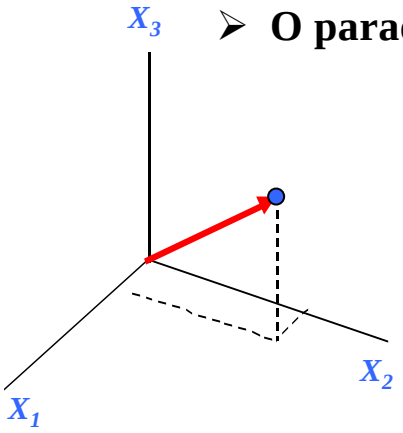
Reconhecimento Estatístico de Padrões



Espaço de características

- **O paradigma pode ser sumarizado da seguinte forma:**

- **Cada padrão é representado por um vector de características**

$$x = (x_1, x_2, \ldots, x_N)$$

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \ldots \\ x_d \end{bmatrix}$$

- **Um dado padrão deve ser classificado em uma de *C* categorias, w_1, w_2, …, w_C, com base no seu vector de características.**

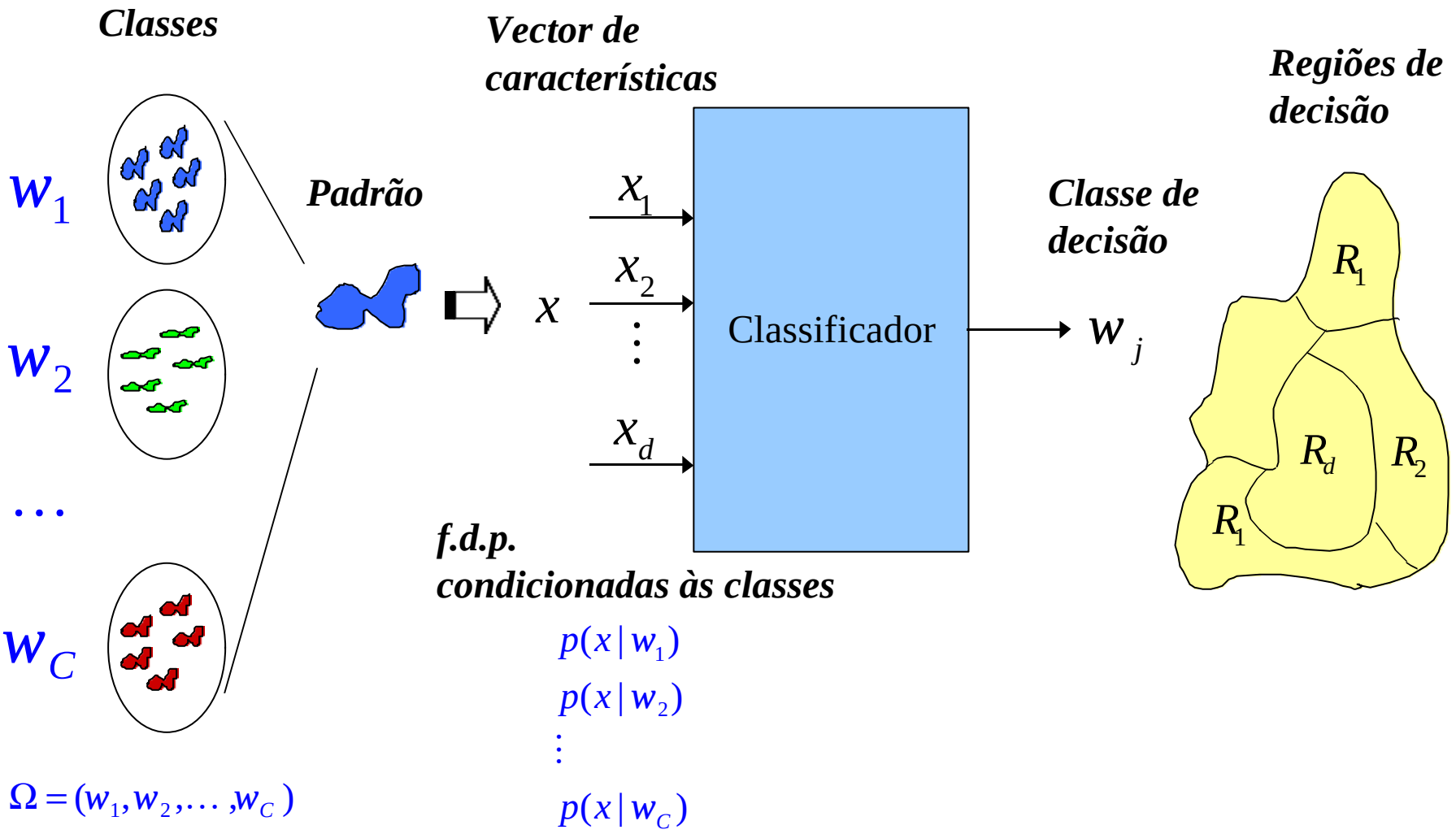
$$\Omega = (w_1, w_2, \ldots, w_C)$$

- **Assume-se que o vector de características possui uma f.d.p. típica da sua classe. Um vector *x* pertencente à classe w_i é visto como uma observação gerada aleatoriamente de acordo com a f.d.p. condicionada à classe**

$$p(x \mid w_i)$$

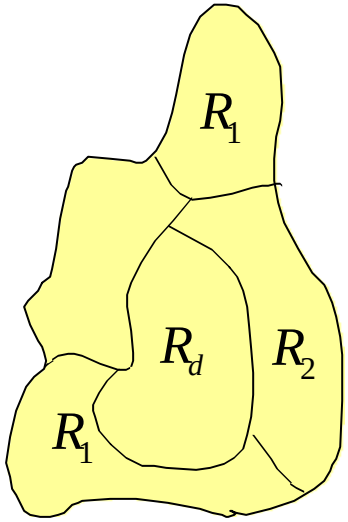
- **Conceitos da teoria estatística de decisão são usados para estabelecer fronteiras de decisão entre as classes**

