Universidade de São Paulo
Instituto de Ciências Matemáticas e de Computação

Métodos de redução de dimensionalidade e seleção de atributos

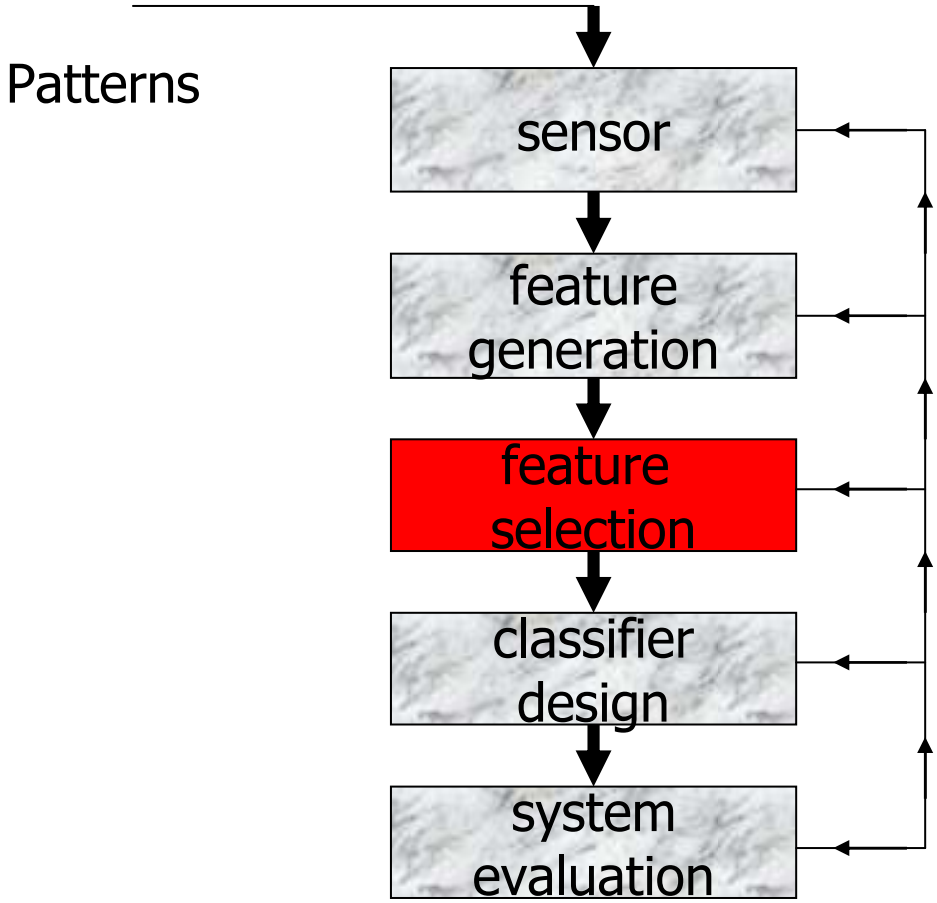
Francisco A. Rodrigues

Departamento de Matemática Aplicada e Estatística - SME

Tópicos

1. Seleção de atributos
2. Redução de dimensionalidade

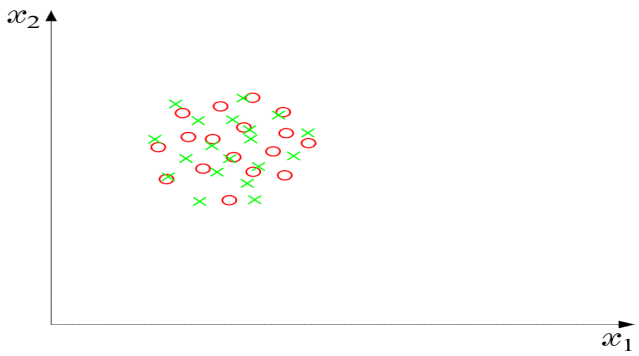
Seleção de atributos



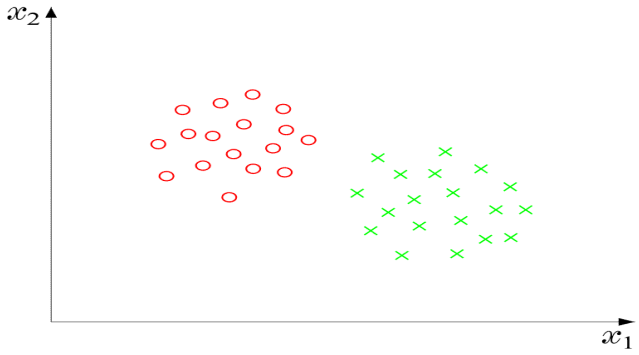
Motivação

- Correlação entre atributos: Apesar de dois atributos oferecerem boa classificação separados, há pouco ou nenhum ganho quando tratados em conjunto.
- Maldição da dimensionalidade

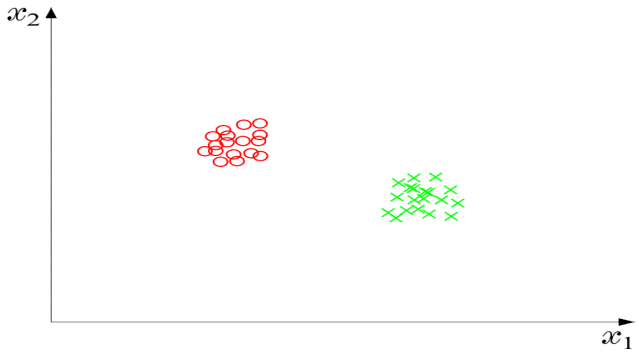
Motivação



Bad choice



Not bad choice



Good choice

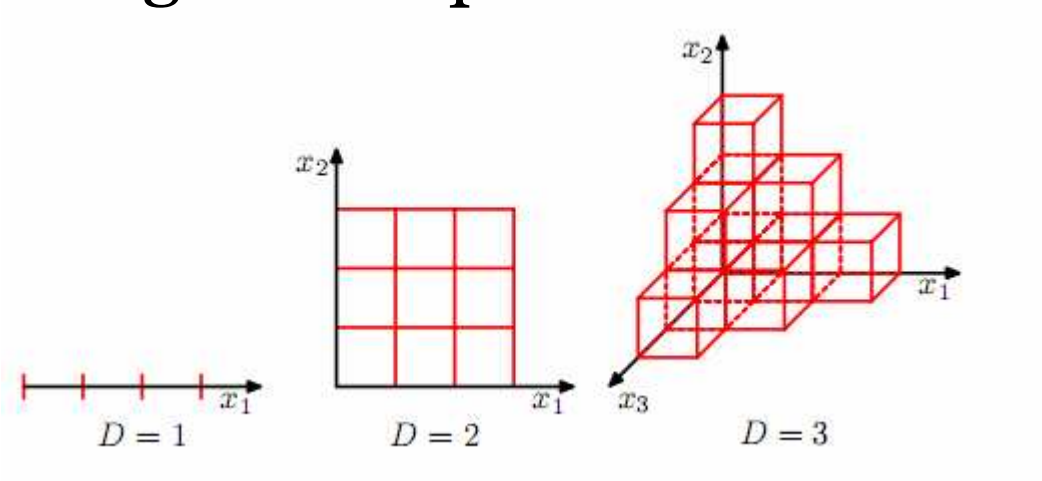
A maldição da dimensionalidade

- Termo introduzido por Bellman refere-se ao problema causado pelo aumento exponencial no volume associado com a adição de dimensões extras a um espaço matemático.

