**CS 4824 / ECE 4424 Homework 2 (Written Portion) Due: March 10, 2020**

Note: If you feel a question is ambiguous, please state your assumptions in your answer. Please upload your solution as a single PDF file on Canvas page.

**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 k-NN and decision trees, which one would you prefer and why?

b. If you had to use logistic regression for classification, which transformation, if any, would you require to apply on the features before learning the ANN model?

**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?

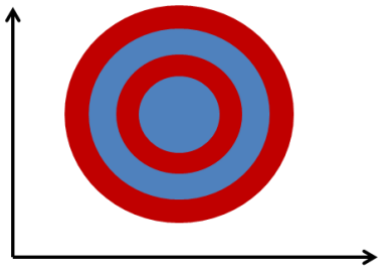


Figure 1

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

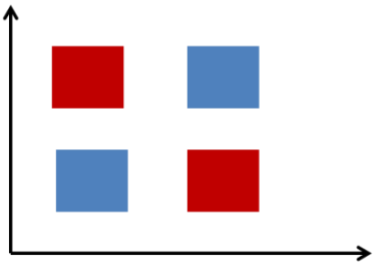


Figure 2

**Question 3 (Comparing Classifiers) [5 points]**

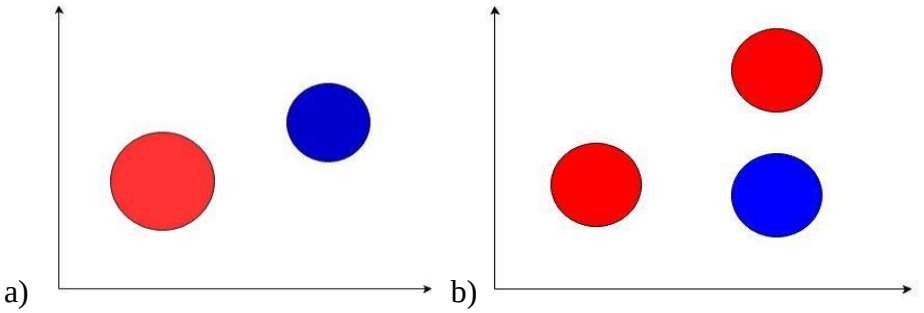
If the training set is such that every combination of attribute values is present in the training data and each combination is either labeled positive or negative, which of the following classification techniques can be used to learn a model with perfect classification (zero errors) on the training set?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

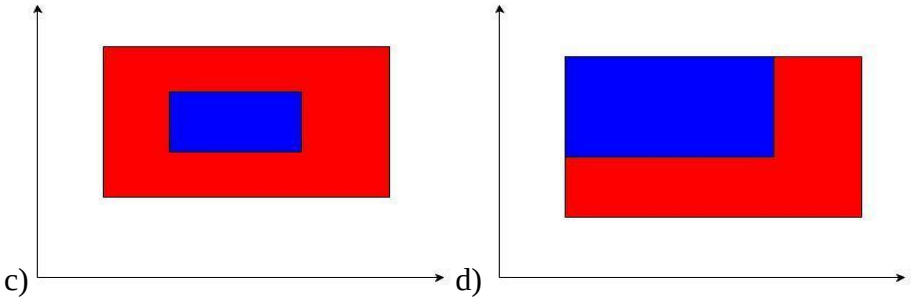
a. Decision Trees
b. Logistic Regression
c. Naïve Bayes
d. Multi-layer ANN
e. Perceptron

**Practice Questions (will not be graded)**

**Question 4 (Naïve Bayes)**

Determine whether the Naïve Bayes assumption holds for each of the datasets shown in the figures below. Give brief explanations for your answers. The two axes in each of the figures represent the two attributes, and the target binary classes are red and blue.






**Question 5 (ANN)**

State whether the following statements are true/false, giving a one-line justification for your answer.

a) In the back-propagation algorithm for training ANN models, the gradients of weights at the $k+1_{th}$ layer can be computed using the gradients of weights at the $k_{th}$ layer.

b) While applying an ANN model on a test instance, the activations at nodes at the $k+1_{th}$ layer can be computed using the activations at nodes at the $k_{th}$ layer.

c) If, at a given iteration of the back-propagation algorithm, the ANN model perfectly classifies all training instances, then the gradients of weights at all layers will be 0.

**Question 6 (ANN and Logistic Regression)**

State two similarities and two differences between perceptron and logistic regression.

**Question 7 (Naïve Bayes)**

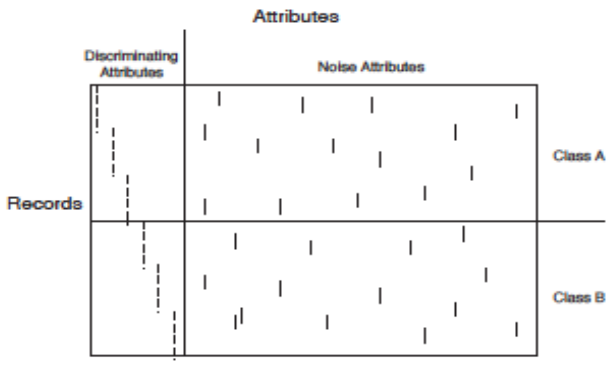
Given a dataset with Y as the target label and (A, B) as the set of features, the Naïve Bayes assumption requires that:

$$P(A, B \mid Y) = P(A \mid Y) \times P(B \mid Y)$$

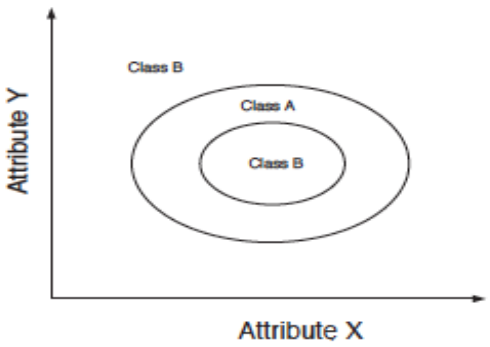
State whether the following statement is true or false: under the Naïve Bayes assumption, we also require that $P(A, B) = P(A) \times P(B)$. Provide a one-line justification.

**Question 8 (Comparing Classifiers)**

Given the data sets shown in Figure below, explain how the decision tree, Naïve Bayes (NB), and k-nearest neighbor (k-NN) classifiers would perform on these data sets.




**(a) Synthetic data set 1.**




**(b) Synthetic data set 2.**