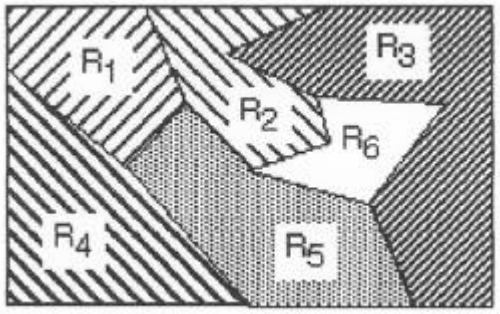
Reconhecimento Estatístico de Padrões

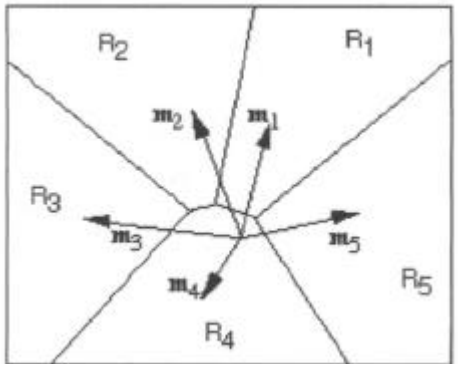


Regiões de Decisão e Superfícies de Separação

- **Em geral, um classificador particiona o espaço de características em volumes designados <u>regiões de decisão.</u>**
- **Todos os vectores de características no interior de uma região de decisão são atribuídos à mesma categoria.**
- **A região de decisão para uma classe pode ser simplesmente conexa, ou pode consistir em duas ou mais sub-regiões não adjacentes.**
- **As regiões de decisão encontram-se separadas por superfícies designadas <u>superfícies de decisão</u> ou <u>superfícies de separação.</u> Estas superfícies representam pontos onde existem "empates" entre duas ou mais categorias.**







Superfícies de decisão do classificador de distância mínima

$$g_i(x) = \|x - m_i\|$$

Ex: Decisão de MAP

- .

Atribuir x `a classe w_i se

$$p(w_i \mid x) \geq p(w_j \mid x) \quad \forall_{j \neq i}$$

com $p(w_i \mid x)$ **probabilidade *a posteriori* da classe** w_i **, definida em termos das f.d.p. condicionadas Às classes e as probabilidades *a priori* das classes** $p(w_i)$ **através de**

$$p(w_i \mid x) = \frac{p(x \mid w_i)\, p(w_i)}{\sum_{j=1}^{C} p(x \mid w_j)\, p(w_j)}$$

- **Esta regra de decisão é óptima no sentido em que, para uma dada distribuição *a priori*, não existe uma regra de decisão com menor probabilidade de erro de classificação.**

Ex: Regra de decisão de Máxima Verosimilhança

- **Quando as probabilidades *a priori* das classes são idênticas ($P(w_i)=1/C$) a regra de decisão de Bayes é idêntica à regra de decisão de Máxima Verosimilhança:**

Atribuir x `a classe w_i se

$$p(x \mid w_i) \geq p(x \mid w_j) \quad \forall_{j \neq i}$$

Funções Discriminantes e Superfície de Decisão

- **A noção de <u>função discriminante</u> e <u>superfície de decisão</u> são muitos importantes em R. P. Estatístico.**

- **Funções discriminantes e regras de decisão:**
  - **Cada classe, w_i, tem associada uma função discriminante, $g_i(x)$**
    - **Ex:** $g_i(x) = p(w_i \mid x)$
  - **Regra de decisão usando funções discriminantes:**

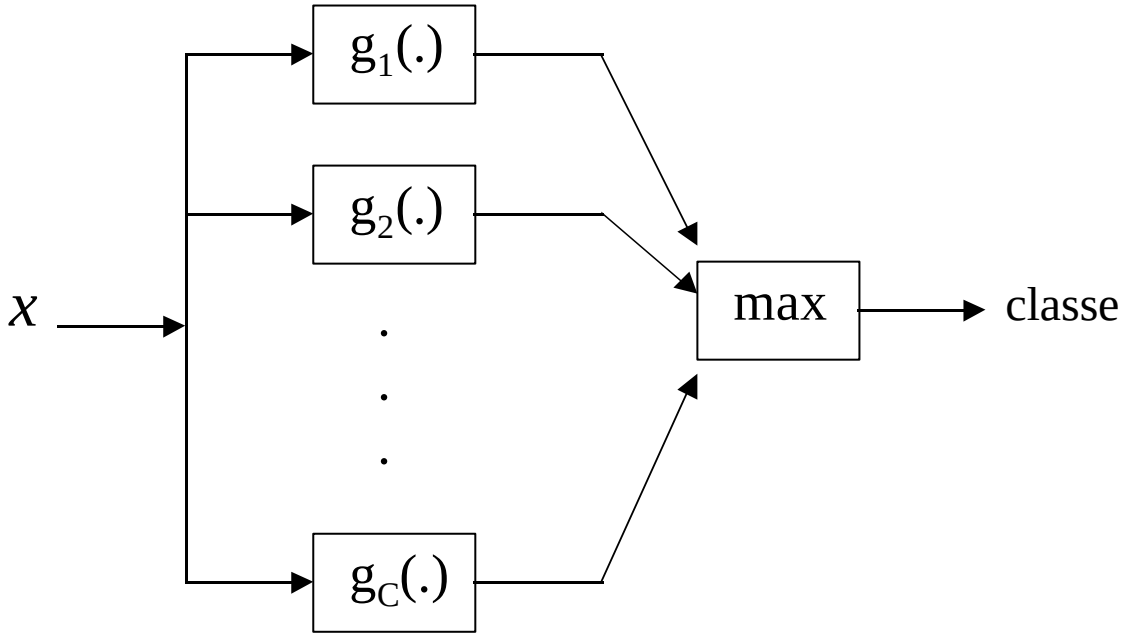
    Atribua x `a classe w_i se $g_i(x) \geq g_j(x) \quad \forall_{j \neq i}$

