Problems 1: Unsupervised Learning

Machine Learning

Grado en Ingeniería Informática y Tecnologías Virtuales

Ingeniería de Datos I

Grado en Ingeniería del Software

Academic year 2025/2026

Antonio M. Durán Rosal

Problems 1: Unsupervised Learning

1. Problem 1 (1)

Two clustering algorithms produced the following assignments for eight samples:

- Algorithm 1: $3, 1, 1, 2, 3, 2, 1, 2$.
- Algorithm 2: $2, 1, 1, 3, 3, 2, 1, 2$.

Assume the ground-truth labels of the samples are known. Compute the external validation metrics *Rand Index (RI)*, *Adjusted Rand Index (ARI)*, and *Mutual Information (MI)* for each algorithm given that the true labels are $1, 1, 2, 2, 3, 3, 1, 3$. Finally, determine which algorithm performs better and justify your choice.

2. Problem 2 (1.5)

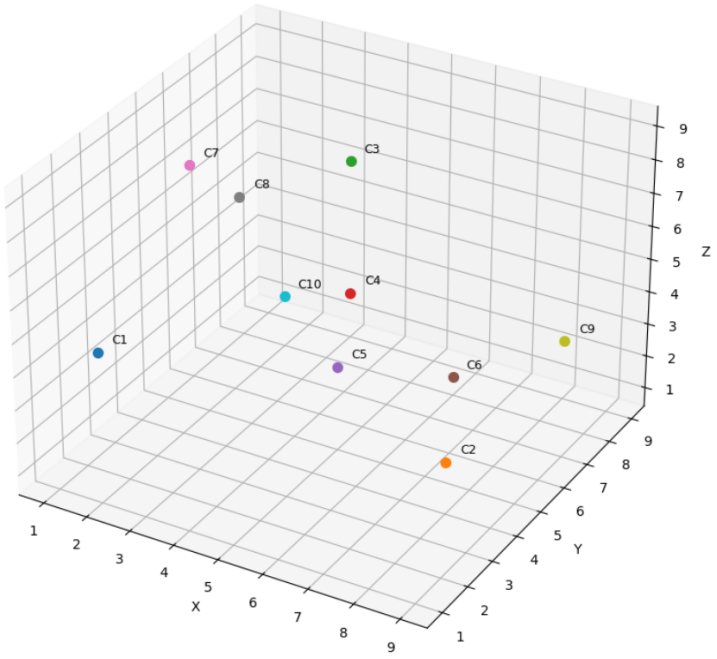
Using k-means, cluster 6 athletes into $k = 2$ groups according to their performance indicators. Scale all features to the same range before clustering. Use the Euclidean distance as the dissimilarity measure. Initialize k-means with the feature vectors of athletes B and C (after scaling) as the two initial centers.

Athlete	Speed	Strength	Endurance	Agility
A	0.2	0.7	0.3	0.4
B	0.9	0.8	0.6	0.9
C	0.4	0.2	0.5	0.3
D	0.8	0.4	0.7	0.6
E	0.3	0.9	0.2	0.5
F	0.6	0.5	0.8	0.7

3. Problem 3 (2)

Solve the following exercise using k-medians:

- A total of 10 three-dimensional patterns are to be grouped into two clusters ($k = 2$). The points are as follows: $C1(2, 1, 5)$, $C2(8, 4, 2)$, $C3(3, 9, 6)$, $C4(7, 2, 8)$, $C5(5, 5, 3)$, $C6(6, 8, 1)$, $C7(1, 6, 7)$, $C8(4, 3, 9)$, $C9(9, 7, 4)$, and $C10(2, 8, 2)$.
- The initial medians are points: $C3$ and $C7$.
- The distance metric used will be the Manhattan distance.
- Don't scale the data.
- It is requested to:
  a) Show the evolution of the procedure.
  b) Compute the value of the SSE metric.



4. Problem 4 (2)

Given the following table of scores in four different subjects for five students:

Student	Physics	Chemistry	Biology	Mathematics
A	20	15	60	42
B	30	18	40	58
C	24	20	50	65
D	15	14	72	38
E	27	16	48	55

... and using the Euclidean distance with prior transformation of the variables, it is requested to:

a) Apply hierarchical clustering with the single linkage method.

b) Apply hierarchical clustering with the complete linkage method.

c) Apply hierarchical clustering with the average linkage method.

d) Apply hierarchical clustering with the centroid linkage method.

e) Plot the dendograms of a), b), c), and d).

5. Problem 5 (1.5)

A dataset contains 12 two-dimensional points representing customer shopping behavior. We want to group them using the **DBSCAN** algorithm. The points are given as follows:

Point	X	Y
P1	1.0	1.2
P2	0.8	1.1
P3	1.2	0.9
P4	8.0	8.5
P5	8.2	8.3
P6	7.9	8.1
P7	5.0	1.0
P8	5.2	1.1
P9	5.1	0.9
P10	3.0	6.0
P11	3.1	6.2
P12	2.9	5.9

Using DBSCAN with parameters $\varepsilon = 0.5$ and $MinPts = 3$, it is requested to:

a) Identify the core points, border points, and noise points.

b) Form the clusters and list which points belong to each cluster.

c) Plot the resulting clusters with different colors, highlighting noise points separately.

6. Problem 6 (2)

A company collected data on 5 products described by four numerical features: production cost, selling price, customer satisfaction, and sales volume. The dataset is given below:

Product	Cost	Price	Satisfaction	Sales
P1	12	20	70	300
P2	15	25	65	280
P3	10	18	80	350
P4	20	30	60	260
P5	14	22	75	320

It is requested to:

a) Standardize the dataset so that all features have mean 0 and variance 1.

b) Compute the covariance matrix of the standardized data.

c) **Once the covariance matrix is obtained, use Python to calculate the eigenvalues and eigenvectors with the command:**

```
eigenvalues, eigenvectors = np.linalg.eig(cov_matrix)
```

d) Select the two principal components that explain most of the variance.

e) Project the original data onto the two-dimensional subspace defined by these components and provide the new coordinates of the products.

f) Plot the 2D representation of the products.