TÉCNICO LISBOA

Aprendizagem 2021/22

Homework IV

Deadline 14/11/2021 23:59 via Fenix as PDF

I. Pen-and-paper [12v]

Consider the following bivariate observations in a Euclidean space:

	y_1	y_2
$\mathbf{x}_1$	2	4
$\mathbf{x}_2$	-1	-4
$\mathbf{x}_3$	-1	2
$\mathbf{x}_4$	4	0

1) [6v] Compute and sketch the clustering solution given by EM assuming considering $\mathbf{x}_1$ and $\mathbf{x}_2$ to be the centroid means and:

$$\mathbf{\Sigma}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \mathbf{\Sigma}_2 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}, \pi_1 = p(c_1 = 1) = 0.7, \pi_2 = p(c_2 = 1) = 0.3$$

E-Step

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

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

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

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

a)-d) $$\gamma(c_{\eta k}) = p(c_k = 1 | \mathbf{x}_\eta) = \frac{\pi_k \cdot \mathcal{N}(\mathbf{x}_\eta | \boldsymbol{\mu}_k, \Sigma_k)}{\sum_{k=1}^{K} \pi_k \cdot \mathcal{N}(\mathbf{x}_\eta | \boldsymbol{\mu}_k, \Sigma_k)}$$

M-Step

$$N_k = \sum_{\eta=1}^{N} \gamma(c_{\eta k})$$

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \cdot \sum_{\eta=1}^{N} \gamma(c_{\eta k}) \cdot \mathbf{x}_\eta$$

$$\Sigma_k = \frac{1}{N_k} \cdot \sum_{\eta=1}^{N} \gamma(c_{\eta k}) \cdot (\mathbf{x}_\eta - \boldsymbol{\mu}_k) \cdot (\mathbf{x}_\eta - \boldsymbol{\mu}_k)^T$$

$$\pi_k = p(c_k = 1) = \frac{N_k}{N}$$

TÉCNICO LISBOA

Aprendizagem 2021/22

Homework IV

Deadline 14/11/2021 23:59 via Fenix as PDF

a)

$p(\mathbf{x}1|c1=1)= 0.159155$
$p(\mathbf{x}2|c1=1)= 2.23909*10^{-17}$
$p(\mathbf{x}3|c1=1)= 0.00023928$
$p(\mathbf{x}4|c1=1)= 7.22562*10^{-6}$

$p(\mathbf{x}1|c2=1)= 9.43878*10^{-10}$
$p(\mathbf{x}2|c2=1)= 0.079577$
$p(\mathbf{x}3|c2=1)= 9.82064*10^{-6}$
$p(\mathbf{x}4|c2=1)= 2.81366*10^{-6}$

b)

$p(\mathbf{x}1,c1=1)= 0.111408$
$p(\mathbf{x}2,c1=1)= 1.56736*10^{-17}$
$p(\mathbf{x}3,c1=1)= 0.000167496$
$p(\mathbf{x}4,c1=1)= 5.05794*10^{-6}$

$p(\mathbf{x}1,c2=1)= 2.83163*10^{-10}$
$p(\mathbf{x}2,c2=1)= 0.0238732$
$p(\mathbf{x}3,c2=1)= 2.94619*10^{-6}$
$p(\mathbf{x}4,c2=1)= 2.94619*10^{-6}$

c)

$p(\mathbf{x}1)= 0.111408$
$p(\mathbf{x}2)= 0.0238732$
$p(\mathbf{x}3)= 0.000170442$
$p(\mathbf{x}4)= 5.90203*10^{-6}$

d)

$\gamma(c_{11})= p(c1=1|\mathbf{x}1)= 1$
$\gamma(c_{21})= p(c1=1|\mathbf{x}2)= 6.56535*10^{-16}=0$
$\gamma(c_{31})= p(c1=1|\mathbf{x}3)= 0.982714$
$\gamma(c_{41})= p(c1=1|\mathbf{x}4)= 0.856982$

