Aprendizagem 2022/23

Homework IV – Group 71

(ist1103471, ist1104139)

I. Pen-and-paper

1) $a_1 = 1$ $a_2 = 0$ $a_3 = 0$ $a_4 = 1$

$b_1 = \begin{pmatrix} 0,6 \\ 0,1 \end{pmatrix}$ $b_2 = \begin{pmatrix} -0,4 \\ 0,8 \end{pmatrix}$ $b_3 = \begin{pmatrix} 0,2 \\ 0,5 \end{pmatrix}$ $b_4 = \begin{pmatrix} 0,4 \\ -0,1 \end{pmatrix}$

E-step $p(x_m | c_k) = p_k(a_m | c_k) \cdot N(b_m | u_k, \Sigma_k)$

$p(x_1|c_1) = 0,3\,(1-0,3)^{1-1} \times N\left(\begin{pmatrix} 0,6 \\ 0,1 \end{pmatrix} \middle| u_1, \Sigma_1\right) = 0,01997$

$p(x_2|c_1) = 0,03503$

$p(x_3|c_1) = 0,04786$

$p(x_4|c_1) = 0,01771$

$p(x_1|c_2) = 0,08373$

$p(x_2|c_2) = 0,02046$

$p(x_3|c_2) = 0,0388\,7$

$p(x_4|c_2) = 0,08715$

- Posteriors: $\gamma(c_{mk}) = p(c_k|x_m) = p(c_k, x_m) \setminus p(x_m)$

joint prob.: $p(c_k, x_m) = \pi_k \cdot p(x_m|c_k)$

$p(x_m)$: $\sum_{k=1}^{K} p(c_k, x_m)$

$\gamma(c_{11}) = \pi_1 \cdot p(x_1|c_1) \setminus (\pi_1 \cdot p(x_1|c_1) + \pi_2 \cdot p(x_1|c_2)) =$

$= 0,5 \times 0,01997 \setminus (0,5 \times 0,01997 + 0,5 \times 0,08373) =$

$= 0,19259$ $\gamma(c_{12}) = 0,80741$

$\gamma(c_{21}) = 0,63135$ $\gamma(c_{22}) = 0,36865$

$\gamma(c_{31}) = 0,55181$ $\gamma(c_{32}) = 0,44819$

$\gamma(c_{41}) = 0,16892$ $\gamma(c_{42}) = 0,83108$

TÉCNICO LISBOA

Aprendizagem 2022/23

Homework IV – Group 71

(ist1103471, ist1104139)

1) M-step

$$N_1 = \sum_{m=1}^{4} \gamma(c_{m1}) = 1,54$$

$$N_2 = \sum_{m=1}^{4} \gamma(c_{m2}) = 2,46$$

$$\pi_1 = \frac{N_1}{N} = \frac{1,56}{1,56+2,46} = 0,39$$

$$\pi_2 = \frac{N_2}{N} = 0,61$$

$$p_1 = \frac{1}{N_1} \cdot \sum_{m=1}^{4} (a_m \times \gamma(c_{m1})) = \frac{1}{0,39} \times (1\times 0,19 + 0\times 0,63 + 0\times 0,55 + 1\times 0,17)$$

$$* = 0,239$$

$$p_2 = \frac{1}{N_2} \cdot \sum_{m=1}^{4} (a_m \times \gamma(c_{m2})) = 0,667$$

$$u_1 = \frac{1}{N_1} \cdot \sum_{m=1}^{4} (\gamma(c_{m1}) \times b_m) = \begin{pmatrix} 0,02651 \\ 0,50713 \end{pmatrix}$$

$$u_2 = \frac{1}{N_2} \cdot \sum_{m=1}^{4} (\gamma(c_{m2}) \times b_m) = \begin{pmatrix} 0,30914 \\ 0,21042 \end{pmatrix}$$

$$\Sigma_1 = \frac{1}{N_1} \cdot \sum_{m=1}^{4} (\gamma(c_{m1}) \times (b_m - u_1) \times (b_m - u_1)^T) = \begin{bmatrix} 0,14137 & -0,10541 \\ -0,10541 & 0,09605 \end{bmatrix}$$

$$\Sigma_2 = \frac{1}{N_2} \cdot \sum_{m=1}^{4} (\gamma(c_{m2}) \times (b_m - u_2) \times (b_m - u_2)^T) = \begin{bmatrix} 0,10829 & -0,08865 \\ -0,08865 & 0,10412 \end{bmatrix}$$

