CSE 482: Big Data Analysis (Spring 2020) Homework 3 Part 2
Submit a PDF file containing the solutions for all the questions below.

1. Suppose you are given the following training set for predicting the risk score of a person having a certain disease based on their blood pressure and height attributes.

Blood Pressure	Height	Risk Score
110	5.7	3.0
145	6.0	8.5
105	5.0	2.0
150	5.9	9.0
135	6.2	4.0

Consider the following regression model, M:

$$\text{Risk Score} = 0.2\ \text{BP} - 2.4\ \text{Height} - 5 \tag{1}$$

(a) Calculate the predicted value for each of the 5 training points above using model M.

(b) Calculate the root-mean-square-error of the model predictions given in part (a).

(c) Use the coefficients (parameter values) of the regression model given in Equation (1) to identify the most important attribute for predicting risk score.

(d) Compute the mean and standard deviation values for blood pressure and height.

(e) Using the mean and standard deviation values you've found in part (c), derive the equivalent regression formula for the model given in Equation (1) if we use the standardized values of blood pressure (Z_{BP}) and height (Z_H) instead of their original values. In other words, derive the values for w_0, w_1, and w_2 in the equation below:

$$\text{Risk Score} = w_2 Z_{\text{BP}} + w_1 Z_H + w_0$$

(f) Based on your answer in part (e), identify which attribute is most important for predicting risk score.

(g) Does your answer for part (f) consistent with the answer in part (c)? If not, which answer is better and state your reason clearly.

2. Consider the training set given below for determining whether a loan application should be approved or rejected. Draw the full decision tree obtained using entropy as the impurity measure. Show your steps clearly (i.e., the computation of entropy for every candidate attribute must be shown - see lecture notes as example). Compute the training error of the decision tree.

Long-Term Debt	Unemployed	Credit Rating	Class
No	No	Good	Approve
No	No	Bad	Approve
No	No	Bad	Approve
No	No	Bad	Approve
Yes	No	Good	Approve
No	Yes	Good	Reject
Yes	No	Bad	Reject
Yes	No	Bad	Reject
Yes	No	Bad	Reject
Yes	Yes	Bad	Reject

3. Consider the problem of predicting how well a particular baseball player will bat against different pitchers. The training set contains ten positive and ten negative examples, based on the previous performance of the player against 20 different pitchers. Assume there are two attributes: ID (which is unique for every pitcher) and Handedness (left- or right-handed). Among the left-handed pitchers, nine of them are assigned to the positive class and one to the negative class. On the other hand, among the right-handed pitchers, only one of them is from the positive class, while the remaining nine are from the negative class.

   Suppose we apply a decision tree classifier to the given training set. We need to choose which attribute to use as splitting criterion of the decision tree. Assume the classifier uses gini index as its impurity measure.

   (a) Compute the overall gini if we use ID as splitting criterion.

   (b) Compute the overall gini if we use Handedness as splitting criterion.

   (c) Based on your answers in parts (a) and (b), which attribute will be chosen as splitting criterion?

   (d) Explain whether the answer in part (c) is reasonable.