$\gamma(c_{12})= p(c2=1|\mathbf{x}1)= 2.54167*10^{-9}=0$
$\gamma(c_{22})= p(c2=1|\mathbf{x}2)= 1$
$\gamma(c_{32})= p(c2=1|\mathbf{x}3)= 0.0172856$
$\gamma(c_{42})= p(c2=1|\mathbf{x}4)= 0.143018$

M-Step

$N_1= 2.8397$
$N_2= 1.1603$

Means

$$\mu 1 = \frac{1}{2.8397}\left(1\begin{pmatrix}2\\4\end{pmatrix} + 0\begin{pmatrix}-1\\-4\end{pmatrix} + 0.982714\begin{pmatrix}-1\\2\end{pmatrix} + 0.856982\begin{pmatrix}4\\0\end{pmatrix}\right) = \begin{pmatrix}1.56538\\2.10073\end{pmatrix}$$

$$\mu 2 = \frac{1}{1.1603}\left(0\begin{pmatrix}2\\4\end{pmatrix} + 1\begin{pmatrix}-1\\-4\end{pmatrix} + 0.0172856\begin{pmatrix}-1\\2\end{pmatrix} + 0.143018\begin{pmatrix}4\\0\end{pmatrix}\right) = \begin{pmatrix}-0.383704\\-3.41758\end{pmatrix}$$

Covariance matrices

$$\Sigma 1 = \begin{pmatrix}4.13282 & -1.16337\\-1.16337 & 2.6056\end{pmatrix},\ \Sigma 2 = \begin{pmatrix}2.70166 & 2.10624\\2.10624 & 2.16924\end{pmatrix}$$

Mixing parameter equal to N/N_k

$\pi 1 = 0.709924$
$\pi 2 = 0.290076$

TÉCNICO LISBOA

Aprendizagem 2021/22

Homework IV

Deadline 14/11/2021 23:59 via Fenix as PDF

2) [3v] Compare the quality of the produced clustering solutions using silhouette.

$$C1 = \{\mathbf{x}1, \mathbf{x}3, \mathbf{x}4\}, C2 = \{\mathbf{x}2\}$$

a)

if a < b we can use the simplified version **s = 1 – a/b**

(for the case a >= b s = b/a – 1 that is not present here, seldom present)

$$a(\mathbf{x_1}) = \frac{1}{2}(\|\mathbf{x}_1 - \mathbf{x}_3\|_2 + \|\mathbf{x}_1 - \mathbf{x}_4\|_2) = 4.03884 \qquad b(\mathbf{x_1}) = \|\mathbf{x}_1 - \mathbf{x}_2\|_2 = 8.544$$

$$s(\mathbf{x_1}) = 1 - \frac{a(\mathbf{x_1})}{b(\mathbf{x_1})} = 0.527289$$

$$a(\mathbf{x_3}) = \frac{1}{2}(\|\mathbf{x}_3 - \mathbf{x}_1\|_2 + \|\mathbf{x}_3 - \mathbf{x}_4\|_2) = 4.49536 \qquad b(\mathbf{x_3}) = \|\mathbf{x}_3 - \mathbf{x}_2\|_2 = 6$$

$$s(\mathbf{x_3}) = 1 - \frac{a(\mathbf{x_3})}{b(\mathbf{x_3})} = 0.250774$$

$$a(\mathbf{x_4}) = \frac{1}{2}(\|\mathbf{x}_4 - \mathbf{x}_1\|_2 + \|\mathbf{x}_4 - \mathbf{x}_3\|_2) = 4.92865, \qquad b(\mathbf{x_1}) = \|\mathbf{x}_4 - \mathbf{x}_2\|_2 = 6.40312$$

$$s(\mathbf{x_4}) = 1 - \frac{a(\mathbf{x_4})}{b(\mathbf{x_4})} = 0.230274$$

$$s(C1)=((\ s(\mathbf{x_1}) + s(\mathbf{x_3}) + s(\mathbf{x_4})\ )/3 = 0.336133$$

b)

$$a(\mathbf{x_2}) = 0$$

$$s(\mathbf{x_2}) = 1$$

s(C2)=1 *or s(C2)=0 according to some other definitions*

c)

s(Cl)=(1+0.336133)/2=0.668056

the quality of the produced clustering solutions using silhouette is good since the value is close to one.

