In the Name of God

Statisstical Pattern Recognition

Homework V

Unsupervised learning

Assignment Date: **6 Dey** Submission Deadline: **18 Dey**

Contents

1 Objectives and Precautions

In this homework you will learn:

- What are the differences between supervised and unsupervised learning.
- Implement K-means and GMM to cluster datasets.
- How to combine clustering algorithms with a feature extraction technique(PCA).

Keep in mind that you can only use these python libraries in your implementations, unless mentioned in the homework instructions:

- Pandas, Numpy and Matplotlib.pyplot

Also note that:

- You have 2 weeks to complete this homework.
- The home work instructions and datasets will be shared with you in your **Quera class**.
- Save your plots, results, and answers in any format you want, this will be your report.
- Save your code files and your report as a zip file and upload in Quera. (naming format is "your names.zip" for example "Achraf_Hakimi_Mohamed_Salah.zip")
- Late submission strategy is: 70 percent score for one day delay and 50 percent for 2 days delay. Submissions after 2 days will not be graded.
- Use only the python programming language.
- Feel free to ask your questions in the telegram channel.
- Do not copy other works, write your own code.

Thank you. Good Luck.

2 K-means

K-means is a popular machine learning and data mining algorithm that discovers potential clusters within a dataset. Finding these clusters in a dataset can often reveal interesting and meaningful structures underlying the distribution of data. K-means clustering has been applied to many problems in science and still remains popular today for its simplicity and effectiveness.

1. Load the given datasets (“blobs.csv , banana.csv , dartboard2.csv , twenty.csv , elliptical.csv“).
2. Perform K-means clustering on all given datasets using euclidean distance.
3. Visualize each dataset and report which datasets you think can be clustered well using K-means.
4. How do we choose k? implement the below methods:
5. Use the elbow method and plotting the total within-cluster variation against the number of clusters for k-means clustering with k in (2, . . . , 20).
6. Use Davies Bouldin Index(DBI) and plotting the score against the number of clusters for k in (2, . . . , 20). (Implementing this part from scratch will ensure you receive a bonus.)
7. Plot the clusters for each dataset after you performed the k-means algorithm on them with different k’s
8. After analyzing the plots produced by elbow method and DBI method, discuss the number of clusters that you feel is the best fit for each of the given datasets. Defend your answer with evidence from these two parts and their produced plots, and what you surmise about these datasets.
9. after clustering, use the real labels and clustered label and report the accuracy of clustering.

3 GMM

The Gaussian mixture model (GMM) is well-known as an unsupervised learning algorithm for clustering. Here, “Gaussian” means the Gaussian distribution, described by mean and variance; mixture means the mixture of more than one Gaussian distribution. GMM uses Expectation Maximization (EM) to train a GMM model. A GMM model can be employed to estimate the PDF of some samples (like a parametric density estimator)

1. Load the given datasets (“blobs.csv , banana.csv , dartboard2.csv , twenty.csv , elliptical.csv“).
2. Perform GMM clustering method on all given datasets.
3. Split datasets into train and test parts, then visualize each training and testing data before clustering.
4. Construct a GMM for clustering, with K = 1,5,10,15 Gaussian components, and train on Train Data.
5. For each k, plot the test and train data clustered by the GMM algorithm.
6. How do we choose the number of gaussian components? Search various ways of evaluating the quality of a clustering assignment for the GMM algorithm. Explain at least two metrics.
7. Report the best k based on the two metrics you found.
8. Compare the results you’ve gotten from k-means with GMM. For each dataset determine which method works better? Why?
9. after clustering, use the real labels and clustered label and report the accuracy of clustering.

4 PCA and Clustering algorithms

1. Load the given dataset (“Spellman.csv“).

2. for PCA components in range (2,10) and number of clusters (3,4), find the best configuration in term of DBI. (you will have 4 best configuration at last.)

3. the configuration can be like (PCA component , number of clusters(K-means and GMM) , DBI score)

4. Visualize your best configurations in 3D.