MÉTODOS ESTATíSTICOS EM DATA MINING
- Folha 2: Análise de Clusters -

11. Uma forma de generalizar o método das K-médias, que minimiza o erro quadrático médio, consiste em definir o critério

$$J_T = \sum_{i=1}^{K} \sum_{x \in \mathcal{X}_i} (x_i - \mathbf{m_i})^t S_T^{-1} (x_i - \mathbf{m_i}),$$

onde $\mathbf{m_i}$ é a média das n_i amostras de $\mathcal{X}_i$ e S_T é a matriz de dispersão total.

(a) **(2)** Mostre que J_T é invariante para transformações lineares não singulares dos dados.

(b) Suponha que se transfere uma amostra $\hat{x}$ de $\mathcal{X}_i$ para $\mathcal{X}_j$. Mostre que:

i. **(1)** $\mathbf{m_j}$ passa a ser $\mathbf{m_j^*} = \mathbf{m_j} + \frac{\hat{x} - \mathbf{m_j}}{n_j + 1}$

ii. **(1)** $\mathbf{m_i}$ passa a ser $\mathbf{m_i^*} = \mathbf{m_i} - \frac{\hat{x} - \mathbf{m_i}}{n_i - 1}$

iii. **(1)** J_T passa a ser
$J_T^* = J_T + \left[\frac{n_j}{n_j+1} (\hat{x} - \mathbf{m_j})^t S_T^{-1} (\hat{x} - \mathbf{m_j}) - \frac{n_i}{n_i - 1} (\hat{x} - \mathbf{m_i})^t S_T^{-1} (\hat{x} - \mathbf{m_i}) \right]$

(c) **(3)** Sugira um algoritmo iterativo para minimizar J_T.

12. A semelhança para comparar dois objectos i e j é tal que $0 < s_{ij} \leq 1$. Mostre que $d_{ij} = 1 - s_{ij}$ e $d_{ij}^* = -log(s_{ij})$ são dissemelhanças.

13. Para comparar as espécies animais Tigre, Cão, Golfinho, Tubarão, Homem, Macaco, foram considerados os atributos binários:

- come outros animais
- come vegetais
- desloca-se sobre quatro patas
- vive na água
- tem pêlo

Construa as matrizes de semelhanças entre os animais com base no coeficiente de Jacard ($s_{ij} = \frac{a}{a+b+c}$) e no coeficiente de concordância simples ($s'_{ij} = \frac{a+d}{a+b+c+d}$). Construa depois um dendrograma para cada um dos casos com base no método da ligação completa e compare os resultados.

14. Mostre que a desigualdade ultramétrica implica a desigualdade triangular mas que a implicação inversa não é verdadeira.

15. Uma análise de clusters sobre cinco objectos é realizada a partir da matriz de dissemelhanças

$$D = \begin{bmatrix} 0.00 & & & & \\ 18.03 & 0.00 & & & \\ 20.62 & 14.14 & 0.00 & & \\ 22.36 & 11.18 & 5.00 & 0.00 & \\ 8.60 & 17.00 & 25.08 & 25.15 & 0.00 \end{bmatrix}$$

e no último passo do processo o grupo (15) é unido com o grupo (234).

(a) Apresente a matriz de dissemelhanças actualizada, referente ao último passo, considerando que são usados os métodos da ligação simples, completa e média.

(b) Ao tentar efectuar a análise de clusters um analista argumentou que os grupos (23) e (45) devem ser aglutinados já que a sua dissemelhança é pequena.

- Calcule essa dissemelhança.
- Diga porque é que essa aglutinação não foi considerada pelos métodos da alínea (a).

16. Num problema de Análise Classificatória, suponha que a classe ("cluster") $\mathcal{X}_i$ contém n_i amostras e designe por d_{ij} a distância entre duas classes $\mathcal{X}_i$ e $\mathcal{X}_j$. Suponhamos que juntamos estas duas classes numa só, $\mathcal{X}_k$. Seja agora d_{kh} a distância entre esta nova classe $\mathcal{X}_k$ e uma outra classe qualquer, $\mathcal{X}_h$. Em geral, é de esperar que a relação entre d_{kh} e as duas distâncias antes do agrupamento, d_{ih} e d_{jh}, não seja simples. Considere no entanto a equação seguinte (fórmula de Lance-Williams), que exprime d_{kh} à custa de d_{ih} e d_{jh} :

$$d_{kh} = \alpha_i d_{ih} + \alpha_j d_{jh} + \beta d_{ij} + \gamma |d_{ih} - d_{jh}|.$$

Mostre que algumas distâncias usuais em Análise Classificatória, como as que se encontram nas alíneas que se seguem, se podem escrever usando a equação acima:

(a) d_{min} (minimo de todas as distâncias entre um elemento de uma classe e um elemento de outra classe): $\alpha_i = \alpha_j = 0.5$, $\beta = 0$, $\gamma = -0.5$.