  - **Outra forma de escrever a regra:**

    $x \in w_k$: k=arg $\max_{i=1,\ldots,C} g_i(x)$

  - **A superfície de separação no espaço d-dimensional de características entre as classes w_i e w_j é definida pela equação** $g_i(x) - g_j(x) = 0$

Estrutura do Classificador Baseado em Funções Discriminantes



Região de decisão: $\Re_j = \{x : g_j(x) \geq g_i(x), \forall_i\}$

Ex. Classificador de MAP: Caso em que as f.d.p. condicionadas às classes são Gaussianas $N(m_i, I)$

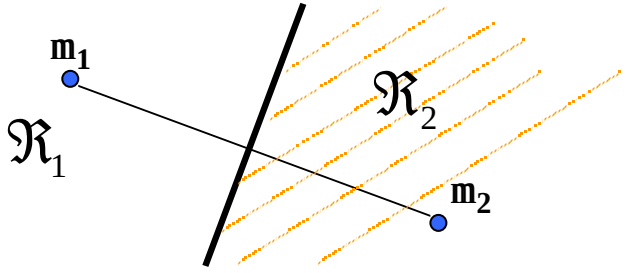
$$P(x|w_i) = N(m_i, I) = \frac{1}{(2p)^{1/d}} e^{-\frac{\|x-m_i\|^2}{2}}$$

- **Se as probabilidades *a priori* das classes são idênticas, então a função discriminante para a classe w_i pode ser escrita como**

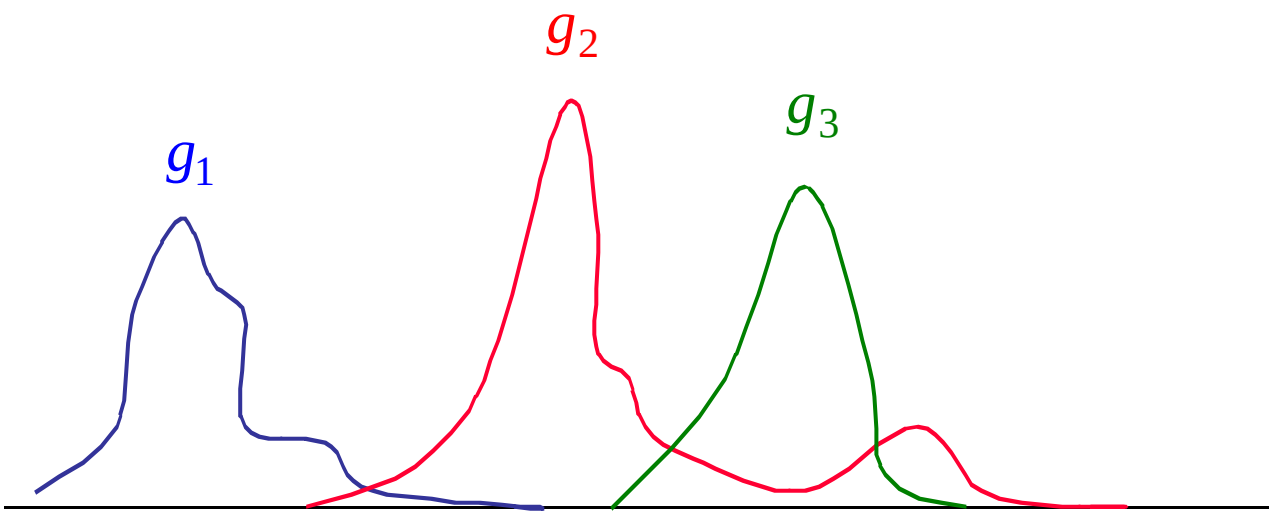
$$g_i(x) = -\frac{\|x-m_i\|^2}{2}$$

com ||.|| a denotar a norma euclideana.

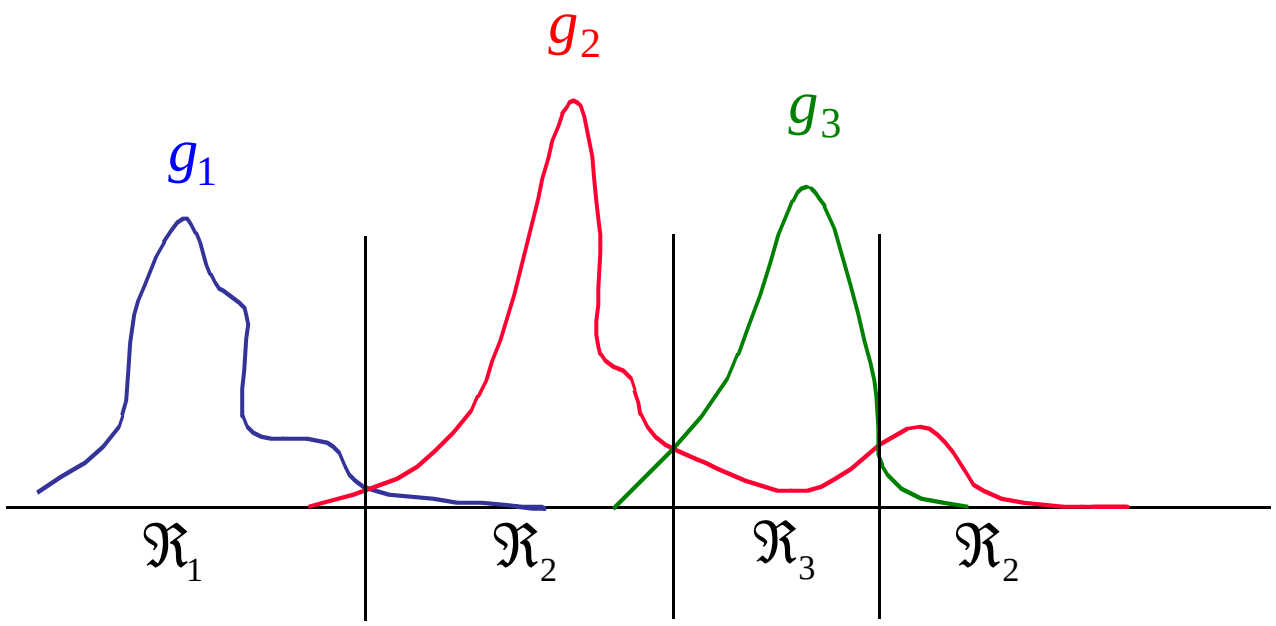
- **Este classificador atribui *x* à classe cujo vector da média esteja mais próximo - classificador de distância mínima.**
- **A superfície de separação entre as classes w_i e w_j é o hiperplano perpendicular à linha que une mi a mj, passando pelo ponto (mi + mj)/2**



