Biomedical Data Science & AI

Exercise sheet 5 - Introduction

SoSe2021

Prof. Dr. Holger Fröhlich, Mohamed Aborageh, Vinay Bharadhwaj, Yasamin Salimi

Due date: May 25th

Questions

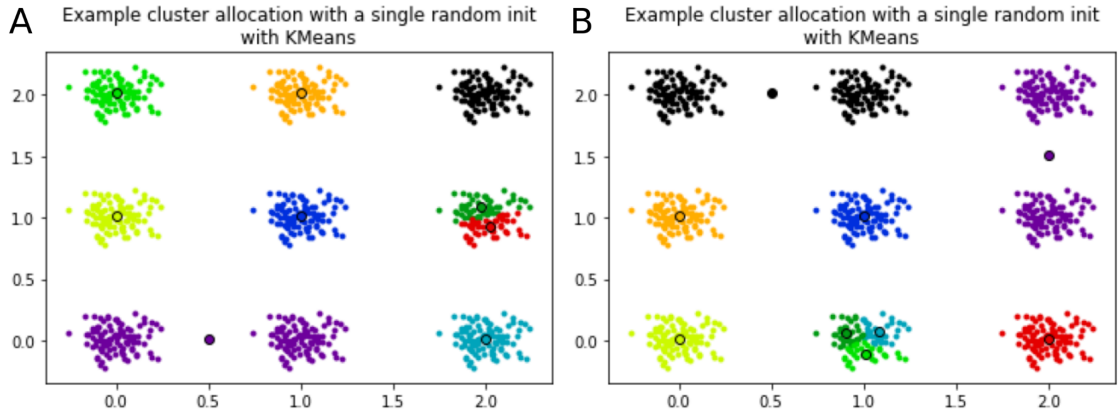
Exercise 1 - k-means clustering (11 points)

1. Use the K-means algorithm and Euclidean distance to cluster the 10 data points into K = 3 clusters. The coordinates of the data points are given in Table 1. Use the data points a4, a5, and a8 as initialization and perform 2 iteration steps. You can do the cluster assignment also visually without computing the exact distances. **(2 point)**

	a1	a2	a3	a4	a5	a6	a7	a8	a9	a10
(x, y)	(2,1)	(5,7)	(3,2)	(4,8)	(3,1)	(7,4)	(4,6)	(6,4)	(3,7)	(6,3)

Table 1: Coordinates of the data points.

2. Shown are the results of a k-means clustering with three different initializations:



   I. How does the choice of the initial starting points affect the clustering? **(1 point)**
   II. How can you avoid getting a clustering result that is dependent on the initialization? **(1 point)**
   III. What are the pros and cons of the k-means clustering? **(1 point)**

3. Use the provided breast cancer data (cancer.csv) to perform a k-means clustering. Perform the clustering for a range of clusters between 2 and 10. Set the random_state to 20 to keep reproducibility. **(2 point)**

    a. For each clustering plot the cluster assignment within a scatter plot for the features “mean radius” and “mean concavity”. **(1 point)**
    b. For each clustering create silhouette plots and print out the score. **(1 point)**
    c. Which is the best choice for the number of clusters? Why? **(1 point)**
4. Explain the difference between k-means and k-medoids. **(1 point)**

Exercise 2 - Gaussian mixture models (11 points)

1. Explain the EM-Algorithm in your own words, without using any formula. **(2 points)**
2. The complexity of the Gaussian mixture model can be controlled by restricting how the covariance matrices are allowed to vary. Assume your data has three features and you want to cluster it into 2 clusters. **(3 points)**
    a. How many parameters (depending on the number of clusters) need to be estimated in the most general model (no restrictions on the covariances)?
    b. Assuming that there is no correlation between the variables for each Gaussian, how many parameters does this model need to estimate?
    c. Assuming that there is neither correlation nor does the variation for each feature change. How many parameters does the model have to estimate now?
3. Cluster the breast cancer dataset (on the entire dataset: *cancer_all.csv*) with the help of a Gaussian mixture model. Perform the clustering for a range of clusters between 2 and 10 and for all possible assumptions for the covariance matrices. Plot the BIC of each clustering. **(2 points)**
    a. Which is the best choice for the clustering? Why?
    b. Plot the data (features “mean radius” and “mean compactness”), the cluster assignment and ellipses (to show the Gaussian component) for your selected model.
4. How does the k-means model differ from the GMM model? Which model would you prefer for the given data and why? **(1 point)**
5. Generate the K-Means model for the entire dataset and visualise both K-Mean and GMM models using PCA. **(2 points)**
6. What are the advantages of GMMs over k-means? **(1 point)**

Exercise 3 - Consensus clustering (3 points)

1. Perform (k-means) consensus clustering of samples for the given gene expression data *allData.csv*. Take minimum clusters as 2, maximum clusters as 6, resampling proportion as 80% and number of iterations as 10. Find the following:
    a. Best number of clusters **(1 point)**

b. Change in area under CDF **(1 point)**
c. Best cluster from the consensus matrix for each sample **(1 point)**