LEIC-T 2023/2024
Aprendizagem - Machine Learning
Homework 4
Deadline 30/10/2024 20:00
Submit on Fenix as pdf

I) (7 pts) Clustering

Assuming the original announcement in which the given covariance matrices are not symmetrical, however despite the fact assuming random initialization the algorithms works since there are inverse covariance matrices and after the adaptation step the covariance matrices become symmetrical.

Given the data

$$\mathbf{x}_1 = \begin{pmatrix} 2.5 \\ 2.5 \end{pmatrix}, \mathbf{x}_2 = \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \mathbf{x}_3 = \begin{pmatrix} 0.5 \\ 0.55 \end{pmatrix},$$

$$\pi_1 = 0.6, \pi_2 = 0.4$$

$$C_1\left(\mathbf{u}_1 = \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}\right),\ C_2\left(\mathbf{u}_2 = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}, \boldsymbol{\Sigma}_2 = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}\right).$$

i) (6 pts)

Perform one iteration of EM clustering algorithm step by step and determine the new parameters. Indicate all the calculations step by step. (To make the calculation easier for each step you can use a computer, however you should be able to do it by hand)
Solution:

E-Step

a) Likelihood

$$p(\mathbf{x}_\eta | c_k = 1) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \Sigma_k) = \frac{1}{(2\cdot\pi)^{D/2}} \cdot \frac{1}{|\Sigma_k|^{1/2}} \cdot \exp\left(-\frac{1}{2}\cdot(\mathbf{x}_\eta - \boldsymbol{\mu}_k)^T \Sigma_k^{-1} \cdot (\mathbf{x}_\eta - \boldsymbol{\mu}_k)\right)$$

b) Joint distribution

$$p(c_k = 1, \mathbf{x}_\eta) = \pi_k \cdot \mathcal{N}(\mathbf{x}_\eta|\boldsymbol{\mu}_k, \Sigma_k)$$

c) Data

$$p(\mathbf{x}_\eta) = \sum_{k=1}^{K} p(c_k = 1, \mathbf{x}_\eta)$$

d) Posterior probability

LEIC-T 2023/2024
Aprendizagem - Machine Learning
Homework 4
Deadline 30/10/2024 20:00
Submit on Fenix as pdf

$$\gamma(c_{\eta k}) = p(c_k = 1|\mathbf{x}_\eta) = \frac{p(c_k = 1, \mathbf{x}_\eta)}{p(\mathbf{x}_\eta)}$$

a)-d)

a) Likelihood
p(**x**1|c1=1)= 0.1404537443096252
p(**x**2|c1=1)= 0.15915494309189535
p(**x**3|c1=1)= 0.05357744689112038

p(**x**1|c2=1)= 0.021539279301848634
p(**x**2|c2=1)= 0.05167004496706156
p(**x**3|c2=1)= 0.1589561237010377

b) Joint distribution

p(**x**1,c1=1)= 0.08427224658577512
p(**x**2,c1=1)= 0.09549296585513721
p(**x**3,c1=1)= 0.03214646813467223

p(**x**1,c2=1)= 0.008615711720739454
p(**x**2,c2=1)= 0.020668017986824626
p(**x**3,c2=1)= 0.06358244948041508

c) Data

p(**x**1)= 0.09288795830651457
p(**x**2)= 0.11616098384196183
p(**x**3)= 0.09572891761508731

d) Posterior probability

$\gamma(c_{11})$= p(c1=1|**x**1)= 0.9072461933945295
$\gamma(c_{21})$= p(c1=1|**x**2)= 0.8220743548888698
$\gamma(c_{31})$= p(c1=1|**x**3)= 0.33580728723925113

$\gamma(c_{12})$= p(c2=1|**x**1)= 0.09275380660547043
$\gamma(c_{22})$= p(c2=1|**x**2)= 0.17792564511113015
$\gamma(c_{32})$= p(c2=1|**x**3)= 0.6641927127607489

M-Step

N_1= 2.0651278355226506
N_2= 0.9348721644773494

New Means

$$\mu 1=\begin{pmatrix}1.97574589170702 08\\1.9838763154382533\end{pmatrix}$$

$$\mu 2=\begin{pmatrix}0.9839122374881629\\1.019435421191684\end{pmatrix}$$

covariance matrices are symmetrical

$$\Sigma 1=\begin{pmatrix}0.4751101079583842 & 0.4631116685391803\\0.4631116685391803 & 0.4514536465164886\end{pmatrix}$$

$$\Sigma 2=\begin{pmatrix}0.5909127998290468 & 0.5737226965203728\\0.5737226965203728 & 0.5570468558164406\end{pmatrix}$$

Mixing Parameter equal to N_k/N

$\pi 1=2.0651278355226506/3{=}0.6883759451742169$
$\pi 2=0.9348721644773494/3{=}0.31162405482578315$

Clustering solution assuming covariance matrices are identity matrices (symmetrical)

Given the data

$$\mathbf{x}_1=\begin{pmatrix}2.5\\2.5\end{pmatrix},\mathbf{x}_2=\begin{pmatrix}2\\2\end{pmatrix},\mathbf{x}_3=\begin{pmatrix}0.5\\0.55\end{pmatrix},$$

$$\pi_1=0.6,\pi_2=0.4$$

$$C_1\left(\mathbf{u}_1=\begin{pmatrix}2\\2\end{pmatrix},\mathbf{\Sigma}_1=\begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}\right),\ C_2\left(\mathbf{u}_2=\begin{pmatrix}0.5\\0.5\end{pmatrix},\mathbf{\Sigma}_2=\begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}\right).$$

i) (6 pts)