Ex. Funções Discriminantes em $d=1$, $C=3$



Ex. Funções Discriminantes em $d=1$, $C=3$



- **De acordo com a definição de classificador baseado em funções discriminantes, <u>não existe um única sequência de funções discriminantes</u> para um dado classificador**

- **Por exemplo, se compusermos uma função**

$f : \mathbb{R} \rightarrow \mathbb{R}$ **monótona crescente**

com cada função discriminante, obtém-se uma nova sequência de funções discriminantes equivalentes

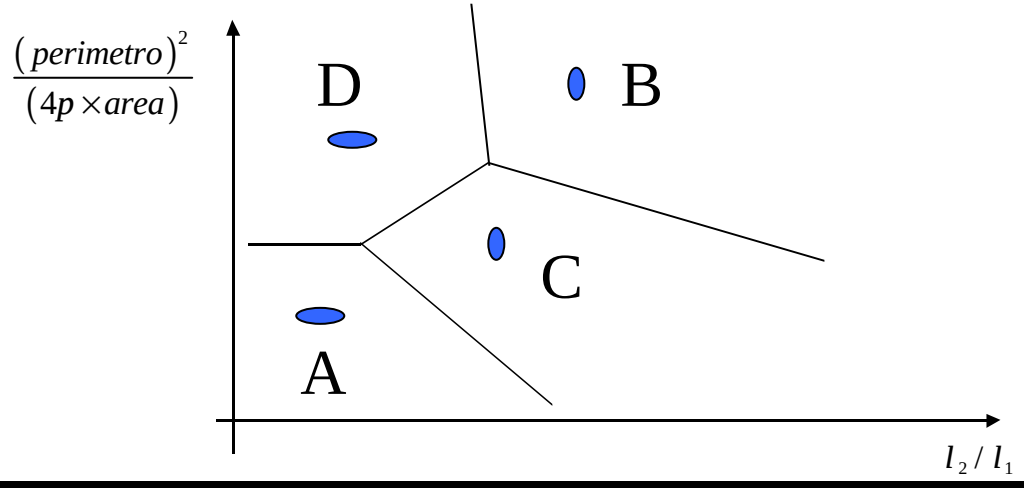
$$\overline{g}_i(x) = f(g_i(x))$$

- **Uma função que é frequentemente usada com este fim é o logaritmo natural**

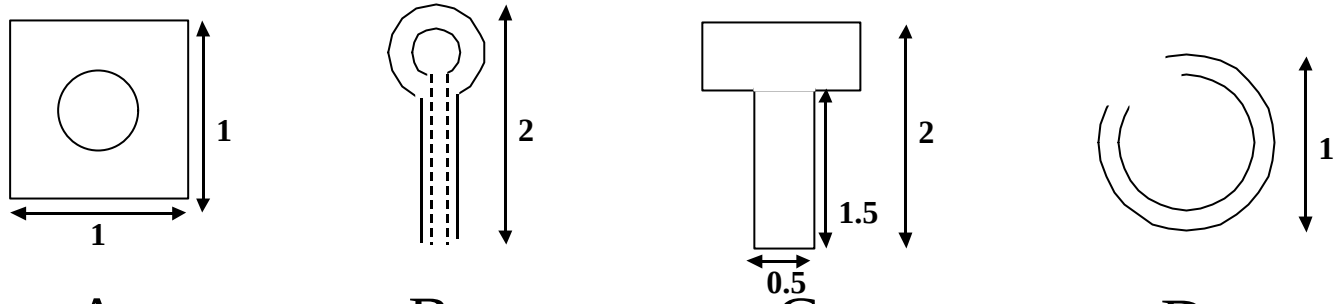
$$f(.) = \log(.)$$

Classificadores de Estadio único *vs* Classificadores Hierárquicos

- **<u>Classificadores de estádio único</u>:**
  - **A atribuição de uma classe a um padrão é feita num único passo**
  - **Desvantagens:**
    - **Para um elevado número de classes, a classificação num único passo requer um <u>número elevado de características</u> com o correspondente aumento do número de amostras de treino por forma a evitar o problema de "*curse of dimensionality*"**
    - **O conjunto de características global pode não ser o óptimo para pares específicos de classes**
    - **Elevado custo computacional**



Ex: Problema de classificação de objectos



- **0 – Obter as imagens dos objectos**
- **1 – Extracção dos contornos**
- **2 – Definição de características a partir dos contornos. Características típicas:**
  - Momentos
  - Descritores de Fourier
  - Morfologia
  - Medidas de compactacidade
- **Vamos definir:**
  1. $(perimetro)^2 / (4p \times area)$
  2. l_2 / l_1 com l_i o $i$ésimo valor próprio do conjunto de pontos 2D pertencentes ao objecto

1- mede o carácter circular do objecto (~1=> forma circular)

2- medida de elongação (valores elevados => o objecto possui uma dimensão que é muito menor do que a outra)

