distintas.

Agrupamento de Dados

Definição

- Uma definição para agrupamento de dados considera o particionamento de n vetores de características em m grupos tal que cada vetor pertença apenas a um grupo, tendo como objetivo atribuir os vetores apresentando características semelhantes a um mesmo grupo.
- Seja $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$ um conjunto de amostras não-identificadas. Um m-agrupamento é uma partição de $\boldsymbol{X}$ em m conjuntos $\boldsymbol{C}_1, \ldots, \boldsymbol{C}_m$ que satisfaz as seguintes condições:
  a) $C_i \neq \emptyset,\ i = 1, \ldots, m$
  b) $\cup_{i=1}^{m} C_i = \boldsymbol{X}$
  c) $C_i \cap C_j = \emptyset, i \neq j, i, j = 1, \ldots, m$

Agrupamento de Dados

Exemplo

- Pontos na figura representam componentes de vetores de características com dois elementos.
- Vetores que apresentam características semelhantes, representadas pela localização espacial, são atribuídos a um mesmo grupo.

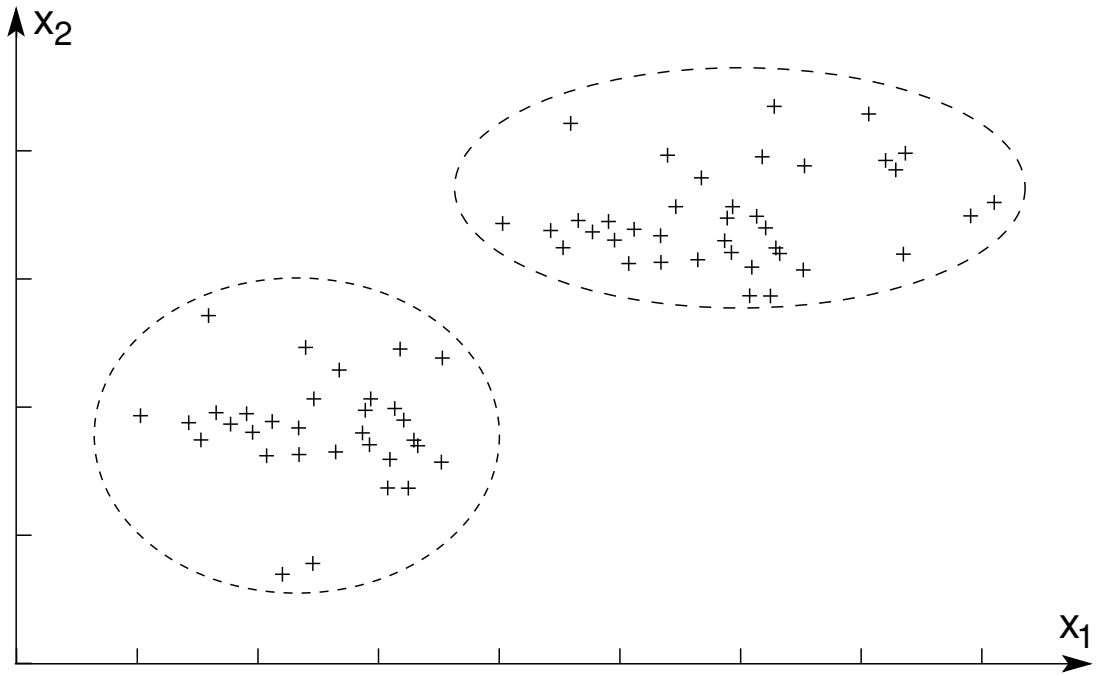


Figura: Agrupamento obtido a partir de pontos dispostos em um espaço de características bidimensional.

- Como métodos de agrupamento de dados utilizam a semelhança entre as amostras para atribuí-las a um mesmo grupo, torna-se necessária a escolha de medidas de proximidade que quantifiquem a distância entre amostras distintas.
- A medida de proximidade deve ser criteriosamente escolhida, considerando tanto a heterogeneidade quanto a diferença de escala presente nas medidas que compõem o vetor de características.

- métrica de Minkowski: $d(\boldsymbol{x}, \boldsymbol{y}) = \left(\sum_{i=1}^{N} |x_i - y_i|^k \right)^{1/k}$
- medida Euclidiana: $d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_{i=1}^{N} (x_i - y_i)^2}$
- medida de Manhattan: $d(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{N} |x_i - y_i|$
- medida de Chebyshev: $d(\boldsymbol{x}, \boldsymbol{y}) = \max_{1 \leq i \leq N} |x_i - y_i|$
- distância de Mahalanobis: $d(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x}_i - \boldsymbol{y}_i)\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{y}_i)^T$
  (considerando que os dados seguem a distribuição Gaussiana multivariada e que $\boldsymbol{\Sigma}$ denota a matriz de covariância obtida a partir das amostras)

- produto interno: $s(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}^T \boldsymbol{y}$
- coeficiente de correlação: $s(\boldsymbol{x}, \boldsymbol{y}) = \dfrac{\sum_{i=1}^{D}(x_i - \bar{x})(y_i - \bar{y})}{\left[\sum_{i=1}^{D}(x_i - \bar{x})^2 \sum_{i=1}^{D}(y_i - \bar{y})^2\right]^{1/2}}$

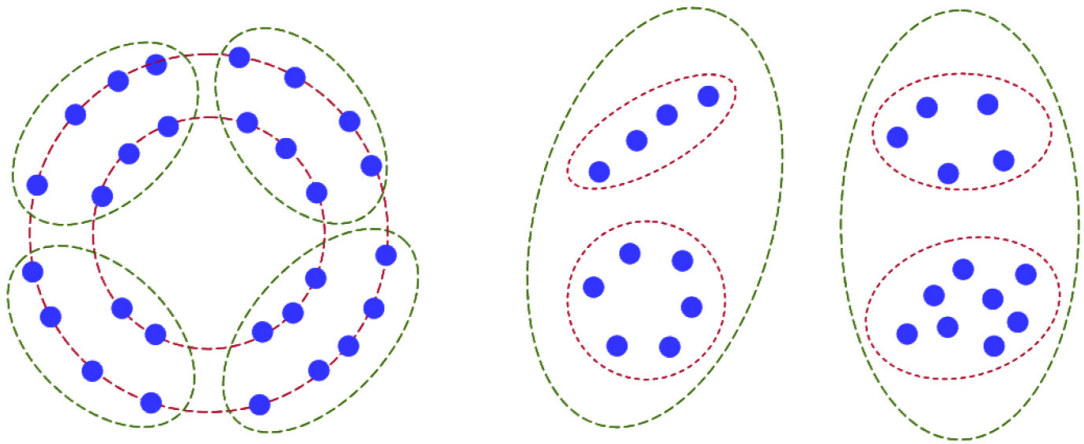


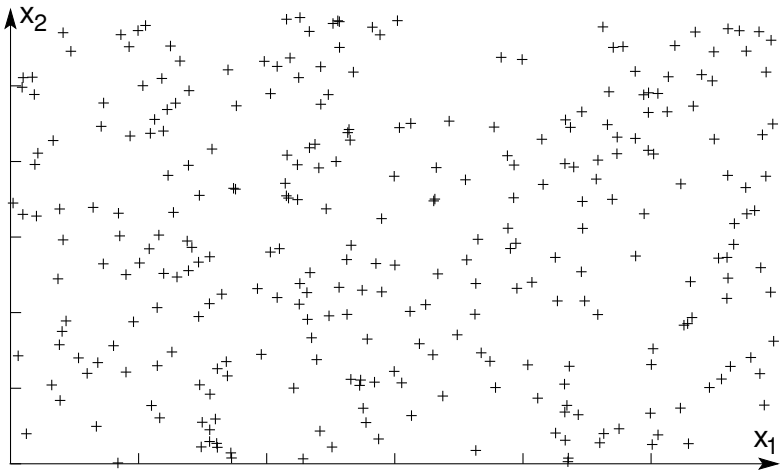
Figura: Quais grupos são mais representativos para o problema? Quantos grupos devem ser considerados?

- Após a definição da medida de distância, executa-se um algoritmo para obtenção do agrupamento, escolhido conforme a aplicação e o volume de dados presente no experimento.
- O agrupamento resultante provê um particionamento no qual amostras pertencentes a um mesmo grupo são consideradas como pertencentes à mesma classe. Portanto, o número de classes é igual ao número de grupos distintos.
- A validação do agrupamento é responsável por analisar os resultados obtidos com o particionamento.
- A tendência de agrupamento utiliza algumas medidas estatísticas para determinar se os dados apresentados podem ser agrupados ou estão uniformemente distribuídos no espaço, assim, não há a possibilidade de criar um particionamento para os dados.

