CO544 : Machine Learning and Data Mining
Nawarathna K.G.I.S. : E/17/219
Lab 07

(a) The completion of the first task (the four labs)
State in your report that you have completed all the four labs and one takeaway point for each lab in your own words.

Lab	Goal of the Lab	What I have learned(takeaways)
Lab 1	Understanding of basic numpy and matplotlib libraries and Concepts of Uncertainty of an estimation, Bivariate Gaussian Distribution and Sampling from a Gaussian Distribution	- When the number of samples is increased in a sample , the corresponding will more likely to be a uniform distribution - When increasing the number of bins, the number of data points that are in the margin of bins might be lost due to approximations. Also the number of points in a bin increases and generally this is not an accurate distribution of the data.
Lab 2	Implementation of the perceptron algorithm and Learning more of Gaussian Densities	- If the data set used does not fall into a linearly separable data set then the perception will not categorise input data appropriately. This mainly happens because perception is not a linear classifier.
Lab 3	Implementation of the linear regression model, Meaning and implementation of Regularisation and Sparse regression	- This is one of the best methods of predicting the y value when the x values are given. - Linear regression is based on probability and more specifically conditional probability.
Lab 4	Most useful functions in matplotlib, Seaborn, and pandas libraries when it comes to Machine Learning and data mining.	- Boxplot is a good measurement of dataset parameters such as locality, skewness groups and spread. - Whiskers in boxplot is helpful to get an idea of variability.

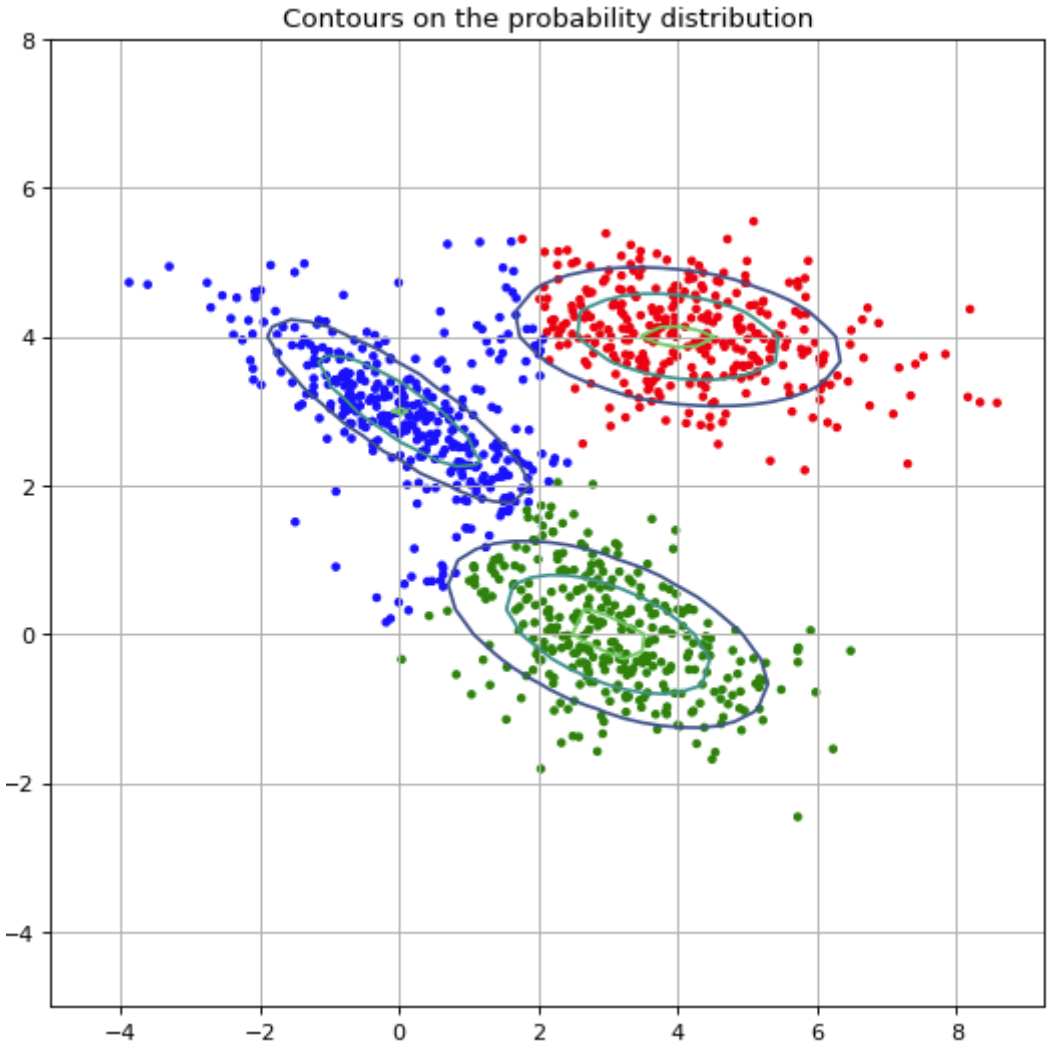
(b) any one of the remaining three defined

K-means clustering

1. Sample data from a mixture Gaussian density and implement K−means clustering algorithm. Snippet of code is provided in Appendix to randomly set means, covariances and proportions to define a model, and to sample from it (for K = 3). You have to write code for the K−means iterations yourself. The implementation of your algorithm should produce results similar to that in Fig. 1.

