Aprendizagem 2023/24

Homework IV – Group 001

(ist1102082, ist1103252)

I. Pen-and-paper

1)

1.

- Observations: $X = \left\{ \begin{pmatrix} 1 \\ 0.6 \\ 0.1 \end{pmatrix}, \begin{pmatrix} 0 \\ -0.4 \\ 0.8 \end{pmatrix}, \begin{pmatrix} 0 \\ 0.7 \\ 0.5 \end{pmatrix}, \begin{pmatrix} 1 \\ 0.4 \\ -0.1 \end{pmatrix} \right\}$

Most of our calculations were made using programming functions as a helpful resource.

- $y_1 \perp\!\!\!\perp \{y_2, y_3\}$ → Gaussian
  → Bernoulli
- $p_1 = P(y_1 = 1) = 0.3$
- $\pi_1 = 0.5$
- $N_1\left(\mu_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \Sigma_1 = \begin{pmatrix} 2 & 0.5 \\ 0.5 & 2 \end{pmatrix}\right)$
- $p_2 = P(y_1 = 1) = 0.7$
- $\pi_2 = 0.5$
- $N_2\left(\mu_2 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma_2 = \begin{pmatrix} 1.5 & 1 \\ 1 & 1.5 \end{pmatrix}\right)$

E-Step:

Cluster 1

- $p(x_1 | C=1) = p_1 \times N(x_1 | \mu_1, \Sigma_1) = p_1 \times \left(\frac{1}{\sqrt{2\pi \cdot |\Sigma_1|}} \cdot \exp\left(-\frac{1}{2} \cdot (x_1 - \mu_1)^T \cdot \Sigma_1^{-1} \cdot (x_1 - \mu_1) \right) \right) = 0.01997$
- $p(x_2 | C=1) = (1 - p_1) \times N(x_2 | \mu_1, \Sigma_1) = 0.03503$
- $p(x_3 | C=1) = (1 - p_2) \times N(x_3 | \mu_1, \Sigma_1) = 0.04786$
- $p(x_4 | C=1) = p_1 \times N(x_4 | \mu_1, \Sigma_1) = 0.01771$

X_n variable will only use y_2, y_3 (normally distributed)

Cluster 2

- $p(x_1 | C=2) = p_2 \times N(x_1 | \mu_2, \Sigma_2) = p_2 \times \left(\frac{1}{\sqrt{2\pi \cdot |\Sigma_2|}} \cdot \exp\left(-\frac{1}{2} \cdot (x_1 - \mu_2)^T \cdot \Sigma_2^{-1} \cdot (x_1 - \mu_2) \right) \right) = 0.08373$
- $p(x_2 | C=2) = (1 - p_2) \times N(x_2 | \mu_2, \Sigma_2) = 0.02046$
- $p(x_3 | C=2) = (1 - p_2) \times N(x_3 | \mu_2, \Sigma_2) = 0.03887$
- $p(x_4 | C=2) = p_2 \times N(x_4 | \mu_1, \Sigma_1) = 0.08715$

Aprendizagem 2023/24

Homework IV – Group 001

(ist1102082, ist1103252)

Joint Probabilities

Cluster 1

- $p(c=1, x_1) = \pi_1 \times N(x_1 \mid \mu_1, \Sigma_1) = 0.00997$
- $p(c=1, x_2) = 0.01752$
- $p(c=1, x_3) = 0.02393$
- $p(c=1, x_4) = 0.00886$

Cluster 2

- $p(c=2, x_1) = \pi_2 \times N(x_1 \mid \mu_2, \Sigma_2) = 0.04187$
- $p(c=2, x_2) = 0.01023$
- $p(c=2, x_3) = 0.01944$
- $p(c=2, x_4) = 0.04358$

Cluster 1

- $\gamma(c_{11}) = p(c=1 \mid x_1) = \frac{p(c=1, x_1)}{p(x_1)^*} = 0.19259$
- $\gamma(c_{21}) = 0.63135$
- $\gamma(c_{31}) = 0.55181$
- $\gamma(c_{41}) = 0.16892$

