APA: Aprenentatge Automàtic (TEMES 2 i 3)

Grau en Enginyeria Informàtica - UPC (2017/18)

Lluís A. Belanche, belanche@cs.upc.edu

Entrega: 23 Octubre 2017

Els problemes marcats **[G]** són de grup; els problemes/apartats marcats **[R]** són per fer-se en R

Objectius:

1. Comprendre l'anàlisi de components principals (PCA) i saber-la calcular
2. Comprendre l'anàlisi discriminant d'en Fisher (FDA) i saber-la calcular
3. Saber quan cal usar PCA i quan FDA i què aporta cadascuna
4. Comprendre el model de barreja de Gaussianes (i el seu cas particular k-means) per a tasques de *clustering* i saber-lo aplicar
5. Saber derivar algorismes de *clustering* probabilístics per barreges de distribucions, com a cas particular de l'algorisme E-M

Problema 1 L'anàlisi de components principals en dues variables

Siguin X_1 i X_2 dues variables aleatòries estandarditzades i amb correlació $\rho > 0$. Construirem un PCA pas a pas a partir de la matriu de correlació teòrica R. Es demana:

1. Expresseu els dos valors propis λ_1 i λ_2 de R
2. Expresseu els dos vectors propis corresponents $\boldsymbol{a}_1$ i $\boldsymbol{a}_2$
3. Expresseu els nous eixos de coordenades, és a dir, doneu les dues components principals Y_1 i Y_2

. . . .

Problema 2 L'anàlisi de components principals en acció [G, R]

Considerem un problema amb $N = 8$ dades bidimensionals:

$$\{(1,2),(3,3),(3,5),(5,4),(5,6),(6,5),(8,7),(9,8)\}$$

1. Calculeu la matriu de covariança mostral de les dades $\hat{\Sigma}$
2. Calculeu els dos valors propis de $\hat{\Sigma}$
3. Calculeu els dos vectors propis corresponents $\boldsymbol{a}_1$ i $\boldsymbol{a}_2$
4. Dibuixeu les dades i les dues components principals
5. Quin és el percentatge de variança explicada per la primera component principal?

. . . .

Problema 3 L'anàlisi de components principals en acció [R]

Genereu $N = 1000$ dades Gaussianes tridimensionals amb mitjana $\mu = (0, 5, 2)^\top$ i matriu de covariances

$$\Sigma = \begin{pmatrix} 25 & -1 & 7 \\ -1 & 4 & -4 \\ 7 & -4 & 10 \end{pmatrix}.$$

1. Feu un *plot* de les dades
2. Apliqueu PCA; reporteu els 3 components principals, i les seves variances (absolutes i acumulades)
3. Feu 3 *plots* de les noves dades, usant els components principals de dos en dos i comenteu els resultats

. . . .

Problema 4 L'anàlisi discriminant d'en Fisher en acció [G,R]

Considerem un problema amb dades bidimensionals i dues classes:

$$C_1 = \{(4, 1), (2, 4), (2, 3), (3, 6), (4, 4)\}$$
$$C_2 = \{(9, 10), (6, 8), (9, 5), (8, 7), (10, 8)\}$$

1. Calculeu les dues mitjanes de classe $\boldsymbol{m}_1$ i $\boldsymbol{m}_2$.
2. Calculeu les dues matrius de dispersió (*scatter*) intra-classe S_1 i S_2 i la matriu de dispersió intra-classes total $S_W = S_1 + S_2$.
3. Calculeu la matriu de dispersió inter-classes S_B.
4. Trobeu la direcció de projecció òptima $\boldsymbol{w}^*$ de dues maneres:
   (a) Directament amb la fòrmula $\boldsymbol{w}^* = S_W^{-1}(\boldsymbol{m}_2 - \boldsymbol{m}_1)$.
   (b) Resolent el problema de vectors propis $(S_W^{-1} S_B)\boldsymbol{w} = \lambda \boldsymbol{w}$.
5. Representeu gràficament el resultat: dibuixeu les dades, la direcció de projecció òptima $\boldsymbol{w}^*$ i la projecció de les dades

. . . .

Problema 5 Obtenció del criteri d'en Fisher

Usant les definicions vistes a classe per les matrius de dispersió (*scatter*) intra-classes S_W i inter-classes S_B, demostreu que el criteri d'en Fisher es pot escriure com:

$$J(\boldsymbol{w}) = \frac{\boldsymbol{w}^\top S_B \boldsymbol{w}}{\boldsymbol{w}^\top S_W \boldsymbol{w}}$$

1. Demostreu primer que $s_1^2 + s_2^2 = \boldsymbol{w}^\top S_W \boldsymbol{w}$
2. Demostreu que $(\mu_2 - \mu_1)^2 = \boldsymbol{w}^\top S_B \boldsymbol{w}$

. . . .

