[CSC 5825 Fall 2024]

Due. Oct 9 11:59pm, 2024 Homework 2

Full credit: 100 points

September 18, 2024

Question. (100 points) Programming question: Generative Classifiers

In this question, you are asked to train two classifiers (Naive Bayes and k-NN) to predict the presence of heart disease 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. Therefore, you can use this 14 variables as features for your training. You can find more details of the data on the Kaggle website below.

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 920 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/datasets/redwankarimsony/heart-disease-data`.
- Train your Naive Bayes and k-NN Classifiers on the training set. (70 points)
- After training/validation, test them on the test set, construct confusion matrices for the testing set results, and show these confusion matrices calculate accuracy, precision, recall, and F-score. (20 points)
- Compare the results between the two classifiers. Which 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 the Canvas website on or before the due date and time. All homework must be typed using LaTeX or Word. Along with the .tex (or Word) file, please also submit a PDF version. Code can be submitted as a .py file or an .ipynb file. Late homework will not be accepted unless there is a legitimate excuse supported by documentation. Please do not use functions from the scikit-learn package directly in the homework.