ML BACKGROUND EXAM - AGENT COMPANY

START HERE: Instructions

- **Written:** For written problems such as short answer, multiple choice, derivations, proofs, or plots, please use the provided template. Submissions can be handwritten onto the template, but should be labeled and clearly legible. If your writing is not legible, you will not be awarded marks. Alternatively, submissions can be written in LaTeX. Each derivation/proof should be completed in the boxes provided. You are responsible for ensuring that your submission contains exactly the same number of pages and the same alignment as our PDF template. If you do not follow the template, your assignment may not be graded correctly by our AI assisted grader and there will be a **2% penalty** (e.g., if the homework is out of 100 points, 2 points will be deducted from your final score). The 'select all that apply' questions will be graded as full score only if all the correct options are selected and no wrong options are selected. They will be awarded 0 otherwise.

Written Questions (15 points)

1 LaTeX Point and Template Alignment (1 points)

1. (1 point) **Select one:** Did you use LaTeX for the entire written portion of this homework?

   ● Yes

   ○ No

2. (0 points) **Select one:** I have ensured that my final submission is aligned with the original template given to me in the handout file and that I haven't deleted or resized any items or made any other modifications which will result in a misaligned template. I understand that incorrectly responding yes to this question will result in a penalty equivalent to 2% of the points on this assignment.
   Note: Failing to answer this question will not exempt you from the 2% misalignment penalty.

   ● Yes

2 Decision Trees (7 points)

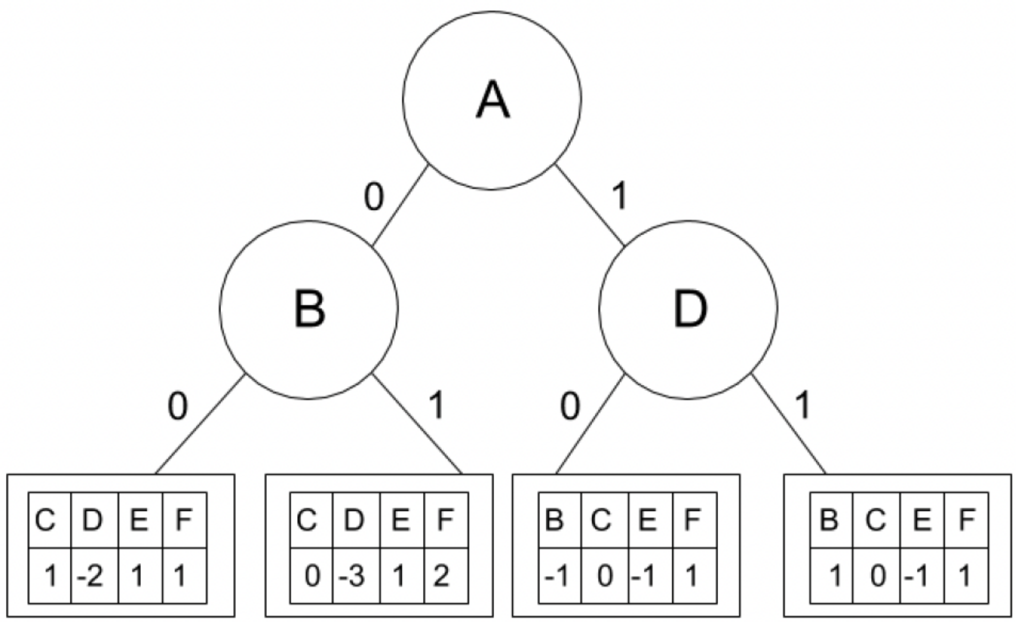


Figure 1: Perceptron Tree of depth 2

1. To exploit the desirable properties of decision tree classifiers and perceptrons, Adam came up with a new algorithm called the "perceptron tree" that combines features from both. Perceptron trees are similar to decision trees, but each leaf node contains a perceptron rather than a majority vote. To create a perceptron tree, the first step is to follow a regular decision tree learning algorithm (such as ID3) and perform splitting on attributes until the specified maximum depth is reached. Once maximum depth has been reached, at each leaf node, a perceptron is trained on the remaining attributes which have not yet been used in that branch. Classification of a new example is done via a similar procedure. The example is first passed through the decision tree based on its attribute values. When it reaches a leaf node, the final prediction is made by running the corresponding perceptron at that node. Assume that you have a dataset with 6 binary attributes A, B, C, D, E, F and two output labels -1, 1. A perceptron tree of depth