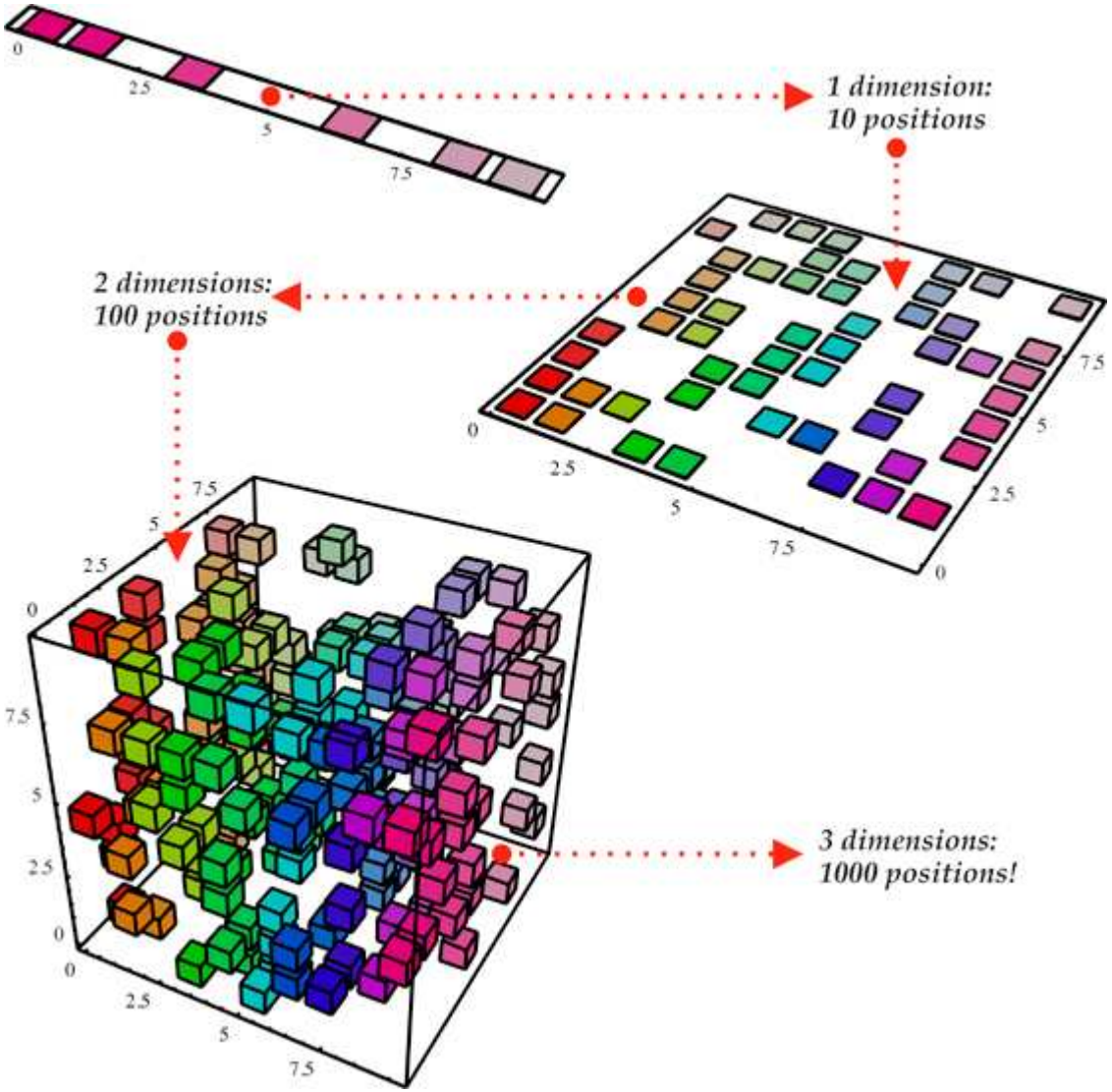
A maldição da dimensionalidade

• Se dividirmos uma região do espaço em células regulares, o número células cresce exponencialmente com a dimensão do espaço.

• Assim, o número de amostras deve crescer exponencialmente para garantir que nenhuma célula fique vazia.



A maldição da dimensionalidade



A maldição da dimensionalidade

• Na prática, a maldição da dimensionalidade implica que para um dados tamanho de amostras, existe um número máximo de características a partir do qual o desempenho do classificador irá degradar, ao invés de melhorar.

• **Solução**: Reduzir a dimensão através de seleção de características ou métodos de redução de dimensionalidade.

Seleção X projeção

- A seleção de atributos não implica na modificação dos atributos de um objeto.
- Já a projeção transforma os dados e conseqüentemente resulta em perda de informação.

Seleção de atributos

Objetivo

- Encontrar um número l de características de forma ótima.
- Selecionar as l "melhores" características.

Motivação

- A utilização de poucos atributos tem vantagens:

1. **Complexidade computacional**

2. Quanto maior a razão entre o número de padrões de treinamento N e o número de parâmetros do classificador, melhor a **generalização** das propriedades do respectivo classificador.

3. Um número grande características pode resultar em um **número maior de parâmetros** (e.g. pesos sinapticos em uma rede neural, pesos em um classificador linear, etc.).

Motivação

- A utilização de poucos atributos tem vantagens:

1. Quanto maior a razão entre o número de elementos no conjunto de treinamento e o número de características (dimensão), N/l, melhor a estimativa do erro na classificação.

2. Na prática, $l < N/3$ tem demonstrado ser uma boa escolha para muitos casos.

Problema:

- Dado um número de características, como podemos selecionar as mais importantes delas de tal modo a reduzir esse número e ao mesmo tempo reter o máximo possível de sua informação discriminatória?
- Solução: Selecionar as características que levam a uma alta distância entre classes e pequena distância entre os elementos da mesma classe.

Problema:

- Soluções possíveis:

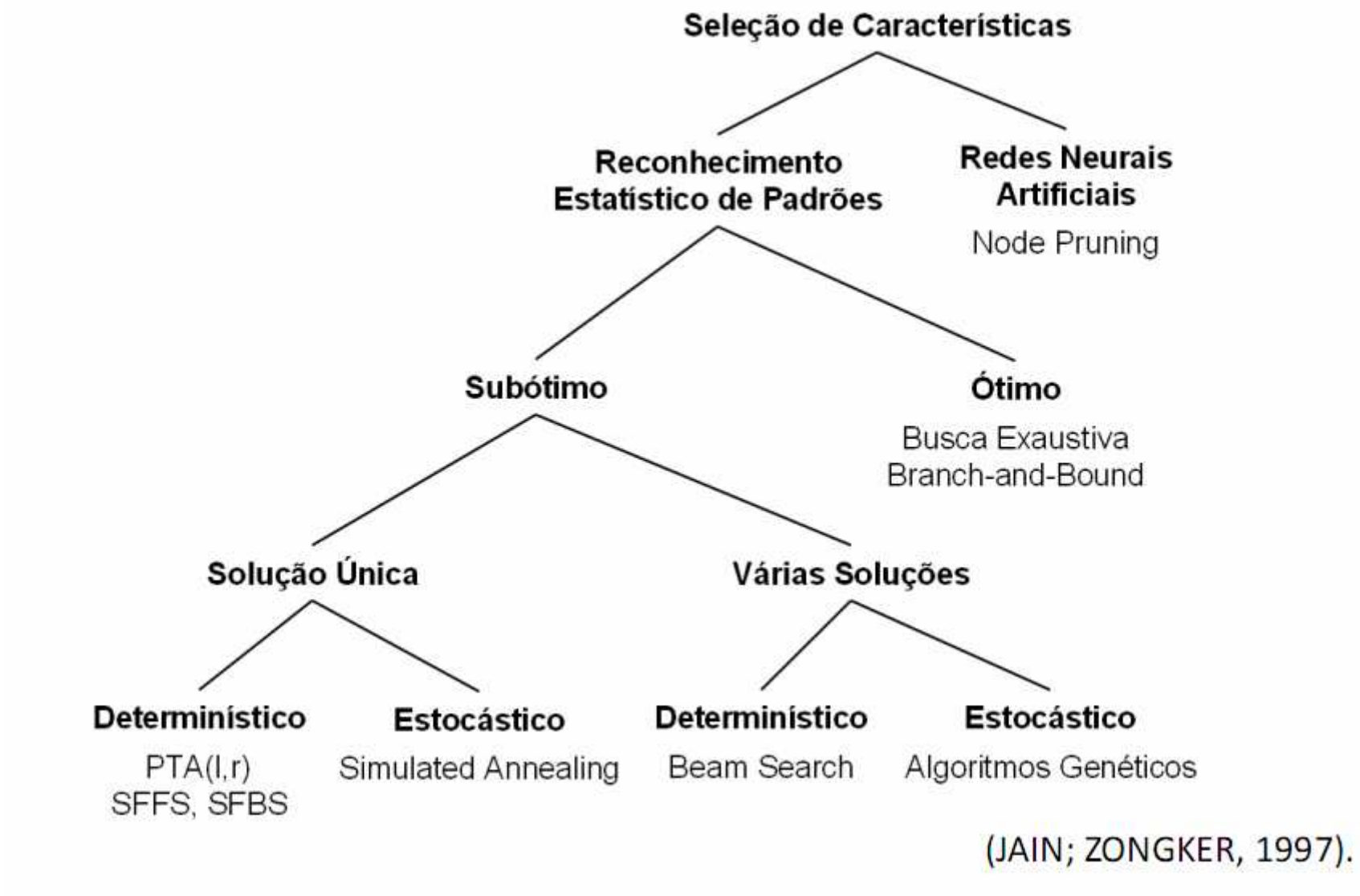
1. Examinar os atributos individualmente e descartar aqueles com pequena capacidade discriminatória.