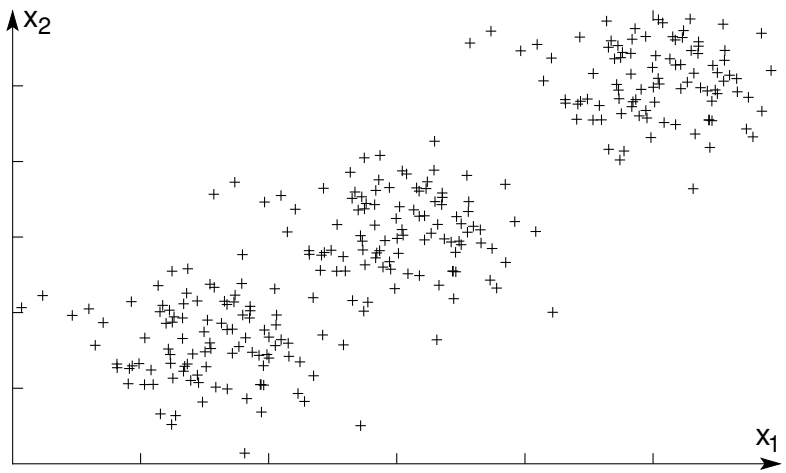
Agrupamento de Dados

Exemplo de distribuição dos dados

- Ilustração de amostras uniformemente distribuídas, impossibilitando a criação de um particionamento que agrupe aquelas com características semelhantes, bem como de distribuição na qual três grupos são facilmente perceptíveis.



dados uniformemente distribuídos



presença de grupos distintos

- Para determinar o nível de aleatoriedade presente nos dados, pode-se utilizar o teste de Hopkins, por meio da aplicação de um teste de hipóteses.
- Testa-se a hipótese nula, H_0, indicando que os dados estão uniformemente distribuídos, contra a hipótese que atesta a existência de possíveis particionamentos para os dados contidos no espaço de características.
- O objetivo do teste de Hopkins é determinar se um conjunto de dados dispostos espacialmente difere significativamente de dados que seguem a distribuição uniforme.
- Para isso, compara-se a distância entre os pontos contidos no conjunto e seus vizinhos mais próximos pertencentes a um conjunto de pontos que seguem a distribuição uniforme.

- Seja o conjunto $\boldsymbol{X}$ formado pelos vetores de características de cada amostra utilizada em um experimento, $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$. Dado um conjunto $\boldsymbol{Y}$, formado por M vetores de características, geralmente $M \approx 0.1n$, as medidas que os compõem seguem a distribuição uniforme no mesmo intervalo e escala daquelas que compõem o conjunto de vetores $\boldsymbol{X}$. Finalmente, seja $\boldsymbol{X}_1$ um subconjunto de $\boldsymbol{X}$ composto de M elementos escolhidos aleatoriamente.
- O teste de Hopkins é definido como

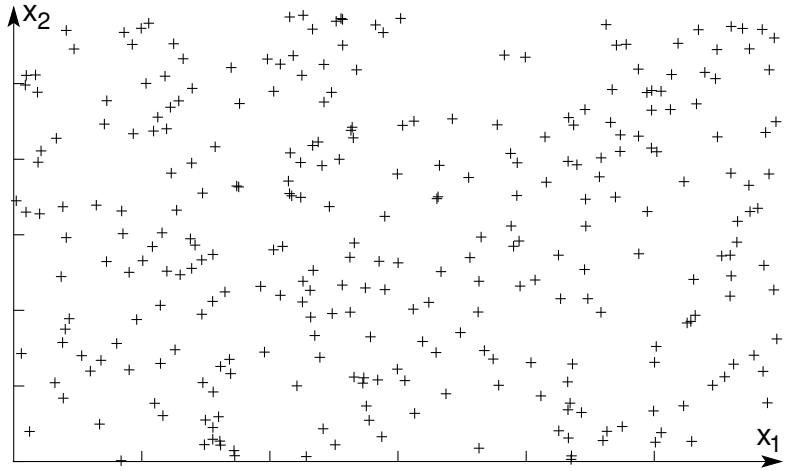
$$H = \frac{\sum_{j=1}^{M} d_j^l}{\sum_{j=1}^{M} d_j^l + \sum_{j=1}^{M} e_j^l}$$

em que l é a dimensionalidade dos dados e d_j denota a distância entre o j-ésimo componente de $\boldsymbol{Y}$ e o elemento $\boldsymbol{x}_k \in \boldsymbol{X}_1$, tal que seja mínima para todos os componentes de $\boldsymbol{X}_1$, e_j contém a menor distância entre $\boldsymbol{x}_k$ e os elementos de $\boldsymbol{X}_1 - \{\boldsymbol{x}_k\}$.

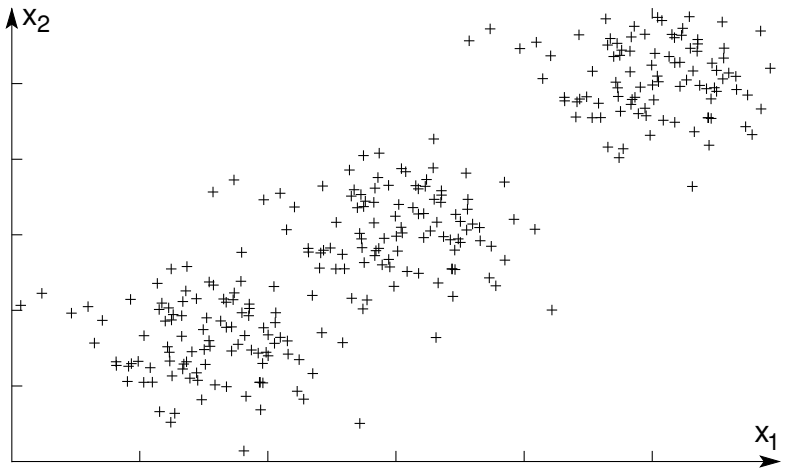
- Após sucessivas execuções do cálculo do teste de Hopkins, determinando a cada iteração novos $\boldsymbol{Y}$ e $\boldsymbol{X}_1$, se a média dos valores de H estiver acima de 0.75, a hipótese H_0 pode ser rejeitada com mais de 90% de confiança, ou seja, os dados podem ser particionados em grupos distintos. Caso contrário, os dados estão uniformemente distribuídos no espaço bidimensional de características, impossibilitando o particionamento e, consequentemente, a utilização de métodos de agrupamento de dados.
- Quando existe uma possível partição para os dados, a distância entre os elementos do conjunto $\boldsymbol{X}_1$ tende a ser pequena, portanto, o somatório dos elementos e_j resulta em valores baixos e leva H a assumir valores próximos de 1. Por outro lado, se a distribuição espacial dos dados for uniforme, a distância entre os elementos de $\boldsymbol{X}_1$ será maior, acarretando em valores expressivos para os termos que compõem o somatório de e_j, localizado no denominador da equação.

Agrupamento de Dados

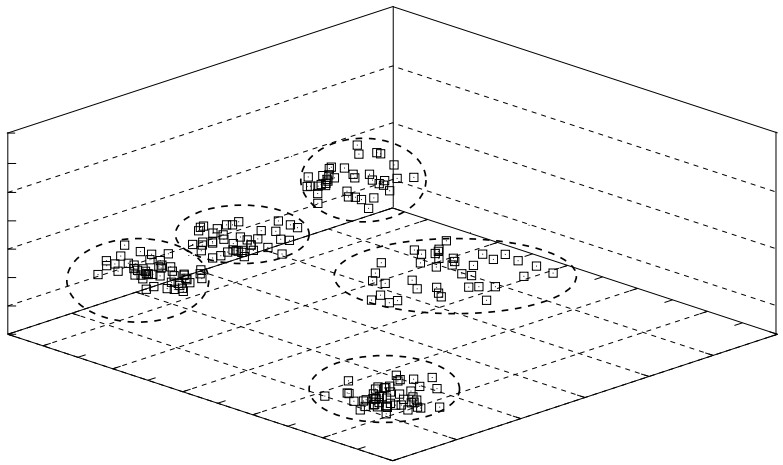
Exemplos de tendência de agrupamento



H = 0.467



H = 0.933



H = 0.998

Abordagens para Agrupamento de Dados

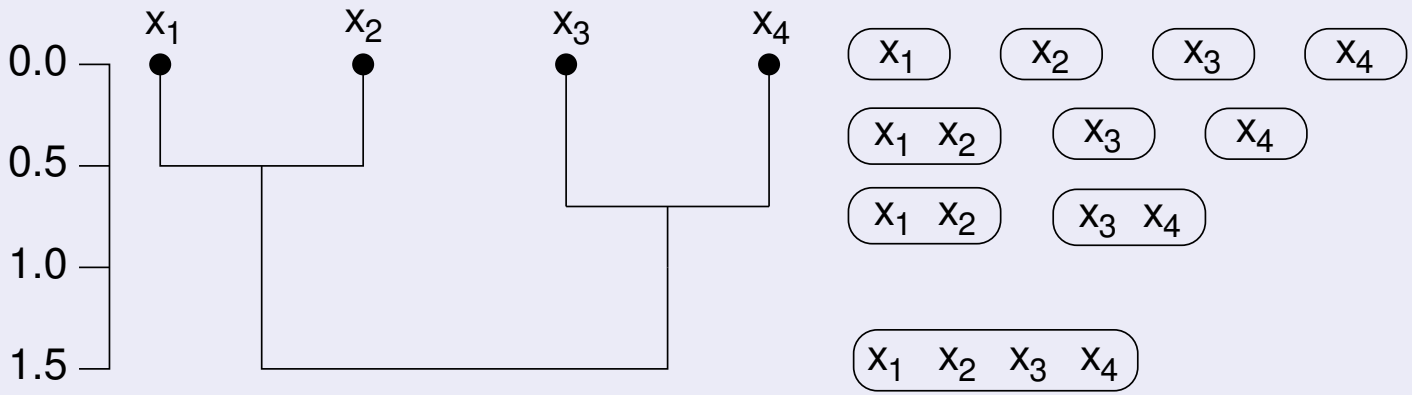
- Os algoritmos de agrupamento de dados são divididos em duas categorias:
  - hierárquicos: geram partições em diversos níveis.
    - ligação simples
    - ligação completa
    - média do grupo
    - centroide
  - particionais: geram um particionamento em nível único, com um número específico de grupos.