Classificadores de Estadio único *vs* Classificadores Hierárquicos

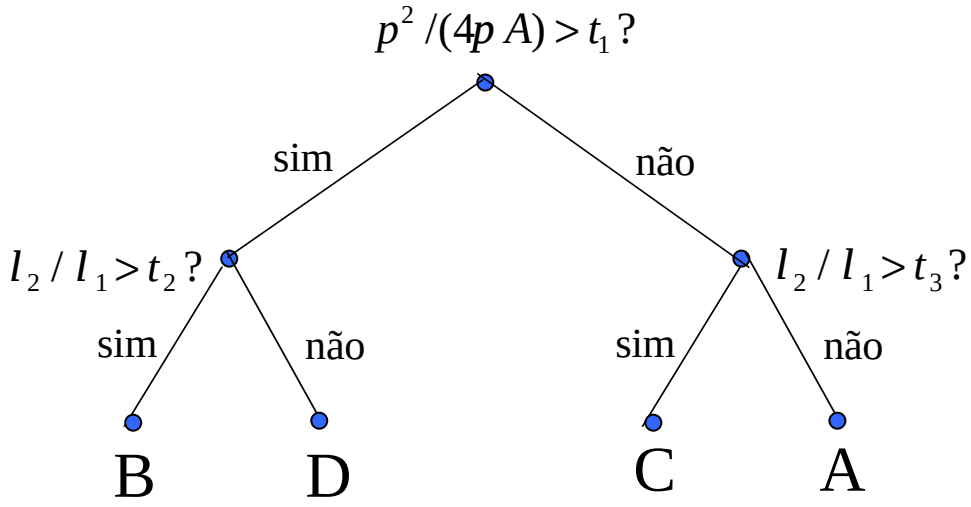
Classificadores hierárquicos ou em árvore:

- **O problema de classificação é partido em diversos problemas mais simples**
- **As discriminações mais óbvias são realizadas primeiro; as distinções mais subtis entre padrões são feitas em níveis posteriores**
- **Vantagens:**
  - **Maior rapidez de processamento**
  - **Redução do nº médio de características usadas em cada fase ou nó da árvore de decisão**
- **Dificuldades:**
  - **No projecto do classificador. Isto requer a definição de**
    - **Esqueleto da árvore de decisão ou uma ordenação hierárquica dos rótulos das classes**
    - **Quais as características a avaliar em cada nó não terminal**
    - **Regra de decisão em cada nó**

  Estes problemas são difíceis porque o nº de estruturas de árvores possíveis para cada classificação em C classes é muito elevado.

  Na prática são usadas heurísticas no projecto de um classificador hierárquico

Classificador Hierárquico



Abordagem Paramétrica *vs* Não Paramétrica

<u>Técnicas Paramétricas:</u> são usadas quando a forma das densidades condicionadas às classes são conhecidas ou seja razoável uma dada aproximação

Ex: Assume-se muitas vezes que estas densidades são Gaussianas multivariadas

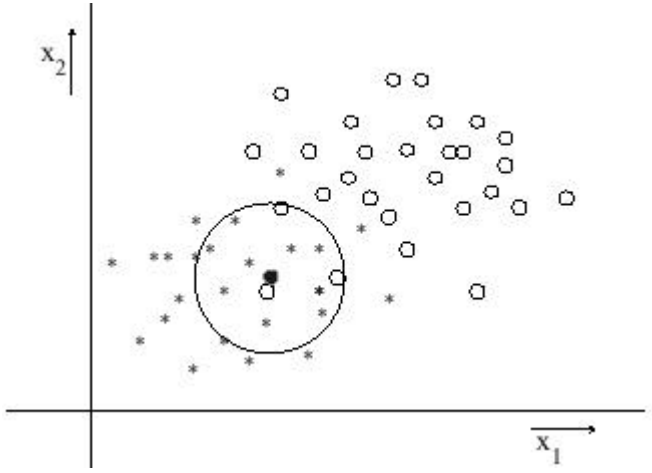
- **É difícil verificar se os dados multivariável têm uma distribuição Gaussiana**
- **No entanto esta hipótese conduz a superfícies de decisão com formas simples (linear ou quadrática)**
- **A hipótese de gaussianidade tem a vantagem adicional que a regra de decisão resultante é robusta, ie, se a hipótese for violada, a degradação do desempenho do classificador é gradual.**
- **Se os parâmetros das distribuições são conhecidos, usa-se a regra de decisão de Bayes óptima; caso contrário, estes parâmetros são estimados a partir dos dados de treino. Os parâmetros estimados substituem os valores exactos nas funções discriminantes.**

Abordagem Paramétrica *vs* Não Paramétrica

<u>Técnicas Não Paramétricas:</u> são usadas quando não existe qualquer base para assumir uma forma paramétrica para a função densidade de probabilidade

Estratégias para desenho de um classificador não paramétrico

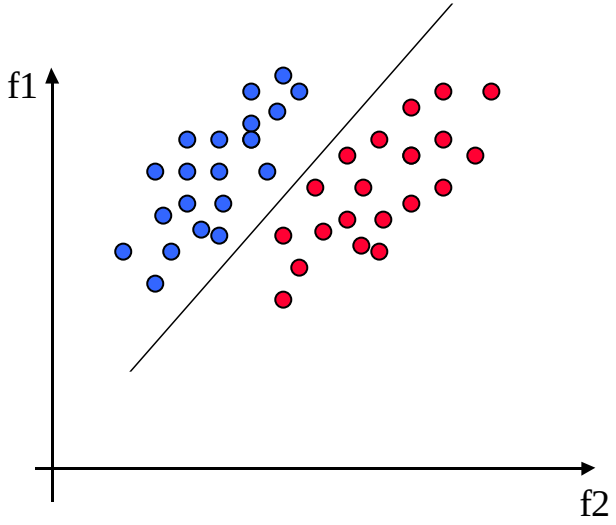
- **Estimar as distribuições *p(x|wi)* através de janelas de Parzen, usando estas estimativas numa regra de decisão de Bayes ou de Máxima Verosimilhança**
- **Evitar a estimação das densidades usando a regra de decisão k-NN (k- vizinhos mais próximos). Nesta abordagem procura-se, no conjunto de treino, os k-vizinhos mais próximos da amostra de teste; esta é classificada na classe maioritária dentre os k-vizinhos. O valor de k é dependente dos dados e do problema**



- **Embora a escolha entre as abordagens paramétrica ou não paramétrica dependa da credibilidade do modelo paramétrico, estudos recentes mostram que, se o número de amostras de treino é baixo, então as técnicas não paramétricas conduzem a melhores desempenhos do que as técnicas paramétricas, mesmo quando o modelo é correcto.**

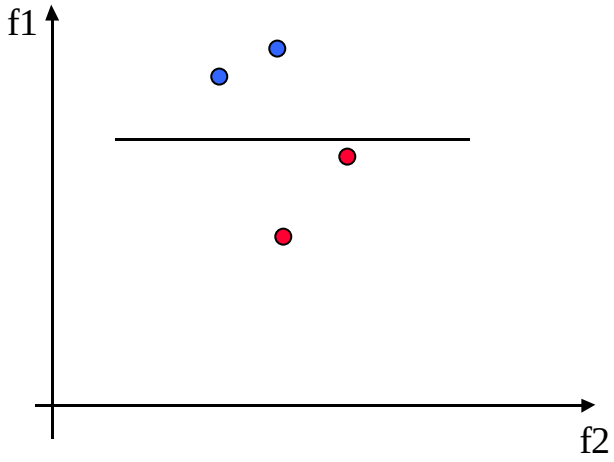
Relação entre Dimensionalidade e Conjuntos Amostra

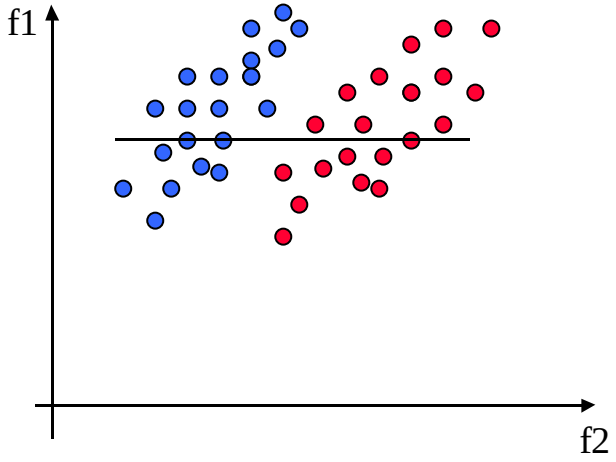
- **Questão: Quantas características devem ser usadas no classificador?**

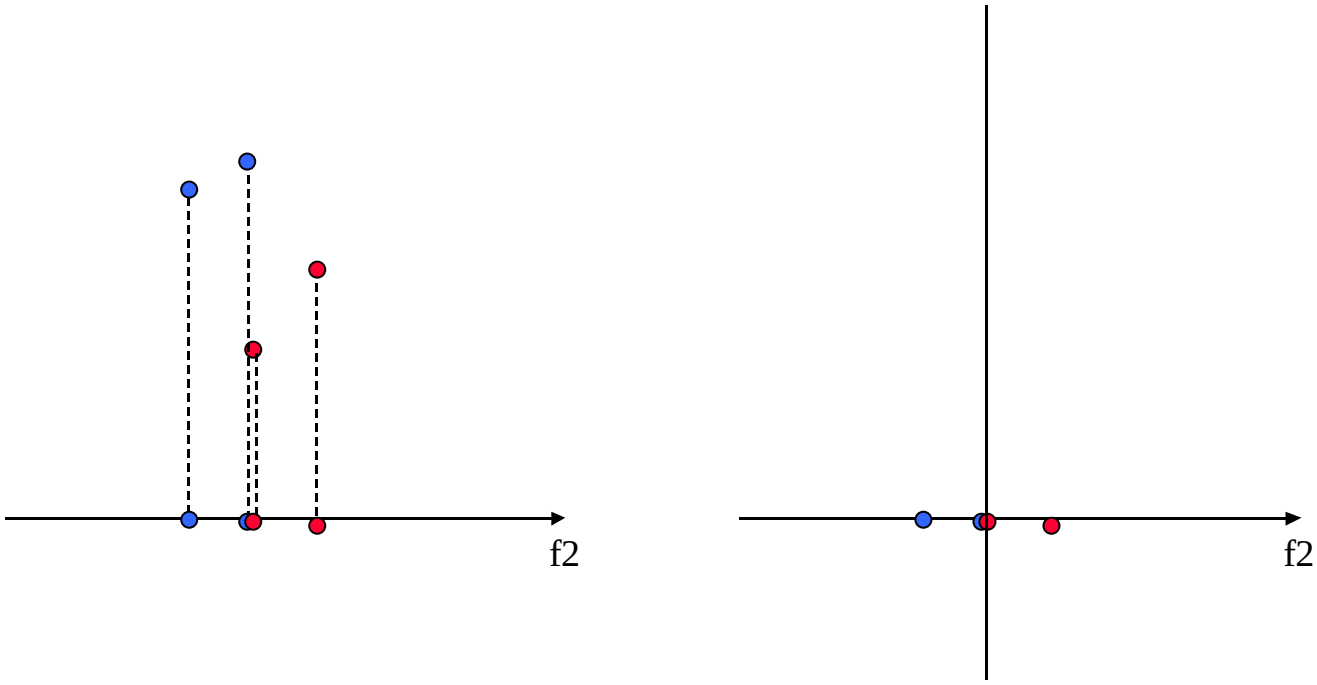


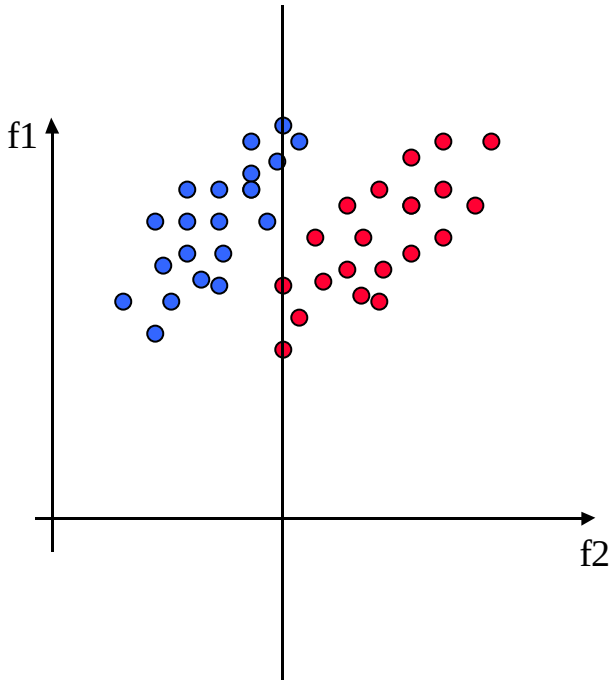
Relação entre Dimensionalidade e Conjuntos Amostra