2. Examinar os atributos em conjunto.

Seleção de características

1. Pré-processamento
2. Teste de hipóteses
3. Filtros
4. Wrappers
5. Medidas de separação

Abordagens



(JAIN; ZONGKER, 1997).

Pré-processamento

- **Remoção de outliers:** outliers são observações que ficam longe da média de uma dada variável aleatória.
- Se o número de outliers é pequeno, eles podem ser descartados.
- Técnicas para tratamento de outliers:

P.J.Huber, Robust Statistics, J. Wiley e Sons, 1981.

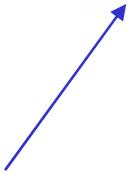
Pré-processamento

- **Normalização dos dados: Linear**

$$\bar{x}_k = \frac{1}{N}\sum_{i=1}^{N} x_{ik}, \quad k = 1, 2, \ldots, l$$

$$\sigma_k^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_{ik} - \bar{x}_k)^2$$

$$\hat{x}_{ik} = \frac{x_{ik} - \bar{x}_k}{\sigma_k}$$



Novos dados com média igual a 0 e variância igual a 1.

Pré-processamento

- **Normalização dos dados: Não-linear**

$$y = \frac{x_{ik} - \bar{x}_k}{r\sigma_k}, \quad \hat{x}_{ik} = \frac{1}{1 + \exp(-y)}$$

Softmax scaling
Limita os dados entre 0 e e1.

Seleção baseada em testes de hipóteses

- **Estratégia:**
- Descartar características com pouca discriminabilidade individual.
- Selecionar as características restantes e examiná-las conjuntamente como vetores.

Seleção baseada em testes de hipóteses

- **Objetivo**:
- Para cada característica individual, verificar se os valores que as características apresentam em diferentes classes **diferem significantemente**.
- Isto é:
- Os valores diferem significantemente: → $H_1: \theta_1 \neq \theta_0$
- Os valores não diferem significantemente: → $H_0: \theta_1 = \theta_0$
- Se eles não diferem significantemente, rejeite a característica nos estágios futuros.

Seleção baseada em testes de hipóteses

- **Teste de hipóteses: Os passos**
- N medidas são conhecidas: $x_i, i = 1,2,\ldots, N$
- Defina uma função dessas medidas:

$$q = f(x_1, x_2, \ldots, x_N):$$

- De tal modo que $p_q(q;\theta)$ possa ser facilmente parametrizada em termos de θ.
- Seja D um intervalo onde q tem uma alta probabilidade de cair sob a hipótese H_0, isto é, $p_q(q|\theta_0)$.
- Seja $\overline{D}$ o complementar de D. Então:

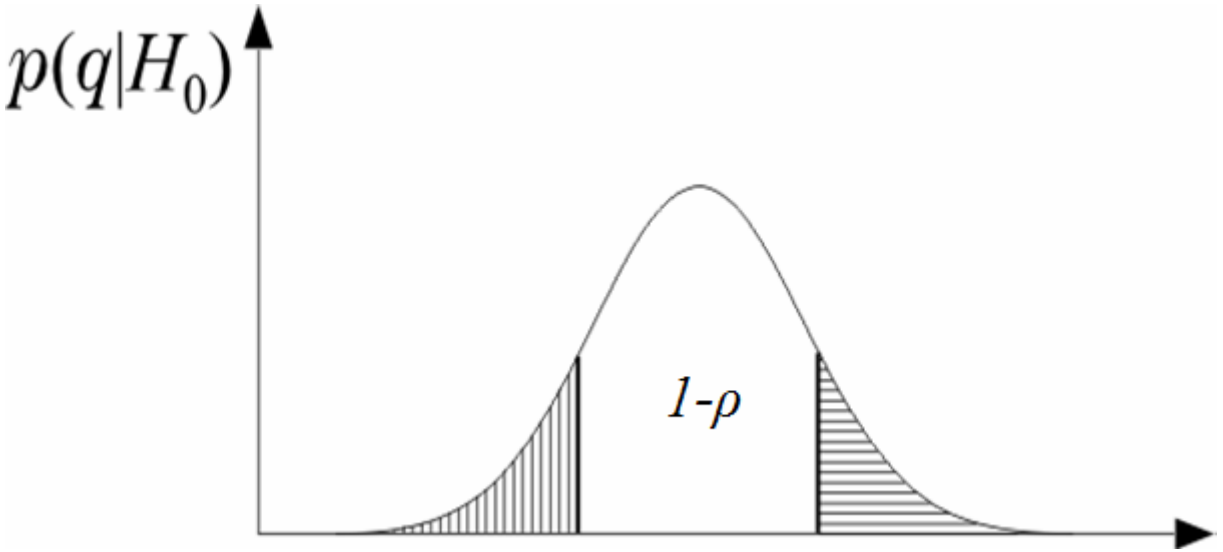
$D \longrightarrow$ Acceptance Interval
$\overline{D} \longrightarrow$ Critical Interval

- Se q, resultante de $x_1, x_2, \ldots, x_N$, cai no intervalo D, então aceitamos H_0, caso contrário, o rejeitamos.

Seleção baseada em testes de hipóteses

- Probabilidade de erro:

$$p_q(q \in \overline{D} | H_0) = \rho$$



ρ é pré-selecionado e é conhecido como nível de significância do teste.

