APA: Aprenentatge Automàtic (TEMA 2)

Grau en Enginyeria Informàtica - UPC (2018/19)

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

Entrega: 23 Octubre 2018

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

Objectius:

1. Comprendre el model de barreja de Gaussianes (i el seu cas particular k-means) per a tasques de *clustering* i saber-lo aplicar
2. Saber derivar algorismes de *clustering* probabilístics per barreges de distribucions Gaussianes, com a cas particular de l'algorisme E-M

Problema 1 Descomposició de barreja de Gaussianes

Considereu el model de barreja de Gaussianes:

$$p(\boldsymbol{x}) = \sum_{c=1}^{k} \pi_c \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_c, \Sigma_c)$$

A classe hem vist que podem treballar amb un vector de variables (anomenades latents) $\boldsymbol{z}$, on $z_c \in \{0,1\}$ i $\sum_{c=1}^{k} z_c = 1$, de manera que $p(z_c = 1) = \pi_c$. 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 2 Convergència de k-means

Demostreu o argumenteu (de manera convincent i breu) 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 $\{r_{ic}\}$ és finit.

. . . .

Problema 3 Simplificació de la barreja de Gaussianes 1

Considereu el model de barreja de Gaussianes:

$$p(\boldsymbol{x}) = \sum_{c=1}^{k} \pi_c \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_c, \Sigma_c)$$

Preneu el cas uni-dimensional ($d = 1$), és a dir, el problema es redueix a estimar $\pi_1, \ldots \pi_k, \sigma_1^2, \ldots, \sigma_k^2, \mu_1, \ldots \mu_k$.

1. Construïu la funció de log-versemblança negativa.

2. Deriveu les equacions que en resulten i escriviu l'algorisme de *clustering* complet.
3. Genereu una mostra de dades de mida $n = 200$ (que realment vingui d'una BdG) i executeu E-M en funció de varis números de *clusters*; comenteu-ne els resultats. **[R]**

. . . .

Problema 4 Simplificació de la barreja de Gaussianes 2 [G]

Considereu el model de barreja de Gaussianes:

$$p(\boldsymbol{x}) = \sum_{c=1}^{k} \pi_c \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_c, \Sigma_c)$$

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 $\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_c, \Sigma_c)$ que en resulta.
3. Construïu la funció de log-versemblança negativa.
4. Deriveu les equacions que en resulten 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 5 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 6 Simplificació de la barreja de Gaussianes 3 [G]

Considereu el model de barreja de Gaussianes:

$$p(\boldsymbol{x}) = \sum_{c=1}^{k} \pi_c \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_c, \Sigma_c)$$

Preneu el cas que totes les matrius de covariança són iguals i proporcionals a una variança comuna, és a dir, $\Sigma_1 = \ldots = \Sigma_k = \Sigma = \sigma^2 \mathbf{I}$, on $\mathbf{I}$ 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 $\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_c, \Sigma_c)$ que en resulta.
3. Construïu la funció de log-versemblança negativa.
4. Deriveu les equacions que en resulten i escriviu l'algorisme de *clustering* complet.
5. Argumenteu convincentment perquè, en fer $\sigma^2 \to 0$, l'algorisme esdevé k-means.

. . . .

Problema 7 *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 $\sigma^2 \mathbf{I}$, on $\mathbf{I}$ 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 8 *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 9 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?

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

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).

. . . .

Problema 10 Clustering per densitat

Volem construïr unes dades de mida $n = 3\nu$ en el pla ($d = 2$). Una tercera part estan distribuïdes *uniformement* en un cercle de radi 1 centrat a l'origen; una altra tercera part en un cercle de radi 10 centrat a $(11, 11)$, i l'altra tercera part en un cercle de radi 20 centrat a $(22, 22)$. Si usem k-means per trobar un *clustering* per una certa k:

1. Com creieu que es localitzaran els centres donats per l'algorisme? (per igual, més en el *cluster* més dens, més en el *cluster* menys dens, ...)
2. Dissenyeu un experiment per comprovar-ho, executant l'algorisme diverses vegades per la mateixa k i fent variar k. Com en depèn tot plegat del valor de ν? **[R]**

Nota: per generar les dades, és útil fer-ho en polars: generar un angle aleatori i un radi aleatori; aquest darrer es pot fer com $r\sqrt{\rho}$, on $\rho \sim U(0, 1)$ i r és el radi dessitjat.

. . . .