- Os métodos de agrupamento hierárquicos dividem o conjunto de dados iterativamente, produzindo múltiplos níveis de particionamento.
- Esses métodos são divididos em duas categorias:
  - *aglomerativos*: caracterizam-se por conter inicialmente n grupos, ou seja, cada classe contém apenas um elemento, sendo representada pelo vetor de características do respectivo elemento.
  - *divisivos*: efetuam o processo inverso, inicia-se com um único grupo, particionando-o até a obtenção de n grupos, ao final da execução.
- Uma característica em comum dos algoritmos divisivos e aglomerativos está no fato de que, uma vez efetuada a fusão ou partição, as amostras que foram agrupadas ou separadas não voltarão à condição anterior, até o final da execução.

- A execução dos algoritmos de agrupamento hierárquicos pode ser representada por meio de um diagrama conhecido com *dendrograma*, o qual ilustra as fusões ou os particionamentos efetuados a cada uma das etapas do algoritmo.
- Ao lado esquerdo do dendrograma, apresenta-se uma medida que especifica a distância em que cada novo grupo é formado.
- Tal medida pode ser interpretada como a similaridade entre os grupos, ou seja, quanto mais distante ocorrer uma fusão, os grupos apresentam-se menos similares.

Exemplo:

O dendrograma a seguir indica que o agrupamento $\boldsymbol{X} = \{\{\boldsymbol{x}_1, \boldsymbol{x}_2\}, \{\boldsymbol{x}_3\}, \{\boldsymbol{x}_4\}\}$ é obtido quando a distância entre os grupos está entre 0.5 e 0.6, e que os elementos x_1 e x_2 apresentam maior similaridade que x_3 e x_4.



- O processo de fusão entre dois agrupamentos comumente é dirigido pela matriz $\boldsymbol{D}$, $n \times n$, denominada *matriz das distâncias*, em que n denota o número de amostras consideradas pelo algoritmo de agrupamento e o elemento $d_{i,j}$ representa o valor da medida de similaridade $d(\boldsymbol{x}_i, \boldsymbol{x}_j)$.

Algoritmos Hierárquicos Aglomerativos

Definição

- Os métodos de agrupamentos de dados aglomerativos atribuem iterativamente as amostras para um mesmo grupo quando possuem características semelhantes, estas definidas segundo uma medida de similaridade específica.
- Conforme a execução desses métodos evolui, várias partições são produzidas, a primeira dessas consiste em n grupos compostos de uma amostra cada, enquanto a última é formada por um único grupo contendo todas as amostras consideradas no experimento.
- Como procedimento geral para os algoritmos aglomerativos, tem-se: a cada ciclo, fundem-se os dois grupos apresentando maior similaridade e suas medidas são atualizadas de modo a obter uma representação que considere os novos elementos contidos na classe.

Algoritmos Hierárquicos Aglomerativos

Algoritmo

Sejam n amostras descritas pelos vetores de características $\boldsymbol{x}_i$. Inicialmente, há n grupos distintos contendo apenas uma amostra. A cada ciclo, fundem-se os dois grupos que apresentam maior similaridade; repete-se o processo até a obtenção de um agrupamento composto de apenas uma classe.

1: // inicialização
2: $R_0 = \{C_i = \{\boldsymbol{x}_i\}, i = 1, \ldots, n\}$
3: $t = 0$
4: // definição dos agrupamentos
5: enquanto existir mais que um grupo faça
6: $\quad t = t + 1$
7: $\quad$ escolher C_i e C_j que apresentem maior similaridade
8: $\quad R_t = (R_{t-1} - \{C_i, C_j\}) \cup (C_i \cup C_j)$

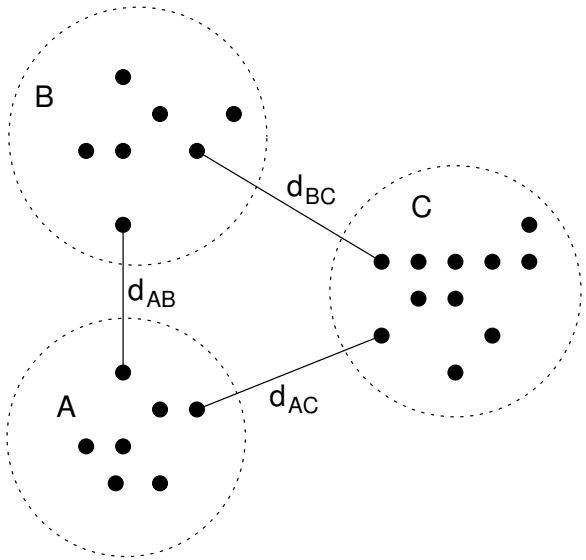
Algoritmos Hierárquicos Aglomerativos

Ligação Simples

- Considera a distância entre dois grupos como sendo a menor entre todos os possíveis pares de amostras que compõem tais grupos, em que A e B denotam dois grupos distintos, i e j seus respectivos elementos e os valores de $d_{i,j}$ são obtidos a partir da matriz de distâncias.

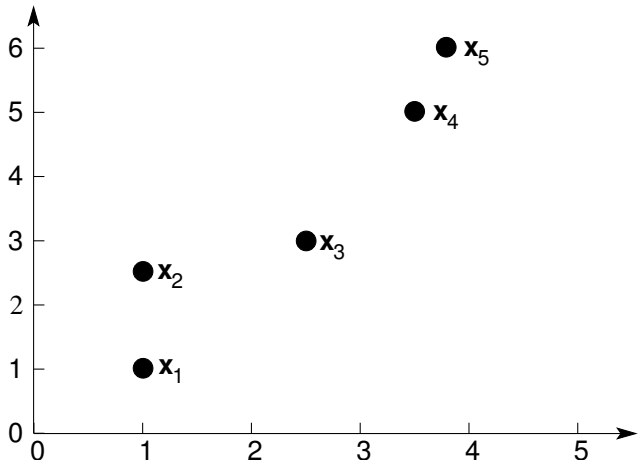
$$d_{AB} = \min_{i \in A, j \in B} d_{i,j}$$

- Ilustração da abordagem de ligação simples para determinar a distância entre dois grupos quaisquer. A fusão deve ser considerada para os grupos A e B, pois a distância mínima ocorre entre seus elementos.