2) $a_{new} = 1$

$$b_{new} = \begin{pmatrix} 0,3 \\ 0,7 \end{pmatrix}$$

$$p(x_{new} | c_1) = 0,239 (1-0,239)^{1-1} \times N\left(\begin{pmatrix} 0,3 \\ 0,7 \end{pmatrix} \middle| u_1, \Sigma_1\right) = 0,00634$$

$$p(x_{new} | c_2) = 0,667 (1-0,667)^{1-1} \times N\left(\begin{pmatrix} 0,3 \\ 0,7 \end{pmatrix} \middle| u_2, \Sigma_2\right) = 0,04567$$

Cluster memberships:

$$\gamma(c_{new\,1}) = 0,39 \times 0,006 \setminus (0,39 \times 0,006 + 0,61 \times 0,046) = 0,080//$$

$$\gamma(c_{new\,2}) = 0,61 \times 0,046 \setminus (0,39 \times 0,006 + 0,61 \times 0,046) = 0,920//$$

3) New likelihoods:

$p(x_1|c_1) = 0,231$ $\quad p(x_1|c_2) = 0,950$

$p(x_2|c_1) = 1,266$ $\quad p(x_2|c_2) = 0,089$

$p(x_3|c_1) = 1,438$ $\quad p(x_3|c_2) = 0,454$

$p(x_4|c_1) = 0,021$ $\quad p(x_4|c_2) = 0,723$

clusters $= \{c_1 = (x_2, x_3), c_2(x_1, x_4)\}$

$d(x_1, x_2) = |1-0| + |0,6+0,4| + |0,1-0,8| = 2,7$

$d(x_1, x_3) = |1-0| + |0,6-0,2| + |0,1-0,5| = 1,9$

$d(x_1, x_4) = |1-1| + |0,6-0,4| + |0,1+0,1| = 0,4$

$d(x_2, x_3) = |0-0| + |-0,4-0,2| + |0,8-0,5| = 0,9$

$d(x_2, x_4) = |0-1| + |-0,4-0,4| + |0,8+0,1| = 2,7$

$d(x_3, x_4) = |0-1| + |0,2-0,4| + |0,5+0,1| = 1,8$

For cluster c_1:

$a(x_2) = d(x_2, x_3) = 0,9$

$b(x_2) = \frac{d(x_1, x_2) + d(x_2, x_4)}{2} = 2,7$

$s(x_2) = 1 - \frac{a(x_2)}{b(x_2)} \approx 0,667 \quad s(x_3) = 1 - \frac{a(x_3)}{b(x_3)} \approx 0,514$

$s(c_1) = \frac{s(x_1) + s(x_2)}{2} = 0,591$

For cluster c_2:

$s(x_1) = 1 - \frac{a(x_1)}{b(x_1)} = 1 - \frac{0,4}{2,3} = 0,826$

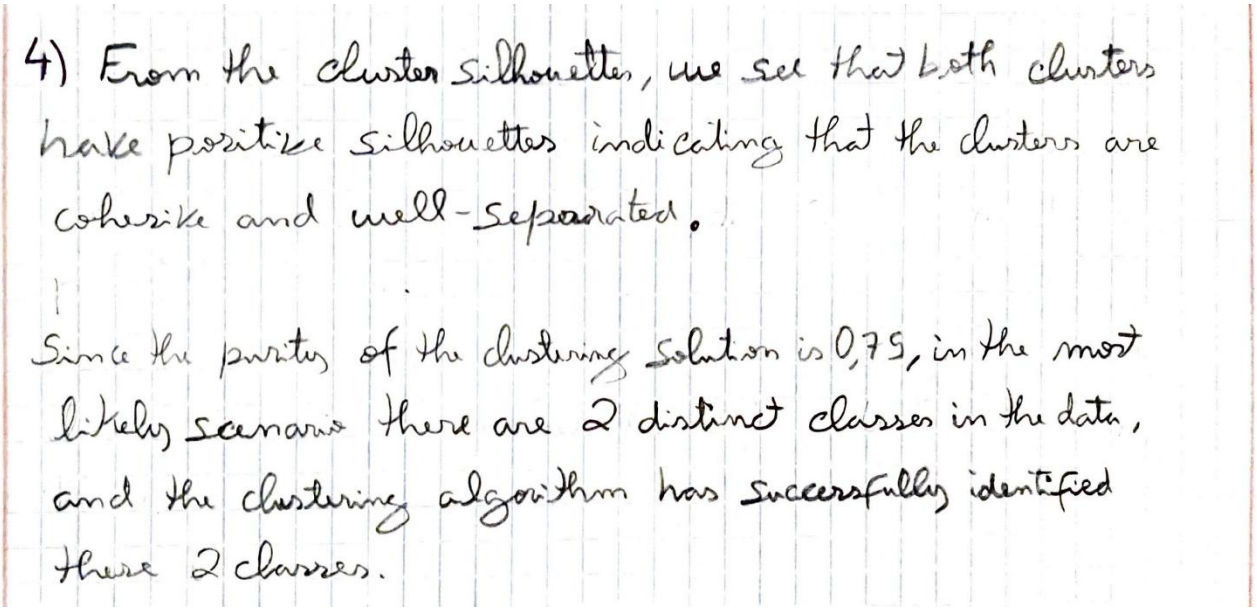
$s(x_4) = 1 - \frac{a(x_4)}{b(x_4)} = 1 - \frac{0,4}{2,25} = 0,822$