2 on this dataset is given below. Weights of the perceptron are given in the leaf nodes. Assume bias b = 1 for each perceptron.

(a) (1 point) **Numerical answer**: What would the given perceptron tree predict as the output label for the sample x = [1, 1, 0, 1, 0, 1]?

Label
1

(b) (1 point) **True or False:** The decision boundary of a perceptron tree will always be linear.

○ True

● False

(c) (1 point) **True or False:** For small values of max depth, decision trees are more likely to underfit the data than perceptron trees.

● True

○ False

2. (1 point) **Select all that apply:** Given an input feature vector x, where $x \in R^n$, you are tasked with predicting a label for y, where $y = 1$ or $y = -1$. You have no knowledge about the distributions of x and of y. Which of the following methods are appropriate?

   □ Perceptron

   ■ k-Nearest Neighbors

   ■ Linear Regression

   ■ Decision Tree with unlimited depth

   □ None of the above.

3. (1 point) **True or False:** The ID3 algorithm is guaranteed to find an optimal decision tree.

   ○ True

   ● False

4. (1 point) **True or False:** One advantage of decision trees is that they are not easy to overfit.

   ○ True

   ● False

5. (1 point) **True or False:** Neural the Narwhal flips two fair coins. Let Y denote the number of heads that are facing up. Then, the entropy of Y is 1.

   ○ True

   ● False

3 K Nearest Neighbors (7 points)

1. (1 point) **Select one:** A k-Nearest Neighbor model with a large value of k is analogous to...
   - ● A short Decision Tree with a low branching factor
   - ○ A short Decision Tree with a high branching factor
   - ○ A long Decision Tree with a low branching factor
   - ○ A long Decision Tree with a high branching factor

2. (1 point) **Select one:** Imagine you are using a k-Nearest Neighbor classifier on a dataset with lots of noise. You want your classifier to be less sensitive to the noise. Which of the following is likely to help and with what side effect?
   - ● Increase the value of $k \implies$ Increase in prediction time
   - ○ Decrease the value of $k \implies$ Increase in prediction time
   - ○ Increase the value of $k \implies$ Decrease in prediction time
   - ○ Decrease the value of $k \implies$ Decrease in prediction time

3. Consider the following training dataset for a regression task:

$$\mathcal{D} = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(N)}, y^{(N)})\}$$

