HOMEWORK 3: DECISION TREES, K-NN, PERCEPTRON, REGRESSION [*]

10-301 / 10-601 INTRODUCTION TO MACHINE LEARNING (SPRING 2025)
http://www.cs.cmu.edu/~mgormley/courses/10601/

OUT: Monday, February 3
DUE: Monday, February 10
TAs: Bhargav, Changwook, Maxwell, Santiago, Zachary, Neural the Narwhal

Summary It's time to practice what you've learned! In this assignment, you will answer questions on topics we've covered in class so far, including Decision Trees, K-Nearest Neighbors, Perceptron, and Linear Regression. This assignment consists of a written component split into four sections, one for each topic. These questions are designed to test your understanding of the theoretical and mathematical concepts related to each topic. For each topic, you will also apply your understanding of the concept to the related ideas such as overfitting, error rates, and model selection. This homework is designed to help you apply what you've learned and solve a few concrete problems.

START HERE: Instructions

- **Collaboration Policy**: Please read the collaboration policy here: http://www.cs.cmu.edu/~mgormley/courses/10601/syllabus.html
- **Late Submission Policy:** See the late submission policy here: http://www.cs.cmu.edu/~mgormley/courses/10601/syllabus.html
- **Submitting your work:** You will use Gradescope to submit answers to all questions.
  - **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).

[*] Compiled on Wednesday 3rd September, 2025 at 01:40

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

□ Noam Chomsky

□ 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

☒ Noam Chomsky

☒ 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

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

3. (0 points) **Select one:** Did you fill out the Exit Poll for the previous HW? Completing the exit poll will count towards your participation grade.

   ● Yes

2 Decision Tree (Revisited) (13 points)

1. Consider the following $\{a, b, c, d, e, f, g, h\} \times \{1, 2, 3, 4, 5, 6, 7, 8\}$ chessboard, with a white rook on d4. Suppose our goal is to perfectly classify the positions that the rook can get to in one move, knowing that the rook cannot stay in place. (For those who aren't chess-heads, a rook may move any number of squares horizontally or vertically.) All the squares accessible in one move are labelled as $+1$, and all the squares not accessible in one move are -1. Let the horizontal axis denote feature x_1 and vertical axis denote feature x_2.

   NOTE: For the purposes of these questions, you may consider the alphabet as a totally ordered set such that, for example, $a < d < h$.

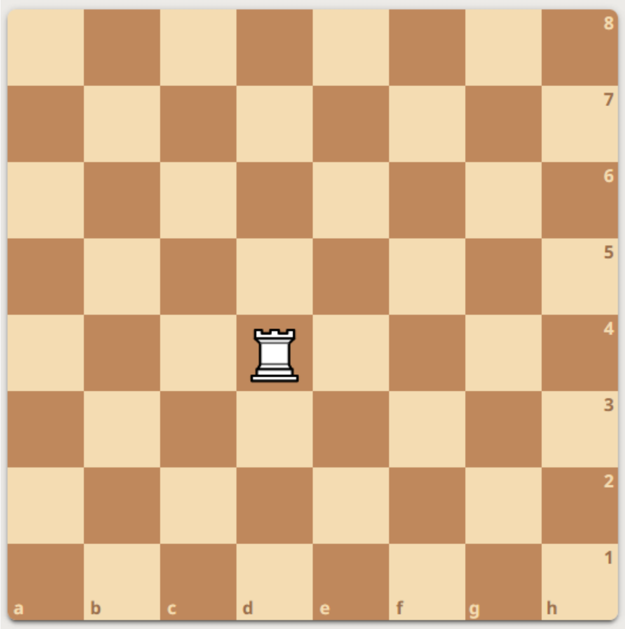


Figure 1: Rook on a chessboard

   (a) (2 points) What is the minimum depth of a binary decision tree that perfectly classifies the squares our rook can move to in one move in Figure 1, using *only* splits of the form $x_1 < C$ or $x_2 < C$ for different values C?