Seleção baseada em testes de hipóteses

❖ Application: The known variance case:

➢ Let x be a random variable and the experimental samples, $x_i = 1,2,\ldots,N$, are assumed mutually independent. Also let

$$E[x] = \mu$$

$$E[(x-\mu)^2] = \sigma^2$$

➢ Compute the sample mean

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

➢ This is also a random variable with mean value

$$E[\bar{x}] = \frac{1}{N}\sum_{i=1}^{N} E[x_i] = \mu$$

That is, it is an Unbiased Estimator

Seleção baseada em testes de hipóteses

- The variance $\sigma_{\bar{x}}^2$

$$E[(\bar{x}-\mu)^2] = E[(\frac{1}{N}\sum_{i=1}^{N} x_i - \mu)^2]$$

$$= \frac{1}{N^2}\sum_{i=1}^{N} E[(x_i-\mu)^2] + \frac{1}{N^2}\sum_i\sum_j E[(x_i-\mu)(x_j-\mu)]$$

Due to independence

$$\sigma_{\bar{x}}^2 = \frac{1}{N}\sigma_x^2$$

That is, it is Asymptotically Efficient

- Hypothesis test

$$H_1 : E[x] \neq \hat{\mu}$$
$$H_0 : E[x] = \hat{\mu}$$

Seleção baseada em testes de hipóteses

- Test Statistic: Define the variable

$$q = \frac{\bar{x} - \hat{\mu}}{\sigma / \sqrt{N}}$$

- Central limit theorem under H_0

$$p_{\bar{x}}(\bar{x}) = \frac{\sqrt{N}}{\sqrt{2\pi}\sigma} \exp\left(- \frac{N(\bar{x} - \hat{\mu})^2}{2\sigma^2} \right)$$

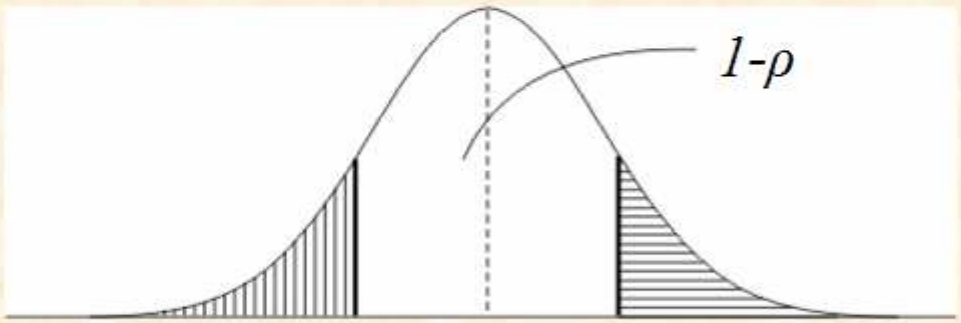
- Thus, under H_0

$$p_q(q) = \frac{1}{\sqrt{2\pi}} \exp\left(- \frac{q^2}{2} \right) \quad q \approx N(0,1)$$

Seleção baseada em testes de hipóteses

➢ The decision steps

- Compute q from x_i, $i=1,2,\ldots,N$
- Choose significance level ρ
- Compute from $N(0,1)$ tables $D=[-x_\rho, x_\rho]$



- if $q \in D$ accept H_0

  if $q \in \overline{D}$ reject H_0

Seleção baseada em testes de hipóteses

- An example: A random variable x has variance $\sigma^2=(0.23)^2$. $N=16$ measurements are obtained giving $\bar{x}=1.35$. The significance level is $\rho=0.05$.

Test the hypothesis

$H_0 : \mu = \hat{\mu} = 1.4$

$H_1 : \mu \neq \hat{\mu}$

- Since σ^2 is known, $q = \dfrac{\bar{x} - \hat{\mu}}{\sigma / 4}$ is $N(0,1)$.

From tables, we obtain the values with acceptance intervals $[-x_\rho, x_\rho]$ for normal $N(0,1)$

$1-\rho$	0.8	0.85	0.9	0.95	0.98	0.99	0.998	0.999
x_ρ	1.28	1.44	1.64	1.96	2.32	2.57	3.09	3.29

Seleção baseada em testes de hipóteses

- Thus

$$\text{Prob}\left\{-1.967 < \frac{\bar{x} - \hat{\mu}}{0.23/4} < 1.967\right\} = 0.95$$

or

$$\text{Prob}\left\{-0.113 < \bar{x} - \hat{\mu} < 0.113\right\} = 0.95$$

or

$$\boxed{\text{Prob}\left\{1.237 < \hat{\mu} < 1.463\right\} = 0.95}$$

- Since $\hat{\mu} = 1.4$ lies within the above acceptance interval, we accept H_0, i.e.,

$$\mu = \hat{\mu} = 1.4$$

The interval [1.237, 1.463] is also known as confidence interval at the $1\text{-}\rho = 0.95$ level.

We say that: There is no evidence at the 5% level that the mean value is not equal to $\hat{\mu}$

Seleção baseada em testes de hipóteses

- The Unknown Variance Case
  - Estimate the variance. The estimate

$$\hat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

is unbiased, i.e.,

$$E[\hat{\sigma}^2] = \sigma^2$$

  - Define the test statistic

$$q = \frac{\bar{x} - \mu}{\hat{\sigma}/\sqrt{N}}$$

Seleção baseada em testes de hipóteses

- This is no longer Gaussian. If x is Gaussian, then q follows a t-distribution, with N-1 degrees of freedom

- An example:

x is Gaussian, $N = 16$, obtained from measurements, $\bar{x} = 1.35$ and $\hat{\sigma}^2 = (0.23)^2$. Test the hypothesis

$H_0 : \mu = \hat{\mu} = 1.4$

at the significance level $\rho = 0.025$.

Seleção baseada em testes de hipóteses

- Table of acceptance intervals for t-distribution

Degrees of Freedom	1-ρ	0.9	0.95	0.975	0.99
12		1.78	2.18	2.56	3.05
13		1.77	2.16	2.53	3.01
14		1.76	2.15	2.51	2.98
15		1.75	2.13	2.49	2.95
16		1.75	2.12	2.47	2.92
17		1.74	2.11	2.46	2.90
18		1.73	2.10	2.44	2.88

- $$\text{Prob}\left\{-2.49 < \frac{\bar{x} - \hat{\mu}}{\hat{\sigma}/4} < 2.49\right\}$$

$$1.207 < \hat{\mu} < 1.493$$

Thus, $\hat{\mu} = 1.4$ is accepted

Seleção baseada em testes de hipóteses

- **Seleção de atributos:**

  - The goal here is to test against zero the difference μ_1-μ_2 of the respective means in ω_1, ω_2 of a single feature.
  - Let x_i i=1,…,N , the values of a feature in ω_1
  - Let y_i i=1,…,N , the values of the same feature in ω_2
  - Assume in both classes $\sigma_1^2 = \sigma_2^2 = \sigma^2$ (unknown or not)
  - The test becomes

$$H_0 : \Delta\mu = \mu_1 - \mu_2 = 0$$

$$H_1 : \Delta\mu \neq 0$$

Seleção baseada em testes de hipóteses

- Define

  $z=x-y$

- Obviously

  $E[z]=\mu_1-\mu_2$

- Define the average

$$\bar{z}=\frac{1}{N}\sum_{i=1}^{N}(x_i-y_i)=\bar{x}-\bar{y}$$

- Known Variance Case: Define

$$q=\frac{(\bar{x}-\bar{y})-(\hat{\mu}_1-\hat{\mu}_2)}{\sigma\sqrt{\frac{2}{N}}}$$

- This is $N(0,1)$ and one follows the procedure as before.

Seleção baseada em testes de hipóteses

- Unknown Variance Case:
  Define the test statistic

$$q = \frac{(\bar{x} - \bar{y}) - (\mu_1 - \mu_2)}{S_z \sqrt{\frac{2}{N}}}$$

$$S_z^2 = \frac{1}{2N-2}\left(\sum_{i=1}^{N}(x_i - \bar{x})^2 + \sum_{i=1}^{N}(y_i - \bar{y})^2\right)$$

  - q is t-distribution with $2N$-2 degrees of freedom,
  - Then apply appropriate tables as before.

