COMS 4701 Artificial Intelligence

Homework 4 - Conceptual

Due date: **SUNDAY** November 21st, 2021

1. **Decision Trees**

   Consider the task of spam prediction using this dataset of four training examples (emails), each described by two features, *word1* and *word2*, and a label *spam* to denote whether the email is a spam or not.

word1	word2	spam
T	F	T
T	F	T
F	T	T
F	F	F

   1. Calculate the gini index at the root.

$$Gini = 1 - p_{\oplus}^2 - p_{\ominus}^2$$

   2. What feature is picked at the root? Justify your work.

2. **Linear regression**

   Consider the problem of predicting student scores in the final exam in function of their scores in the midterm exam. Let feature x denote the midterm scores and label y denote the final scores. Consider the training set pairs (x, y) as follows: $(55, 67), (60, 63), (66, 72), (72, 90), (85, 93), (90, 92)$.

   1. Suppose we learned a linear classifier and the weights are $\beta_0 = -8$ and $\beta_1 = 1.2$. what is the predicted final score if the midterm grade is 80? Show your work.
   2. Recall the cost function R seen in class that depends on the regression weights. Calculate the cost for $\beta_0 = -8$ and $\beta_1 = 1.2$.
   3. Suppose we managed to train a linear regression on the training data and we found β_0 and β_1 such that $R = 0$. Which of the following is correct/incorrect. Explain each answer.
      (a) We must have $\beta_0 = 0$ and $\beta_1 = 0$.
      (b) We have found a linear regressor that perfectly fit the data.
      (c) We will do a perfect prediction in the test set.

3. **Naive Bayes classifier**

   Consider the following dataset with three binary features taking their values in {0,1}, and the label taking its values in {TRUE, FALSE}.

A	B	C	Label
0	1	1	TRUE
1	1	0	TRUE
1	0	1	FALSE
1	1	1	FALSE
0	1	1	TRUE
0	0	0	TRUE
0	1	1	FALSE
1	0	1	FALSE
0	1	0	TRUE
1	1	1	TRUE

Using a Naive Bayes classifier, predict the label of the new example: $(A = 1, B = 0, C = 1)$.

No smoothing needed and no need to calculate all probabilities of the NB classifier, calculate only the probabilities you need to make this prediction.