Fall 2022 5710 Machine Learning: Assignment 4

Alekhya Nalla

700734522

Programming elements: Linear Regression, K-Means Clustering and Data Analysis

In class programming:

1. Apply Linear Regression to the provided dataset using underlying steps.

a. Import the given “Salary\_Data.csv”

Graphical user interface, text, application, email

Description automatically generated

b. Split the data in train\_test partitions, such that 1/3 of the data is reserved as test subset.

Graphical user interface, text, application, email

Description automatically generated

c. Train and predict the model.

Graphical user interface, text, application

Description automatically generated

d. Calculate the mean\_squared error

Graphical user interface, text, application, email

Description automatically generated

e. Visualize both train and test data using scatter plot.

\*\*\*Train plot\*\*\*

Chart, scatter chart

Description automatically generated

\*\*\*\*\*Test data plot\*\*\*\*\*\*\*

Chart, scatter chart

Description automatically generated

2. Apply K means clustering in the dataset provided:

a. Remove any null values by the mean.

Graphical user interface, text, application

Description automatically generated

b. Use the elbow method to find a good number of clusters with the K-Means algorithm

Graphical user interface, text, application, email

Description automatically generated

Chart, line chart

Description automatically generated

C. Calculate the silhouette score for the above clustering

Graphical user interface, text, application

Description automatically generated

3. Try feature scaling and then apply K-Means on the scaled features. Did that improve the Silhouette score? If Yes, can you justify why

Graphical user interface, text, application, email

Description automatically generated

The silhouette score is poor after applying the K-Means on the scaled features of the given dataset.

Reasons:

1. The value is near to zero ,that indicates there is no significant difference between the clusters and it denotes there would be overlapping of clusters, where a high value indicates that the object is well matched to its own cluster.
2. Also, the data is not accurate and does not follow normal distribution.