METU Computer Engineering Department

CENG 514 Data Mining

Spring 2020-2021

Assignment II

In this assignment we will work on Classifiers and Classfier Accuracy Calculations. We will elaborate on two related sub tasks.

Q1. Implementing the k-NN algorithm.

In this part of the assignment, you will implement the well-known k-NN algorithm yourself in Python. Remember that it is an instance based learning algorithm including the following main steps:

- Given the instance whose label to be predicted (x_q), find k nearest neighbors of x_q within the training data set. At the step, you are not expected to implement an index structure, just calculate the similarites and find k nearest neighbors.
- Decide for the label of x_q according to the majority voting/mean/weighted mean.

The important point in your implementation will be to make the similarity function parametric so that you can use your k-NN implementation with different similarity metrics.

Key points:

- For the experiments, we will use "*Productivity Prediction of Garment Employees*" Data Set https://archive.ics.uci.edu/ml/datasets/Productivity+Prediction+of+Garment+Employees
- You will predict *actual_productivity* of the workers. Use weighted mean for the prediction.
- Implement two different similarity metrics suitable for the data set and use it together with your k-NN implementation.

Reporting:

- Describe the similarity metrics you devised.
- Report the result for k values from 2 to 10, under 3-fold cross validation for both of the similarity metrics. Since your prediction results are numeric values, report the prediction performance in terms of MSE, RMSE and MAPE.
- Report the prediction time for all cases.

Q2. Comparing with the classifiers in the scikit-learn library.

In this part of the assignment, compare the best performance of your implementation with the following supervised learning methods in scikit-learn library: KNeighborsRegressor, naive_Bayes and DecisionTreeRegressor. Note that you may need to adapt the domains of the attributes according to the classifier. You do not need to optimize the parameters for these classifiers. Just use the default settings.

Reporting:

- Report the results under 3-fold cross validation in terms of MSE, RMSE and MAPE.
- Report the prediction time for all cases.

Submission: Codes and the report presenting the analysis result.