- **Questão: Quantas características devem ser usadas no classificador?**



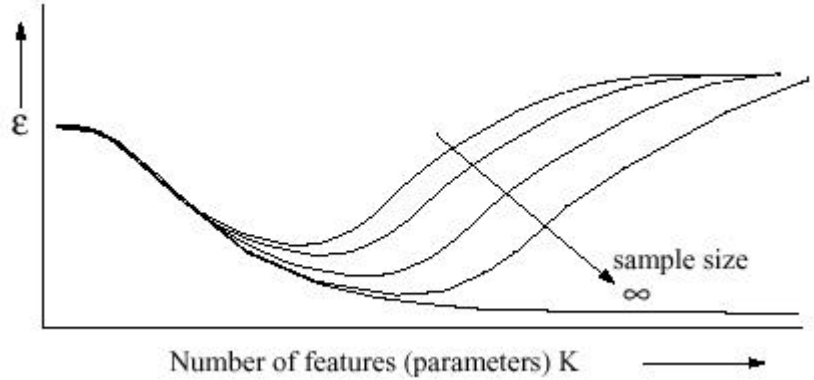

f1
f2


f2
f2


f1
f2

Relação entre Dimensionalidade e Conjuntos Amostra

- **Questão: Quantas características devem ser usadas no classificador?**
  - Ideia errada: quantas mais melhor
  - Prática: o desempenho começa por melhorar mas vai deteriorando à medida que mais características são consideradas



  - Os erros ocorrem devido ao uso "não óptimo" da informação adicional, que supera a vantagem da informação extra.

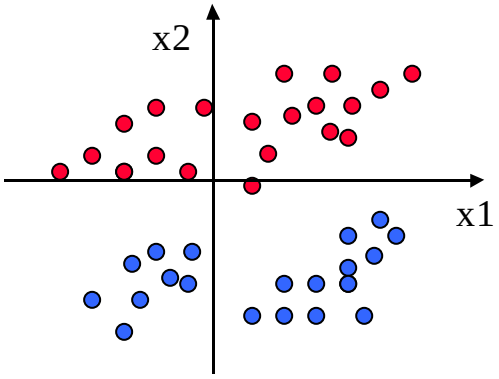
  - É portanto necessário limitar o nº de características para uma dada dimensão do conjunto de treino
  - Regra empírica: $\frac{\#caracteristicas}{\#amostras\_de\_treino}$ baixo $\boxed{\frac{\#N}{\#d} \approx 5-10}$

Selecção ou Extracção de Características

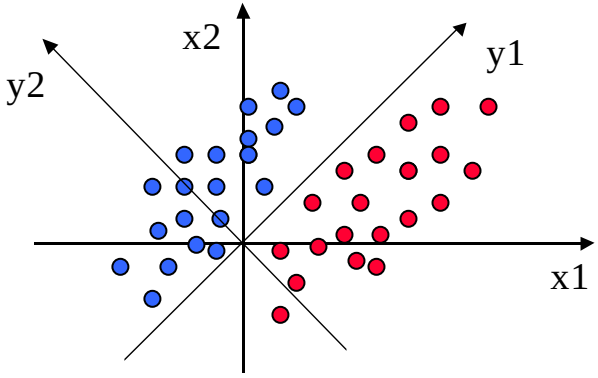
- **Problema: como representar um objecto ou padrão em termos de um conjunto reduzido de atributos**

1. **<u>Selecção</u> de características: processo de <u>escolha</u> de um sub-conjunto das características originais**
2. **<u>Extracção</u> de características: <u>definição</u> de <u>novas características</u> que podem ser função das características originais**

1 e 2 são muito importantes, influenciando o desempenho e a simplicidade do classificador



Selecção: escolhe-se x2 pois separa facilmente as classes



Extracção: é mais apropriada pois após rotação dos eixos de coordenadas, torna-se evidente que é apenas necessária umas das características

Selecção de Características

- **<u>Selecção</u> de características:**
  - **<u>Objectivo</u>: encontrar o melhor subconjunto de dimensão d das D características existentes ou potenciais.**
    - **O critério geralmente usado é a probabilidade de classificação errada**
  - **<u>Facto</u>: a melhor solução só pode ser encontrada através de uma procura exaustiva em todos os conjuntos possíveis de dimensão *d*:**
    - C_d^D **computacionalmente impraticável!**
    - **Uso de heurísticas em detrimento da optimalidade**
      - **Heuristica 1: (errada!) – escolher as *d* características que produzem individualmente melhores resultados**
        - **Mas: o melhor subconjunto pode não conter a melhor característica individual**
      - **Heuristica 2: técnica de selecção sequencial:**
        - **Suponhamos que seleccionamos k características. Então, a (k+1)ésima característica é aquela que, em combinação com as k existentes, proporciona o melhor desempenho**
    - **Outras soluções:**
      - **Algoritmos genéticos**
      - **Algoritmos de procura em grafos**

Extracção de Características

- **Extracção de características:**
  - **Objectivo: aplicar algum tipo de transformação sobre o conjunto original de características, de forma a que as classes estejam mais separadas no novo espaço**
    - **Vantagem adicional: o problema de selecção de características no novo espaço é mais simples**
  - **Técnicas mais usadas:**
    - **Transformações lineares, derivadas dos vectores próprios das matrizes de dispersão**
      - **Ex: Componentes principais ou expansão de Karhunen-Loeve –usa os vectores da matriz de covariância de todos os dados (matriz de dispersão total) pelo que não usa informação sobre as classes dos padrões**

