10-601 Machine Learning **Name:**
Spring 2025 **Andrew ID:**
Practice Problems **Room:**
Updated: February 7, 2025 **Seat:**
Time Limit: N/A **Exam Number:**

Instructions:

- Verify your name and Andrew ID above.
- This exam contains 24 pages (including this cover page).
  The total number of points is 0.
- Clearly mark your answers in the allocated space. If you have made a mistake, cross out the invalid parts of your solution, and circle the ones which should be graded.
- Look over the exam first to make sure that none of the 24 pages are missing.
- No electronic devices may be used during the exam.
- Please write all answers in pen or *darkly* in pencil.
- You have N/A to complete the exam. Good luck!

Instructions for Specific Problem Types

For "Select One" questions, please fill in the appropriate bubble completely:

Select One: Who taught this course?

● Matt Gormley

○ Marie Curie

○ Noam Chomsky

If you need to change your answer, you may cross out the previous answer and bubble in the new answer:

Select One: Who taught this course?

● Henry Chai

○ Marie Curie

⊗ Noam Chomsky

For "Select all that apply" questions, please fill in all appropriate squares completely:

Select all that apply: Which are instructors for this course?

■ Matt Gormley

■ Henry Chai

☐ I don't know

Again, if you need to change your answer, you may cross out the previous answer(s) and bubble in the new answer(s):

Select all that apply: Which are the instructors for this course?

■ Matt Gormley

■ Henry Chai

☒ I don't know

For questions where you must fill in a blank, please make sure your final answer is fully included in the given space. You may cross out answers or parts of answers, but the final answer must still be within the given space.

Fill in the blank: What is the course number?

10-601 10-~~6~~301

Instructions for Specific Problem Types

For "Select One" questions, please fill in the appropriate bubble completely: **Select One:** Who taught this course? ● Henry Chai

○ Marie Curie

○ Noam Chomsky

If you need to change your answer, you may cross out the previous answer and bubble in the new answer:

Select One: Who taught this course?

● Henry Chai

○ Marie Curie
⊗ Noam Chomsky

For "Select all that apply" questions, please fill in all appropriate squares completely:

Select all that apply: Which are scientists?

■ Stephen Hawking

■ Albert Einstein

■ Isaac Newton

□ I don't know

Again, if you need to change your answer, you may cross out the previous answer(s) and bubble in the new answer(s):

Select all that apply: Which are scientists?

■ Stephen Hawking

■ Albert Einstein

■ Isaac Newton
⊠ I don't know

For questions where you must fill in a blank, please make sure your final answer is fully included in the given space. You may cross out answers or parts of answers, but the final answer must still be within the given space.

Fill in the blank: What is the course number?

10-601 10-~~7~~601

1 Decision Trees (0 points)

1.1. 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.

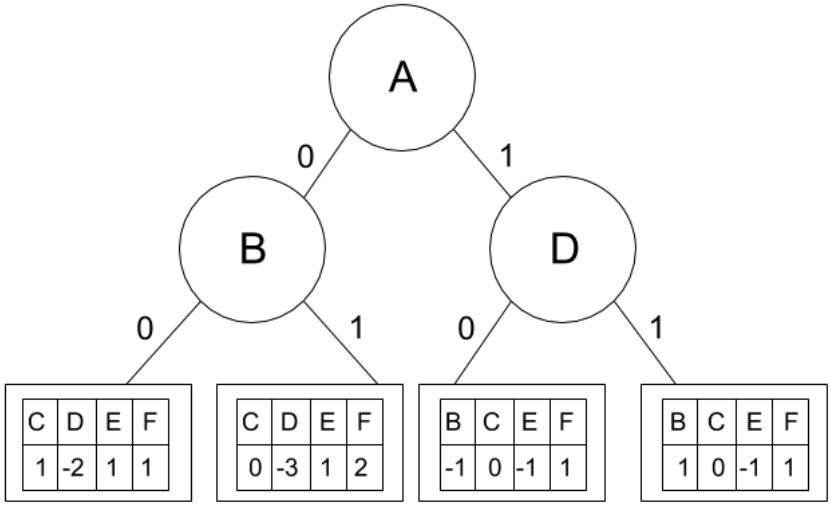


Figure 1: Perceptron Tree of depth 2

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

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

◯ True

◯ False

(c) **True or False:** For small values of max depth (e.g., 2 or 3), decision trees are *more* likely to underfit the data than perceptron trees.

