**Problem 1 (10 points). Consider the following dataset:**

**Classify a new object O = <A1 = Cool, A2 = East, A3 = Low> using the Naïve Bayes algorithm we discussed in the class.**

P(Y) = 5/11

P(N) = 6/11

For A1 = Cool: P(Cool | Y) = 0, P(Cool | N) = 2/6

For A2 = East: P(East | Y) = 3/5, P(East | N) = 3|6

For A3 = Low: P(Low | Y) = 2/5, P(Low | N) = 3/6

P(Y∣O) ∝P(Y)×P(Cool∣Y)×P(East∣Y)×P(Low∣Y)

P(N∣O) ∝P(N)×P(Cool∣N)×P(East∣N)×P(Low∣N)

P(Y∣O)∝(5/11)×0×(3/5)×(2/5)

P(Y∣O) = 0

P(N∣O)∝(6/11)×(2/6)×(3/6)×(3/6)

P(N∣O)∝(6/11)×1/3×1/2×1/2

P(N∣O)∝1/33 > 0

So it should be classify in N.

**problem 2 (10 points). Info gain calculation Consider the following dataset:**

**(1). Calculate the information gain of attribute A1.**

Y = 5, N = 6

P(Y) = 5/11, P (N) = 6/11

Entropy(S) = -(5/11) log2(5/11) – (6/11) log2(6/11)

=0.994

For A1 = Hot, Y = 2, N = 3

Entropy (A1 = Hot) = 0.971

For A1 = Mild:

Y =2, N = 2

Entropy (A1 = Mild) = 1

For A1 = Cool:

Y = 0, N = 3

Entropy (A1 = Cool) = 0

IG(A1) = 0.994 – ((5/11 \* 0.971) + (4/11\*1) + (3/11 \* 0))

= 0.151

**(2). Calculate the information gain of attribute A2.**

For A2 = East:

Y = 2, N = 3

Entropy (A2 = East) = 0.971

For A2 = West:

Y = 3, N = 3

Entropy (A2 = West) = 1

IG(A2) = 0.994 – ((5/11 \* 0.971) + (6/11 \* 1))

= 0.022

**(3). Which is better as the test attribute at the root level?**

0.151>0.022 So A1 is better than A2.

**Problem 3 (10 points). Consider a dataset that has two predictor variables, age and bp, and the class attribute class. The class attribute values are Yes and No. Suppose that you ran a logistic regression algorithm and obtained the following coefficients for the class Yes:**

**intercept = -5.5461, age = 0.0418, bp = 0.0517**

**Calculate the probability that a new object O = <age = 68, bp = 145> belongs to the class Yes using the method we discussed in the class. You must show all intermediate steps and calculations.**

z = - 5.5461 + (0.0418 \* 68) + (0.0517 \* 145)

z = -5.5461 + 2.8424 + 7.4965

z = 4.7928

P = 1/(1+e^-4.7928)

P = 0.9918

So, the probability for yes is 0.9918.

**Problem 4 (10 points). This question is about the discriminant analysis method that we discussed in the class. Suppose that you have a dataset with two classes, Class 1 and Class 2, and that you are trying to classify an object O using the method and you calculated the distance between O and the centroids of the two classes and obtained the following:**

**Squared distance to Class 1: 1.3871**

**Squared distance to Class 2: 3.7342**

**Calculate the probability that the object O belongs to Class 1 and the probability that the object O belongs to Class 2.**

P(Class 1) = 1/(1 + e^(3.7342 – 1.3871))

= 1/(1 + e^2.3471)

P(Class 1) = 1/(1 + 10.453)

= 0.0873

P(Class 2) = 1 – P(Class 1)

P(Class 2) = 0.9127

The probability that object O belongs to Class 1 is 0.0873.

The probability that object O belongs to Class 2 is 0.9127.

**Problem 5 (10 points). Use the heart\_failure.csv dataset and use R. The dataset was downloaded from the UCI Machine Learning Repository and modified for this assignment. In the dataset:**

**Each tuple represents a patient.**

**DEATH\_EVENT is the class attribute; 1 means the patient died and 0 means survived.**

**(1). Convert the data type of the class attribute to factor.**

**(2). Split the dataset into training and test sets with the 66%-34% ratio. Make sure that you use a stratified splitting method.**

**(3). Build a Naïve Bayes model from the training dataset.**

**(4). Test the model on the test dataset.**

**(5). In your submission file, include the confusion matrix and the prediction accuracy of each class.**

**(6). Build a decision tree model using the rpart algorithm from the training dataset. Use information gain as the purity measure.**

**(7). Plot the tree of the model.**

**(8). Test the model on the test dataset.**

**(9). In your submission file, include the screenshot of the tree, confusion matrix and the prediction accuracy of each class.**

