[CSC 5825 Fall 2021]

Due. Before Class of Oct 11, 2021 Homework 2

Full credit: 100 points

September 22, 2021

Question. (100 points) Programming question: Bayesian Classifier

In this question, you are asked to train two Bayesian classifiers (Naive Bayes and k-NN) to predict the presence of heart disease in the patient based on the Cleveland database. This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them, such as age, sex and several medical predictor variables. You can find more details of the data on the Kaggle website.

Tasks:

- Download the data. Create an account with Kaggle (if you have not previously done so) and download the Heart Disease dataset. You should split the 303 instances into training and test sets (8:2) for Naive Bayes classifier. While for the k-NN classifier, you need to split the instances into training, validation, and test sets as (6:2:2). The validation set is used to fine-tune the hyper-parameter k. Data can be downloaded from `https://www.kaggle.com/ronitf/heart-disease-uci`.
- Train your Naive Bayes and k-NN Classifiers on the training set. (70 points)
- After training, test them on the test set, construct confusion matrices (accuracy, precision, recall, and F-score) for the testing set results, and show these confusion matrices. (20 points)
- Compare the results between the two Bayesian classifiers. Which Bayesian classifier performs better? Why? (10 points)

Guidelines:

- Use Euclidean distance (L2) to compute distances between instances. As the attributes in Heart Disease dataset are either categorical or continuous. In the case of mix of these two, the categorical variables may be mapped to numerical values (through one-hot encoding) before applying the k-NN algorithm.
- Each continuous feature should be normalized separately from all other features. Specifically, for both training and testing instances, each feature should be transformed using function $F(X) = (X - mean)/std$, using the mean and std of the values of that feature on the training data.

Submission Instructions

Homework must be submitted electronically through Canvas website on/before the due date/time. Homework assignments are usually due in class at the beginning of lecture on the due date given. Homework must be typed with LaTeX or Word. The code can be submitted as .py file or .ipynb file. Late homeworks will not be accepted unless with legitimate excuses with documents. Do not use functions in scikit-learn package directly in the homeworks.