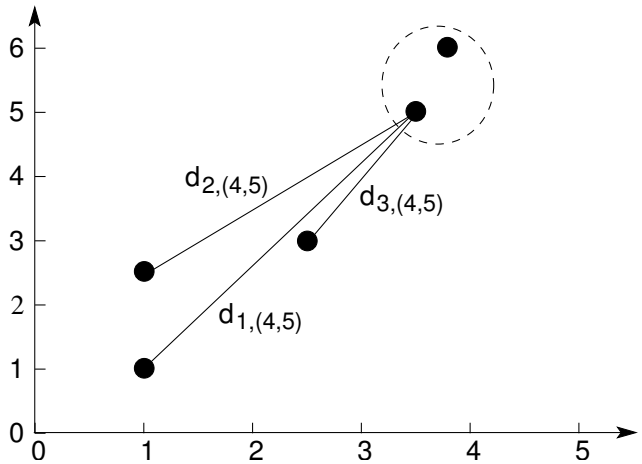


Algoritmos Hierárquicos Aglomerativos

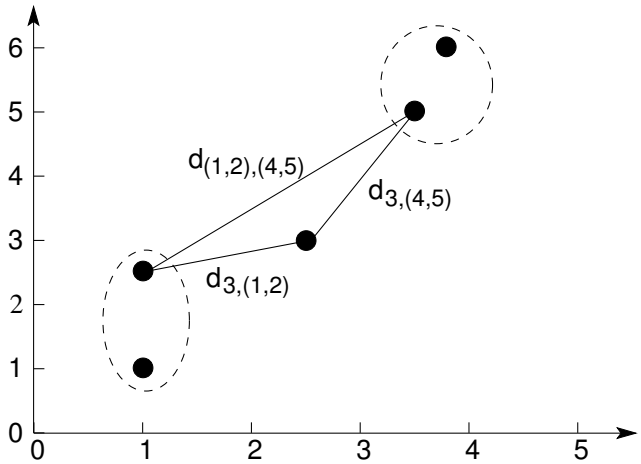
Ligação Simples



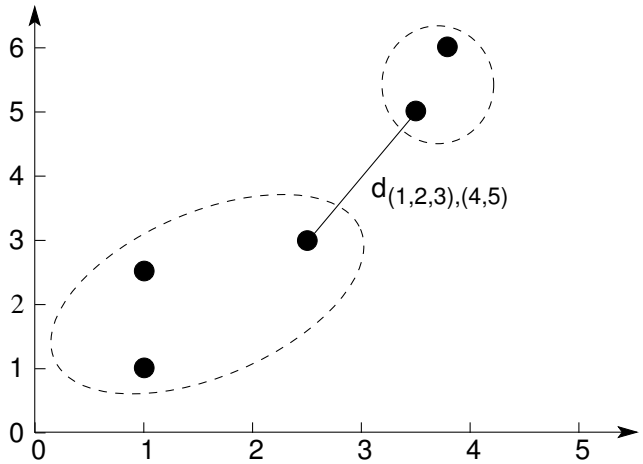
disposição espacial das amostras



fusão entre as amostras $\boldsymbol{x}_4$ e $\boldsymbol{x}_5$



fusão entre as amostras $\boldsymbol{x}_1$ e $\boldsymbol{x}_2$



terceiro nível do algoritmo

Algoritmos Hierárquicos Aglomerativos

Ligação Simples

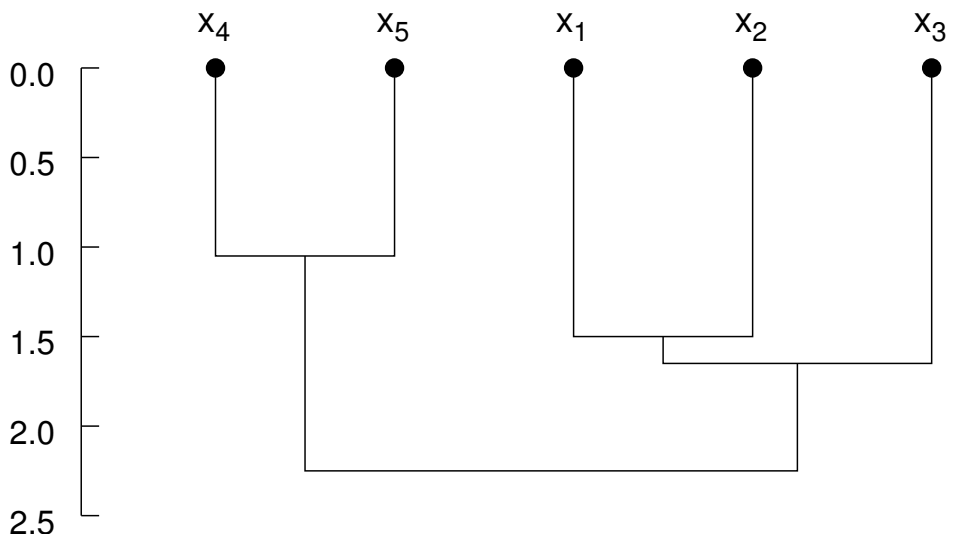


Figura: Dendrograma resultante da aplicação do algoritmo de agrupamento baseado na ligação simples.

- De maneira oposta à ligação simples, esta abordagem utiliza a distância entre os elementos mais distantes pertencentes a cada grupo.

$$d_{AB} = \max_{i \in A, j \in B} d_{i,j}$$

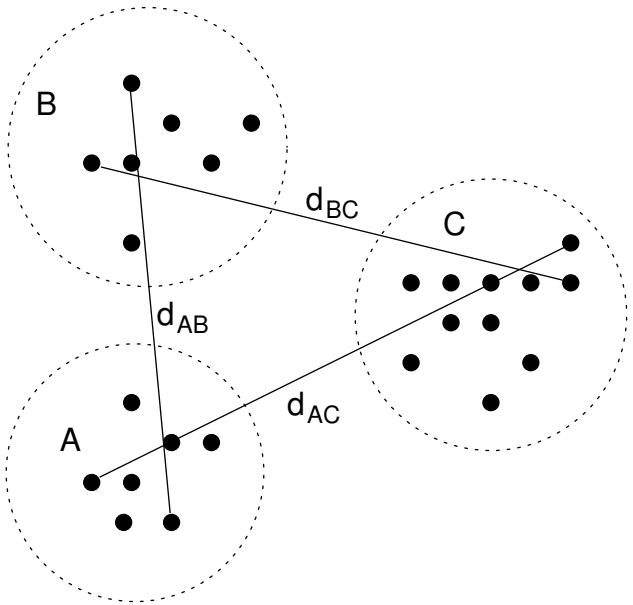
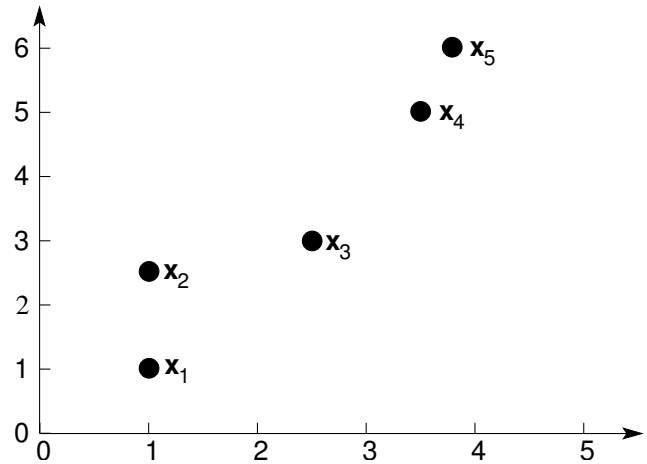


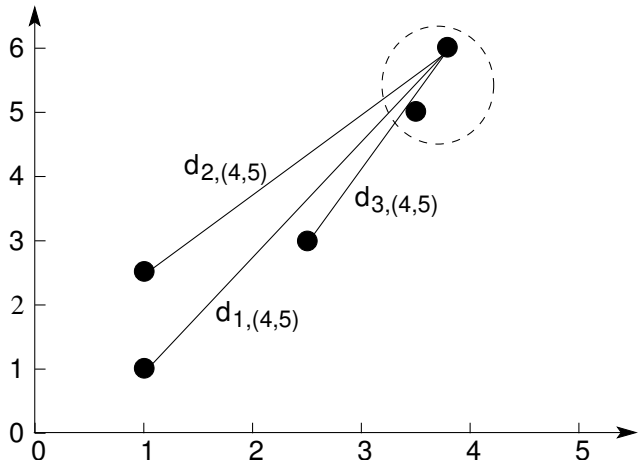
Figura: Determinação das distâncias entre os grupos A, B e C, considerando a ligação completa. A distância entre dois grupos é o máximo das distâncias entre as amostras que os compõem.

