Data Analytics 344 Tutorial - Similarity-Based Learning [26 September 2024]

Instructions

Practice the exercises on this tutorial between 2:00pm and 3:40pm on 26 September 2024. A quiz will open on SUNLearn at 3:45pm and will require you to apply the functions you have written in this tutorial to a dataset. You will have 5 minutes to provide the answers to the quiz. It is therefore important that you are very familiar with the code you have written for this tutorial. The quiz will only be available between 3:45pm and 4:00pm. You must write the quiz from the tutorial venue as per the official university timetable. You must submit the code you used to generate the answers for the quiz, and it must correspond to the quiz answers you provided. Name the file that contains your code as **DA344TutSBL2024_???????.r** where the question marks should be replaced with your student number. Only submissions made on SUNLearn will be assessed. This is an open book tutorial; however, you may not collaborate with anyone on the tutorial questions and must acknowledge any sources you consulted using comments in the code file. During the tutorial you may request for assistance from the Demis, with respect to understanding foundational concepts on either R or similarity-based learning. Carefully consider the assessment criteria before working on this tutorial.

Exercises

a) Write a function (**my_csv_reader**) to read a csv file and return a dataframe of all the instances in the dataset represented by the csv file. The csv file path should be specified by a parameter to the function.
b) Write a function (**my_separate_X_y**) to separate a dataframe D into two dataframes X and y then return a list that can be indexed to get X and y. X should contain descriptive features of all instances in D and y should contain the target features of all instances in D. The dataframe D should be specified by a parameter of the function.
c) Explore the FNN library which contains an implementation of KNNs using KDTrees. Write a function (**my_KNN_classifications**) that calls the knn function from FANN and uses it to display classification results using the KDTree algorithm. The function should accept X, q, y, and k as parameters. Where X is a dataframe of training instance descriptive features, q is a dataframe of test instance descriptive features, y is a factor of training instance target features and k is the number of neighbours to use.
d) Test your functions on a csv file representing a classification dataset with no missing values. Use 5 as your random number seed before calling any functions in your code.

Example Quiz Question

What is the testing set accuracy of a KDTree-based KNN model (K=3) after training it on the given training set.

Assessment Criteria

The following factors will be considered in assessing your work for this tutorial

- Correctness of provided answers.
- Validity of provided answers – do they correspond to the code you uploaded and is the code relevant to the questions asked? Your code should be error free. It is your responsibility to submit the correct code file and name it appropriately.
- Uniqueness – an automated evaluation of how similar your uploaded code is to code uploaded by other class members.

The weighting of these factors will be at the lecturer's discretion. All the best!