◯ True

◯ False

1.2. **Select all that apply:** Given a training dataset, $\mathcal{D}$, suppose you train two decision trees, one using mutual information as the splitting criterion and the other using training error rate. If both trees are trained until they achieve zero training error, which of the following statements must be true?

□ Both trees split on the same feature at the root node.

□ Both trees have the same depth

□ Both trees make the same predictions for each data point in $\mathcal{D}$

□ Both trees use all of the features in $\mathcal{D}$

□ None of the above.

1.3. **True or False:** The ID3 algorithm is guaranteed to find an "optimal" decision tree i.e. a decision tree that achieves zero training error while making the fewest possible splits.

◯ True

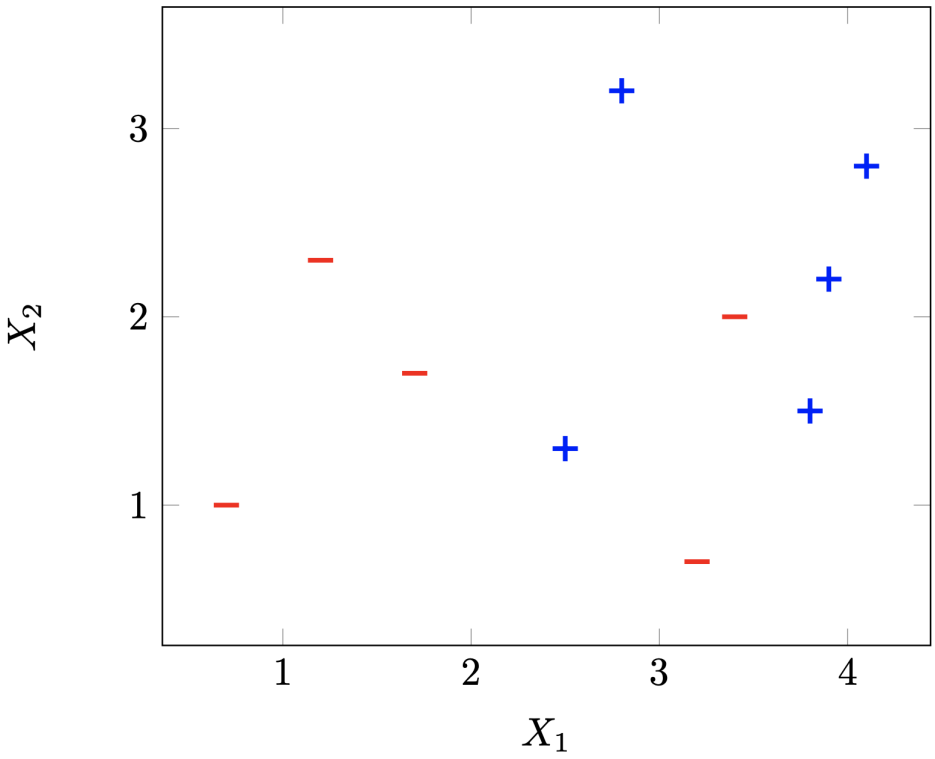
◯ False

1.4. **True or False:** One advantage of the ID3 algorithm for training decision trees is that it is not susceptible to overfitting because it uses recursion.

◯ True

◯ False

1.5. Consider the following training data. The red '−' marks represent $Y = 0$ and the blue '+' marks represent $Y = 1$.



(a) **Numerical answer:** What is the entropy of Y in bits?

(b) **Numerical answer:** What is the mutual information of splitting on $X_2 < 2.5$? You may write your answer as an arithmetic expression that includes $\log_2$ operations, but does not include symbolic information such as H.

(c) **Select one:** Using *training error rate* as the splitting criteria, which of the following is the best choice for the first binary split?

- ◯ $X_1 < 2$
- ◯ $X_1 < 3$
- ◯ $X_2 < 1.5$
- ◯ $X_2 < 2.5$

(d) **Numerical answer:** If we restrict the tree to have a maximum depth of 1 (i.e. a decision stump) and we restrict the split to be of the form $X_d < C$ for some constant C, what is the lowest attainable training error rate?

