Biomedical Data Science & AI

Exercise sheet 6

SoSe2025

Prof. Dr. Holger Fröhlich, Jonas Botz, Dr. Shammi More, Diego Valderrama

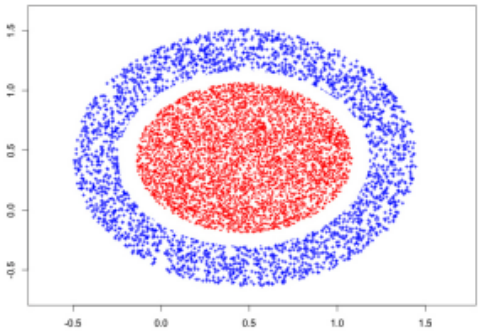
Due date: May 20th

Questions

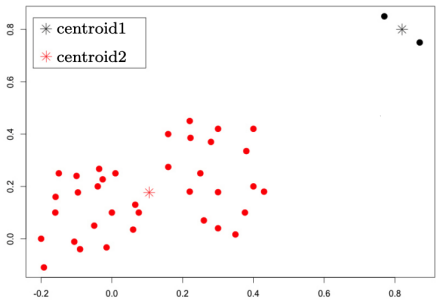
Exercise 1 – Clustering Algorithms (15P)

Theoretical questions

1. **K-Means (5P)**
    a. What are the advantages of k-means? (1P)
    b. Limitations of K-Means
        i. How can you counteract that K-Means depends on the initial conditions? (1P)
        ii. Why is K-Means not optimal if you have class overlap? (1P)
        iii. Consider this dataset. Why would K-Means have problems detecting the classes? (1P)



        iv. Consider this dataset. Why would K-Means have problems detecting classes? Which other algorithm would be better to use in this case? (1P)



2. **GMM for Classification (5P)**
    a. Explain the EM-Algorithm in your own words, without using any formula. Why do we need this algorithm and how does it work? (1P).
    b. Describe how to avoid the problem of getting stuck in local minima when using the EM algorithm. Write a pseudo-algorithm (no coding needed) to describe how to find the best set of clusters AND to reduce the local minima problem. (1P)
    c. The complexity of the GMM can be controlled by restricting how the covariance matrices are allowed to vary. Assume your data has p features and you want to cluster it into k clusters. (3P each one 1P)
        i. How many parameters (depending on the number of clusters) need to be estimated in the most general model (no restrictions on the covariances)?
        ii. Assume there is no correlation between the variables for each Gaussian. How many parameters does this model need to estimate?
        iii. Assume there is neither correlation nor does the variation for each feature change. How many parameters does the model have to estimate now?

3. **Consensus clustering and Non-Negative Matrix Factorization (5P)**
    a. Consensus clustering is used to address statistical instability in clustering. Briefly explain, in your own words, the steps of implementing consensus clustering to a dataset using any clustering method, and how it addresses statistical instability in clustering. (2P).
    b. What is the main constraint of applying non-negative matrix factorization (NMF) as a clustering technique? How does the algorithm work when used for clustering data? (1P).
    c. What are the advantages and drawbacks of NMF? (1P).
    d. How does one choose the appropriate number of clusters for a model based on silhouette index? (1P).

Exercise 2 – Programming task (10P)

Load the iris dataset from sklearn

Please use the random_seed = 2782 for all the questions.

1. Apply K-Means to the iris dataset with different K values i.e. 2-6. Do the following questions for all the different clusters. (3P)
    a. For each clustering, plot the cluster assignment and color-code accordingly within a scatter plot for the features “petal width” and “petal length”.
    b. For each clustering create silhouette plots and print out the score. You can make use of the sklearn library
2. Apply GMM to the iris dataset. (3P)
    a. Select the optimal number of clusters via BIC.
    b. For each clustering, plot the cluster assignment and color-code accordingly within a scatter plot for the features “petal width” and “petal length”.
    c. For each clustering create silhouette plots and print out the score. You can make use of the sklearn library.
3. Compare your results of the two different approaches and interpret them. Which of these methods is the best one to apply in this case? (1P)
4. Dimensionality Reduction Before Clustering (3P)
    a. Apply PCA to reduce the iris dataset to 2 dimensions before applying clustering
    b. Repeat both K-Means and GMM clustering on the PCA-transformed data using cluster numbers from 2 to 6
    c. Plot the cluster assignments using the PCA components
    d. Compute and plot the silhouette scores, but always the silhouette scored based on the original feature space
    e. Compare the clustering results with and without PCA and discuss whether PCA improves or degrades the clustering performance. Explain why this may be the case