   Your Answer

   4

(b) (2 points) What is the minimum depth of a binary decision tree that perfectly classifies the squares our rook can move to in one move in Figure 1, where each split can be a function of either x_1 or x_2 but not both x_1 and x_2?

Since this is a binary decision tree, we can only use features that split into two branches e.g., $b < x_1 < f, \quad x_2 < 1, \quad \text{or} \quad x_2 > 3$

Your Answer

2

(c) (2 points) What is the minimum depth of a binary decision tree that perfectly classifies the squares our rook can move to in one move in Figure 1, using *any* splits that involve x_1, x_2, or both?

Hint: Consider using combinations of logical operators such as AND, OR and NOT.

Your Answer

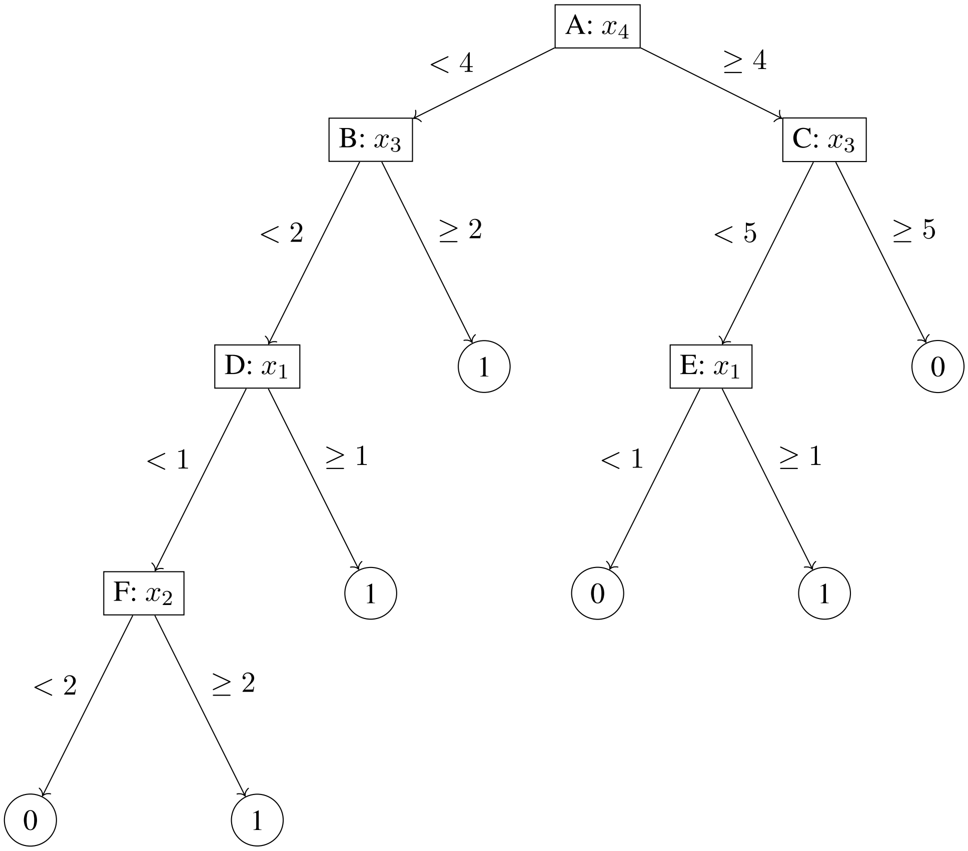
1

(d) (1 point) **True or False:** A decision tree can *perfectly* classify the interior and only the interior of a (2-dimensional) circle of radius $r + 1$ using a finite amount of splits of the form $x_1 < C$ or $x_2 < C$ for different values C.

○ True

● False

2. Consider the following decision tree:



We want to perform reduced error pruning on the decision tree using the following validation dataset. In the case of a tie between two nodes to be pruned, break ties in favor of the alphabetically earlier node (e.g., prune node C before node F)

x_1	x_2	x_3	x_4	y
1	3	3	5	1
1	1	5	6	0
0	2	0	3	0
2	2	1	1	0
0	3	4	4	1

The following table specifies the majority vote for all the training data points under each node of the tree:

Node	A	B	C	D	E	F
Majority Vote	0	1	0	0	1	0

(a) (1 point) **Select one:** Which of the following splits would be the **first** to be removed?