$$* \; p(x_n) = \sum_{k=1}^{K} p(c_{k=1}, x_n)$$

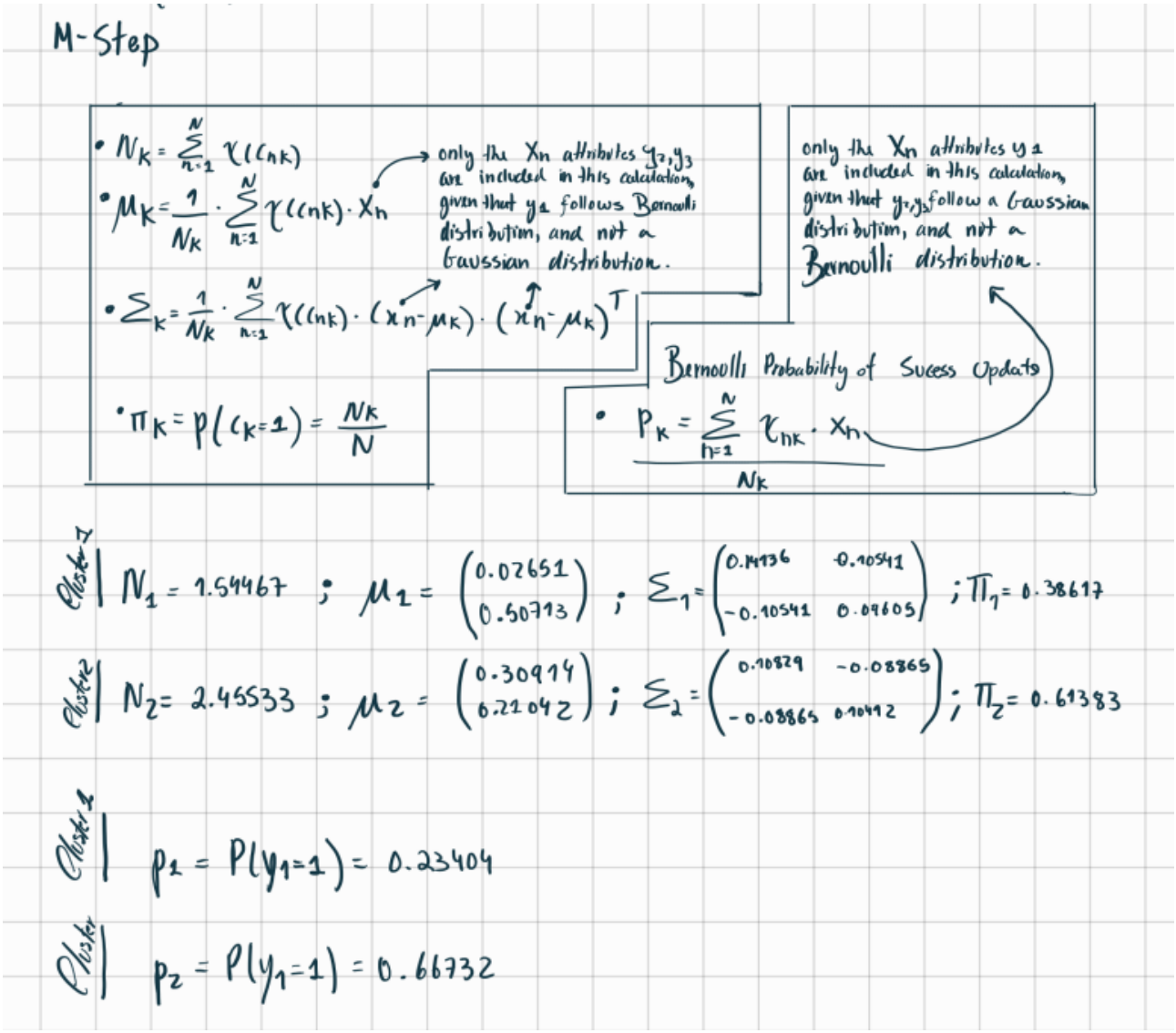
Cluster 2

- $\gamma(c_{12}) = p(c=2 \mid x_1) = \frac{p(c=2, x_1)}{p(x_1)^*} = 0.80741$
- $\gamma(c_{22}) = 0.36865$
- $\gamma(c_{32}) = 0.44819$
- $\gamma(c_{42}) = 0.83108$

Aprendizagem 2023/24

Homework IV – Group 001

(ist1102082, ist1103252)



Homework IV – Group 001

(ist1102082, ist1103252)

2)

2. $X_{new} = \begin{pmatrix} 1 \\ 0.3 \\ 0.7 \end{pmatrix}$ In this exercise, we will be using the updated values from Exercise 1!!