(e) **True or False:** Using only binary splits that involve either X_1 or X_2 but not both, it is possible to have a decision tree with zero training error rate for this dataset.

◯ True

◯ False

1.6. **Fill in the blank:** Complete the following paragraph about pruning decision trees by circling the best of the provided options for each of blanks:

When pruning a decision tree, each split is replaced with the most common label among the training / validation data points that end up at each node. Splits are compared based on their reduction in training / validation error rate and if two splits are tied, we prune the node closer to / farther from the root.

2 K Nearest Neighbors (0 points)

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

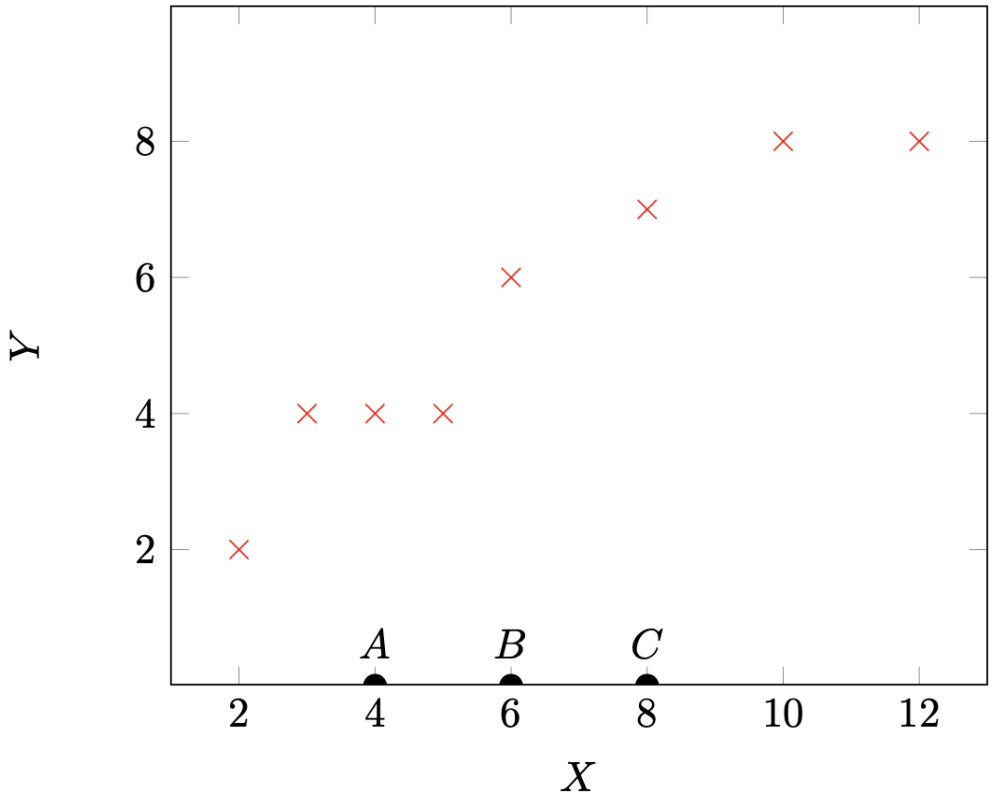
$$\mathcal{D} = \left\{ \left(x^{(1)}, y^{(1)}\right), \left(x^{(2)}, y^{(2)}\right), \cdots, \left(x^{(N)}, y^{(N)}\right) \right\}$$

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



In the dataset shown above, the red $\times$'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. Assume ties are broken by selecting the point with the lower x value.