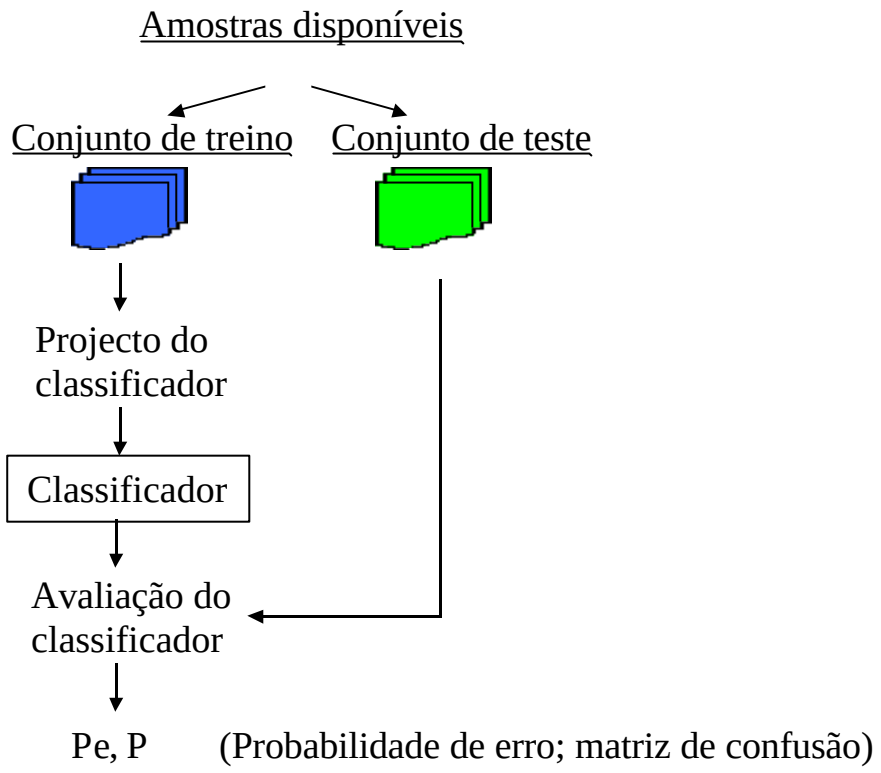
Estimação da Probabilidade de Erro

- **A probabilidade de classificação errada é útil para prever o desempenho do classificador em padrões futuros, comparar classificadores e como critério para selecção de características**

- **Na maioria das aplicações é muito difícil obter uma expressão analítica para a probabilidade de erro em função dos parâmetros do projecto (nº de características, nº de amostras de treino, f.d.p, etc)**

=>A probabilidade de erro é estimada experimentalmente

Estimação da Probabilidade de Erro

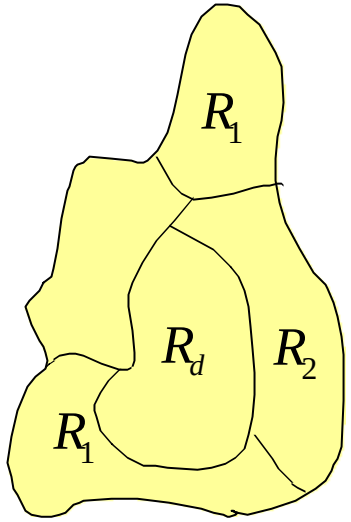


•Os conjuntos de treino e de teste devem ser estatisticamente independentes, ou pelo menos diferentes

Matriz de Confusão

- **P: matriz quadrada, de dimensão igual ao número de classes, em que Pij representa a probabilidade de um padrão gerado pela classe i ser classificado na classe j**

$$P_{ij} = P\{\hat{w} = w_j \mid w = w_i\}$$



– Se *x* é uma v.a discreta: $P_{ij} = \sum_{x \in \Re_j} P(x \mid w_i)$

– Se *x* é uma v.a contínua: $P_{ij} = \int_{\Re_j} p(x \mid w_i) dx$

$P(x \mid w_i)$ -função de probabilidade das observações geradas pela classe ωi

$p(x \mid w_i)$ -função densidade de probabilidade das observações geradas pela classe ωi

Matriz de Confusão

- **A matriz P é uma matriz estocástica pois verifica as condições:**

$$P_{ij} \geq 0 \quad i, j = 1, \ldots, C$$

$$\sum_{j=1}^{C} P_{ij} = 1$$

- **Classificador ideal ⇔ não há erros de classificação**
  - P = I (matriz identidade)

- A matriz P permite identificar os tipos de erros que ocorrem com maior probabilidade para um dado classificador

- Estimação da matriz de confusão a partir da classificação do conjunto de teste:

$$\hat{P}_{ij} = \frac{n_{ij}}{n_i}$$

$n_{ij} \equiv$ nº de padrões gerados pela classe *i* classificados na classe *j*

$n_i \equiv$ nº de padrões da classe *i* pertencentes ao conjunto de teste

Matriz de Confusão

- **Casos particulares de critérios de avaliação a partir da matriz de confusão P:**

$$P_{e_i} = \sum_{j \neq i} P_{ij} \quad \text{Prob. de erro da classe } w_i$$

$$P_e = \sum_i P_{e_i} P(w_i) = \sum_i \sum_{j \neq i} P_{ij} \; P(w_i) \quad \text{Prob. de erro}$$

- **Ex.: Sistema de detecção de defeitos em peças. Assume-se dois tipos de defeitos, havendo 3 classes para as peças: boa, defeituosa tipo A, defeituosa tipo B**

Peças detectadas

Peças produzidas	**Boa**	**A**	**B**
Boa	**.95**	**.04**	**.01**
A	**.01**	**.9**	**.09**
B	**.05**	**.02**	**.93**