Cluster 1 | $p(x_{new} | C=1) = p_1 \cdot N(x_{new} | \mu_1, \Sigma_1) = 0.00634$

Cluster 2 | $p(x_{new} | C=2) = 0.04567$

Cluster 1 | $p(C=1, x_{new}) = \pi_1 \cdot N(x_{new} | \mu_1, \Sigma_1) = 0.00245$

Cluster 2 | $p(C=2, x_{new}) = \pi_2 \cdot N(x_{new} | \mu_2, \Sigma_2) = 0.02803$

$\rightarrow P(X_{new}) = P(C=1, X_{new}) + P(C=2, X_{new}) = 0.03048$

Cluster Memberships (Posteriors) for X_{new}:

Cluster 1 | $\gamma(C_{new1}) = \frac{P(C=1, X_{new})}{P(X_{new})} = 0.08029$

Cluster 2 | $\gamma(C_{new2}) = \frac{P(C=2, X_{new})}{P(X_{new})} = 0.91971$

3)

3. We will be using the updated values!

- $p(x_1|C=1) = p_1 \times N(x_1|\mu_1, \Sigma_1) = 0.23147$
- $p(x_1|C=2) = p_2 \times N(x_1|\mu_2, \Sigma_2) = 0.94954$

→ x_1 belongs to $C=2$

- $p(x_2|C=1) = (1-p_1) \times N(x_2|\mu_1, \Sigma_1) = 1.26633$
- $p(x_2|C=2) = (1-p_2) \times N(x_2|\mu_2, \Sigma_2) = 0.08874$

→ x_2 belongs to $C=1$

- $p(x_3|C=1) = (1-p_1) \times N(x_3|\mu_1, \Sigma_1) = 1.43811$
- $p(x_3|C=2) = (1-p_2) \times N(x_3|\mu_2, \Sigma_2) = 0.45417$

→ x_3 belongs to $C=1$

- $p(x_4|C=1) = p_1 \times N(x_4|\mu_1, \Sigma_1) = 0.02077$
- $p(x_4|C=2) = p_2 \times N(x_4|\mu_1, \Sigma_1) = 0.72331$

→ x_4 belongs to $C=2$

$C_1 = \{X_2, X_3\}$, $C_2 = \{X_1, X_4\}$

Since both clusters have the same number of observations, we will calculate the Silhouette for both of them.

Silhouette: $s(x) = \begin{cases} 1 - \frac{a(x)}{b(x)}, & b(x) > a(x) \\ \frac{b(x)}{a(x)} - 1, & c.c \end{cases}$

- a equals to medium distance from X_i to the points in the same cluster
- b equals to minimum medium distance from X_i to the points of another cluster, in this case, we only have two cluster.
- Important to know we will be using manhattan distance as our distance measurement

Manhattan Distance

$a(x_1) = d(x_1, x_4) = |1-1| + |0.6-0.4| + |0.1-(-0.1)| = 0.40000$

$a(x_2) = d(x_2, x_3) = 0.90000$

$a(x_3) = d(x_3, x_2) = 0.90000$

$a(x_4) = d(x_4, x_1) = 0.40000$

Homework IV – Group 001

(ist1102082, ist1103252)

$$b(x_1) = \frac{d(x_1,x_2) + d(x_1,x_3)}{2} = 2.25000$$
$$b(x_2) = \frac{d(x_2,x_1) + d(x_2,x_4)}{2} = 2.70000$$
$$b(x_3) = \frac{d(x_3,x_1) + d(x_3,x_4)}{2} = 1.80000$$
$$b(x_4) = \frac{d(x_4,x_2) + d(x_4,x_3)}{2} = 2.25000$$

Using the Silhouette function: ⟶

- $s(x_1) = 0.82222$
- $s(x_2) = 0.66667$
- $s(x_3) = 0.50000$
- $s(x_4) = 0.82222$

The silhoutte of a cluster corresponds to the mean of the silhouttes of the observations in that cluster.

- $s(C_1) = \frac{s(x_2) + s(x_3)}{2} = 0.58333$

- $s(C_2) = \frac{s(x_1) + s(x_4)}{2} = 0.82222$

4)

4.- The purity can be defined as a measurement that indicates the clustering's quality:

- A purity score closer to 1 reveals that our clustering is highly adequate.
- Since we are given a purity value of 0.75, we can conclude that 75 % of the observations were correctly assigned to their clusters.
- Thus, 25 % of our observations were assigned to the wrong cluster.
- Since we have 4 observations, only one of them (25 %) was incorrectly classified.
- Given that we don't know if that observation belongs to either cluster 1 or cluster 2, in the worst case scenario, it doesn't belong to neither of them.
- Therefore, we would need at most 3 classes – 1 for each cluster and a new class for the observation that was incorrectly classified.

