Lab Assignment 1

Vasant Honavar
DS 310 - Machine Learning for Data Science

Due October 3, 12:00 PM on Canvas

In all of the following exercises, if there is a need for a random seed, set it to 1234. Using different sklearn libraries are permitted as long as usage is well-understood and explained in the code. In case you would need to interpret your results, do so in your Ipython Notebook by changing the cell type and writing your interpretation immediately below the code and its result so that the interpretation can be matched with the result and the code. Submit a single Ipython Notebook in which all of the answers are organized in a way that can be run and evaluated.

1. **Regression**. In this exercise, we will use linear regression for estimating the number of hours a person would be absent from work given their available information. For this, we use the "absenteesim at work" dataset obtained from the UCI repository, which is attached to the assignment in Canvas. In the dataset, you will find information about individuals, such as their age, education, reasons for absence, etc., as well as the target variable, which is absenteesim time in hours.

   (a) How many data points does the dataset include? How many features does each data point have?

   (b) Randomly split the data into train and test with the ratio 80/20, that is, use 80% of the data to fit the line, and the remaining 20% for testing, with the pre-specified random seed. Train a linear regression model on the training data. Then, use the trained model to estimate hours of absence in the test data. Report the average root mean squared error (RMSE) *on the test data*.

   (c) Perform 10-fold cross validation and report the RMSE obtained from each fold as well as their average.

   (d) How does the result in (b) compare to the result in (c)? Based on your analysis, does it suffice to avoid using cross-validation and simply train and test with a random split of the data? Interpret.

2. **K Nearest Neighbors (KNN) Classification.** Bob, our collaborator at Penn State Hershey Medical School is interested in predicting whether a cell is cancerous or not. Specifically, he is researching on the Breast Cancer dataset in which each cell is represented by a feature vector of its characteristics such as its size, etc., and is

classified as being benign, or malignant. Load the Breast Cancer dataset from sklearn datasets.

(a) How many data points does the dataset include? How many features does each data point have?

(b) Randomly split the data into train and test with the ratio 80/20, that is, use 80% of the data to fit the line, and the remaining 20% for testing, with the pre-specified random seed. Train a 5 nearest neighbor classifier on the training data. Then, use the trained model on the test data and report the area under the ROC curve (AUC).

(c) Perform 5-fold cross validation. Plot the ROC curve of each fold and compute the AUC of each fold. Then, report the average AUC of all 5 folds. Compare the obtained AUCs. Does it suffice to stick with the strategy in item (b), or is it more meaningful to perform cross validation? Please explain.

3. **K Neighbors Regression.** Return to the dataset in Problem 1. This time, we are interested in K neighbors regression instead of a regression on the whole dataset and we would like to analyze what would be reasonable number of neighbors and what distance to use based on the data. To do so, perform 10-fold cross validation. In each fold, fit a "weighted" linear regression in the following manner: For a given test data point, we would like to estimate its outcome based on its $k \in \{1, \ldots, 10\}$ nearest neighbors and a regression line weighted by the inverse of the distance of the neighbors of the test point. We would like to use the Minkowski distance with degree $p \in \{1, \ldots, 10\}$. For each fold, report the k and p at which we obtain the lowest RMSE. Do you get the same k and p from each fold? If yes, what does this mean? If no, why not? Then, report the average RMSE across all folds. How different is this obtained RMSE than the one you obtained from Problem 1? Explain.