$s(c_2) = \frac{s(x_1) + s(x_4)}{2} = 0,824$

Aprendizagem 2022/23

Homework IV – Group 71

(ist1103471, ist1104139)



II. Programming and critical analysis

1)

```
from sklearn import datasets, metrics, cluster, mixture
k_means = [2,3,4,5]
silhouette_scores = []
purities = []
y_preds = []


def purity_score(y_true, y_pred):
    # compute contingency/confusion matrix
    confusion_matrix = metrics.cluster.contingency_matrix(y_true, y_pred)
    return np.sum(np.amax(confusion_matrix, axis=0)) / np.sum(confusion_matrix)_


for k in k_means:
    kmeans_algo = cluster.KMeans(n_clusters=k, random_state=0)
    kmeans_model = kmeans_algo.fit(X)
    y_pred = kmeans_model.labels_
    y_preds.append(y_pred)

    silhouette = metrics.silhouette_score(X, y_pred)
    silhouette_scores.append(silhouette)

    purity = purity_score(y, y_pred)
    purities.append(purity)

print("Silhouettes per k:", silhouette_scores)
print("Purities per k:", purities)
```

```
Silhouettes per k: [0.36044124340441114, 0.29579055730002257, 0.27442402122340176, 0.23823928397844843]
Purities per k: [0.632258064516129, 0.667741935483871, 0.6612903225806451, 0.6774193548387096]
```

The highest silhouette value is for k = 2, this could mean that the data is better separated in two clusters. Also, with the increase of the value of k, the silhouette score decreases, which means that the clusters are getting less distinct from each other.

As for the purity scores, these are relatively close to each other across all values of k. This suggests that regardless of the number of clusters chosen, the clusters tend to contain a variety of different classes within themselves.

2) i)

```
from sklearn.decomposition import PCA

pca = PCA(n_components=2)
X_pca = pca.fit_transform(X)

explained_variance = pca.explained_variance_ratio_
total_explained_variance = round(sum(explained_variance), 4) * 100
print('Explained Variance for first component:', explained_variance[0])
print('Explained Variance for second component:', explained_variance[1])
print(f'Total variability explained by the top two principal components: {total_explained_variance}%')
```

```
Explained Variance for first component: 0.5618144484299212
Explained Variance for second component: 0.20955952591361887
Total variability explained by the top two principal components: 77.14%
```

ii)

```
loadings = np.abs(pca.components_)

top_vars_pc1 = np.argsort(loadings[0])[::-1]
top_vars_pc2 = np.argsort(loadings[1])[::-1]
variable_names = df.columns

print("Ranked variables for first component:")
for idx in top_vars_pc1:
    print('-', variable_names[idx])

print("\nRanked variables for second component:")
for idx in top_vars_pc2:
    print('-', variable_names[idx])
```