Homework IV – Group 001

(ist1102082, ist1103252)

II. Programming and critical analysis

1)

Exercise 1

Loading the data from the column_diagnosis.arf file

```
import numpy as np
import pandas as pd
import warnings
from sklearn.preprocessing import MinMaxScaler
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score, completeness_score
from sklearn import metrics
from scipy.io.arff import loadarff

warnings.filterwarnings('ignore')

# Load the data
data = loadarff('./column_diagnosis.arff')
df = pd.DataFrame(data[0])
df['class'] = df['class'].str.decode('utf-8')
X = df.drop('class', axis=1)
y = df['class']
```

Homework IV – Group 001

(ist1102082, ist1103252)

```python
# Normalize the data using MinMaxScaler
scaler = MinMaxScaler()
X_normalized = scaler.fit_transform(X)

# Define a list of k values to try
k_values = [2, 3, 4, 5]

# Initialize lists to store silhouette scores and purity scores for each k
silhouette_scores = []
purity_scores = []

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)

# Apply k-means clustering for each k value
for k in k_values:
    kmeans = KMeans(n_clusters=k, random_state=0)
    kmeans.fit(X_normalized)

    # Get cluster labels
    cluster_labels = kmeans.labels_

    # Calculate silhouette score
    silhouette = silhouette_score(X_normalized, cluster_labels)
    silhouette_scores.append(silhouette)

    # Calculate purity score (completeness score since we have ground truth labels)
    purity = purity_score(y, cluster_labels)
    purity_scores.append(purity)

# Print the results
for k, silhouette, purity in zip(k_values, silhouette_scores, purity_scores):
    print(f'K={k}: Silhouette Score = {silhouette:.4f}, Purity Score = {purity:.4f}')
```

```
K=2: Silhouette Score = 0.3604, Purity Score = 0.6323
K=3: Silhouette Score = 0.2958, Purity Score = 0.6677
K=4: Silhouette Score = 0.2744, Purity Score = 0.6613
K=5: Silhouette Score = 0.2382, Purity Score = 0.6774
```

2)

```python
from sklearn.decomposition import PCA

# Fit a PCA model on the normalized data
pca = PCA(n_components=2)
pca.fit(X_normalized)

# i. Variability explained by top two principal components
explained_variance = pca.explained_variance_ratio_
print("Explained Variance for Top Two Principal Components:")
print(f"Component 1: {explained_variance[0]:.4f}")
print(f"Component 2: {explained_variance[1]:.4f}")

# ii. Sort input variables by relevance in top two principal components
# Extract the weights of the input variables for the top two components
component_weights = pca.components_

# Sort input variables by absolute loading for component 1
sorted_variables_component1 = [X.columns[i] for i in
                               np.argsort(np.abs(component_weights[0]))[::-1]]

# Sort input variables by absolute loading for component 2
sorted_variables_component2 = [X.columns[i] for i in
                               np.argsort(np.abs(component_weights[1]))[::-1]]


# Print the sorted input variables for each component
print("\nTop Two Principal Components and respective Absolute Weights:")
print(f"\nTop Two Variables for Component 1:")
for variable in sorted_variables_component1:
    print(f" - {variable}: {np.abs(component_weights[0][X.columns.get_loc(variable)]):.5f}")

print(f"\nTop Two Variables for Component 2:")
for variable in sorted_variables_component2:
    print(f" - {variable}: {np.abs(component_weights[1][X.columns.get_loc(variable)]):.5f}")
```