Algoritmos Hierárquicos Aglomerativos

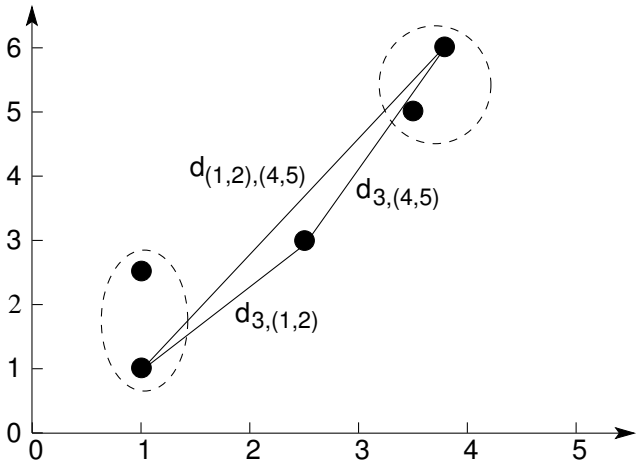
Ligação Completa



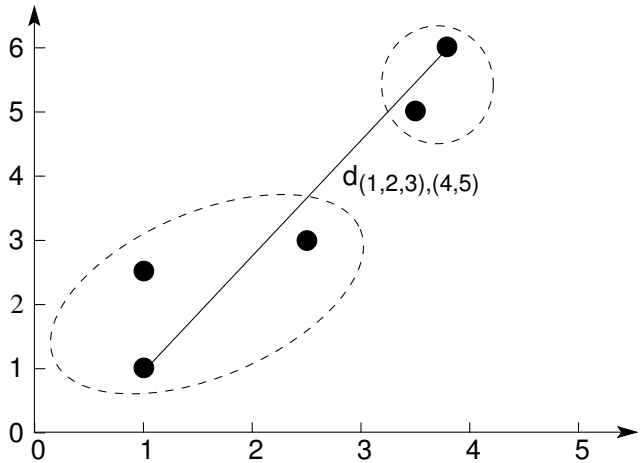
disposição espacial das amostras



fusão entre as amostras $\boldsymbol{x}_4$ e $\boldsymbol{x}_5$



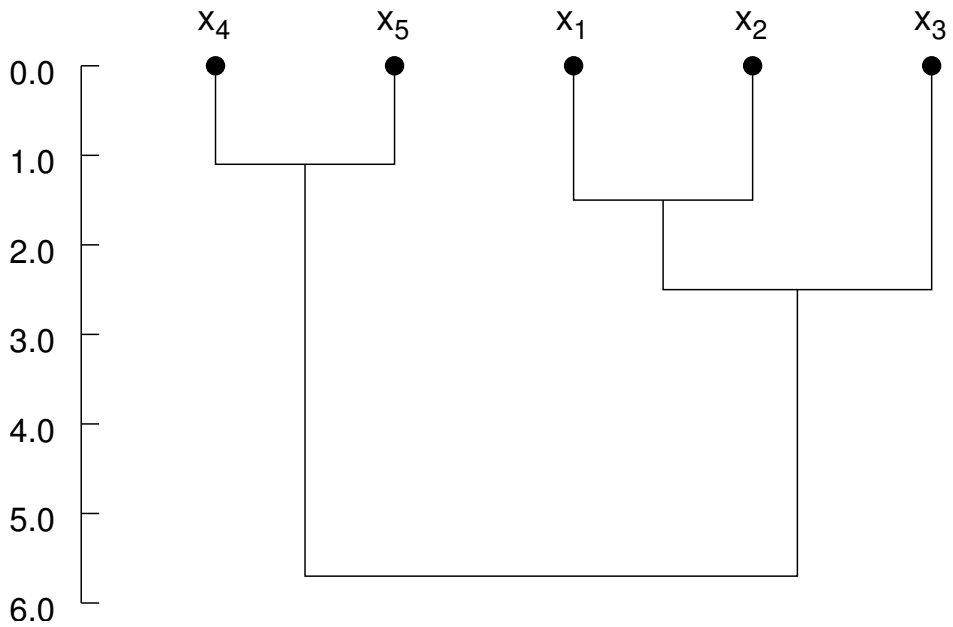
fusão entre as amostras $\boldsymbol{x}_1$ e $\boldsymbol{x}_2$



terceiro nível do algoritmo

Algoritmos Hierárquicos Aglomerativos

Ligação Completa



- Enquanto as duas abordagens apresentadas anteriormente consideram apenas uma amostra para determinar a distância entre dois grupos, esta considera a média aritmética entre todas as amostras que os compõem, conforme equação

$$d_{AB} = \frac{1}{n_A n_B} \sum_{i \in A, j \in B} d_{i,j}$$

em que n_A e n_B denotam o número de amostras contidas nos grupos A e B, respectivamente.

Algoritmos Hierárquicos Aglomerativos

Centroide

- Esta abordagem considera as coordenadas dos vetores de características para determinar o centroide, a cada iteração.
- Ilustração das distâncias entre os grupos, bem como a localização de seus respectivos centroides.

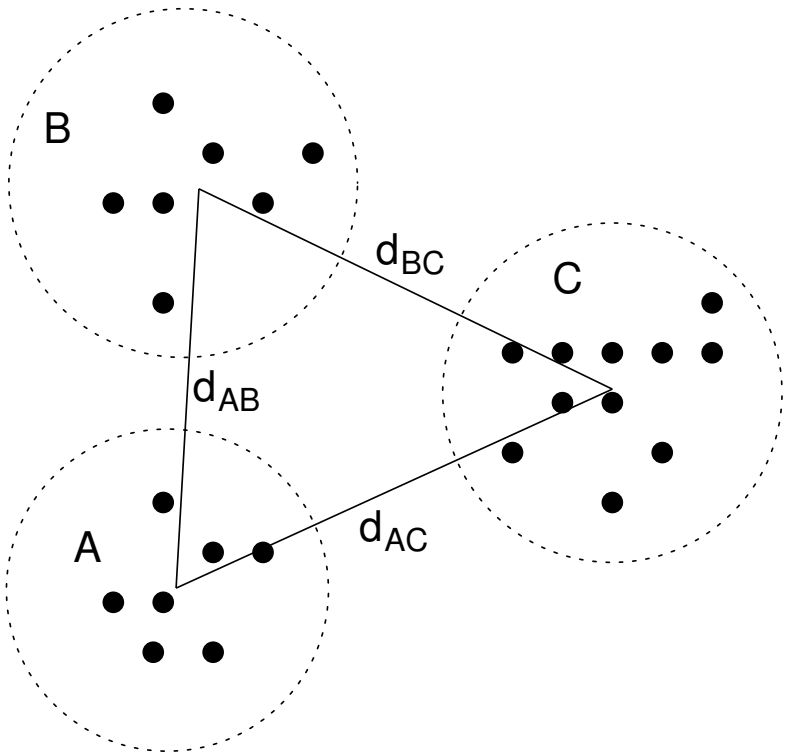


Figura: Determinação das distâncias entre os grupos A, B e C considerando o centroide dos dados como representante de cada grupo.

- A partir de um agrupamento composto de uma classe contendo todas as amostras, os métodos divisivos efetuam sucessivos particionamentos até a obtenção de n classes, ao final da execução.

  Considerando n amostras descritas pelos vetores de características $\boldsymbol{x}_i$. Inicialmente, todas as amostras pertencem a um único grupo, o qual é sucessivamente particionado até restar apenas uma amostra em cada grupo, ou seja, n classes distintas.

1: // inicialização
2: // $R_0 = \{X\}$
3: // $t = 0$
4: // definição dos grupos
5: // enquanto o número de grupos for menor que n faça
6: $t = t + 1$
7: selecionar o grupo C que será particionado
8: particionar C em C_i e C_j
9: $R_t = (R_{t-1} - \{C\}) \cup \{C_i, C_j\}$

- Visando à criação de um agrupamento único para os dados, os algoritmos particionais atribuem a uma mesma classe as amostras que apresentam propriedades semelhantes conforme uma medida de similaridade.
- O fato de efetuar apenas um particionamento caracteriza-se como a principal diferença da abordagem aplicada pelos algoritmos hierárquicos.
- Os algoritmos particionais apresentam como vantagem a inexistência da necessidade de se criar o dendrograma para controle da estrutura do particionamento sendo efetuado.
- No entanto, tem-se como principal desvantagem a necessidade de determinar o número de classes existentes no agrupamento, já que esses algoritmos geram apenas um particionamento.

- Os algoritmos de agrupamento contidos nesta categoria apresentam baixo custo computacional em razão da necessidade de se considerar cada amostra apenas um número reduzido de vezes durante a criação do agrupamento.
- Entretanto, o resultado obtido é altamente dependente da ordem com que as amostras são apresentadas ao algoritmo, podendo acarretar variações tanto no número de classes resultantes quanto na distribuição das amostras dentre as classes.
- O algoritmo sequencial básico utiliza cada amostra apenas uma vez, considerando o seguinte. Se a distância entre o vetor de características e o representante de cada classe existente for maior que um limiar previamente definido e o número de classes for menor que a constante k, uma nova classe é criada. Ao final de sua execução, as n amostras consideradas pelo algoritmo estarão particionadas em no máximo k classes.

- A escolha do parâmetro k, definido experimentalmente antes da execução do algoritmo, afeta diretamente o resultado obtido. Por exemplo, caso esse parâmetro receba um valor abaixo do número de classes existentes no conjunto de dados não será possível a obtenção de um resultado final satisfatório.
- Outro problema está no fato de que os representantes de cada classe não têm seus valores atualizados quando uma nova amostra é rotulada. Dessa maneira, o resultado final está altamente relacionado com a escolha inicial dos representantes de cada classe.
- Como a ordem em que as amostras são apresentadas para o algoritmo é relevante, a atualização do representante da classe torna-se importante para evitar que um elemento localizado em regiões distantes de onde deveria estar o centroide da classe seja considerado como representante.

- Os métodos de agrupamento pertencentes a esta classe efetuam o particionamento dos dados visando à minimização ou maximização de algum critério.
- Mais especificamente, seja um conjunto composto de n pontos pertencentes a um espaço de características, e um inteiro k, o objetivo está na determinação de um particionamento do espaço de características em k agrupamentos, cada um representado por um centro, visando à minimização ou maximização de um critério específico. Por exemplo, a minimização da distância entre os pontos e o centro mais próximos.
- O algoritmo mais conhecido baseado na otimização de uma função de custo é o k-médias, descrito a seguir.

- O algoritmo k-médias é um procedimento de agrupamento que busca minimizar a distância das amostras de dados ao grupo, que pode ser expressa pela soma dos erros quadráticos

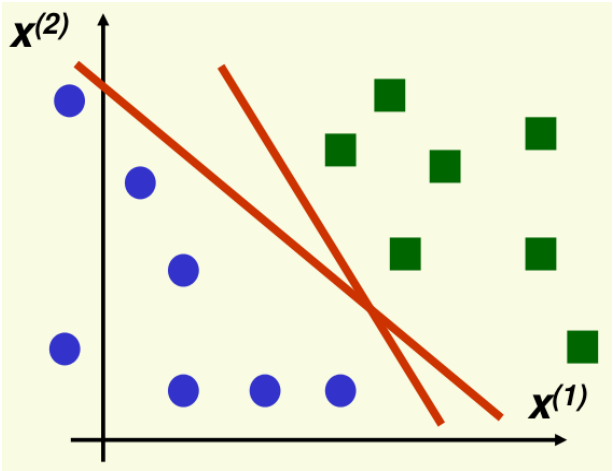
$$J_{\text{erro}} = \arg\min_{C} \sum_{i=1}^{m} \sum_{\boldsymbol{x} \in C_i} \|\boldsymbol{x} - \mu_i\|^2, \qquad \text{tal que } \mu_i = \frac{1}{n_i} \sum_{\boldsymbol{x} \in C_i} \boldsymbol{x}$$

em que m é o número de grupos, n_i é o número de amostras no grupo C_i e μ é a média dessas amostras.

