CS 4824 / ECE 4424, Homework 2 (Written Portion), Due: Feb. 26, 2021

Question 1 (Comparing Classifiers) [6 points]

Consider a scenario where you are supposed to determine if a person has heart diseases or not based on the following attributes: blood pressure, body weight, age, number of cigarettes consumed in a week, and type of job (which can take four values: business, healthcare, engineering, or education).

a. If you had to choose between KNN and decision trees, which one would you prefer and why? I would prefer decision trees. KNN model is not suitable for proximity measure problem in continuous attributes.

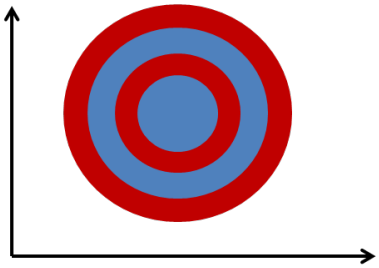
b. Assume you are using logistic regression for this task. How would you modify the features (as given above) before learning the model? (note: not all features need modification)

Change the type of job with 4 attributes into 4 categories, which could be evaluated by 0 or 1 to show whether or not in the specific job.

Question 2 (Comparing Classifiers) [8 points]

Answer the following questions based on different datasets that are provided with each question. Give a brief explanation for your choice.

1. Figure 1 shows a dataset of two classes whose points are shown in blue and red color on a 2-dimensional Cartesian coordinate axis (which are the two features). Among kNN, Naïve Bayes and decision trees, which classifier would have the **best** performance? Justify your answer.



Naive Bayes is limited in variable interactions.

Decision tree is limited in rectilinear split method.

Figure 1

KNN would have the best performance since it works with neighborhood and could distinguish different classes.

2. For the dataset in Figure 2, among KNN, Naïve Bayes and Perceptron, which classifier would have the **worst** performance?

Naïve Bayes may have the worst performance since attributes are not conditionally indepent.

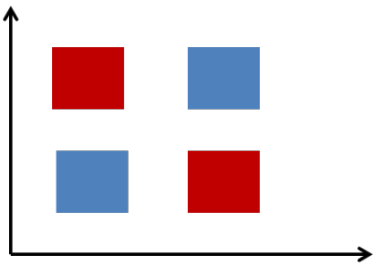


Figure 2

KNN and Perception have no influence with that and could solve interactions between variables.

Question 3 (Comparing Classifiers) [5 points]

If the training set is such that every combination of attribute values is present in the training data and each combination is either labeled C_1 or C_2 (e.g., positive or negative), which of the following classification techniques can be used to learn a model with perfect classification (zero errors) on the training set? Briefly explain your answer.

a. Decision Trees
b. Logistic Regression
c. Naïve Bayes
d. Perceptron

Decision tree can be used to learn the model with perfect classification. It applies the universal approximators with unique class label for each attributes, which could classify different training instance.