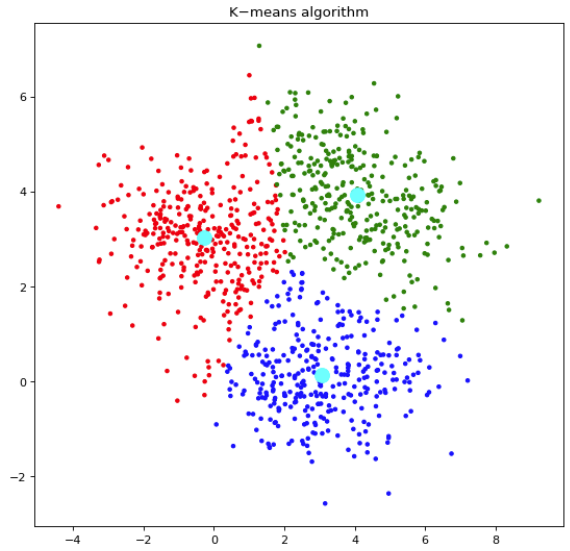
2. Draw contours on the probability density you have used and compare with regions associated with each cluster.

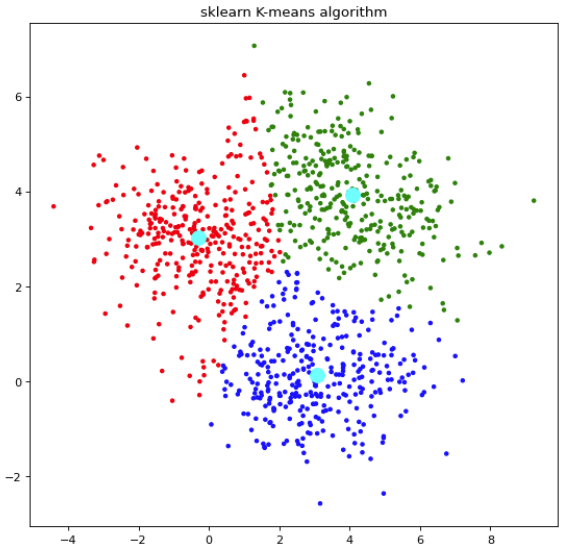
- Each cluster fits perfectly into the corresponding cluster.



3. Compare your results with the K−means clustering algorithm in sklearn.

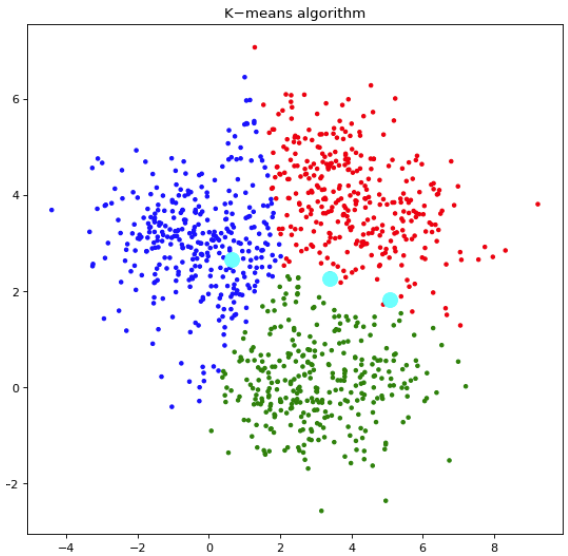
- There is no noticeable difference between the two plots.

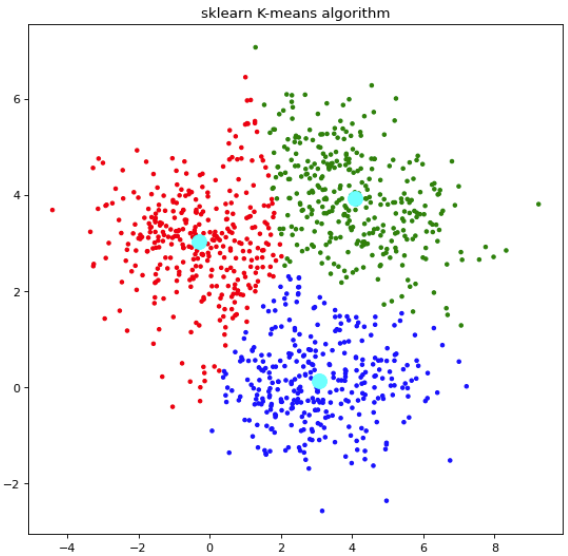




4. It is said that the K−means algorithm is sensitive to the initial guess of the cluster centres and the choice of K. Is this the case in your implementation? Show an example of the algorithm failing.

- The initial guess of centroids are very important for accurate clustering. It is also a good rule of thumb to choose one of the values in the dataset as the initial guess. This value can be chosen as some conditions or randomly. Below plots show a perfect showcase of algorithm filling where centroids are way out of their corresponding clusters.





5. Select a K−class classification dataset from the UCI repository, cluster the input data using K−means clustering and check how well the clusters relate to the targets defined in the dataset.