Omkar Chavan

ME698A Roll No. 210280

HW 1

In this problem, we are dealing with a dataset of propellant samples and we are using k-Nearest-Neighbours (kNN) to classify these samples as either 'pass' or 'fail'.

Given dataset: (converted into a csv file)
3 columns: 1> Propellant Age
2> Storage temperature
3> Pass/fail for application

Aim: To plot a pass/fail decision boundary using k-Nearest Neighbours (k=5) for the dataset.

Steps followed:

1. Converted the dataset into a csv file.
2. Read the csv dataset into a pandas dataframe.
3. Extracted the feature columns (propellant age, storage temperature) & label column (pass/fail for application) into 2 variables X & y.
4. Python lambda fuction was used to convert pass → 1 & fail → 0 value.
5. The kNN model was fitted on X & y. For this "sklearn.neighbours.Kneighboursclassifier" ~~lib~~ was used to declare the model.
6. Function to plot the decision boundary:
   - (i) A colormap of 2 colors - red & green was created
   - (ii) A meshgrid with propellant age (in X) & storage temp. (in y) & step size of 0.02 was created.
   - (iii) Then a variable 'z' was declared which contains

the predictions.

(iv) Finally, all the points were plotted with red color (if they fail) & green color (if they pass) using matplotlib library.

Discussion on possible validation procedure

To validate that our kNN model works properly we can carry out LOOCV (leave one out cross-validation) & it also helps in finding out optimum value of 'k'.

LOOCV is a special case of k-fold cross-validation where k = no. of datapoints in dataset.

In LOOCV, data is split into 2 subsets:

a) Training set: It contains all data points except the current one. (i.e. N-1 data points are used for the training)

b) Validation: The current left out data point is then used for validation.

In this dataset we will have 10 pairs of training & validation set. For each pair, the kNN model is trained on training subset and the prediction is done on validation set. If prediction is correct accuracy is incremented else it is decremented. For each value of k, the avg. accuracy is calculated which is then used to find out optimum value of k.

To find optimum value of k, we iterate 'k' over (1 to 'no. of samples-1'). For each k, LOOCV is conducted & avg LOOCV accuracy is stored. The k value with highest avg LOOCV accuracy is the possible optimum value of k.

In this example, [1, 2, 4, 5, 6] → these have highest accuracy of 0.9.