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

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

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

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

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

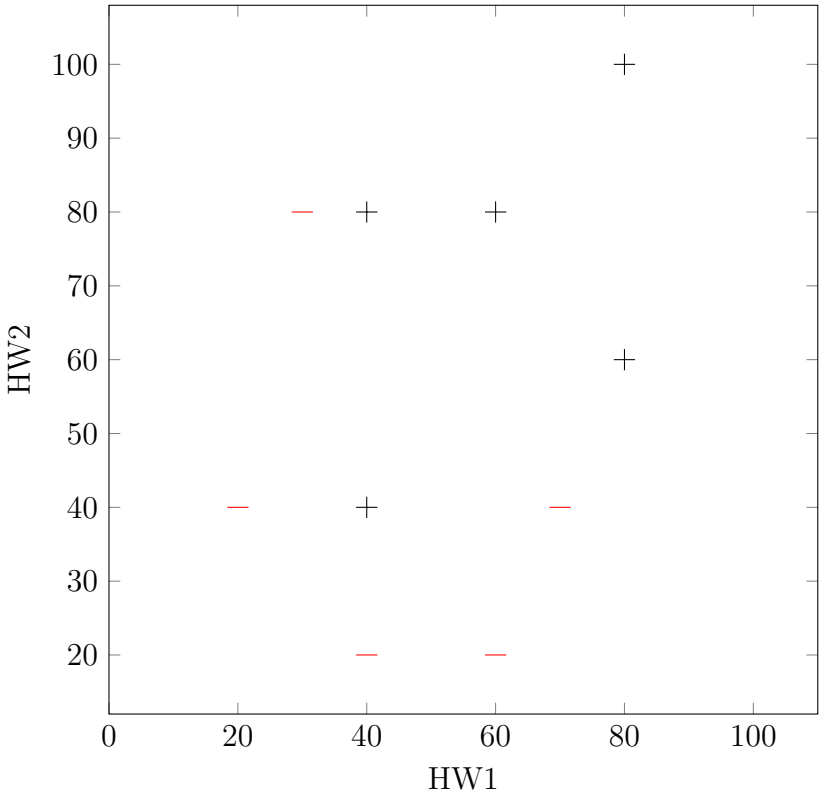
(f) **Math:** With $k = N$, for any dataset $\mathcal{D}$ with the form specified in the beginning of this question, write down a mathematical expression for the predicted value $\hat{y} = h(x)$. Your response shouldn't include a reference to the neighborhood function $\mathcal{N}()$.

2.2. **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 changes is likely to help achieve this goal and what is the corresponding effect on the prediction time?

- ◯ Increase the value of $k \rightarrow$ Increase in prediction time
- ◯ Decrease the value of $k \rightarrow$ Increase in prediction time
- ◯ Increase the value of $k \rightarrow$ Decrease in prediction time
- ◯ Decrease the value of $k \rightarrow$ Decrease in prediction time

2.3. You have just enrolled into your favourite course at CMU - Introduction to Machine Learning 10-301/601 - but you have not yet decided if you want to take it for a grade or as pass/fail. You want to use your performance in HW1 and HW2 to make this decision. You follow a general rule that if you can get at least a B in the course, you will take it for a grade, and if not, you will take it as pass/fail.

You have just learned the new classification technique, k-NN, and wish to employ it to make your decision. You start by collecting data from 10 prior students: you measure their performance on HW1 and HW2, along with their final letter grades. You then create a binary label based on the final grades such that you assign a label of $+$ if the final grade is at least a B and $-$ otherwise. Next, you train a k-NN model on this data set using the **Euclidean distance metric**.



(a) **Numerical answer:** What is the training error rate on this dataset for a k-NN model where $k = 1$?

(b) **Numerical answer:** What is the training error rate on this dataset for a k-NN model where $k = 3$?

2/10

(c) **True or False:** Using Euclidean distance as the distance measure, the decision boundary of a k-NN model with $k = 1$ is a piece-wise straight line, that is, it contains only straight line segments. **Justify your answer.**

◯ True

◯ False

(d) **Drawing:** In the image above, draw a rough decision boundary for k = 1. Clearly label the + and - sides of the decision boundary.

(e) You have scored 60 in HW1 and 40 in HW2, and you now want to predict if your final grade would be at least a B.

i. **Numerical answer:** What would be the predicted class when $k = 1$?

ii. **Numerical answer:** What would be the predicted class when $k = 3$?

(f) **Short answer:** Looking at the training errors, you choose the model with k = 1 as it has the lowest training error. Do you think this is the right approach to select a model? Why or why not?

2.4. **Select all that apply:** Which of the following statements about k-NN models is/are *always* true?