○ A

○ B

○ C

○ D

○ E

○ F

(b) (1 point) **Select one:** Which of the following splits would be the **last** to be removed?

○ A

○ B

○ C

○ D

○ E

○ F

3. (2 points) **Select all that apply:** Which of the following are valid ways to avoid overfitting?

■ Prune the tree so that cross-validation error is minimal.

□ Increase the tree depth.

■ Set a threshold for a minimum number of examples required to split at an internal node.

□ Decrease the training set size.

□ None of the above.

4. A discrete hyperparameter is a hyperparameter that can only take on a finite set of values e.g., in the ID3 algorithm, the minimum number of data points needed to split a node is a discrete hyperparameter as it can only take on integer values between 1 and the number of training data points (inclusive). Suppose you have a machine learning model with two discrete hyperparameters: α, which can take on 10 possible values and β, which can take on 20 possible values. Unfortunately, training your model is computationally expensive: you only have enough time to try B different combinations of the hyperparameters. You are considering using either random search or grid search to find the best setting of the hyperparameters.

   (a) (1 point) If B $\geq$ 200, would you expect random search to perform better than grid search in terms of finding a better setting of the hyperparameters? Why or why not?

   Your answer:

   No, because with $B \geq 200$, we can perform exhaustive grid search to find the best combination of hyperparameters.

   (b) (1 point) Based on the intuition presented in Lecture 5, if B = 50, would you expect random search or grid search to perform better in terms of finding a higher quality setting of the hyperparameters? Explain **why** your chosen strategy performs better. It does not suffice to only cite lecture.

   Your answer:

   Random search would perform better in terms of finding a higher quality setting of the hyperparameters. This is because with $B = 50$, grid search can find only the first 50 combinations of α and β and can't the later combinations. With random search, we have some chance to find this combination.

3 k-Nearest Neighbors (28 points)

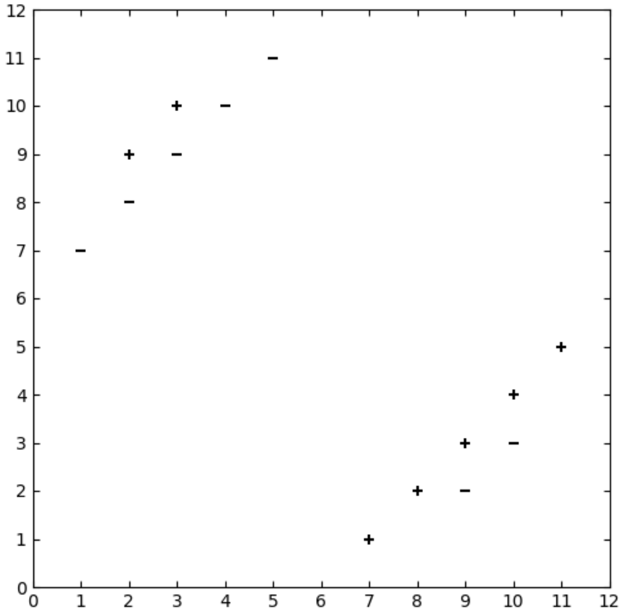


Figure 2: k-NN Dataset

1. Consider a k-nearest neighbors (k-NN) binary classifier which assigns the class of a test point to be the class of the majority of the k-nearest neighbors in the training dataset, according to the Euclidean distance metric. Assume that ties are broken by selecting one of the labels uniformly at random.

   NOTE: An example tie scenario can occur when the classes of the 6 nearest neighbors are {+, +, +, -, -, -} i.e. the number of neighbors belonging to each class type is equal. In this case, you can assume the test point's class to be + or - randomly.