According to the other definition we punish the clusters with one element

s(Cl)=(0+0.336133)/2=0.1681

the quality of the produced clustering solutions using silhouette averge since the value is close to zero.

TÉCNICO LISBOA

Aprendizagem 2021/22

Homework IV

Deadline 14/11/2021 23:59 via Fenix as PDF

3) [3v] Identify the VC dimension of the following two-class/binary classifiers: **i)** MLP with three hidden layers with as much nodes as the number of input variables; **ii)** decision tree assuming input variables are discretized using three bins; and **iii)** Bayesian classifier with a multivariate Gaussian likelihood.

(a) Assume the data dimensionality is five.

N=5

MLP with one output unit:

$$(n * n + n) * 3 + n + 1 = 1 + 4 * n + 3 * n^2 = \mathbf{96}$$

Or using two units to represent two classes:

$$(n * n + n) * 3 + 2 * (n + 1) = 2 + 5 * n + 3 * n^2 = \mathbf{102}$$

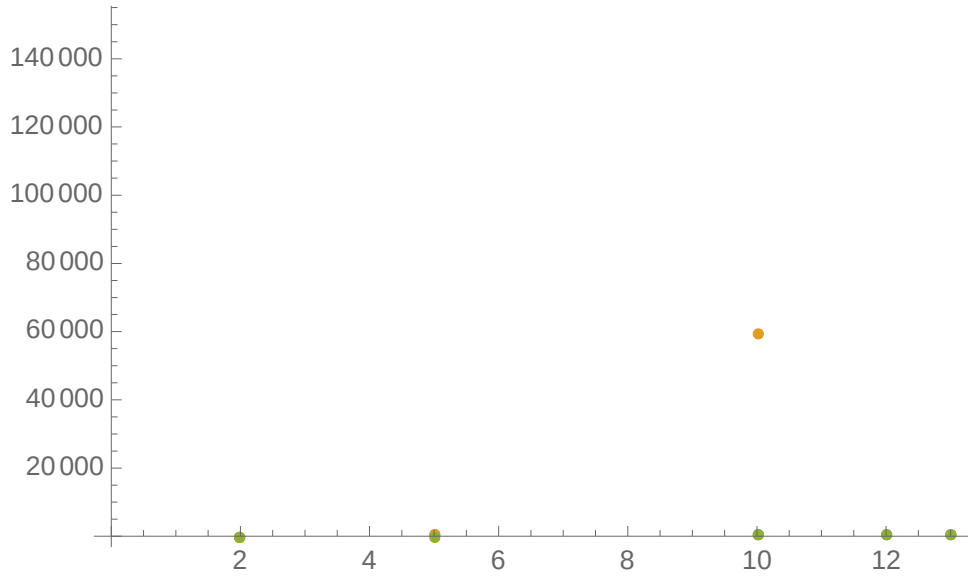
DT: $3^n = \mathbf{243}$

Bayesian classifier:

Prior 1, two Gaussians with n means and covariance matrix $n + (n * (n - 1)/2$

$$1 + 2 * (n + (n + n * (n - 1)/2) = 1 + 3 * n + n^2 = \mathbf{41}$$

(b) Plot in a single chart how the VC dimension varies with data dimensionality for $m \in \{2,5,10,12,13\}$. What can you conclude (one sentence, English or Portuguese)?,