- ☐ Decreasing k makes a k-NN model have simpler or less complex decision boundaries.
- ☐ Increasing k makes a k-NN model less sensitive to outliers.
- ☐ k-NNs can be applied to classification problems but not regression problems.
- ☐ k-NNs can be applied to datasets with real-valued features but not categorical features.
- ☐ None of the above

2.5. Consider a binary (two classes) classification problem using k-nearest neighbors. We have n 1-dimensional training points $\{x^{(1)}, x^{(2)}, ..., x^{(n)}\}$ with $x^{(i)} \in \mathbb{R}$, and their corresponding labels $\{y^{(1)}, y^{(2)}, ..., y^{(n)}\}$ with $y^{(i)} \in \{-, +\}$. Assume the data points $x^{(1)}, x^{(2)}, ..., x^{(n)}$ are sorted in ascending order and we use Euclidean distance as the distance metric.

True or False: We can build a decision tree (with splits of the form "$x \geq t$" and "$x < t$", for $t \in \mathbb{R}$) that behaves *exactly* the same (i.e. makes the same predictions for every possible input) as the 1-nearest neighbor classifier on this dataset.

- ◯ True
- ◯ False

3 Model Selection and Errors (0 points)

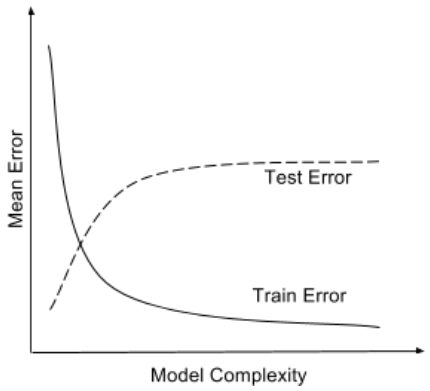
3.1. **Train and test errors:** In this problem, we will see how you can debug a classifier by looking at its train and test errors. Consider a classifier trained until convergence on some training data $\mathcal{D}^{\text{train}}$, and tested on a separate test set $\mathcal{D}^{\text{test}}$. You look at the test error, and find that it is very high. You then compute the training error and find that it is close to 0.

(a) **Short Answer:** What is this scenario called?

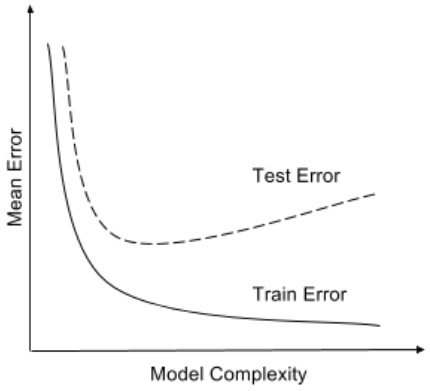
(b) **Select all that apply:** Which of the following actions are likely to address the issue you identified in the previous question?

- ☐ Increasing the training data size.
- ☐ Decreasing the training data size.
- ☐ Increasing model complexity (For example, if your classifier is a decision tree, increase the depth).
- ☐ Decreasing model complexity.
- ☐ Training on a combination of $\mathcal{D}^{\text{train}}$ and $\mathcal{D}^{\text{test}}$ and test on $\mathcal{D}^{\text{test}}$
- ☐ None of the above

(c) **Select one:** Say you plot the train and test errors as a function of the model complexity. Which of the following two plots is your plot expected to look like?



(a)



(b)

- ◯ Plot A
- ◯ Plot B

3.2. What are the effects of the following on overfitting? Choose the best answer.

(a) Increasing the max depth a decision tree is allowed to grow to.

◯ Less likely to overfit

◯ More likely to overfit

(b) Increasing the minimum mutual information threshold required to add a split to a decision tree.

◯ Less likely to overfit

◯ More likely to overfit

(c) Increasing the minimum number of data points required to add a split to a decision tree.

◯ Less likely to overfit

◯ More likely to overfit

(d) Increasing k in k-nearest neighbor.

◯ Less likely to overfit

◯ More likely to overfit

(e) Increasing the training data size for decision trees. Assume that training data points are drawn independently from the true data distribution.

◯ Less likely to overfit

◯ More likely to overfit

(f) Increasing the training data size for 1-nearest neighbor. Assume that training data points are drawn independently from the true data distribution.

◯ Less likely to overfit

◯ More likely to overfit

3.3. Consider a learning algorithm that uses two hyperparameters, γ and ω, and it takes 1 hour to train *regardless* of the size of the training set.