1: Inicializar k centroides dos grupos (por exemplo, de forma aleatória).
2: Atribuir cada amostra ao grupo mais próximo.
3: Computar a média das amostras de cada grupo.
4: Atualizar os centroides como a média de seus grupos.
5: Repetir passos 2 a 4 até que os centroides não sofram alterações.

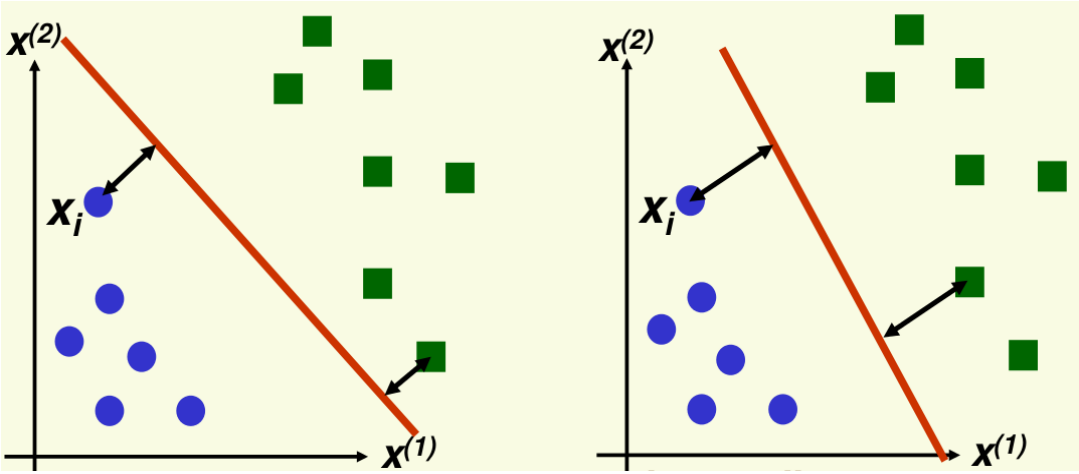
- Inicialização dos centroides.
  - Convergência mais rápida a partir de conhecimento prévio dos centroides.
  - Execuções múltiplas do algoritmo podem reduzir impacto da inicialização aleatória.

- Seja um conjunto de treinamento com n amostras
- O objetivo é encontrar um hiperplano para separar amostras em diferentes regiões.



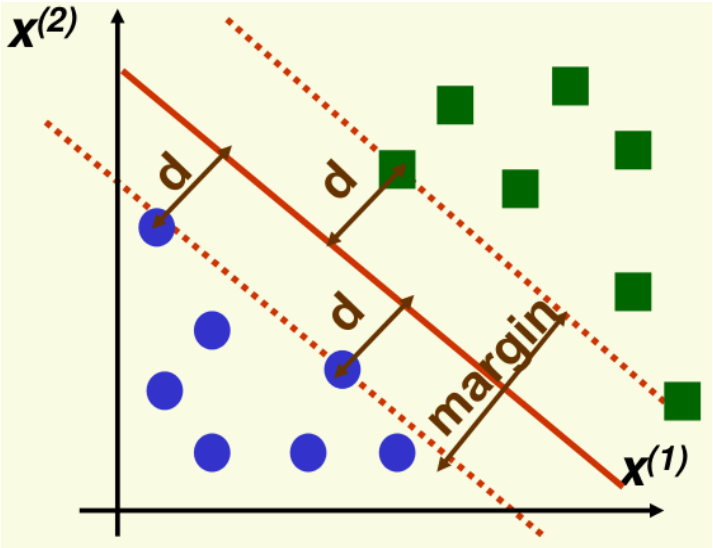
como escolher o hiperplano?

- ideia: maximizar a distância do hiperplano em relação à amostra mais próxima.



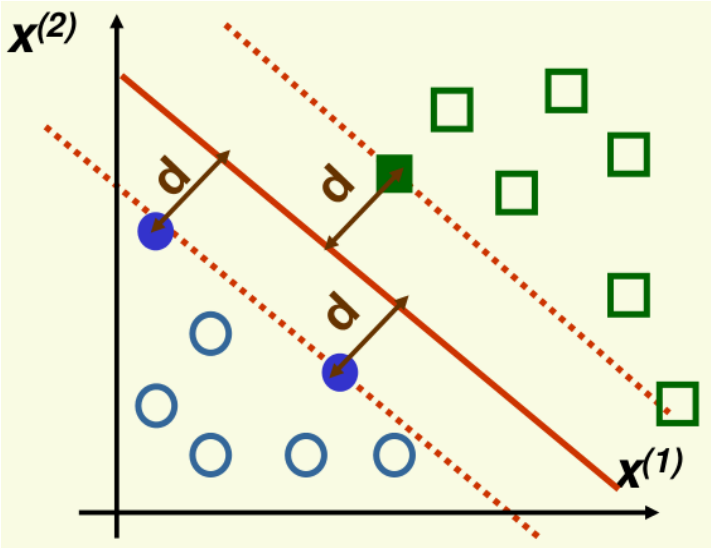
menor distância $\times$ maior distância

- SVM: procura maximizar a margem, que é o dobro do valor absoluto da distância d das amostras mais próximas ao hiperplano de separação.



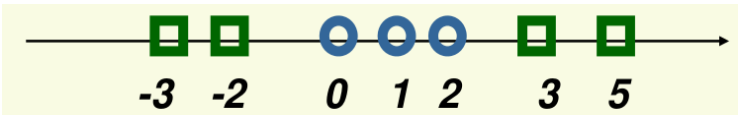
margem

- Vetores de suporte: são as amostras mais próximas ao hiperplano de separação.
  - essas amostras são os padrões mais difíceis de classificar.
  - hiperplano ótimo é definido pelos vetores de suporte.

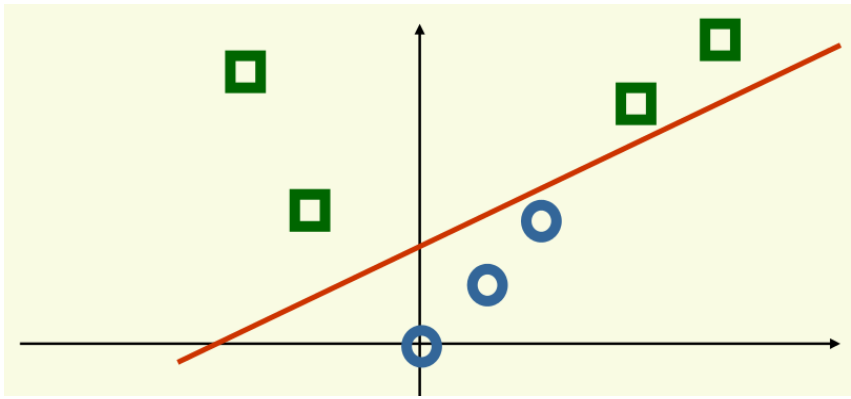


vetores de suporte

- Teorema de Cover: problemas de classificação de padrões em um espaço de alta dimensionalidade são mais prováveis de serem linearmente separáveis do que em um espaço de baixa dimensionalidade.



não separável linearmente em 1D



separável linearmente em 2D

- Qualquer classificador linear poderia ser utilizado em um espaço de alta dimensionalidade.
- Entretanto, deve-se lidar com a "maldição da dimensionalidade":
  - generalização pobre aos dados de teste
  - caro computacionalmente
- SVM trata do problema
  - garantir as maiores margens permite boa generalização.
  - computação em casos de alta dimensionalidade é realizada por meio de funções de *kernel*.

- Vantagens
  - baseadas em formulações matemáticas poderosas.
  - exibem propriedades de generalização.
  - podem ser utilizadas para encontrar funções discriminantes não lineares.
  - complexidade do classificador é descrita pelo número de vetores de suporte ao invés da dimensionalidade do espaço transformado.
- Desvantagens
  - seleção da função de *kernel* não é intuitiva.
  - tendem a ser lentas em comparação com outras abordagens (caso não linear).

- O desempenho dos métodos de classificação está estreitamente relacionado com o conjunto de dados no qual estes serão aplicados.
- Os métodos de treinamento apresentam como objetivo encontrar uma função que descreva de maneira satisfatória as amostras contidas no conjunto de treinamento.
- Em outras palavras, dado um conjunto de dados formado pelas tuplas $< \boldsymbol{x}_i, y_i >$, em que $\boldsymbol{x}_i$ e y_i representam o vetor de características e o rótulo atribuído à i-ésima amostra que compõe o conjunto de treinamento, tem-se como objetivo encontrar uma função $f(\boldsymbol{x}_i) \rightarrow y_i, \ \forall i$.
- Como a correta rotulação está disponível apenas para o conjunto de treinamento, as amostras nele contidas são utilizadas para estimar o desempenho de um classificador para amostras que não possuam uma rotulação conhecida, o qual se apresenta como objetivo primário da classificação de padrões.