DT grows exponentially

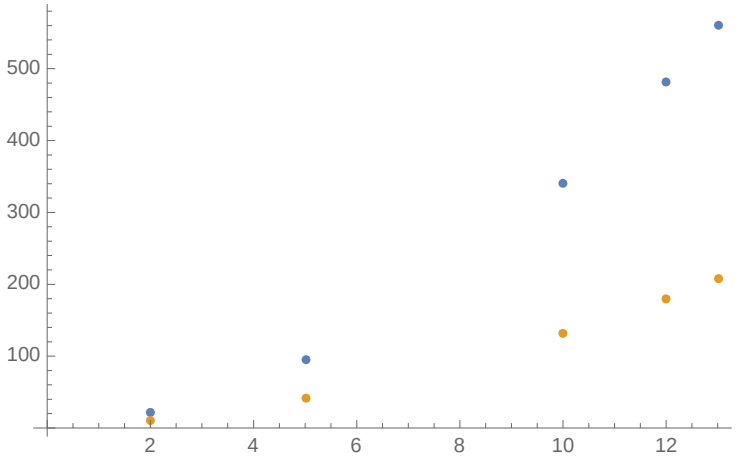
In detail comparing MLP and for one unit and Bayesian classifier

TÉCNICO LISBOA

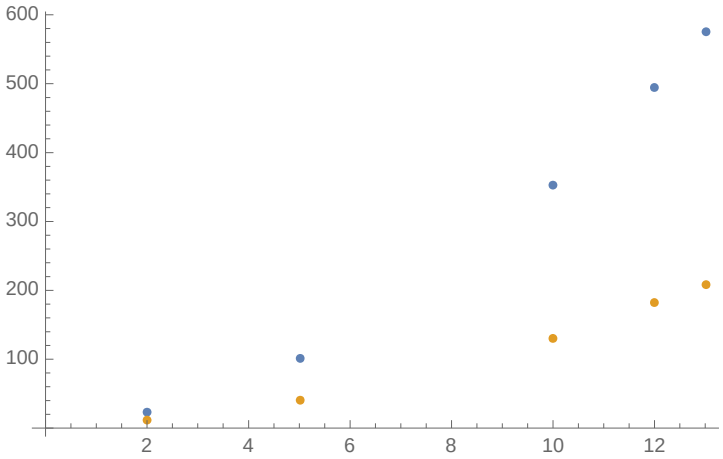
Aprendizagem 2021/22

Homework IV

Deadline 14/11/2021 23:59 via Fenix as PDF

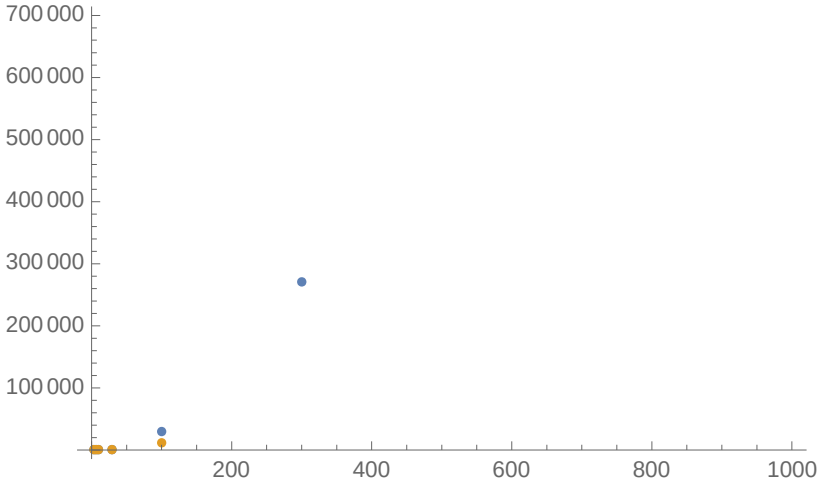


And for two units comparing MLP and for one unit and Bayesian classifier



MLP grows quadratic.

(c) Plot in a single chart how the VC dimension of **i)** and **iii)** with data dimensionality for $m \in \{2,5,10,30, 100,300,1000\}$. What can you conclude (one sentence, English or Portuguese)?,



MLP grows quadratic, much faster than Bayesian classifier

TÉCNICO LISBOA

Aprendizagem 2021/22

Homework IV

Deadline 14/11/2021 23:59 via Fenix as PDF

II. Programming and critical analysis [8v]

Recall the `breast.w.arff` dataset from previous homeworks.