   (a) (2 points) Using Figure 2 shown above to train the classifier and choosing $k = 6$, what is the classification error on the training set? **Report your answer either as a fraction or as a decimal with 4 decimal places after the decimal point.**

Your answer:

$\frac{2}{7}$

(b) (2 points) **Select all that apply:** Let's say that we have a new test point (not present in our training data) $\mathbf{x}^{\text{new}} = [3, 11]^T$ (where 3 is the horizontal component and 11 the vertical component) that we would like to apply our k-NN classifier to, as seen in figure 3.

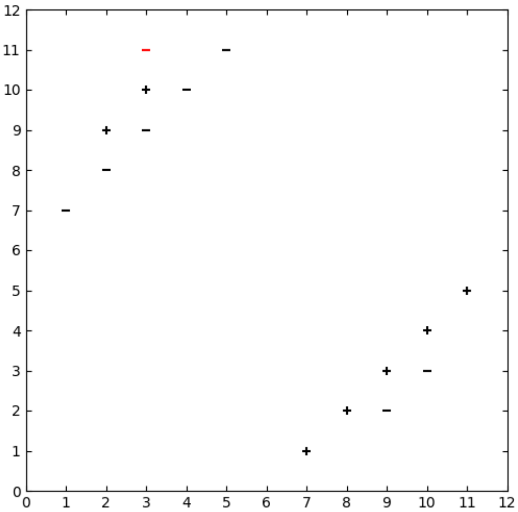


Figure 3: k-NN Dataset with Test Point

For which values of k is this test point always correctly classified by the k-NN algorithm?

□ k = 1

■ k = 5

■ k = 9

□ k = 12

□ None of the above

2. **Select one:** Assume we have a large labeled dataset that is randomly divided into a training set and a test set, and we would like to classify points in the test set using a k-NN classifier.

(a) (1 point) In order to minimize the classification error on this test set, we should always choose the value of k which minimizes the training set error.

○ True

● False

(b) (2 points) **Select one:** Instead of choosing the hyperparameters by merely minimizing the training set error, we instead consider splitting the training-all data set into a training and a validation data set, and choose the hyperparameters that lead to lower validation error. Is choosing hyperparameters based on validation error better than choosing hyper-parameters based on training error?

● Yes, lowering validation error instead of training error is better because lowering training error will not always help generalize our model and may lead to overfitting.

○ Yes, lowering validation error is better for the model because cross-validation guarantees a better test error.

○ No, lowering training error instead of validation error is better because lowering validation error will not help generalize our model and may lead to overfitting.

○ No, lowering training error is better for the model because we have to learn the training set as well as possible to guarantee the best possible test error.

(c) (2 points) **Select one:** Your friend Sally suggests that instead of splitting the original training set into separate training and validation sets, we should instead use the test set as the validation data for choosing hyperparameters. Is this a good idea? Justify your opinion with no more than 3 sentences.

○ Yes

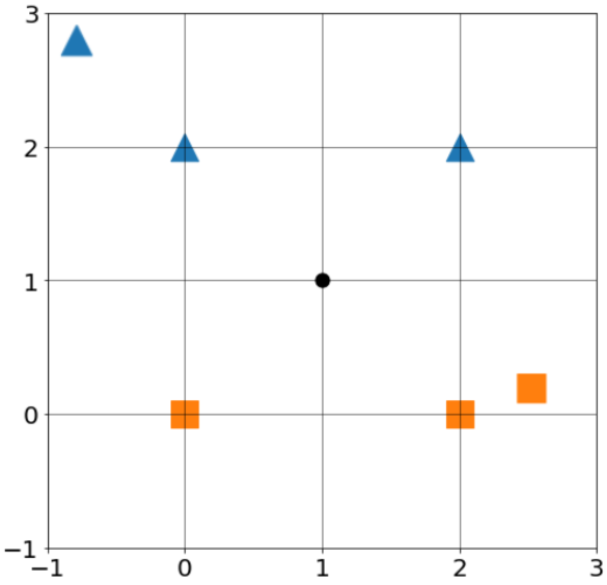
● No

Your answer:

If we chose the test set as the validation data for choosing hyperparameters, we would not have any data for measuring the model accuracy. We may overfit the model to the dataset.