(b) d_{max} (máximo de todas as distâncias entre um elemento de uma classe e um elemento de outra classe): $\alpha_i = \alpha_j = 0.5$, $\beta = 0$, $\gamma = 0.5$.

(c) d_{media} (média de todas as distâncias entre um elemento de uma classe e um elemento de outra classe): $\alpha_i = \frac{n_i}{n_i+n_j}$, $\alpha_j = \frac{n_j}{n_i+n_j}$, $\beta = \gamma = 0$.

17. Considere de novo o conjunto dos caranguejos da aula teórica e utiliza cores e/ou símbolos diferentes para fazer um gráfico dos dados. De seguida aplique o K-médias com 2 grupos e faça também um gráfico com cores diferentes. De seguida "arredonde" os dados, faça um gráfico e aplique K-médias de novo com 2 grupos. Os resultados são diferentes? Finalmente aplique K-médias com 4 grupos e compare.

18. Considere o conjunto de dados leukemia (ver página web da cadeira em http://www.fc.up.pt/pessoas/jpcosta/MEDM.html), que contém valores de expressão de 7129 genes em 38 amostras de leucemia; há dois tipos de leucemia nestes dados : ALL e AML (coluna 7130). Escolha, usando a estatística t, os 400 genes mais discriminantes entre as duas classes de leucemia. Estandardize esses genes. Use o comando image.plot (library(fields)) para fazer um gráfico destes dados, os pacientes em linhas e os genes em colunas. Faça classificação hierárquica usando a ligação média e desenhe o dendrograma. Ordene os pacientes de acordo com o dendrograma. Use de novo o comando image.plot, mas agora com esta ordem dos pacientes. Faça agora classificação hierárquica mas dos genes. Pode agora usar de novo o image.plot com as ordens dos genes e dos pacientes obtidos nos dois dendrogramas; o mais fácil é usar o comando heatmap. Isto permite-lhe verificar os dois tipos de leucemia existentes nos dados ou eventualmente descobrir novos tipos.

17) Looking at the Crabs data again.

```
rm(list = ls())
library(MASS)
library(lattice)
data(crabs)
splom( log(crabs[,4:8]),col=as.numeric(crabs[,1]),pch=as.numeric(crabs[,2]),
main="circle/triangle is gender, black/red is species")
```

Apply kmeans with 2 clusters and plot results.

```
cl= kmeans( log(crabs[,4:8]), 2, nstart=1, iter.max=10)
splom( log(crabs[,4:8]),col=cl$cluster+2, main="blue/green is cluster finds big/small")
```

Discovers large/small crabs...

Sphere the data.

```
pcp= princomp( log(crabs[,4:8]) )
spc= pcp$scores %*% diag(1/pcp$sdev)
splom(  spc[,1:3],col=as.numeric(crabs[,1]),pch=as.numeric(crabs[,2]),
main="circle/triangle is gender, black/red is species")
```

And apply kmeans again.

```
cl= kmeans(spc, 2, nstart=1, iter.max=20)
splom(  spc[,1:3],col=cl$cluster+2, main="blue/green is cluster")
```

Discovers gender difference...
Results depends crucially on sphering the data first.

---

18)

```
dados=read.table("leukemia.data",sep=",")
s=c(rep(0,7129))
for(i in 1:7129) s[i]=t.test(dados[dados[,7130]=="AML",i],dados[dados[,7130]=="ALL",i],
var.equal=TRUE)$statistic
b=order(abs(s))
X=dados[,b[6730:7129]]
X = scale(X)
library(fields)
image.plot(1:ncol(X),1:nrow(X),t(X),col=tim.colors(200),xlab="GENES",
ylab="PATIENTS", cex.lab=1.4)
dd = dist(X)
hh = hclust(dd,method="average")
plot(hh) ou então plot(hh, cex.lab=1.3, xlab="", ylab="HEIGHT", sub="")
ord = hh$order
image.plot(1:ncol(X),1:nrow(X),t(X),col=tim.colors(200),xlab="GENES",
ylab="PATIENTS", cex.lab=1.4)
dd1=dist(t(X))
hh1 = hclust(dd1,method="average")
plot(hh1)
heatmap(X)
```

package FactorMine R (hierarchical clustering on PCA (HCPC); chooses also number of PCs; also important contributions of the variables to PCs; HCPC suggests a natural cut of the hierarchical tree).

also for number of clusters: silhouette, GAP statistic, Hopkin's statistic (this one to see if there are clusters or not)

package Nbclust (tests 26 different indices to determine the most appropriate number of clusters)

packages corrplot, rcorr (cophenetic correlation, matrix of correlations, colors on correlations, p-values...)

package pvclust (provides p-values for each cluster; the function pvrect emphasizes clusters with high p-values; see also rect.hclust)