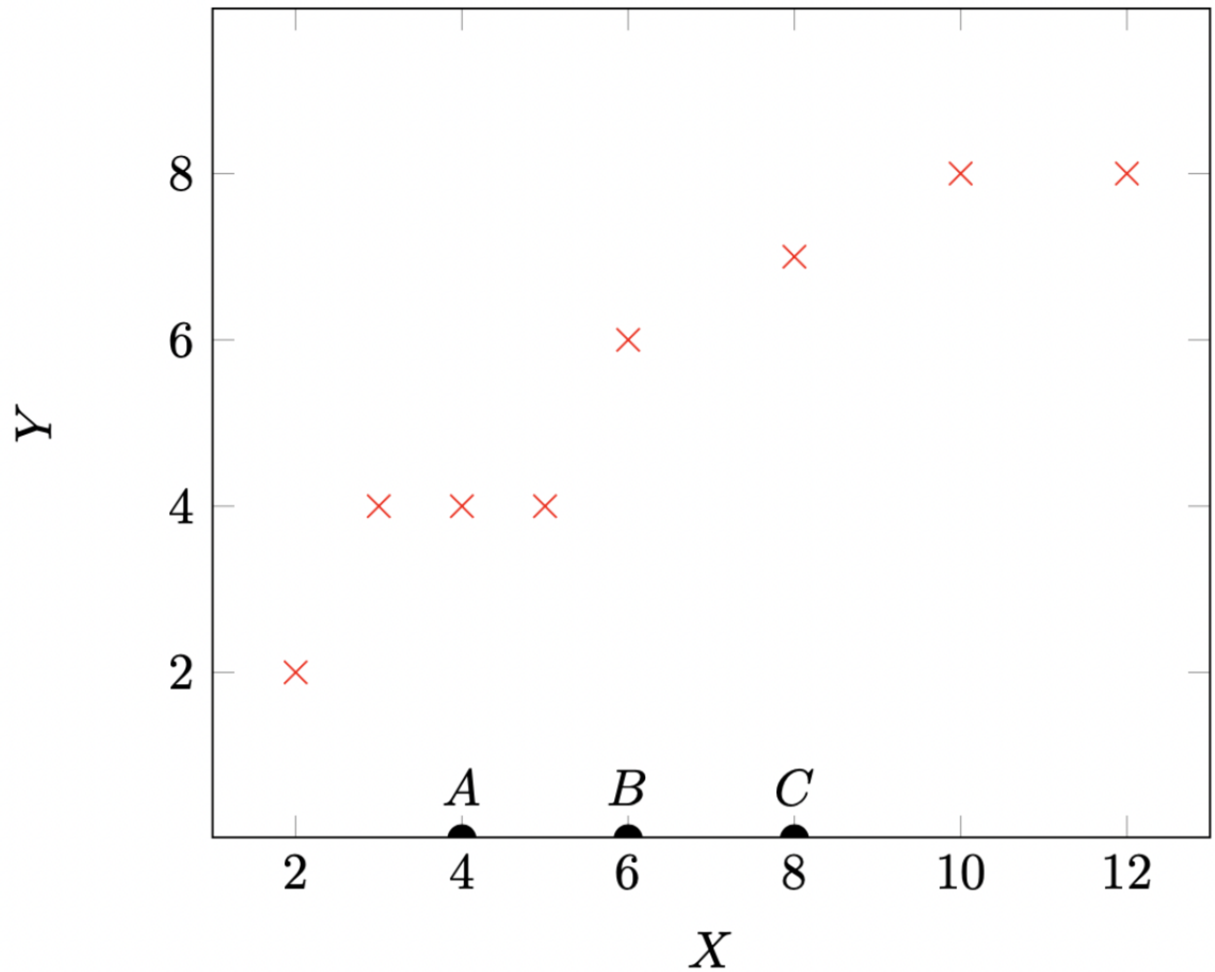
with $x^{(i)} \in \mathbb{R}$ and $y^{(i)} \in \mathbb{R}$.

For regression with k-nearest neighbors, we make predictions on unseen data points similar to the classification algorithm, but instead of a majority vote, we take the mean of the output values of the k nearest points to some new data point x. That is,

$$h(x) = \frac{1}{k} \sum_{i \in \mathcal{N}(x,\mathcal{D})} y^{(i)}$$

where $\mathcal{N}(x, \mathcal{D})$ is the set of indices of the k closest training points to x.

In the above dataset, the red ×'s denote training points and the black semi-circles A, B, C denote test points of unknown output values. For convenience, all training data points have integer input and output values. Any ties are broken by selecting the point with the lower x value.



(a) (1 point) **Numerical answer**: When k = 1, what is the mean squared error on the training set?

Answer
0

(b) (1 point) **Numerical answer**: When k = 2, what is the predicted value at A?

Answer
5

(c) (1 point) **Numerical answer**: When k = 2, what is the predicted value at B?

Answer

5

(d) (1 point) **Numerical answer**: When k = 3, what is the predicted value at C?

Answer

7

(e) (1 point) **Numerical answer**: When k = 8, what is the predicted value at C?

Answer

5.475