Homework 5

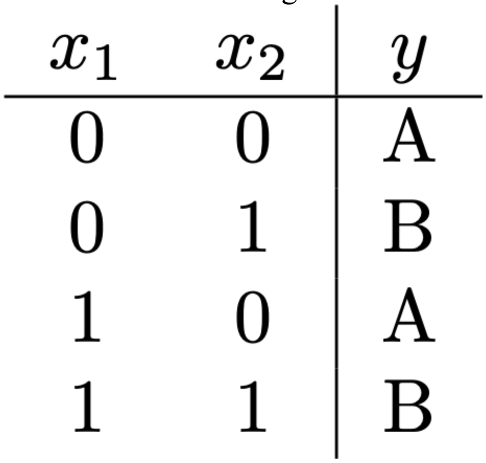
Instructions
This homework contains **4** concepts and **7** programming questions. In MS word or a similar text editor, write down the problem number and your answer for each problem. Combine all answers for concept questions in a single PDF file. Export/print the Jupyter notebook as a PDF file including the code you implemented and the outputs of the program. Make sure all plots and outputs are visible in the PDF.

Combine all answers into a single PDF named andrewID_hw5.pdf and submit it to Gradescope before the due date. Refer to the syllabus for late homework policy. Please assign each question a page by using the "Assign Questions and Pages" feature in Gradescope.

Question	**Points**
Concept 1	3
Concept 2	3
Concept 3	3
Concept 4	3
M5_L1_P1	6
M5_L1_P2	6
M5_L1_P3	9
M5_L2_P1	6
M5_L2_P2	9
M5_HW1	36
M5_HW2	36
Total	**120**
Bonus	6

Problem 1

Consider the following dataset with features x_1 and x_2 and labels y.



x_1	x_2	y
0	0	A
0	1	B
1	0	A
1	1	B

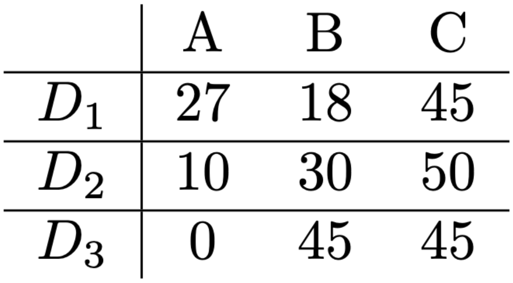
Which of the following features should be used in the first node of the decision tree?
Multiple choice (choose one)

1. x_1
2. x_2
3. It doesn’t matter which is used

X2

Problem 2

Consider the following 3 datasets which are made up of samples belonging to classes A,B, and C. The following table summarizes how many samples belong to each class in a given dataset.



	A	B	C
D_1	27	18	45
D_2	10	30	50
D_3	0	45	45

Which dataset is most impure (e.g. has the highest Gini score)?
Multiple choice (choose one):

1. D_1
2. D_2
3. D_3

D1

Problem 3

Multiple Choice (select all that are true)
Which of the following functions would a decision tree be able to accurately predict out of range samples for?

1. $f(x) = 4x^2 + 1$
2. $f(x) = 2x$
3. $f(x) = 3$
4. $f(x) = \sin(x) + 5$

Function 2 and 3 are accurate as they are linear.

Problem 4

Multiple choice (choose one)
Let's consider two bootstrap aggregation models trained on the same dataset. Each model is trained using 10 decision trees. Each decision tree in Model 1 trained using 50% of the samples in the dataset, selected at random. Each decision tree in Model 2 is trained using 90% of the samples in the dataset, selected at random. Which model is more likely to accurately predict unseen, in range, test samples?

1. Model 1
2. Model 2

Model 1 is likely to be more accurate as model 2 probably overfits.