Homework 2

MA 506 Probability and Statistical Inference: Fall 2021

Due: October 5 (Tuesday), 11:59pm

100 points

Question 1: (25 points)

Import the Boston housing dataset from sklearn library (load_boston()) in python. Now answer the following questions:

1. (5 points) Based on the information in the description (DESCR) method, explain what is this dataset about ?
2. (5 points) How many rows are in this data set? How many columns? What do the rows and columns represent ?
3. (5 points) Make pairwise scatterplots of the following predictors/features = {CRIM, INDUS, NOX, RM, AGE, DIS, RAD, TAX, PTRATIO} in this dataset with respect to the median house value. What can such a scatterplot be used for ?
4. (5 points) Do any of the areas in Boston appear to have particularly high crime rates? Tax rates? Pupil-teacher ratios? Comment on the range of each of the predictors mentioned in (3).
5. (5 points) What is the median pupil-teacher ratio among the towns in this data set?

Question 2: (75 points)

Import the Iris dataset from sklearn library (load_iris()) in python. Now answer the following questions

1. (10 points) By visualizing the data, comment on whether 'petal width' alone can be used to perfectly classify the 3 classes.
2. (15 points) Based on data visualization find the best 2 features to classify the data into 3 classes ?
3. (10 points) Break the datasets into 2 equal subsets: training data and testing data. While doing this split, make sure that there are equal number of samples from all 3 classes in the the training data. Comment on why balancing the classes in the training data might be important.

4. (10 points) Use K-Nearest Neighbor (KNN) classification with the default parameters in sklearn to fit on training data, and show the performance on testing data in terms of accuracy of prediction.

5. (10 points) Plot accuracy of the KNN model on testing data vs K(number of nearest neighbor) and find the best K (number of nearest neighbors).

6. (10 points) Explain the behavior of the plot in (5) in terms of overfitting/underfitting and less-flexible/more-flexible models.

7. (10 points) Do you think the K found in (5) is the best number of nearest neighbors for this data ? If yes, explain the logic, if no, why not ?