- Example: The values of a feature in two classes are:

ω_1: 3.5, 3.7, 3.9, 4.1, 3.4, 3.5, 4.1, 3.8, 3.6, 3.7

ω_2: 3.2, 3.6, 3.1, 3.4, 3.0, 3.4, 2.8, 3.1, 3.3, 3.6

Test if the mean values in the two classes differ significantly, at the significance level ρ=0.05

Seleção baseada em testes de hipóteses

- We have

$$\omega_1: \ \bar{x} = 3.73, \ \hat{\sigma}_1^2 = 0.0601$$
$$\omega_2: \ \bar{y} = 3.25, \ \hat{\sigma}_2^2 = 0.0672$$

For N=10

$$S_z^2 = \frac{1}{2}(\hat{\sigma}_1^2 + \hat{\sigma}_2^2)$$

$$q = \frac{(\bar{x} - \bar{y}) - 0}{S_z \sqrt{\frac{2}{10}}}$$

$$\boxed{q = 4.25}$$

- From the table of the t-distribution with $2N$-2=18 degrees of freedom and ρ=0.05, we obtain D=[-2.10,2.10] and since q=4.25 is outside D, H_1 is accepted and the feature is selected.

Estratégias

- A ênfase até agora foi em componentes individuais. No entanto, tal metodologia não leva em conta a correlação que pode ocorrer entre as características.
- Se duas características fornecem boa discriminabilidade, mas são altamente correlacionadas, nós não devemos usá-las em conjunto.
- Para evitar tal problema, agrupamos as características em vetores.

Estratégias

- Assim:
- Descartamos as características que apresentam baixa discriminabilidade através do teste de hipóteses.
- Escolhemos um número máximo l de características a serem usadas. Essa escolha pode ser ditada pelo problema (e.g. número N de padrões de treinamento disponíveis e o tipo do classificador adotado).

Estratégias

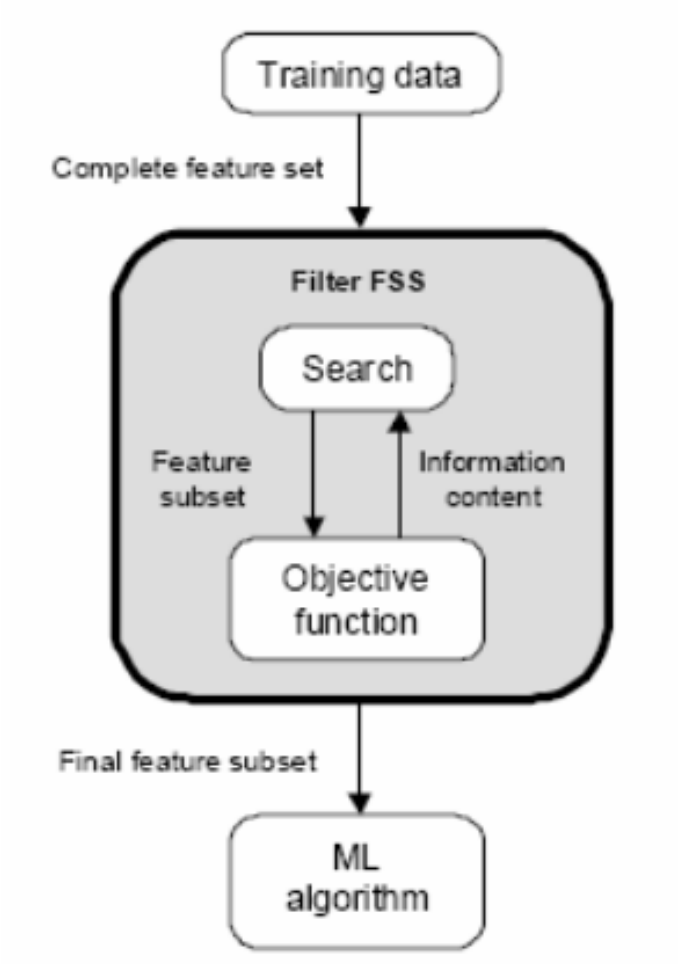
- Combinamos as medidas para selecionar a “melhor” combinação. Para este fim:

1. Use diferentes combinações de características para formar o vetor. Treine o classificador e escolha a combinação que resulta na melhor classificação. Uma desvantagem dessa abordagem é a alta complexidade. Além disso, mínimos locais podem ser obtidos e levar a resultados enganosos.
2. Adote uma medida de separabilidade e escolha a combinação de características que forneça melhor custo.

Funções objetivo

- **Filtros:**
- a função avalia o subconjunto de características a partir da informação contida no conjunto, tipicamente distância interclasse, dependência estatística, etc...

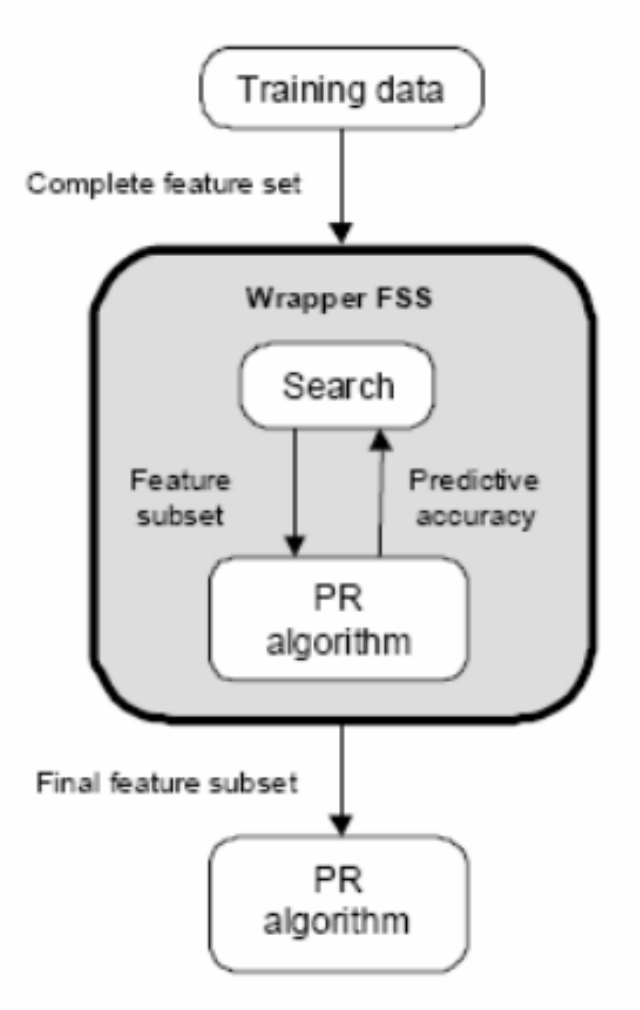
Filtros



Funções objetivo

- **Wrappers:**
- a função objetivo é um classificador de padrões, que avalia o subconjunto a partir de sua capacidade preditiva (taxa de reconhecimento no conjunto de teste) por meio de reamostragem estatística ou cross-validação.

Wrappers



Estratégias

- **Filtros**
- **Vantagens:**
- Execução mais rápida.
- Generalidade: por avaliarem propriedades intrínsecas dos dados, ao contrário de interações específicas de um classificador, seus resultados exibem maior generalidade: a solução tende a ser “boa” para uma família de classificadores.
- **Desvantagens:**
- Tendência em selecionar um subconjunto muito grande: uma vez que são geralmente monotônicas, os filtros tendem a selecionar o conjunto original como a melhor solução. Isso força o usuário a estabelecer um critério de corte no número de características

Estratégias

- **Wrappers**
- **Vantagens**
- Precisão: normalmente provêem melhor taxa de reconhecimento que filtros, uma vez que são projetadas especificamente para um determinado tipo de classificador e conjunto de teste.
- Habilidade de generalização: possuem mecanismo para evitar overfitting.
- Normalmente usam medidas de validação cruzada.
- **Desvantagens:**
- Execução lenta: wrappers tem que treinar um classificador para cada subconjuto.
- Solução muito associada ao classificador utilizado na avaliação.