Problema 6 Comparació entre PCA i FDA [G, R]

Genereu $N = 200$ dades Gaussianes bidimensionals en 4 grups, amb mitjanes $\mu_1 = (0.2, 0.3)^\top, \mu_2 = (0.35, 0.75)^\top, \mu_3 = (0.65, 0.55)^\top$ i $\mu_1 = (0.8, 0.25)^\top$ (50 de cada), totes elles amb matriu de covariances

$$\Sigma = \begin{pmatrix} 1.8 & 0.7 \\ 0.7 & 1.1 \end{pmatrix}.$$

1. Feu un *plot* de les dades
2. Apliqueu PCA i feu un *plot* de les noves dades, usant el primer component principal
3. Apliqueu FDA i feu un *plot* de les noves dades, usant el primer discriminant
4. Repetiu els plots pintant cada grup d'un color. Comenteu els resultats

. . . .

Problema 7 Descomposició de barreja de Gaussianes

Considereu el model de barreja de Gaussianes:

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_k, \Sigma_k)$$

A classe hem vist que podem treballar amb un vector de variables (anomenades latents) $\boldsymbol{z}$, on $z_i \in \{0,1\}$ i $\sum_{k=1}^{K} z_i = 1$, de manera que $p(z_k = 1) = \pi_k$. Demostrar la descomposició alternativa de la barreja:

$$p(\boldsymbol{x}) = \sum_{\boldsymbol{z}} p(\boldsymbol{z}) p(\boldsymbol{x}|\boldsymbol{z}),$$

on $\boldsymbol{z}$ es mou per tots els vectors que tènen una sola component a 1 (i la resta a 0).

. . . .

Problema 8 Convergència de k-means

Demostreu o argumenteu que l'algorisme de k-means convergeix (és a dir, s'atura després d'un número finit de voltes) amb independència de les condicions inicials. Pista: fixeu-vos que el conjunt de valors possibles de les variables indicador $\{r_{nk}\}$ és finit i que, per cadascuna de les configuracions, hi ha un únic òptim pels prototipus $\{\boldsymbol{\mu}_k\}$.

. . . .

Problema 9 Simplificació de la barreja de Gaussianes 1 [G]

Considereu el model de barreja de Gaussianes:

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_k, \Sigma_k)$$

Preneu el cas que totes les matrius de covariança són iguals i diagonals, és a dir, $\Sigma_1 = \ldots = \Sigma_K = \Sigma = diag(\sigma_1^2, \ldots, \sigma_d^2)$.

1. Enraoneu en quin sentit representa una simplificació respecte al cas general (amb matrius de covariança generals), des dels punts de vista *estadístic* i *geomètric*.

2. Expresseu la funció de densitat de probabilitat $\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_k, \Sigma_k)$ que en resulta.
3. Construïu la funció de log-versemblança negativa.
4. Deriveu les equacions de l'algorisme E-M que en resulta i escriviu l'algorisme de *clustering* complet.
5. Enraoneu sobre les implicacions (possibles avantatges/inconvenients) que representa la simplificació respecte el cas general des del punt de vista del *clustering*.

. . . .

Problema 10 Distàncies ponderades

Suposeu que extenem les distàncies Euclidianes

$$d(\boldsymbol{x}, \boldsymbol{y}) = ||\boldsymbol{x} - \boldsymbol{y}|| = \sqrt{\sum_{i=1}^{d} (x_i - y_i)^2}, \qquad \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$$

i considerem distàncies Euclidianes ponderades

$$d_{\boldsymbol{w}}(\boldsymbol{x}, \boldsymbol{y}) = ||\boldsymbol{x} - \boldsymbol{y}||_{\boldsymbol{w}} = \sqrt{\sum_{i=1}^{d} w_i (x_i - y_i)^2}, \qquad \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{w} \in \mathbb{R}^d,$$

on $w_i > 0$.

1. Trobeu vectors $\boldsymbol{z}, \boldsymbol{t} \in \mathbb{R}^d$ tals que $d_{\boldsymbol{w}}(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{z}, \boldsymbol{t})$ (cal que els expresseu en funció de $\boldsymbol{w}, \boldsymbol{x}, \boldsymbol{y}$); interpreteu el resultat.
2. Té algún avantatge usar distàncies Euclidianes ponderades en un *clustering*? Distingiu el cas on $\boldsymbol{w}$ és conegut a priori del cas en què no.

. . . .

Problema 11 Simplificació de la barreja de Gaussianes 2 [G]

Considereu el model de barreja de Gaussianes:

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_k, \Sigma_k)$$

Preneu el cas que totes les matrius de covariança són iguals i proporcionals a una variança comuna, és a dir, $\Sigma_1 = \ldots = \Sigma_K = \Sigma = \sigma^2 I$, on I és la matriu identitat.

1. Enraoneu en quin sentit representa una simplificació respecte al cas general (amb matrius de covariança generals), des d'un punt de vista estadístic i geomètric.
2. Expresseu la funció de densitat de probabilitat $\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_k, \Sigma_k)$ que en resulta.
3. Construïu la funció de log-versemblança negativa.
4. Deriveu les equacions de l'algorisme E-M que en resulta i escriviu l'algorisme de *clustering* complet.
5. Suposant que σ^2 fos coneguda, argumenteu perquè, en fer $\sigma^2 \to 0$, l'algorisme esdevé k-means.