We choose to do random subsampling cross-validation, where we do K runs of cross-validation and for each run, we randomly subsample a fixed fraction αN of the dataset for validation and use the remaining for training, where $\alpha \in (0, 1)$ and N is the number of data points.

(a) **Numerical answer:** In combination with the cross-validation method above, we choose to do grid search on discrete values for the two hyperparameters.

Given $N = 1000$ data points, $K = 4$ runs, and $\alpha = 0.25$, if we have 100 hours to complete the entire cross-validation process, what is the maximum number of discrete values of γ that we can include in our search if we also want to include 8 values of ω? Assume that any computations other than training are negligible.

(b) **Short answer:** In 1-2 concise sentences, describe one advantage of increasing the value of α.

3.4. **Select one:** When implementing k-Fold Cross-Validation for a Linear Regression model with a 1-dimensional input feature, which of the following best describes the correct procedure?

- ◯ The dataset is divided into k contiguous blocks, where each block consists of adjacent data points based on the sorting of the 1-dimensional feature values. The model is trained k times, each time using a different block as the test set and the remaining blocks for training.
- ◯ The dataset is randomly partitioned into k equal-sized subsets. For each fold, the Linear Regression model is trained on k-1 subsets and validated on the remaining subset to assess the model's performance.
- ◯ The model is trained once using the entire dataset as the training set and then tested k times, each time with a different single data point as the test set.
- ◯ We fit a Linear Regression model to every possible subset of size k of the training data (resulting in $\binom{N}{k}$ models) and then average their errors on a held-out test dataset.

4 Perceptron (0 points)

4.1. Suppose you are given the following dataset:

Example Number	X_1	X_2	Y
1	-1	2	-1
2	-2	-2	+1
3	1	-1	+1
4	-3	1	-1

You wish to perform the Batch Perceptron algorithm on this data. Assume you start with initial weights $\theta^T = [0, 0]$ and bias $b = 0$, and that you pass through all of the examples in order of their example number.

(a) **Numerical answer:** What would be the updated weight vector θ after we pass example 1 through the Perceptron algorithm?

(b) **Numerical answer:** What would be the updated bias b after we pass example 1 through the Perceptron algorithm?

(c) **Numerical answer:** What would be the updated weight vector θ after we pass example 2 through the Perceptron algorithm?

(d) **Numerical answer:** What would be the updated bias b after we pass example 2 through the Perceptron algorithm?

(e) **Numerical answer:** What would be the updated weight vector θ after we pass example 3 through the Perceptron algorithm?

(f) **Numerical answer:** What would be the updated bias b be after we pass example 3 through the Perceptron algorithm?

(g) **True or False:** Your friend stops you here and tells you that you do not need to update the Perceptron weights or bias anymore; is this true or false?

◯ True

◯ False

4.2. **Select all that apply:** Let $S = \{(\mathbf{x}^{(1)}, y^{(1)}), \cdots, (\mathbf{x}^{(n)}, y^{(n)})\}$ be n linearly separable points by a separator through the origin in $\mathbb{R}^d$. Let S' be generated from S as: $S' = \{(c\mathbf{x}^{(1)}, y^{(1)}), \cdots, (c\mathbf{x}^{(n)}, y^{(n)})\}$, where $c > 1$ is a constant. Suppose that we would like to run the perceptron algorithm on both data sets separately, with the same initial $\boldsymbol{\theta}$ from class, and that the perceptron algorithm converges on S. Which of the following statements are true?

□ The mistake bound of perceptron on S' is larger than the mistake bound on S

□ The perceptron algorithm when run on S and S' returns the same classifier, modulo constant factors (i.e., if $\mathbf{w}_S$ and $\mathbf{w}'_S$ are outputs of the perceptron for S and S', then $\mathbf{w}'_S = c_1\mathbf{w}_S$ for some constant c_1).

□ The perceptron algorithm converges on S'.

□ None of the above.

4.3. **True or False:** Given a linearly separable dataset, the convergence time of the perceptron algorithm depends on the sample size n.

◯ True

◯ False

4.4. **Select all that apply:** Which of the following are inductive biases of the perceptron algorithm?

□ Most of the cases in a small neighborhood in feature space belong to the same class.

□ The true decision boundary is linear.

□ We prefer to correct the most recent mistakes.

□ We prefer the simplest hypothesis that explains the data.

□ None of the above.

