Name+Email ID Jon-Michael Hoang | hoang339

CSCI 5521: Machine Learning Fundamentals (Spring 2022)

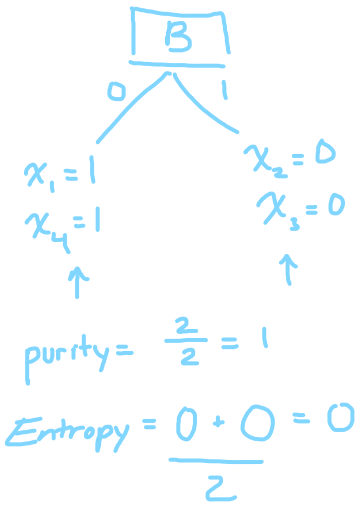
Quiz 4 (Thursday, Apr 21))

Due on Gradescope at 02:00 PM, Friday, Apr 22

Instructions:

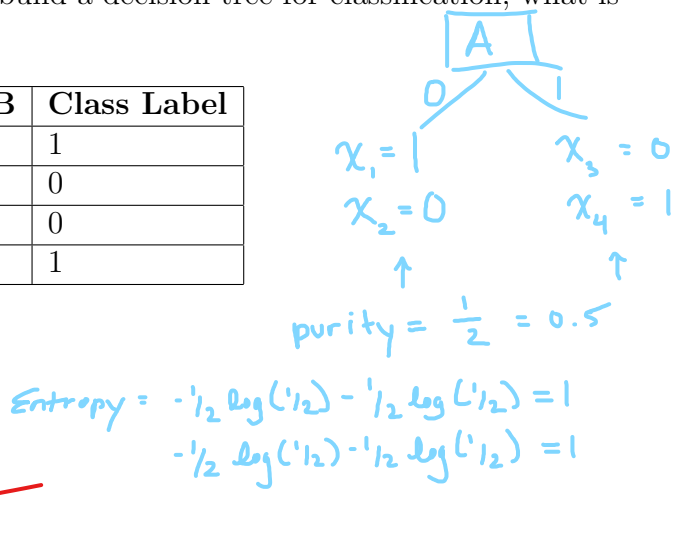
- This quiz has 3+1 questions (1 question for extra credits), 30+2 points, on 2 pages.
- Please **write your name & ID on this cover page**.

1. (**12 points**) Given the data in the figure below, to build a decision tree for classification, what is the attribute you will use to split. Show your work.



	Attribute A	Attribute B	Class Label
x_1	0	0	1
x_2	0	1	0
x_3	1	1	0
x_4	1	0	1

∴ B is the best to split b/c purity = 0 & Entropy = 0 .



2. (**12 points**) Select all correct statement(s) about decision tree and random forest methods:

   (a) Decision tree method is an unsupervised learning method.

   (b) Entropy, Gini Index, and Misclassification Error are all valid options for feature selection in decision tree method.

   (c) Building more decision trees to form a random forest helps alleviate overfitting.

   (d) Decision tree method is more interpretable than random forest method.

3. (**6 points**) List one difference between linear SVM and kernel SVM.

   One difference between a linear SVM and a kernel SVM is that a linear SVM is best used for when the original data itself is linearly seperable, that is, that some decision line could be drawn to classify data whereas a kernel SVM ***transforms*** the original space into a space that allows one to be able to separate the data with a "line" or rather, a "plane".

4. (**2 points, extra credits**) Ada is working in a financial firm. Her team is developing algorithms to approve loans and budgets. They are also interested in understanding and analyzing deciding factors in the process. Ada's team has access to a small set of training data with personal information (e.g., age, income, etc.) and their labels. Suggest a machine learning method that they could use. Briefly explain.

A machine learning method that they could use are Decision Trees because she has [row] sets of data with [col] sets of features with appropriate labels for each [row], that is, whether or not a loan is approved or not. She and her team would be then able to split the data for each certain feature to then build a decision tree s.t. each terminal leaf provides them with a decision on whether or not to approve the loan (provided that each leaf is pure).