TÉCNICO LISBOA

Aprendizagem 2023/24

Homework IV

Deadline 30/10/2023 (Monday) 23:59 via Fenix as PDF

–

I. Pen-and-paper [11v]

Consider the following observations, $\left\{\begin{pmatrix}1\\0.6\\0.1\end{pmatrix},\begin{pmatrix}0\\-0.4\\0.8\end{pmatrix},\begin{pmatrix}0\\0.2\\0.5\end{pmatrix},\begin{pmatrix}1\\0.4\\-0.1\end{pmatrix}\right\}$.

Let y_1 be described by a Bernoulli distribution, y_2 and y_3 by a multivariate Gaussian, $\{y_1\} \perp\!\!\!\perp \{y_2, y_3\}$, and the following mixture

$$\pi_1 = 0.5, \pi_2 = 0.5$$

$$Bin_1(p_1 = P(y_1 = 1) = 0.3),\ \ Bin_2(p_2 = P(y_1 = 1) = 0.7)$$

$$N_1\left(\mathbf{u}_1 = \begin{pmatrix}1\\1\end{pmatrix}, \mathbf{\Sigma}_1 = \begin{pmatrix}2 & 0.5\\0.5 & 2\end{pmatrix}\right),\ \ N_2\left(\mathbf{u}_2 = \begin{pmatrix}0\\0\end{pmatrix}, \mathbf{\Sigma}_2 = \begin{pmatrix}1.5 & 1\\1 & 1.5\end{pmatrix}\right).$$

1) [6v] Perform one epoch of the EM clustering algorithm and determine the new parameters. Hint: you can use a computer, however disclose all the intermediary results step by step.

E-Step

a) $p(\mathbf{x}|c_k) = p(x_1|p_k) \times N(x_2, x_3|\mathbf{u}_k, \mathbf{\Sigma}_k)$

where $N(\mathbf{x}|\boldsymbol{\mu}, \Sigma) = \frac{1}{2\pi\sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{u})^T \cdot \Sigma^{-1} \cdot (\mathbf{x}-\mathbf{u})\right)$

For simplicity we below write $p(\mathbf{x}|N_1)$ to denote $p(x_2, x_3|N_1)$

$$\begin{aligned} p(\mathbf{x}_1|N_1) &= 0.0200,\ \ p(\mathbf{x}_1|N_2) = 0.0837\\ p(\mathbf{x}_2|N_1) &= 0.0350,\ \ p(\mathbf{x}_2|N_2) = 0.0205\\ p(\mathbf{x}_3|N_1) &= 0.0479,\ \ p(\mathbf{x}_3|N_2) = 0.0389\\ p(\mathbf{x}_4|N_1) &= 0.0177,\ \ p(\mathbf{x}_4|N_2) = 0.0872 \end{aligned}$$

b) $p(c_k, \mathbf{x}) = \pi_k p(\mathbf{x}|c_k)$

c) $p(\mathbf{x}) = \sum_k p(c_k, x) = \pi_k p(\mathbf{x}|c_k)$

d) $p(c_k|\mathbf{x}) = \frac{p(c_k,\mathbf{x})}{p(\mathbf{x})}$

$$\begin{aligned} \gamma(c_{11}) &= p(c_1|\mathbf{x}_1) = 0.1926,\ \ \gamma(c_{12}) = p(c_2|\mathbf{x}_1) = 0.8074\\ \gamma(c_{21}) &= p(c_1|\mathbf{x}_2) = 0.6313,\ \ \gamma(c_{22}) = p(c_2|\mathbf{x}_2) = 0.3687\\ \gamma(c_{31}) &= p(c_1|\mathbf{x}_3) = 0.5518,\ \ \gamma(c_{32}) = p(c_2|\mathbf{x}_3) = 0.4482\\ \gamma(c_{41}) &= p(c_1|\mathbf{x}_4) = 0.1689,\ \ \gamma(c_{42}) = p(c_2|\mathbf{x}_4) = 0.8311 \end{aligned}$$