. . . .

Problema 12 Clustering de dades 2D artificials [R]

Volem analitzar un problema d'agrupament amb dades circulars en 2D usant la rutina `mlbench.2dnormals`. Generem dades arranjades circularment en $k = 6$ grups Gaussians amb el codi:

```
library(mlbench)

N <- 1000
k <- 6
sigma2 <- 0.6^2

data.1 <- mlbench.2dnormals (N,k,sd=sqrt(sigma2))
plot(data.1)
```

Veureu que cadascun dels grups és una Gaussiana bivariada. Els centres estàn equiespaiats en un cercle entorn de l'origen de radi $r = \sqrt{k}$. Les matrius de covariança són de la forma $\sigma^2 I$, on I és la matriu identitat i hem pres $\sigma^2 = 0.6^2$ (vegeu `?mlbench.2dnormals`). El `plot` anterior us mostrarà la veritat de les dades (els 6 grups generats). Si ara feu:

```
plot(x=data.1$x[,1], y=data.1$x[,2])
```

veureu les dades en brut (el que rebrà el mètode de *clustering*). Es demana:

1. Decidiu per endavant quin mètode de *clustering* hauria de treballar millor i amb quins paràmetres. Consell: feu una ullada a la forma en què es generen les dades (`?mlbench.2dnormals`)
2. Apliqueu k-means un cert nombre de vegades amb $k = 6$ i observeu els resultats
3. Apliqueu k-means amb una selecció de valors de k al vostre criteri (20 cops cadascun) i monitoritzeu l'índex de Calinski-Harabasz mitjà; quin k es veu millor?
4. Apliqueu l'algorisme E-M amb $k = 6$ i observeu els resultats (mitjanes, coeficients i covariàncies) Comproveu els resultats contra les vostres expectatives (apartat 1).

....

Problema 13 Clustering del geyser 'Old Faithful' [R,G]

Volem analitzar un problema d'agrupament amb dades d'erupcions del geyser 'Old Faithful', al Yellowstone National Park, Wyoming. Les dades corresponen al temps d'espera entre erupcions i la durada de l'erupció (1 al 15 d'Agost, 1985).

```
library(MASS)
help(geyser)
summary(geyser)
plot(geyser)
```

1. Decidiu per endavant quin mètode de *clustering* hauria de treballar millor i amb quins paràmetres (no hi ha pistes, és un problema real).
2. Apliqueu k-means amb una selecció de valors de k al vostre criteri i observeu els resultats
3. Apliqueu k-means 100 cops per aquest valors i monitoritzeu l'índex de Calinski-Harabasz mitjà; quin k es veu millor?
4. Apliqueu l'algorisme E-M amb una família de la vostra elecció ("spherical", "diagonal", etc), amb la millor k lliurada per k-means

5. El criteri BIC s'utilitza sovint per triar el millor model per barrejes de Gaussianes. BIC es defineix com $q\ln(N) - 2l$, sent l el valor de la log-versemblança, q el nombre de paràmetres lliures en el model de barreja, i N el nombre d'observacions. Es tria el model i el nombre de clusters amb el menor BIC. Trobareu aquesta opció al paràmetre `mixmodCluster (..., criterion = "BIC")`. Apliqueu E-M de nou amb una família de la vostra elecció ("spherical", "diagonal", etc), aquesta vegada deixant BIC decidir el millor nombre de clusters[1]. La forma més fàcil d'inspeccionar els resultats finals és amb un `summary` de la vostra crida a `mixmodCluster`. Un cop hagueu acabat, grafiqueu els resultats (baseu-vos en un plot del resultat de `mixmodCluster`).

. . . .

Problema 14 Clustering de les dades artificials Cassini [R]

Volem analitzar un problema d'agrupament amb dades en 2D usant la rutina `mlbench.cassini`. Generem dades en 3 grups amb el codi:

```
library(mlbench)

N <- 2000

data.1 <- mlbench.cassini(N, relsize = c(1,1,0.25))
plot(data.1)
```

Veureu que les estructures externes tenen forma de plàtan i entre elles hi ha un cercle amb menys densitat de dades. El `plot` anterior us mostrarà la veritat de les dades (els 3 grups generats). Si ara feu:

```
plot(x=data.1$x[,1], y=data.1$x[,2])
```

veureu les dades en brut (el que rebrà el mètode de *clustering*). Es demana:

1. Decidiu per endavant quin mètode de *clustering* hauria de treballar millor i amb quins paràmetres.
2. Apliqueu k-means varis amb $k = 3$ i observeu els resultats. Com es comporta?
3. Apliqueu k-means amb una selecció de valors de k al vostre criteri (20 cops per cadascun) i monitoritzeu l'índex de Calinski-Harabasz mitjà; quin k es veu millor?
4. Apliqueu l'algorisme E-M amb una selecció de valors de k al vostre criteri (10 cops cadascun) i observeu els resultats. Comproveu els resultats contra les vostres expectatives (apartat 1).

. . . .

[1] Això es pot fer de forma automàtica amb una crida semblant a `mixmodCluster(geyser, nbCluster=2:6)`