Homework IV – Group 001

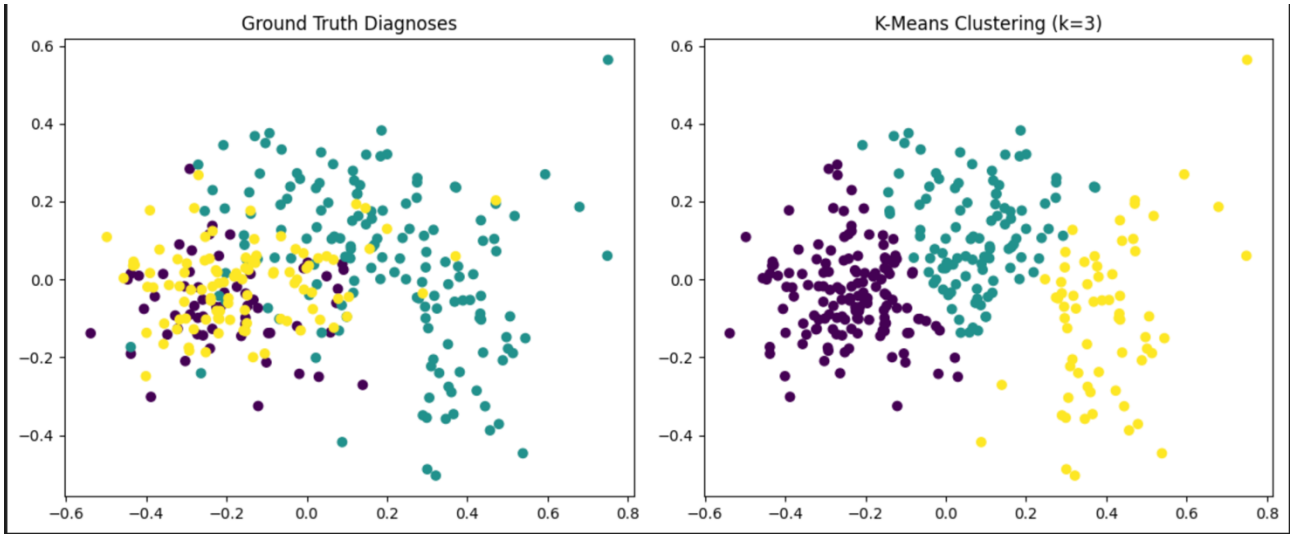
(ist1102082, ist1103252)

```
Explained Variance for Top Two Principal Components:
Component 1: 0.5618
Component 2: 0.2096

Top Two Principal Components and respective Absolute Weights:

Top Two Variables for Component 1:
 - pelvic_incidence: 0.59162
 - lumbar_lordosis_angle: 0.51508
 - pelvic_tilt: 0.46704
 - sacral_slope: 0.32569
 - degree_spondylolisthesis: 0.21693
 - pelvic_radius: 0.11582

Top Two Variables for Component 2:
 - pelvic_tilt: 0.67037
 - pelvic_radius: 0.58107
 - sacral_slope: 0.44330
 - pelvic_incidence: 0.10004
 - lumbar_lordosis_angle: 0.08005
 - degree_spondylolisthesis: 0.00458
```

3)

Exercise 3

```
import matplotlib.pyplot as plt
from sklearn.decomposition import PCA
from sklearn.cluster import KMeans

# Fit a PCA model on the normalized data with 2 components
pca = PCA(n_components=2)
X_2D = pca.fit_transform(X_normalized)

# K-means clustering labels with k=3
kmeans = KMeans(n_clusters=3, random_state=0)
kmeans_labels = kmeans.fit(X_normalized)

# Map diagnoses to 0, 1, 2
codes = {'Hernia': 0, 'Spondylolisthesis': 1, 'Normal': 2}
codes_labels = y.map(codes).tolist()

# Create a scatter plot for ground truth labels
plt.figure(figsize=(12, 5))
plt.subplot(121)
plt.scatter(X_2D[:, 0], X_2D[:, 1], c=codes_labels, label=codes)
plt.title("Ground Truth Diagnoses")

# Create a scatter plot for K-means clustering labels
plt.subplot(122)
plt.scatter(X_2D[:, 0], X_2D[:, 1], c=kmeans.labels_, label='K-Means Clustering')
plt.title("K-Means Clustering (k=3)")

plt.tight_layout()
plt.show()
```



Homework IV – Group 001

(ist1102082, ist1103252)

4)

Exercise 4

Considering the results from question (1) and question (3), clustering can be used to characterize the population of ill and healthy individuals in the following ways:

1. Identifying Subpopulations:
    - Clustering can help identify subpopulations within the dataset based on the similarities of their feature profiles. In the context of the our dataset, these subpopulations could represent different groups of individuals with varying degrees of health or illness.
2. Detection of Anomalies:
    - Clustering offers a useful approach for uncovering outliers or irregular cases within a population. In the realm of healthcare, these anomalies could signify individuals who exhibit uncommon or unexpected health conditions. By cross-referencing the groupings generated by clustering with the known diagnoses (ground truth), as mentioned in question 3, we can identify individuals who find themselves assigned to clusters that don't align with their actual medical conditions. This subset of individuals warrants closer examination, as they may be harboring undiagnosed or rare health issues. Thus, clustering proves to be a valuable tool for the early detection of diseases.

END