Lab Course Machine Learning
Exercise 7

Prof. Dr. Dr. Lars Schmidt-Thieme,
Mofassir ul Islam Arif
Information Systems and Machine Learning Lab
University of Hildesheim
Submission: 13.01.2019 LearnWeb 3112

January 7, 2019

1 Exercise Sheet 7

Datasets

- 1. Classification Datasets: You can use one of the two datasets (or optionally, both datasets).

  (a) Iris dataset D1: Target attribute class:Iris Setosa, Iris Versicolour, Iris Virginica
  https://archive.ics.uci.edu/ml/datasets/Iris
  (b) Wine Quality called D2: (use winequality-red.csv)
  http://archive.ics.uci.edu/ml/ datasets/Wine+Quality

  Note: Dataset D2 can also be used for a regression problem.

You are required to pre-process given datasets.

Exercise 1: Implement K-Nearest Neighbor (KNN) (10 Points)
Your task is to implement KNN algorithm. To implement KNN you have to

- Split data into a train and a test split (70% and 30% respectively).
- Implement a similarity (or a distance) measure. To begin with you can implement the Euclidean Distance
- Implement a function that returns top K Nearest Neighbors for a given query (data point).
- You should provide the prediction for a given query (for a classification task you can use majority voting and for a regression you can use mean).
- Measure the quality of your prediction. [Hint: You have to choose a quality criterion according to the task you are solving i.e. a regression or a classification task **Defend your choice**].

Exercise 2: Optimize and Compare KNN algorithm. (10 Points)

Part A: (5 Points): **Determine Optimal Value of K in KNN algorithm**. In this exercise you have to provide the optimal value of K for given datasets.

- 1. How you can choose value of K for KNN. Give a criterion to choose an optimal value of K
- 2. Implement the criterion for choosing the optimal value of K.
- 3. Experimentally, give evidence that your chosen value is better than other values of K. [Hint: run your experiment with different values of K and plot the error measure for each value].

Part B: (5 Points): **Compare KNN algorithm with Tree based method**. In this task you are allowed to use scikit learn. In particular you have to use Nearest Neighbor and Decision Tree implementation provided by scikit learn.

- You should be able to use Nearest Neighbor and Decision Tree provided by scikit learn to solve classification task for two datasets.
- You have to provide the optimal hyperparameters for both the methods. [Hint: use Grid Search and cross validation and present results for them to support your solution].
- Present the comparison of the two methods using evaluation results on test datasets. [Hint: Better to use cross validation to ascertain your results]

1.1 ANNEX

- Following lecture is relevant this exercise https://www.ismll.uni-hildesheim.de/lehre/ml-16w/script/ml-04-A3-regularization.pdf
- sklearn.model_selection, sklearn.metrics, sklearn.linear_model, sklearn.preprocessing
- Scikit Learn User Guide http://scikit-learn.org/stable/user_guide.html
- You can use matplotlib for plotting.
- sklearn.metrics http://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics