BİL 470/570
HW 3

TOBB ETU
9 Aug 2023
Deadline:
19 Aug 2023 23:59

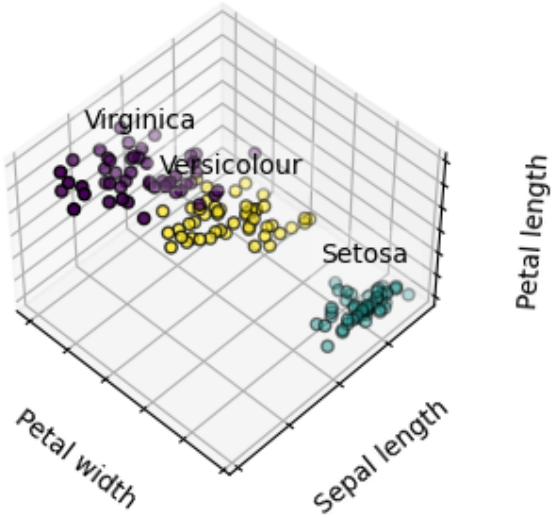

Figure 1: k-means clustering on iris dataset

In this assignment you are going to implement a **k-means clustering for classification**. You will train the model on the iris dataset (same dataset as in HW1) and **interpret the result** of the classification while comparing the results with the decision tree classifier which is the model in HW1.

1 Tasks

1.1 K-means Clustering Classifier

Train the K-means Cluster Classifier you have learnt in this course. To determine k value, use the **elbow method** and plot 'the number of cluster' versus 'sum of squared distances of samples to their closes cluster center'. It should be like Fig. 2. To calculate distance between samples and centroids, use **Euclidean distance**.

The signature of the aforementioned classifier will be as follows:

- `KMeansClusterClassifier(n_cluster)`
  - `fit(X, y)`
  - `predict(X)`

Train the classifier using the first %80 of the data and test it with the remaining data.

You **cannot use any libraries** to implement the `KMeansClusterClassifier`. It should work with builtin types. For vectors you can use lists, and for 2D input you can use list of lists.

BİL 470/570
HW 3

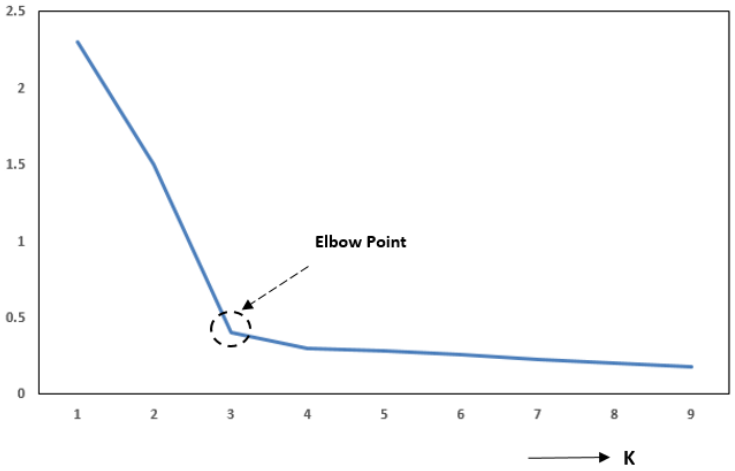

Figure 2: Elbow method for selection of optimal k cluster

1.2 Results

- Plot the 3D cluster plot as shown in Fig. 1
- Display confusion matrix
- Calculate following metrics:
  - F1-Score
  - Accuracy
  - Precision
  - Recall
- Plot the receiver operating characteristic (ROC) curve and calculate area under the ROC curve (AUC)
- Comment on these results [1].
- **Compare these results** with the output of decision tree classifier which is implemented in HW1. Comparison should give the idea of that why one of them is better than the other one, what is the advantages and disadvantages of using these methods, in which situation which one is useful.

You can add additional plots or tables to elaborate results more.
You can use libraries like pandas, numpy, seaborn, matplotlib to show your results in this part. Feel free to add additional libraries if you want.

[1] All results should be given for both training and test data

BİL 470/570
HW 3

TOBB ETU
9 Aug 2023
Deadline:
19 Aug 2023 23:59

2 Submission

You are to submit 3 files:

1. `Python file (kmeans.py)`: Contains the implementation of `KMeansClusterClassifier`. It should have sufficent number of comments to explain the code you wrote. Using docstring is also plus. You **cannot use any libraries in this file**.
2. `Notebook file (report.ipynb)`: Contains 2 part; (1) training of the classifier and, (2) interpretation and comparison of the results. You can use markdown syntax to explain steps, write python code to train the model, plot the graphs and tables.
3. `Report (report.pdf)`: PDF export of the corresponding report.ipynb file. This file should have same content with the notebook file. You can create this file from File > Download as > .pdf from the menu of the jupyter notebook.

3 Notes for development environment

As you know, we will use python3 for this class. You need to setup few things in order to get going:

- Pick a IDE (VSCode is suggested).
- Install python or install anaconda instead because you can use conda environments in your project. Conda also contains python
- Install jupyter notebooks, if you install python via anaconda, this step can be ignored.

Feel free to ask questions on piazza.

Academic Integrity

This assignment is an individual assignment and cannot be done in groups. The originality of your work should not be taken from a person or source. Demo for your assignments may be asked nd your homework grade will be given based on your demo performance. If you need supervision, you can apply to the assistant or the lecturer of the course. The homework grade of students who are found to be cheating is considered 0 and a disciplinary measures will be taken. In order not to put yourself and your friends in a difficult situation, you should take the necessary care in homework.