Filtros: Medidas de separação

- Seja **x** o vetor da combinação de características corrente.
- Vamos considerar o caso de duas classes. Sejam:

$$D_{12} = \int_{-\infty}^{+\infty} p(\underline{x} \mid \omega_1) \ln \frac{p(\underline{x} \mid \omega_1)}{p(\underline{x} \mid \omega_2)} d\underline{x}$$

$$D_{21} = \int_{-\infty}^{+\infty} p(\underline{x} \mid \omega_2) \ln \frac{p(\underline{x} \mid \omega_2)}{p(\underline{x} \mid \omega_1)} d\underline{x}$$

$$d_{12} = D_{12} + D_{21}$$

- Essa medida é conhecida como **divergência** e pode ser usada como uma medida de separabilidade.

Filtros: Medidas de separação

- Para o caso de muitas classes, defina d_{ij} para cada par de classes e a divergência média é definida por:

$$d = \sum_{i=1}^{M} \sum_{j=1}^{M} P(\omega_i) P(\omega_j) d_{ij}$$

- Algumas propriedades:

$$d_{ij} \geq 0$$

$$d_{ij} = 0, \text{ if } i = j$$

$$d_{ij} = d_{ji}$$

- Quanto maior o valor de d, melhor a combinação das características.
- Se as variâncias das classes são iguais:
- dist. Mahalanobis: $d_{ij} = (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T \Sigma^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$

Filtros: Medidas de separação

- Distância de Bhattacharyya:

$$B_{ij} = \frac{1}{8} \cdot (\mathbf{M}_j - \mathbf{M}_i)^T \cdot \left(\frac{\mathbf{C}_i + \mathbf{C}_j}{2} \right)^{-1} \cdot (\mathbf{M}_j - \mathbf{M}_i) + \frac{1}{2} \ln \left(\frac{\left| \frac{\mathbf{C}_i + \mathbf{C}_j}{2} \right|}{\sqrt{|\mathbf{C}_j||\mathbf{C}_i|}} \right),$$

sendo que

$\mathbf{M}_i$ e $\mathbf{M}_j$ são os vetores de médias das classes i e j, respectivamente,

$\mathbf{C}_i$ e $\mathbf{C}_j$ são as matrizes de covariância das classes i e j, respectivamente,

$|\cdot|$ representa o determinante da matriz.

Filtros: Medidas de separação

- Distância de Jeffries-Matusita

$$J_{ij} = 2\left(1 - \exp\left(-B_{ij}\right)\right).$$

- O valor varia de 0 a 2, sendo 2 quando há uma completa separação entre duas classes.

Medidas de separação: Scatter matrix

- Scatter Matrices. These are used as a measure of the way data are scattered in the respective feature space.
  - Within-class scatter matrix

$$S_w = \sum_{i=1}^{M} P_i S_i$$

where

$$S_i = E\left[(\underline{x} - \underline{\mu}_i)(\underline{x} - \underline{\mu}_i)^T\right]$$

and

$$P_i \equiv P(\omega_i) \approx \frac{n_i}{N}$$

n_i the number of training samples in ω_i.

Trace $\{S_w\}$ is a measure of the average variance of the features.

Medidas de separação: Scatter matrix

- Between-class scatter matrix

$$S_b = \sum_{i=1}^{M} P_i \left(\underline{\mu}_i - \underline{\mu}_0\right)\left(\underline{\mu}_i - \underline{\mu}_0\right)^T$$

$$\underline{\mu}_0 = \sum_{i=1}^{M} P_i \underline{\mu}_i$$

Trace $\{S_b\}$ is a measure of the average distance of the mean of each class from the respective global one.

- Mixture scatter matrix

$$S_m = E\left[\left(\underline{x} - \underline{\mu}_0\right)\left(\underline{x} - \underline{\mu}_0\right)^T\right]$$

It turns out that:

$$S_m = S_w + S_b$$

Medidas de separação: Scatter matrix

➢ Measures based on Scatter Matrices.

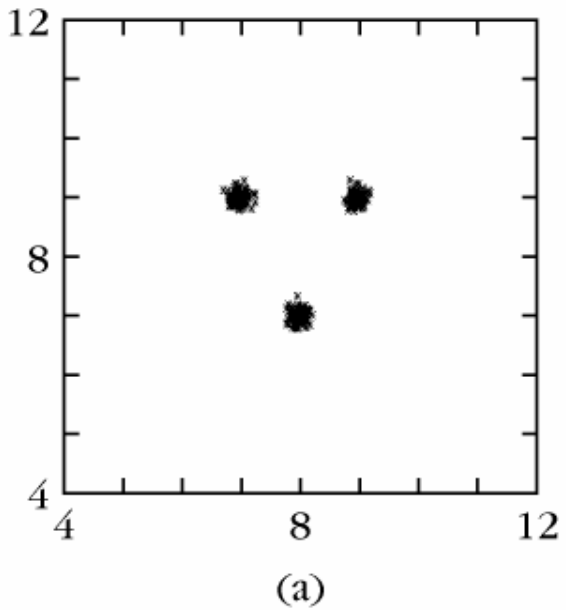
- $J_1 = \dfrac{\text{Trace}\{S_m\}}{\text{Trace}\{S_w\}}$
- $J_2 = \dfrac{|S_m|}{|S_w|} = |S_w^{-1} S_m|$
- $J_3 = \text{Trace}\{S_w^{-1} S_m\}$
- Other criteria are also possible, by using various combinations of S_m, S_b, S_w.

The above J_1, J_2, J_3 criteria take high values for the cases where:

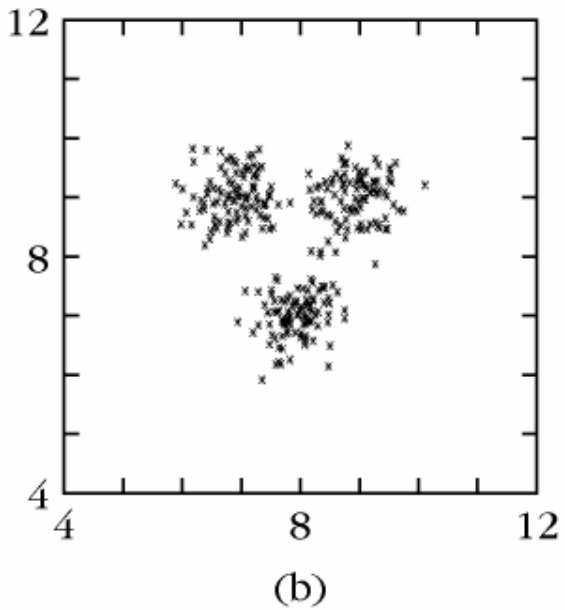
- Data are clustered together within each class.
- The means of the various classes are far.

Medidas de separação: Scatter matrix

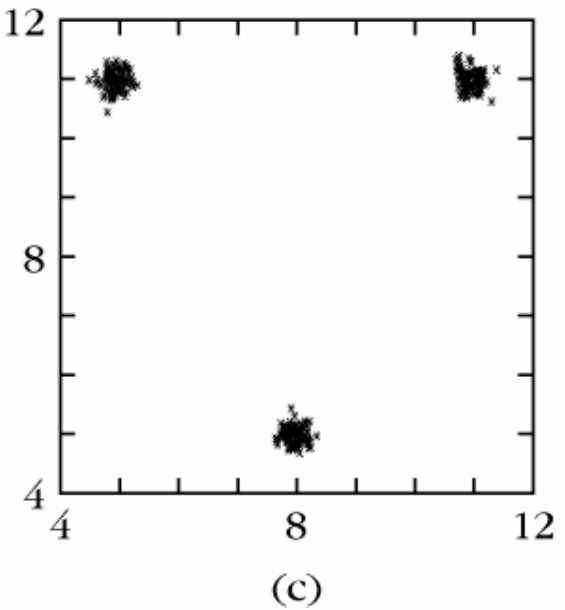
J3 = 164,7



J3 = 12,5



J3 = 620,9



Filtros: Estratégias de busca

- **Seqüencial**
- Estes algoritmos adicionam ou removem características seqüencialmente, e apresentam tendência em se prender a mínimos locais:

1. Sequential forward selection
2. Sequential backward selection
3. Sequential floating selection

Filtros: Estratégias de busca

- **Exponencial**
- Estes algoritmos avaliam um número de subconjuntos que crescem exponencialmente com a dimensão do espaço de busca:

1. Busca exaustiva
2. Branch and Bound
3. Beam search

Filtros: Estratégias de busca

- **Aleatório**
- Estes algoritmos incorporam aleatoriedade em seus procedimentos de busca a fim de escapar de mínimos locais:

1. Simulated Annealing
2. Algoritmos genéticos

Busca seqüencial

- Sequential forward selection. Let x_1, x_2, x_3, x_4 the available features (m=4). The procedure consists of the following steps:
  - Adopt a class separability criterion (could also be the error rate of the respective classifier). Compute its value for ALL features considered jointly $[x_1, x_2, x_3, x_4]^T$.
  - Eliminate one feature and for each of the possible resulting combinations, that is $[x_1, x_2, x_3]^T$, $[x_1, x_2, x_4]^T$, $[x_1, x_3, x_4]^T$, $[x_2, x_3, x_4]^T$, compute the class reparability criterion value C. Select the best combination, say $[x_1, x_2, x_3]^T$.

Busca seqüencial

- From the above selected feature vector eliminate one feature and for each of the resulting combinations, $[x_1, x_2]^T$, $[x_2, x_3]^T$, $[x_1, x_3]^T$ compute C and select the best combination.

The above selection procedure shows how one can start from m features and end up with the "best" ℓ ones. Obviously, the choice is suboptimal. The number of required calculations is:

$$1+\frac{1}{2}\left((m+1)m-\ell(\ell+1)\right)$$

In contrast, a full search requires:

$$\binom{m}{\ell}=\frac{m!}{\ell!(m-\ell)!}$$

operations.

Busca seqüencial

- Sequential backward selection. Here the reverse procedure is followed.
  - Compute C for each feature. Select the "best" one, say x_1
  - For all possible 2D combinations of x_1, i.e., $[x_1, x_2]$, $[x_1, x_3]$, $[x_1, x_4]$ compute C and choose the best, say $[x_1, x_3]$.
  - For all possible 3D combinations of $[x_1, x_3]$, e.g., $[x_1, x_3, x_2]$, etc., compute C and choose the best one.

The above procedure is repeated till the "best" vector with ℓ features has been formed. This is also a suboptimal technique, requiring:

$$\ell m - \frac{\ell(\ell-1)}{2}$$

operations.

Brach and bound

- O algoritmo começa pelo conjunto inteiro e remove características usando uma estratégia de busca em profundidade.
- Nós cuja função objetivo estão abaixo do atual melhor não são explorados, uma vez que o critério de monotonicidade garante que seus filhos não conterão solução melhor.

Brach and bound

- Quando a função critério é monotônica, é possivel usar o método Branch and bound.
- O algoritmo branch and bound garante achar uma solução ótima sob o critério de monotonicidade
- O critério de monotonicidade garante que a adição de novas características sempre aumenta o valor da função objetivo, ou seja

$$J(x_{i_1}) < J(x_{i_1}, x_{i_2}) < J(x_{i_1}, x_{i_2}, x_{i_3}) < \cdots < J(x_{i_1}, x_{i_2}, \cdots, x_{i_N})$$

Brach and bound

$D = 6,\ d = 2$ e $Y = \{1, 2, 3, 4, 5, 6\}$

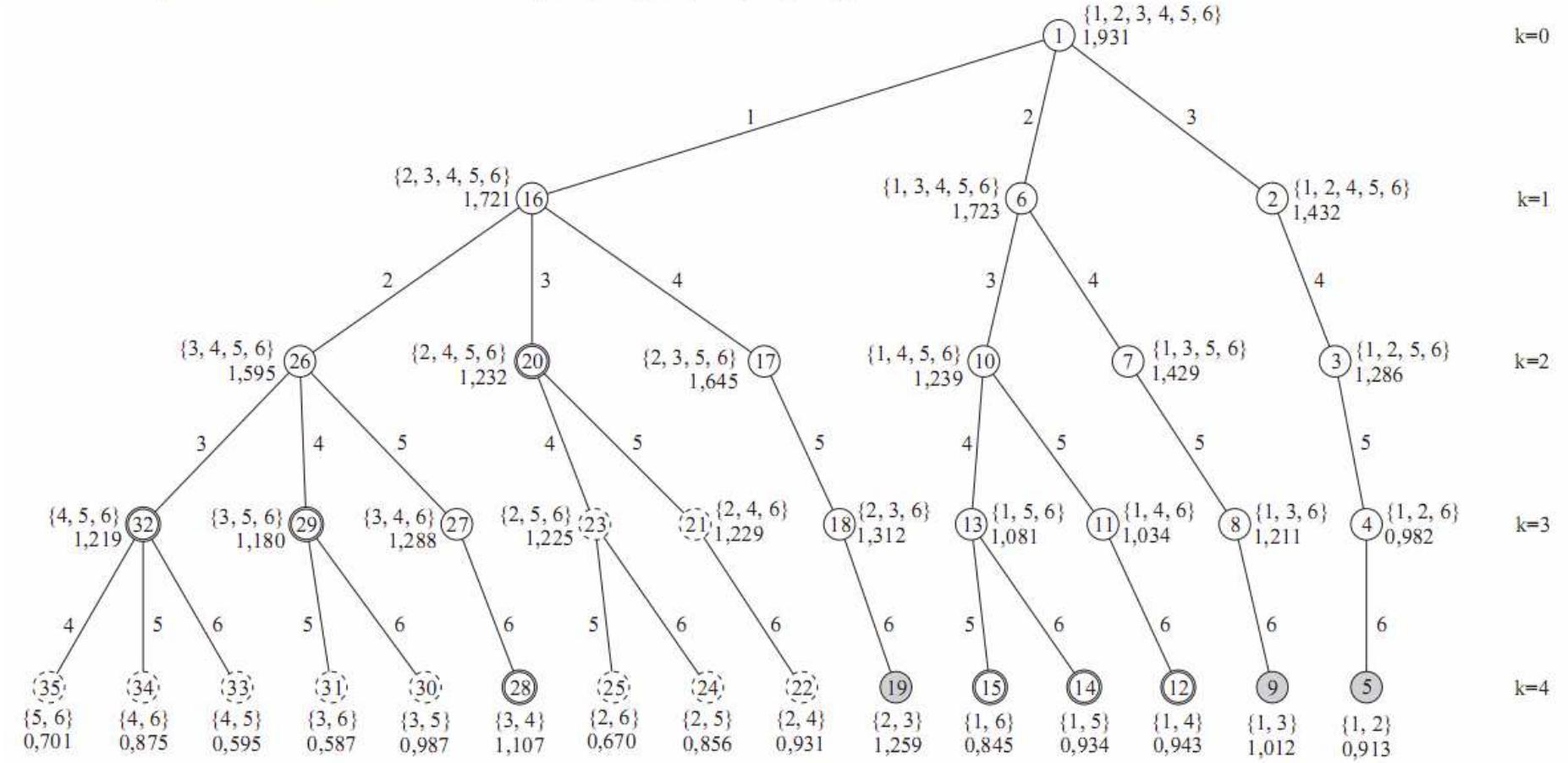


Figura 3.2: Árvore de busca do *branch and bound* básico para $D = 6$ e $d = 2$. A numeração interna dos nós indica o caminho em que o percurso é realizado. O subconjunto de características e o valor de $J(\cdot)$ estão indicados próximo do nó correspondente. O rótulo de cada aresta indica a característica que foi removida na passagem de um nó do nível k para um nó do nível $k + 1$. Os nós preenchidos com cinza indicam que o limite foi atualizado. Os nós com contorno duplo indicam que foi encontrado $J(\cdot) < B$ e, se o nível do nó for $k < 4$, indicam poda. Os nós com contorno tracejado foram eliminados pelas podas.