4) [4v] Apply k-means clustering unsupervised on the original data with $k = 2$ and $k = 3$.

   a. Compare the produced solutions against the ECR (external measure)

   For k-means with k = 2 the ECR is 13.5 and for k-means with k = 3 the ECR is 6.(6).

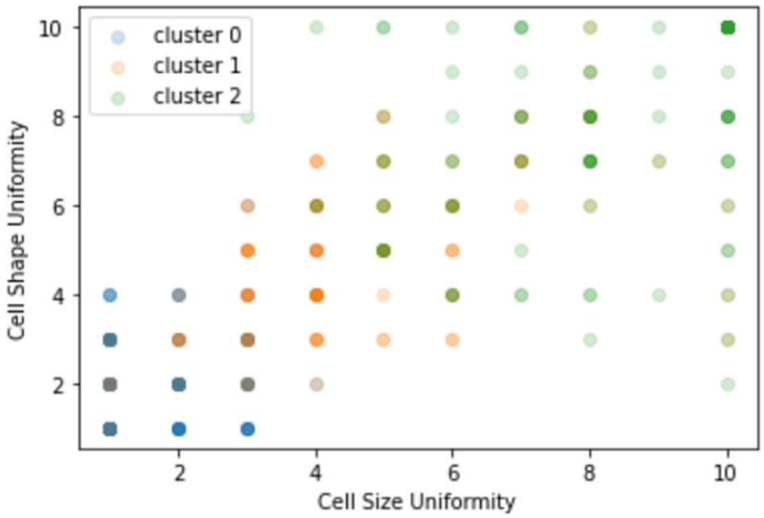
   - ECR (Error Classification Rate) is calculated using prior knowledge of the classes of each observation as it tells us the mean number of observations that are wrongfully classified in a cluster
   - For this dataset, 3-means clustering has a lower error rate
   - However, ECR is not enough to decide which k to choose since it is easy to minimize the error rate by adding more clusters, we can obtain a zero-error rate by having the same number of clusters as the number of samples

   b. Compare the produced solutions against the Silhouette coefficient (internal measure).

   For k-means with k = 2 the silhouette is 0,596798 and for k-means with k = 3 the silhouette is 0.524543.

   - the silhouette is an internal measure and is calculated using the mean intra-cluster distance (a) and the mean nearest-cluster distance (b) for each observation, combining the idea of cohesion and separation
   - the silhouette returns a value in the interval [-1, 1]. The higher the silhouette value, the more separated and compact the clusters are. Despite the small difference, 2-means clustering leads to a solution with more compact and better separated clusters

5) [2v] Visually plot the $k = 3$ clustering solution using the top-2 features with higher mutual information.



TÉCNICO LISBOA

Aprendizagem 2021/22

Homework IV

Deadline 14/11/2021 23:59 via Fenix as PDF

The given plot is not the only accepted (required: x-axis and y-axis legend and colors legend).
The plot can be complemented with a visualization of the reference groups (class)

6) [2v] Using empirical results from (5), comment on the quality of the produced clustering solution.

- We are describing the clusters using a slice of only two features (chosen with higher mutual information) from a 10-dimensional dataset. For this reason, some observations may seem closer to a cluster that isn't the one they weren't assigned to, as well as overlapping phenomena
- Despite the overlapping, we can conclude that while the blue cluster appears to be contained on the left inferior part of the plot, the green and orange clusters seem to be less separated from each other
- Looking at the labels from each observation, we can assume that the blue cluster corresponds to the benign class while the orange and green clusters correspond to the malignant class (a scatterplot with the colors representing the labels of the observations is interesting to complement the analysis)
- Uniformity of the size and shape of cells, important criteria for the cancer diagnostic, are moderately correlated with the gathered clustering solution