M-Step

$$N_1 = 1.545,\ \ N_2 = 2.455$$

updated Bernoulli parameters

$$Bin_1(p_1 = 0.234),\ \ Bin_2(p_2 = 0.667)$$

updated means

$$\mu_1 = \begin{pmatrix}0.141\\-0.105\end{pmatrix},\ \ \mu_2 = \begin{pmatrix}0.309\\0.210\end{pmatrix}$$

updated covariance matrices

$$\Sigma_1 = \begin{pmatrix}0.141 & -0.105\\-0.105 & 0.096\end{pmatrix},\ \ \Sigma_2 = \begin{pmatrix}0.108 & -0.089\\-0.089 & 0.104\end{pmatrix}$$

updated priors: $\pi_1 = 0.386,\ \ \pi_2 = 0.614$

Aprendizagem 2023/24

Homework IV

Deadline 30/10/2023 (Monday) 23:59 via Fenix as PDF

2) [2v] Given the new observation, $\mathbf{x}_{new} = \begin{pmatrix} 1 \\ 0.3 \\ 0.7 \end{pmatrix}$, determine the cluster memberships (posteriors).

$$p(c_k|\mathbf{x}) = \frac{p(x_1|p_k) \times N(x_2, x_3|\mathbf{u}_k, \mathbf{\Sigma}_k)}{p(\mathbf{x})}$$

$$p(c_1|\mathbf{x}) = 0.08, \;\; p(c_2|\mathbf{x}) = 0.92$$

3) [2.5v] Performing a hard assignment of observations to clusters under a ML assumption, identify the silhouette of the larger cluster under a Manhattan distance.

$$\begin{aligned} p(c_1|\mathbf{x}_1) = 0.133, \;\; p(c_2|\mathbf{x}_1) = 0.867 \\ p(c_1|\mathbf{x}_2) = 0.900, \;\; p(c_2|\mathbf{x}_2) = 0.100 \\ p(c_1|\mathbf{x}_3) = 0.666, \;\; p(c_2|\mathbf{x}_3) = 0.334 \\ p(c_1|\mathbf{x}_4) = 0.018, \;\; p(c_2|\mathbf{x}_4) = 0.982 \end{aligned}$$

$$\text{clusters}=\{c_1=\{\mathbf{x}_2, \mathbf{x}_3\}, c_2=\{\mathbf{x}_1, \mathbf{x}_4\}\}$$

$$s(\mathbf{x}_1) = 1 - \frac{a(\mathbf{x}_1)}{b(\mathbf{x}_1)} = 1 - \frac{\|\mathbf{x}_4 - \mathbf{x}_1\|_1}{\frac{\|\mathbf{x}_2 - \mathbf{x}_1\|_1 + \|\mathbf{x}_3 - \mathbf{x}_1\|_1}{2}} = 0.348$$

$$s(\mathbf{x}_2) = 0.469, \;\; s(\mathbf{x}_3) = 0.204, \;\; s(\mathbf{x}_4) = 0.348$$

$$s(c_1) = \frac{s(\mathbf{x}_2) + s(\mathbf{x}_3)}{2} = 0.336, \;\; s(c_2) = 0.348$$

4) [0.5v] Knowing the purity of the clustering solution is 0.75, identify the number of possible classes (ground truth).

Considering only the targets (observable classes), answer is $|C| \in \{2,3\}$

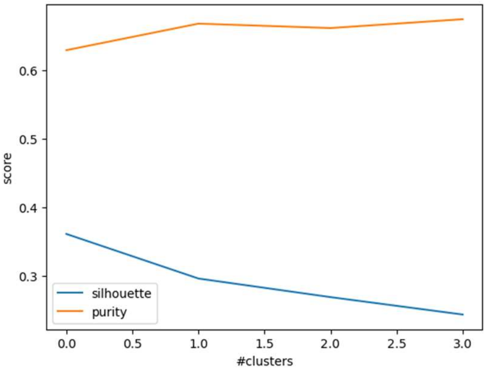
2 classes => e.g. $\{\{\mathbf{x}_2\}, \{\mathbf{x}_1, \mathbf{x}_3, \mathbf{x}_4\}\}$

3 classes => e.g. $\{\{\mathbf{x}_2\}, \{\mathbf{x}_3\}, \{\mathbf{x}_1, \mathbf{x}_4\}\}$

II. Programming and critical analysis [9v]

Recall the `pd_speech.arff` dataset from previous homeworks. For the following exercises, normalize the data using sklearn's MinMaxScaler.