Avaliação dos Classificadores

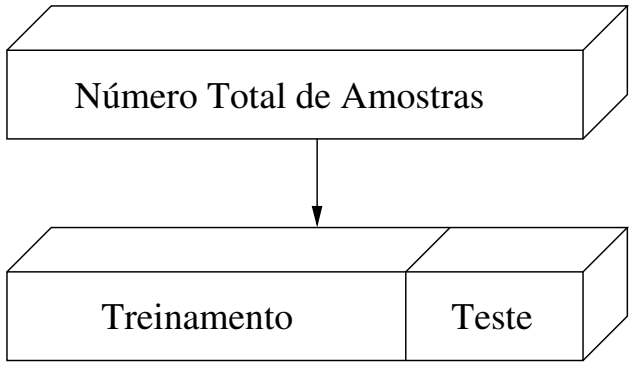
Problemas de ajuste na aprendizagem

- Sobreajuste (*overfitting*): classifica perfeitamente os dados de treinamento, entretanto, generaliza mal.
  - grande quantidade de ruídos.
  - dados de treinamento não representativos.
  - poucas amostras de treinamento.
- Subajuste (*underfitting*): não se ajusta aos dados de treinamento
  - dados não representativos ou esparsos.

Avaliação dos Classificadores

Holdout

- Este estimador divide as amostras em uma porcentagem fixa de exemplos p para treinamento e $(1 - p)$ para teste, considerando normalmente $p > 1/2$.



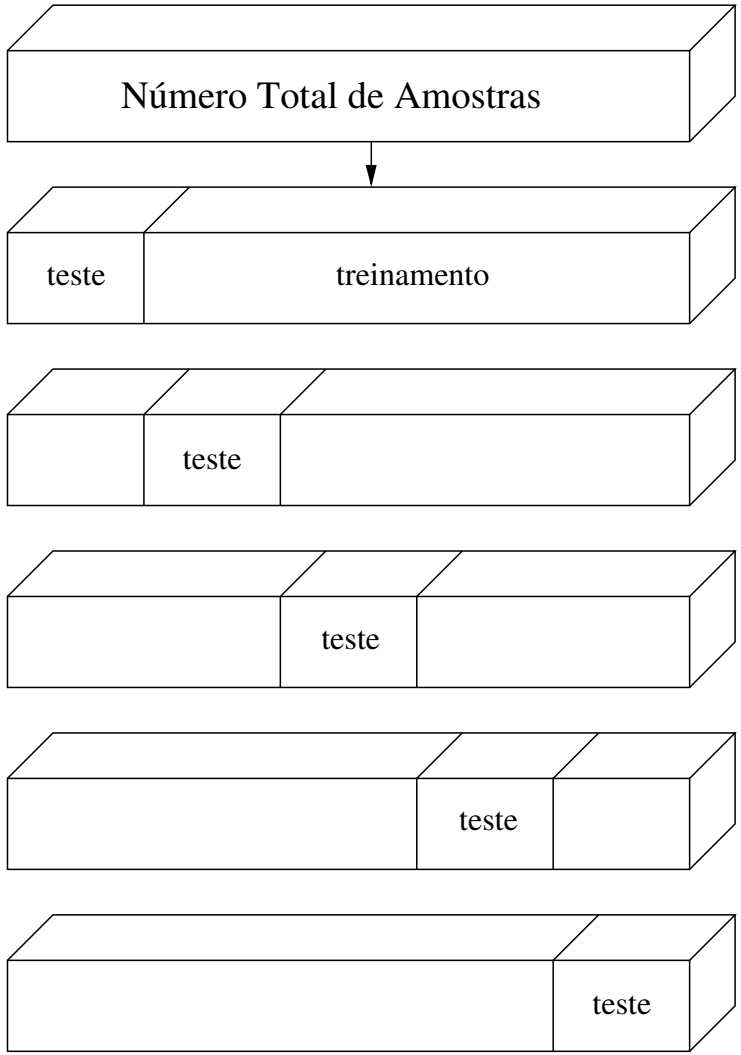
- Uma vez que uma hipótese construída utilizando todas as amostras, em média, apresenta desempenho melhor do que uma hipótese construída utilizando apenas uma parte das amostras, este método tem a tendência de superestimar o erro verdadeiro.
- De forma a tornar o resultado menos dependente da forma de divisão dos exemplos, pode-se calcular a média de vários resultados de *holdout* pela construção de várias partições obtendo-se, assim, uma estimativa média do *holdout*.

- Na amostragem aleatória, L hipóteses ($L << n$) são induzidas a partir de cada um dos L conjuntos de treinamento.
- O erro final é calculado como sendo a média dos erros de todas as hipóteses induzidas e calculados em conjuntos de teste independentes e extraídos aleatoriamente.
- Amostragem aleatória pode produzir melhores estimativas de erro que o estimador *holdout*.

- Este estimador é um meio termo entre os estimadores *holdout* e *leave-one-out*.
- Na validação cruzada com k partições, as amostras são aleatoriamente divididas em k partições mutuamente exclusivas de tamanho aproximadamente igual a n/k exemplos.
- Os exemplos nas $(k-1)$ partições são usadas para treinamento e a hipótese induzida é testada na partição remanescente.
- Este processo é repetido k vezes, cada vez considerando uma partição diferente para teste.
- O erro na validação cruzada é a média dos erros calculados em cada um das k partições.

Avaliação dos Classificadores

Validação Cruzada



- Este procedimento de rotação reduz tanto o *bias* inerente ao método de *holdout* quanto o custo computacional do método *leave-one-out*.
- Entretanto, deve-se observar, por exemplo, que na validação cruzada com 10 partições, cada par de conjuntos de treinamento compartilha 80% de exemplos.
- À medida que o número de partições aumenta, esta sobreposição pode evitar que os testes estatísticos obtenham uma boa estimativa da quantidade de variação que seria observada se cada conjunto de treinamento fosse independente dos demais.

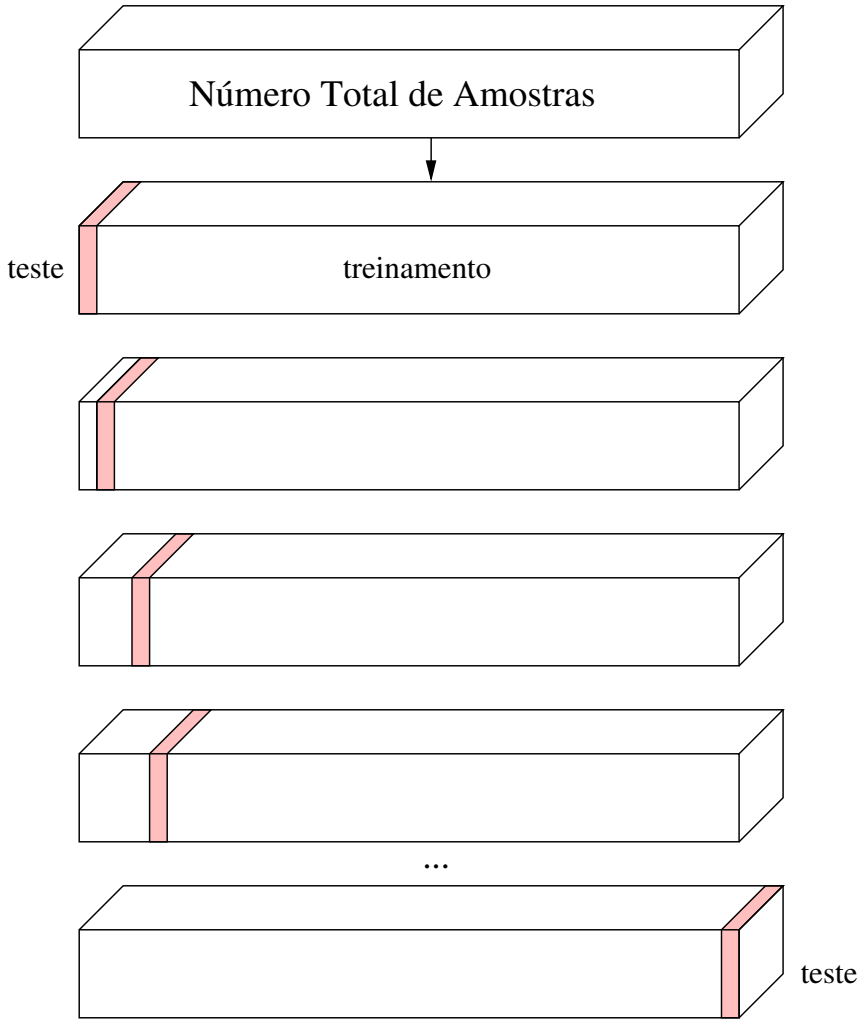
Avaliação dos Classificadores

Leave-One-Out

- O estimador *leave-one-out* é um caso especial de validação cruzada.
- É computacionalmente dispendioso e frequentemente é usado em amostras pequenas.
- Para uma amostra de tamanho n, uma hipótese é induzida utilizando $(n-1)$ exemplos; a hipótese é então testada no único exemplo remanescente.
- Este processo é repetido n vezes, cada vez induzindo uma hipótese deixando de considerar um único exemplo.
- O erro é a soma dos erros em cada teste dividido por n.