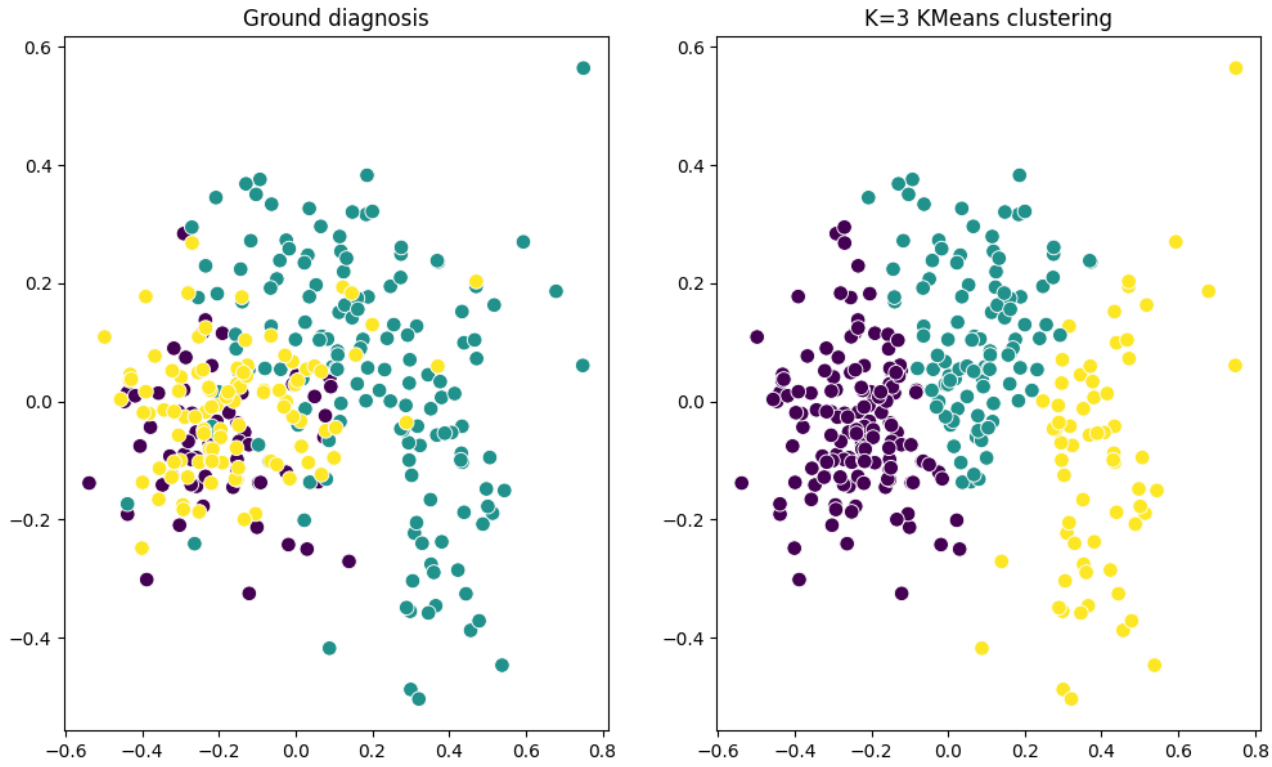
```
Ranked variables for first component:
- pelvic_incidence
- lumbar_lordosis_angle
- pelvic_tilt
- sacral_slope
- degree_spondylolisthesis
- pelvic_radius

Ranked variables for second component:
- pelvic_tilt
- pelvic_radius
- sacral_slope
- pelvic_incidence
- lumbar_lordosis_angle
- degree_spondylolisthesis
```

Homework IV – Group 71

(ist1103471, ist1104139)

3)

```
import matplotlib.pyplot as plt
codes = {'Hernia': 0, 'Spondylolisthesis': 1, 'Normal': 2}
true_y = y.map(codes).tolist()
plt.figure(figsize=(12, 7))
plt.subplot(121)
plt.title("Ground diagnosis")
sns.scatterplot(x=X_pca[:, 0], y=X_pca[:, 1], c=true_y, s=70)
plt.subplot(122)
plt.title("K=3 KMeans clustering")
sns.scatterplot(x=X_pca[:, 0], y=X_pca[:, 1], c=y_preds[1], s=70)
plt.show()
```



4) Since the silhouette values represent how well-separated clusters are, and purity scores measure how well-defined the clusters are in terms of the actual classes, we can conclude that clustering can be used to characterize a population of ill and healthy individuals. More precisely, the silhouette score can tell us if our sample provides a good separation between the health status of individuals, and the purity can measure if the clusters can actually represent the ill and the healthy. This is particularly important, as mistreating any kind of information in this area can result in problematic consequences. Looking at Exercise 1, since the best silhouette value is for k=2, this could mean that the 'column diagnosis' data can be better grouped into two different categories, which could be healthy ill.

By using PCA and clustering, visualizing the data in a reduced 2-dimensional space is possible. This allows us to assess if clusters are well-formed by looking at the points that comprise them specifically, in a way that is easily comprehensible. For example, in Exercise 3, we can see that for KMeans with a k value of 3, the points create clusters that appear quite concise. This implies that the algorithm can effectively differentiate groups from each other, much like characterizing a population by healthy and ill individuals.

END