Brach and bound

- Outros algoritmos mais eficientes:
- Branch and bound ordenado
- Branch and bound rápido
- Branch and bound com previsão parcial
- Branch and bound adaptativo

- Leitura complementar:
- M. A. Roncatti, Avaliação de métodos ótimos e subótimos de seleção de características de textura em imagens, 2008. Dissertação de mestrado.

Redução de dimensionalidade

Análise dos componentes principais

- Suponha que $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_M}$ são vetores N x 1

1. Calcule o vetor média: $\bar{x} = \frac{1}{M}\sum_{i=1}^{M} x_i$
2. Subtraia a média: $\Phi_i = x_i - \bar{x}$
3. Forme a matriz $A = [\Phi_1 \Phi_2 \cdots \Phi_M]$ e calcule a matriz de covariância:

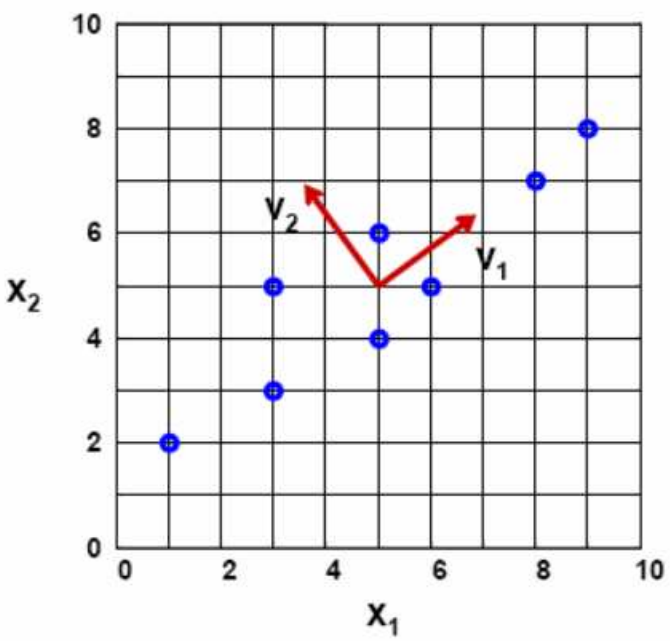
$$C = \frac{1}{M}\sum_{n=1}^{M} \Phi_n \Phi_n^T = AA^T$$

4. Calcule os autovalores e autovetores de C.
5. Ordene os autovetores de acordo com os autores (do maior para o menor) e coloque numa nova matriz U.
6. Redução da dimensionalidade: Qualquer vetor pode ser escrito como combinação linear dos autovetores obtidos:

$$b_i = u_i^T (x_i - \bar{x})$$

PCA: algoritmo

- Calcular os componentes principais para o seguinte conjunto de dados 2D
  - $X=(x_1,x_2)=\{(1,2),(3,3),(3,5),(5,4),(5,6),(6,5),(8,7),(9,8)\}$



- Solução (manual)
  - As variâncias estimadas são:

$$\Sigma_x = \begin{bmatrix} 6.25 & 4.25 \\ 4.25 & 3.5 \end{bmatrix}$$

  - Os autovalores são os zeros da equação:

$$\Sigma_x v = \lambda v \Rightarrow |\Sigma_x - \lambda I| = 0 \Rightarrow \begin{vmatrix} 6.25-\lambda & 4.25 \\ 4.25 & 3.5-\lambda \end{vmatrix} = 0 \Rightarrow \lambda_1 = 9.34;\ \lambda_2 = 0.41;$$

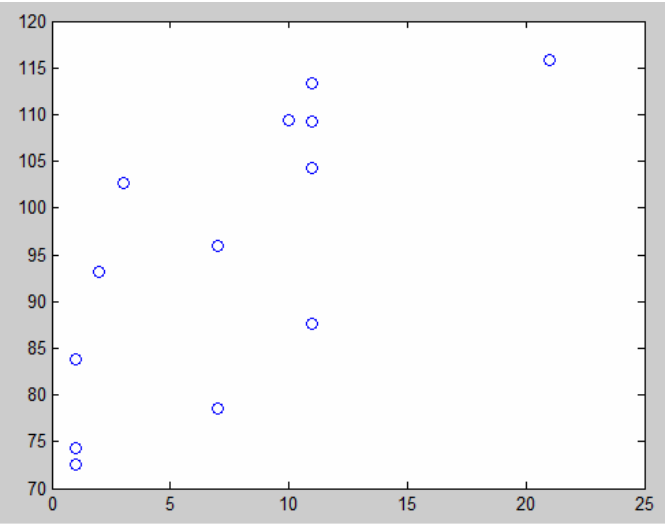
  - Os autovetores são as soluções do sistema:

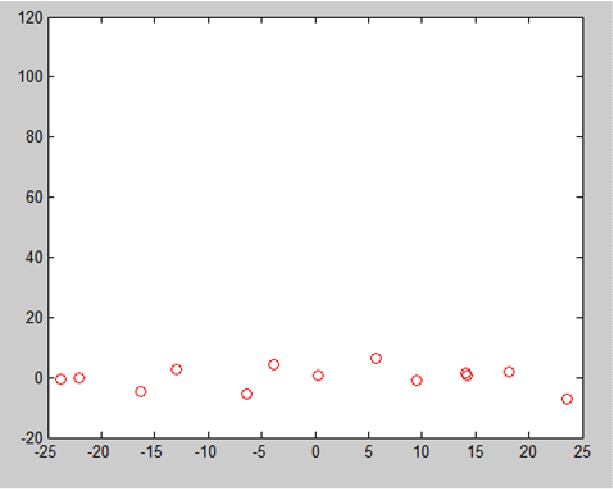
$$\begin{bmatrix} 6.25 & 4.25 \\ 4.25 & 3.5 \end{bmatrix}\begin{bmatrix} v_{11} \\ v_{12} \end{bmatrix} = \begin{bmatrix} \lambda_1 v_{11} \\ \lambda_1 v_{12} \end{bmatrix} \Rightarrow \begin{bmatrix} v_{11} \\ v_{12} \end{bmatrix} = \begin{bmatrix} 0.81 \\ 0.59 \end{bmatrix}$$

$$\begin{bmatrix} 6.25 & 4.25 \\ 4.25 & 3.5 \end{bmatrix}\begin{bmatrix} v_{21} \\ v_{22} \end{bmatrix} = \begin{bmatrix} \lambda_2 v_{21} \\ \lambda_2 v_{22} \end{bmatrix} \Rightarrow \begin{bmatrix} v_{21} \\ v_{22} \end{bmatrix} = \begin{bmatrix} -0.59 \\ 0.81 \end{bmatrix}$$

PCA: algoritmo

- No Matlab:
- [COEFF,SCORE,latent,tsquare] = princomp(X)
- **Exemplo:**
- load hald;
- x=[hald(:,1),hald(:,5)]; plot(x(:,1),x(:,2),'bo');
- [COEFF,SCORE,latent,tsquare] = princomp(X); figure;
- plot(SCORE(:,1),SCORE(:,2),'ro');





Projeto

- **Parte 1:**
- Procure implementações dos métodos de seleção de atributos na web e utilize-os para selecionar os 3 atributos mais importantes.
- Faça a classificação usando o Weka e compare com as classificações obtidas no projeto anterior.
- **Parte 2:**
- Utilize os atributos dos projetos anteriores e projete-os em 2 dimensões usando PCA.
- Faça a classificação usando janelas de Parzen.
- Compare com os resultados obtidos sem projeção (projeto anterior).

Referências

• *C. Bishop,* Pattern Recognition and Machine Learning , Springer (2006)

• *Sergios Theodoridis, Konstantinos Koutroumbas, Pattern Recognition*. Elsevier (2006).