Perform one iteration of EM clustering algorithm step by step and determine the new parameters. Indicate all the calculations step by step. (To make the calculation easier for each step you can use a computer, however you should be able to do it by hand)
Solution:

LEIC-T 2023/2024
Aprendizagem - Machine Learning
Homework 4
Deadline 30/10/2024 20:00
Submit on Fenix as pdf

E-Step

a) Likelihood

$$p(\mathbf{x}_\eta|c_k=1)=\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k,\Sigma_k)=\frac{1}{(2\cdot\pi)^{D/2}}\cdot\frac{1}{|\Sigma_k|^{1/2}}\cdot\exp\left(-\frac{1}{2}\cdot(\mathbf{x}_\eta-\boldsymbol{\mu}_k)^T\Sigma_k^{-1}\cdot(\mathbf{x}_\eta-\boldsymbol{\mu}_k)\right)$$

b) Joint distribution

$$p(c_k=1,\mathbf{x}_\eta)=\pi_k\cdot\mathcal{N}(\mathbf{x}_\eta|\boldsymbol{\mu}_k,\Sigma_k)$$

c) Data

$$p(\mathbf{x}_\eta)=\sum_{k=1}^{K}p(c_k=1,\mathbf{x}_\eta)$$

d) Posterior probability

$$\gamma(c_{\eta k})=p(c_k=1|\mathbf{x}_\eta)=\frac{p(c_k=1,\mathbf{x}_\eta)}{p(\mathbf{x}_\eta)}$$

a)-d)

a) Likelihood

p(**x**1|c1=1)=0.12395
p(**x**2|c1=1)= 0.159155
p(**x**3|c1=1)= 0.01805871172833642

p(**x**1|c2=1)= 0.00291502
p(**x**2|c2=1)= 0.0167748
p(**x**3|c2=1)= 0.1589561237010377

b) Joint distribution

p(**x**1,c1=1)= 0.07437
p(**x**2,c1=1)= 0.095493
p(**x**3,c1=1)= 0.010835

p(**x**1,c2=1)=0.00116601
p(**x**2,c2=1)= 0.00670992
p(**x**3,c2=1)= 0.06358244948041508

c) Data

p(**x**1)= 0.075536
p(**x**2)= 0.102203
p(**x**3)= 0.07441767651741693

d) Posterior probability

$\gamma(c_{11})= p(c1=1|\mathbf{x}1)= 0.9845635235165604$
$\gamma(c_{21})= p(c1=1|\mathbf{x}2)= 0.934347031598555$
$\gamma(c_{31})= p(c1=1|\mathbf{x}3)= 0.14560017920562118$

$\gamma(c_{12})= p(c2=1|\mathbf{x}1)= 0.0154365$
$\gamma(c_{22})= p(c2=1|\mathbf{x}2)= 0.065653$
$\gamma(c_{32})= p(c2=1|\mathbf{x}3)= 0.8543998207943788$

M-Step

$N_1= 2.064510734320737$
$N_2= 0.9354892656792633$

New Means

$$\mu 1 = \begin{pmatrix} 2.132661694801001 \\ 2.136187958355487 \end{pmatrix}$$

$$\mu 2 = \begin{pmatrix} 0.6382724637413377 \\ 0.6839383977151495 \end{pmatrix}$$

New Covariance Matrices

$$\Sigma 1 = \begin{pmatrix} 0.26030753540220 74 & 0.25455033997102466 \\ 0.25455033997102466 & 0.2489570231829106 \end{pmatrix}$$

$$\Sigma 2 = \begin{pmatrix} 0.20479038729429136 & 0.19847604609468314 \\ 0.19847604609468314 & 0.19235962406806495 \end{pmatrix}$$

Mixing Parameter equal to N_k/N

$\pi 1 = 2.064510734320737/3=0.6881702447735789$
$\pi 2 = 0.9354892656792633/3=0.3118297552264211$

LEIC-T 2023/2024
Aprendizagem - Machine Learning
Homework 4
Deadline 30/10/2024 20:00
Submit on Fenix as pdf

ii) (1 pts)

Performing a hard assignment of observations to clusters identify the silhouette of the larger cluster under a Manhattan distance (l_1 distance)

As in the announcement:

$\gamma(c_{11})$= p(c1=1|**x**1)= 0.9072461933945295 -> **x**1$\in C1$
$\gamma(c_{21})$= p(c1=1|**x**2)= 0.8220743548888698 -> **x**2$\in C1$
$\gamma(c_{31})$= p(c1=1|**x**3)= 0.33580728723925113 -> **x**3$\in C2$

With identity matrices:

$\gamma(c_{11})$= p(c1=1|**x**1)= 0.984564 -> **x**1$\in C1$
$\gamma(c_{21})$= p(c1=1|**x**2)= 0.934347. -> **x**2$\in C1$
$\gamma(c_{31})$= p(c1=1|**x**3)= 0.145600. -> **x**3$\in C2$

So, the hard assignment in both cases lead to the same solution (using the Euclidean distance):

s(x1)=1-a(x1)/b(x1)=1-0.7071/2.793=0.7469
s(x2)=1-a(x2)/b(x2)=1-0.7071/2.086=0.6611

s(C1)=(s(x1)+s(x2))/2= 0.70396