1) [4v] Using `sklearn`, apply k-means clustering fully unsupervisedly on the normalized data with $k \in \{2,3,4,5\}$ (`random=0` and remaining parameters as default). Assess the silhouette and purity of the produced solutions.



```
def purity_score(y_true, y_pred):
    confusion_matrix = metrics.cluster.contingency_matrix(y_true, y_pred)
    return np.sum(np.amax(confusion_matrix, axis=0)) / np.sum(confusion_matrix)
```

```
X_scaled = MinMaxScaler().fit_transform(X)
for nclust in [2,3,4,5]:
    kmeans_model = cluster.KMeans(n_clusters=nclust, random_state=0).fit(X_scaled)
    silhouette = metrics.silhouette_score(X_scaled, kmeans_model.labels_, metric='euclidean')
    purity = purity_score(y, kmeans_model.labels_)
    print("k =",k,"\tsilhouette:",silhouette,"\n\tpurity:",purity)
```

2) [2v] Consider the application of PCA after the data normalization:

i. Identify the variability explained by the top two principal components.

Explained variance = 0.77

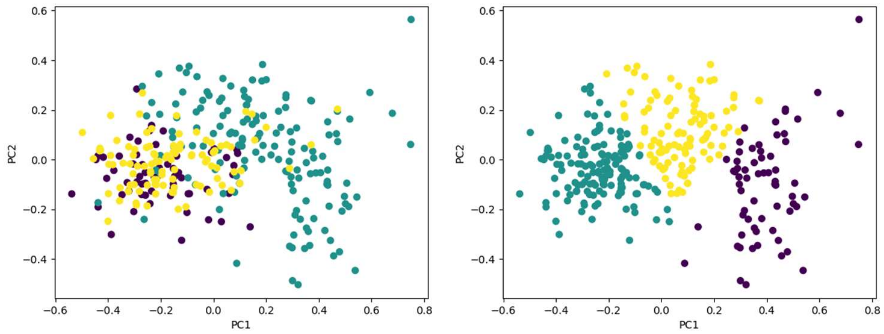
ii. For each one of these two components, sort the input variables by relevance by inspecting the absolute weights of the linear projection.

Feature importance in C1:
[(0.59, 'pelvic_incidence'), (0.51, 'lumbar_lordosis_angle'), (0.47, 'pelvic_tilt'),
(0.33, 'sacral_slope'), (0.22, 'degree_spondylolisthesis'), (0.12, 'pelvic_radius')]

Feature importance in C2:
[(0.67, 'pelvic_tilt'), (0.58, 'pelvic_radius'), (0.44, 'sacral_slope'), (0.10, 'pelvic_incidence'),
(0.08, 'lumbar_lordosis_angle'), (5E-3, 'degree_spondylolisthesis')]

```
pca = PCA(svd_solver='full')
pca.fit(X_scaled)
v1, v2 = pca.explained_variance_ratio_[0], pca.explained_variance_ratio_[1]
sorted(zip(np.abs(pca.components_[0]),X.columns),reverse=True))
sorted(zip(np.abs(pca.components_[1]),X.columns),reverse=True))
```

3) [2v] Visualize side-by-side the data using: i) the ground diagnoses, and ii) the *previously* learned $k = 3$ clustering solution. To this end, projected the normalized data onto a 2-dimensional data space using PCA and then color observations using the reference and cluster annotations.



4) [1v] Considering the results from questions (1) and (3), identify two ways on how clustering can be used to further describe the population of ill and healthy individuals.

Wide range of possibilities, including: 1) identifying heterogeneous groups within a diagnosis-specific populations (risk groups); 2) study populations with similar features yet different diagnoses to study confounding factors; 3) use cluster information to support classification (e.g. learn one classifier per cluster to better handle data heterogeneity; add cluster information to the input variables); 4) detect outlier individuals (e.g. small or sparse clusters) and treat/remove to aid learning; 5) …

END