Machine Learning Assignment - 2

Instructions.

- This coding assignment covers supervised and unsupervised learning basics.
- All implementations must be from scratch by using only **numpy** and **math** libraries, for plotting use **matplotlib**, not following will fetch you **0 marks**
- No late submissions are allowed.
- Submit a single ipynb file with naming convention as name-rollno-assignment-2.ipynb.

Dataset link kaggle.

Supervised Learning

Load dataset Salary.csv and perform regression analysis on this dataset with following conditions, for checking goodness of fit use R-square method.

1. ⟨ **3 Marks** ⟩ Use Least Square method to find 3 best fit lines using "Education" as independent and "Salary" as dependent variable, "Occupation" as independent and "Salary" as dependent variable and "Experience" as independent and "Salary" as dependent variable. Plot the best fit lines for all three case.
2. ⟨ **2 Marks** ⟩ Check goodness of fit by R-square method for all three fits. Comment your observation.
3. ⟨ **2 Marks** ⟩ Use above mentioned dependent and independent variables to perform a multiple linear regression and compare its performance with polynomial regression. Comment your observation
4. ⟨ **3 Marks** ⟩ Identify all relevant data for salary prediction and perform principal component analysis (PCA) on the data, and again perform linear regression on new transformed data and compare results with above two methods. Comment your observation.
5. ⟨ **3 Marks** ⟩ Check for orthogonality in the relevant data matrix if not make them orthogonal and then perform multiple regression and compare the results with above results and comment your observation.
6. ⟨ **2 Marks** ⟩ Plot original "Experience" and orthogonalized "Experience" variable as scatter plot and comment on changes in the variable data after orthogonalization.

Unsupervised Learning

Load Airplane.csv and perform clustering analysis on this dataset with following conditions, for checking correctness of clustering implement silhouette score from scratch and use it.

1. ⟨ **4 Marks** ⟩ Do basic data pre-processing on the data and perform PCA (scratch implementation) and tell what percentage of variance is covered by first principal component, first and second principal component.
2. ⟨ **4 Marks** ⟩ Implement K-means (scratch implementation) clustering for the column "Type" in data with (K=Number of classes) and plot the results.
3. ⟨ **2 Marks** ⟩ Perform above implemented K-means clustering on columns "Location", "Operator" in data and report the results, discuss on your findings.
4. ⟨ **3 Marks** ⟩ Using column "Route" from dataset and above implemented clustering algorithm tell which route is the most dangerous (here route with most crashes), plot and write your observation.
5. ⟨ **2 Marks** ⟩ By using silhouette score report correctness of all above clusterings and write your observations.