EECS 440: Machine Learning Fall 2020 Written Problems Week 2 due 9/18 11:59pm

General Instructions: Write or type your answers neatly and remember to show all relevant work. All questions are worth 10 points. Each answer should be a separate pdf, and you can turn in the pdfs on canvas in the appropriate assignment. Some questions may be very challenging; significant partial credit is available for reasonable attempts at solutions. Since each question is worth the same number of points, do not waste too much time on any one. Ask me or the TAs for help if stuck.

Some of the questions require you to write short programs to simulate things. You can use any language/software to do this, and you do not need to turn in your code.

Upload your answers to Canvas as a pdf file by 11:59pm on the due date specified after the question. You will receive a 10% bonus for a solution turned in a week or more in advance of the due date. You can use one late day each week (up to Saturday 11:59pm) with a penalty of 20%. Submissions after Saturday 11:59pm for any week will not be graded.

Each group must do their own work. Only one submission is needed from each group. Do not use any source other than the lecture notes, textbook(s) and readings on the class website to answer these questions. Only those who contributed equally to a submission should have their names and Case IDs on the submission. Those not listed as contributing will not receive points.

5. Consider a learning problem where the examples are described by *n* Boolean attributes, and the hypothesis space is the space of all Boolean functions on *n* variables. How many distinct examples (feature vectors) can you have in this setting? (9/18)
6. Consider the same setting as above. How many distinct decision trees can you construct in this setting? “Distinct” means that each tree must represent a different hypothesis in the space. Give a rigorous justification for your answer. (9/18)
7. Consider the following table of examples over Boolean attributes, annotated with the target concept's label. Ignore the "Weight" column and use information gain to find the first split in a decision tree (remember that ID3 stops if there is no information gain). (9/18)
8. Now from the same table, find another split using "weighted" information gain. In this case, instead of counting the examples for each label in the information gain calculation, add the numbers in the **Weight** column for each example. (9/18)
9. Is there a difference between the splits for Q7 and Q8? Can you explain what is happening? (9/18)

A1	A2	A3	A4	Label	Weight
F	F	F	F	**0**	1/256
F	F	F	T	**0**	3/256
F	F	T	F	**1**	3/256
F	F	T	T	**1**	9/256
F	T	F	F	**1**	3/256
F	T	F	T	**1**	9/256
F	T	T	F	**0**	9/256
F	T	T	T	**0**	27/256
T	F	F	F	**1**	3/256
T	F	F	T	**1**	9/256
T	F	T	F	**0**	9/256
T	F	T	T	**0**	27/256
T	T	F	F	**0**	9/256
T	T	F	T	**0**	27/256
T	T	T	F	**1**	27/256
T	T	T	T	**1**	81/256