**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.

Problem 1 (2.5 points)

Consider the following dataset with features x1 and x2 and labels y.

A black background with a black square

Description automatically generated with medium confidence

Which of the following features should be used in the first node of the decision tree?

Multiple choice (choose one)

1. x1
2. x2
3. It doesn’t matter which is used

Problem 2 (2.5 Points)

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 black background with a black square

Description automatically generated with medium confidence

Which dataset is most impure (e.g. has the highest Gini score)?

Multiple choice (choose one):

1. D1
2. D2
3. D3

Problem 3 (2.5 Points)

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) = 4x2 + 1
2. f(x) = 2x
3. f(x) = 3
4. f(x) = sin(x) + 5

Problem 4 (2.5 Points)

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