Avaliação dos Classificadores

Leave-One-Out



Avaliação dos Classificadores

Bootstrap

- No estimador *bootstrap*, a ideia básica consiste em repetir o processo de classificação um grande número de vezes.
- Diferentemente da validação cruzada (que usa amostragem sem reposição), a técnica de *bootstrap* usa amostragem com reposição para formar o conjunto de treinamento
- Estimam-se então valores, tais como o erro ou *bias*, a partir dos experimentos replicados, cada experimento sendo conduzido com base em um novo conjunto de treinamento obtido por amostragem com reposição do conjunto original de amostras.

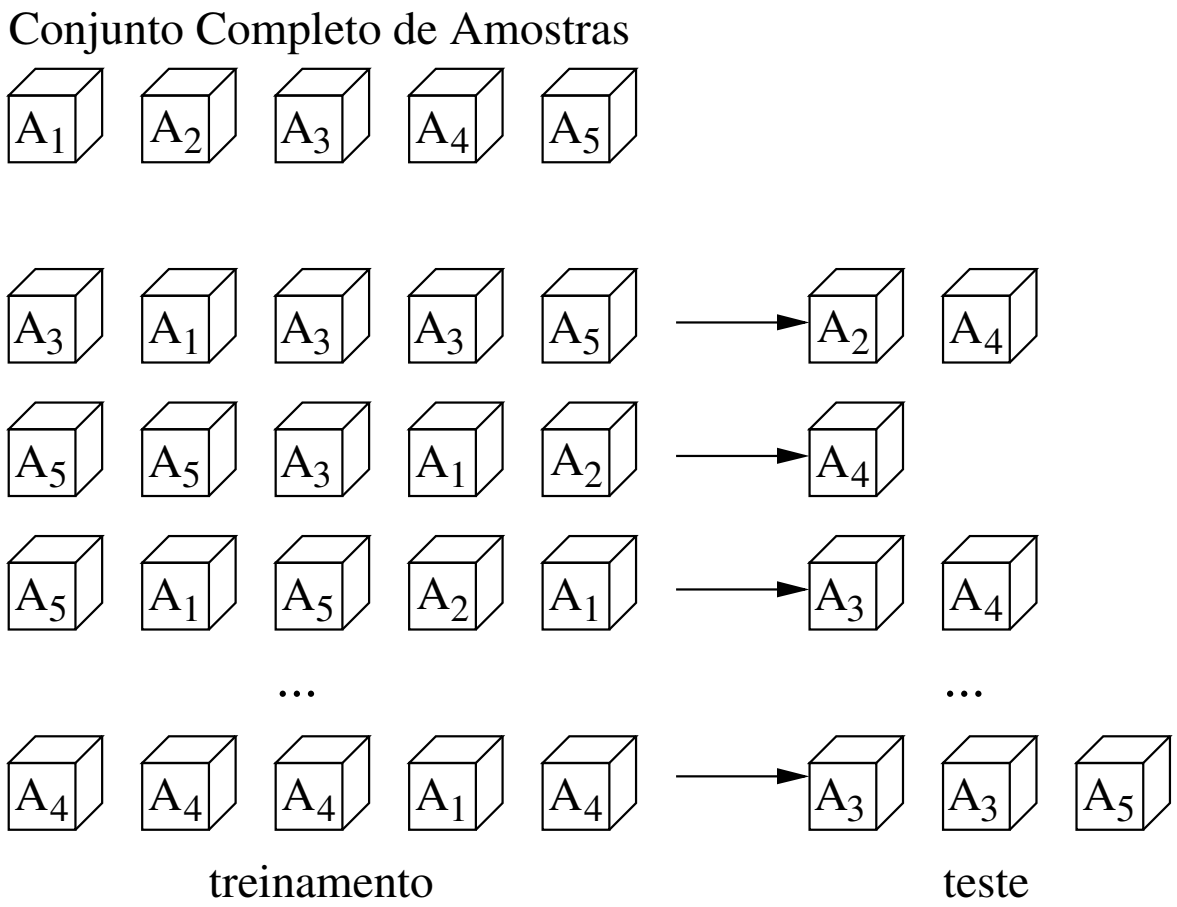
Avaliação dos Classificadores

Bootstrap e0

- Há muitos estimadores *bootstrap*, sendo o mais comum denominado *bootstrap e0*.
- Um conjunto de treinamento *bootstrap* consiste em n amostras (mesmo tamanho do conjunto original) amostradas com reposição a partir do conjunto original de amostras.
- Isto significa que algumas amostras A_i podem não aparecer no conjunto de treinamento *bootstrap* e algumas A_i podem aparecer mais de uma vez.
- As amostras remanescentes (aquelas que não aparecem no conjunto de treinamento *bootstrap*) são usadas como conjunto de teste.

Avaliação dos Classificadores

Bootstrap e0



Avaliação dos Classificadores

Bootstrap e0

- Para uma dada amostra *bootstrap*, uma amostra de treinamento tem probabilidade $1 - (1 - 1/n)^n$ de ser selecionada pelo menos uma vez em cada uma das n vezes nas quais as amostras são aleatoriamente selecionadas a partir do conjunto original de amostras. Para n alto, isto é aproximadamente $1 - 1/e = 0.632$.
- Portanto, para esta técnica, a fração média de amostras não repetidas é 63.2% no conjunto de treinamento e 36.8% no conjunto de teste.
- Geralmente, o processo de *bootstrap* é repetido um número de vezes, sendo o erro estimado como a média dos erros sobre o número de iterações.

Avaliação dos Classificadores

Matriz de Confusão

- A matriz de confusão apresenta o número de linhas e colunas equivalente ao número de classes do problema, em que um elemento a_{ij} indica o número de amostras atribuídas à classe C_i dado que a classe correta é a C_j. Dessa maneira, os elementos contidos na diagonal principal da matriz denotam o número de amostras classificadas corretamente.

$$\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mm} \end{bmatrix}$$

- Exemplo: resultado da classificação de 300 amostras pertencentes a seis classes diferentes.

$$\begin{bmatrix} 40 & 4 & 0 & 0 & 7 & 0 \\ 3 & 45 & 0 & 0 & 1 & 0 \\ 0 & 0 & 29 & 1 & 3 & 18 \\ 0 & 0 & 4 & 43 & 0 & 14 \\ 7 & 1 & 1 & 0 & 39 & 0 \\ 0 & 0 & 16 & 6 & 0 & 18 \end{bmatrix}$$

$$\text{taxa de erro} = 1 - \frac{214}{300} \approx 0.287$$

Avaliação dos Classificadores

Tabela de Contingência

- Estrutura construída com:
  - verdadeiros positivos (VP): valores positivos que o sistema julgou positivos (acerto).
  - falsos negativos (FN): valores positivos que o sistema julgou negativos (erro).
  - verdadeiros negativos (VN): valores negativos que o sistema julgou como negativos (acerto).
  - falsos positivos (FP): valores negativos que o sistema julgou positivos (erro).

		classe predita	
		positivo	**negativo**
classe real	**positivo**	verdadeiro positivo	falso negativo
	negativo	falso positivo	verdadeiro negativo

Avaliação dos Classificadores

Medidas derivadas da tabela de contingência

- Acurácia: proporção de predições corretas, sem levar em consideração o que é positivo e o que é negativo. Esta medida é altamente suscetível a desbalanceamentos do conjunto de dados e pode facilmente induzir a uma conclusão errada sobre o desempenho do sistema.

$$\text{Acurácia} = \frac{\text{Total de Acertos}}{\text{Total de Dados no Conjunto}} = \frac{\text{VP} + \text{VN}}{\text{P} + \text{N}}$$

- Sensibilidade: proporção de verdadeiros positivos, ou seja, a capacidade do sistema em predizer corretamente a condição para casos que realmente a têm.

$$\text{Sensibilidade} = \frac{\text{Acertos Positivos}}{\text{Total de Positivos}} = \frac{\text{VP}}{\text{P}} = \frac{\text{VP}}{\text{VP} + \text{FN}}$$

- Especificidade: proporção de verdadeiros negativos, ou seja, a capacidade do sistema em predizer corretamente a ausência da condição para casos que realmente não a têm.

$$\text{Especificidade} = \frac{\text{Acertos Negativos}}{\text{Total de Negativos}} = \frac{\text{VN}}{\text{N}} = \frac{\text{VN}}{\text{VN} + \text{FP}}$$

Avaliação dos Classificadores

- Exemplo: resultado da classificação de 27 animais (8 gatos, 6 cachorros e 13 coelhos), cuja matriz de confusão é:

		classe predita		
		gato	**cachorro**	**coelho**
classe real	**gato**	5	3	0
	cachorro	2	3	1
	coelho	0	2	11

  - Dos 8 gatos, 3 foram classificados como cachorros;
  - Dos 6 cachorros, 2 foram classificados como gato e 1 como coelho;
  - Dos 13 coelhos, 2 foram classificados como cachorros.
- Matriz de confusão para a classe gato:

5 verdadeiros positivos (gatos que foram corretamente classificados como gatos)	3 falsos negativos (gatos que foram incorretamente classificados como cachorros)
2 falsos positivos (cachorros que foram incorretamente classificados como gatos)	17 verdadeiros negativos (animais restantes, corretamente classificados como não gatos)