4.5. **True or False:** All *data points* $(\mathbf{x}, y)$ that the Perceptron learning algorithm sees during training affect the weights equally.

◯ True

◯ False

5 Linear Regression (0 points)

5.1. **Select all that apply:** For some function $f : \mathbb{R} \to \mathbb{R}$, suppose the unique argmax of f is x^*. Which of the following functions are guaranteed to have the same argmax as f?

- ☐ $g(x) = f(x) + 2$
- ☐ $g(x) = 2f(x)$
- ☐ $g(x) = f(x)^2$
- ☐ $g(x) = 2^{f(x)}$
- ☐ None of the above.

5.2. Consider linear regression on 1-dimensional data i.e., $x, y \in \mathbb{R}$. We apply linear regression in both directions on this data: we first fit a model to predict y using x and get $y = m_1 x + b_1$ as the fitted line that minimizes the mean squared error, then we fit a model to predict x using y and get $x = m_2 y + b_2$.

True or False: It is always the case that $m_2 = \frac{1}{m_1}$.

- ◯ True
- ◯ False

5.3. **Select all that apply:** Which of the following statements about the gradient of a function is/are true?

- ☐ The gradient is the same dimensionality as the function's output.
- ☐ The gradient points in the direction of steepest descent.
- ☐ The gradient of a function always exists everywhere.
- ☐ The gradient is the limit of the slope between two points on the function as the distance between the points approaches ∞.
- ☐ None of the above.

5.4. Instead of mean squared error, suppose instead that you fit a linear regression model by minimizing the mean cubed error:

$$J(\theta) = \frac{1}{N} \sum_{i=1}^{N} \left(\theta^T \mathbf{x}^{(i)} - y^{(i)}\right)^3$$

(a) **Math:** What is the gradient of the mean cubed error with respect to the parameter vector θ?

(b) **True or False:** This loss function is more sensitive to outliers than the mean squared error.

- ◯ True
- ◯ False

(c) **Short answer:** In 1-2 concise sentences, briefly describe the primary issue with this objective function for regression?

5.5. **True or False:** Consider a linear regression model with only one parameter, the intercept, i.e., $y = b$. Then, given N data points $(x^{(i)}, y^{(i)})$ (where $x^{(i)}$ is the feature and $y^{(i)}$ is the target), minimizing the sum of squared errors results in b being the *median* of the $y^{(i)}$ values. Briefly justify your answer

- ◯ True
- ◯ False

5.6. You decide to optimize the function $f(x) = x^2$ with respect to x using gradient descent. You initialize $x^{(0)}$ to -1 and use a step size of $\eta^{(0)} = 1.5$.

(a) **Fill in the blanks:** Fill in the table below with the results you get from running gradient descent in this setting for two iterations; most of the 0^{th} iteration has been filled in on your behalf; you should fill in the missing element in this line along with all the other missing elements.

t	$x^{(t)}$	$f(x^{(t)})$	$\nabla_x f(x^{(t)})$
0	-1	1	_____
1	_____	_____	_____
2	_____	_____	_____

(b) **Select all that apply:** Based on your findings from the previous part, which of the following changes could help you achieve better results?

- ☐ Decreasing the step size.
- ☐ Running gradient descent for more iterations.
- ☐ Moving the initial value $x^{(0)}$ further from the origin.
- ☐ Negating the function and running gradient *ascent*.
- ☐ None of the above.

5.7. **Math:** Your friend accidentally solved linear regression using the wrong objective function for mean squared error; specifically, they used the following objective function that contains two mistakes: 1) they forgot the $1/N$ and 2) they have one sign error.

$$J(\mathbf{w}, b) = \sum_{i=1}^{N} \left(y^{(i)} - \left(\sum_{j=1}^{M} w_j x_j^{(i)} - b \right) \right)^2$$

You realize that you can still use the parameters that they learned, $\mathbf{w}$ and b, to compute the minimizer of the mean squared error. Write the equation that implements this corrected prediction function $h(\mathbf{x}, \mathbf{w}, b)$ using your friend's learned parameters, $\mathbf{w}$ and b.

Do not remove this page! Use this page for scratch work.

Do not remove this page! Use this page for scratch work.

Do not remove this page! Use this page for scratch work.

Do not remove this page! Use this page for scratch work.