3. (2 points) **Select all that apply:** Consider a binary k-NN classifier where $k = 4$ and the two labels are "triangle" and "square". Consider classifying a new point $\mathbf{x} = (1, 1)$, where two of the $\mathbf{x}$'s nearest neighbors are labeled "triangle" and two are labeled "square" as shown below.

   If there is a tie in the distance amongst the nearest neighbors, then break ties in favor of lower values for the horizontal axis first and then lower values for the vertical axis.



   Which of the following methods will guarantee breaking or avoiding ties in the majority vote when classifying x?

   ■ Use k = 2 instead

   □ Flip a coin to randomly assign a label to $\mathbf{x}$ (from the labels of its 4 closest points)

   □ Use $k = 3$ instead

   ■ Use $k = 5$ instead

   □ None of the above.

4. (3 points) **Select all that apply:** Which of the following is/are correct statement(s) about k-NN models?

   ■ A larger k tends to give a smoother decision boundary.

   ■ To reduce the impact of noise or outliers in our data, we should increase the value k.

   □ If we make k too large, we could end up overfitting the data.

   ■ We can use cross-validation to help us select the value of k.

   □ We should never select the k that minimizes the error on the validation dataset.

   □ None of the above.

5. Consider the following data concerning the relationship between academic performance and salary after graduation. High school GPA and university GPA are two numerical features and salary is the numerical target. Note that salary is measured in thousands of dollars per year.

Student ID	**High School GPA**	**University GPA**	**Salary**
1	2.5	3.8	45
2	3.3	3.5	90
3	4.0	4.0	142
4	3.0	2.0	163
5	3.8	3.0	2600
6	3.3	2.8	67
7	3.9	3.8	unknown

(a) (2 points) Among Students 1 to 6, who is the nearest neighbor to Student 7, using Euclidean distance? Answer the Student ID only.

Your answer:

3

(b) (2 points) Now, our task is to predict the salary Student 7 earns after graduation. We apply k-NN to this regression problem: the prediction for the numerical target (salary in this example) is equal to the average of salaries for the top k nearest neighbors. If $k = 3$, what is our prediction for Student 7's salary? Be sure to use the same unit of measure (thousands of dollars per year) as the table above.
Round your answer to the nearest integer.

Your answer:

(c) (2 points) **Select all that apply:** Suppose that the first 6 students shown above are only a subset of your full training data set, which consists of 10,000 students. We apply k-NN regression using Euclidean distance to this problem and we define training loss on this full data set to be the mean squared error (MSE) of salary. Now consider the possible consequences of modifying the data in various ways. Which of the following changes **could** have an effect on training loss on the full data set as measured by mean squared error (MSE) of salary?

■ Rescaling only "High School GPA" to be a percentage of 4.0

■ Rescaling only "University GPA" to be a percentage of 4.0

□ Rescaling both High School GPA and University GPA by the same percentage/scale

□ None of the above.

6. An archaeologist discovers a 242 kilobyte 8-inch floppy disk buried beneath the hedges near Wean Hall. The floppy disk contains a few hundred black and white images of 3x3 pixels. You are asked to classify them as either a photo ($y = +$) or artwork ($y = -$) to aid in the analysis.

   You build a k-Nearest Neighbor (k-NN) classifier trained on a training dataset obtained from the web (converted to similarly small black and white images). Suppose you are informed that each image is represented as a 3×3 matrix $\mathbf{x}$ of binary values and you plan to use Hamming distance to measure the distance between each pair of 3×3 pixel images as follows:

$$d(\mathbf{u}, \mathbf{v}) = \sum_{i=1}^{3} \sum_{j=1}^{3} \mathbb{1}(\mathbf{u}_{i,j} \neq \mathbf{v}_{i,j}) = \text{the number of pixels that differ between } \mathbf{u} \text{ and } \mathbf{v}$$

   While calculating the distance metric, if there is a tie in distance among the points competing for k nearest points, the classifier increases k to include all those tied points in the majority vote. If, in the end, there is a tie in the vote, your classifier returns $\hat{y} =?$. You can try out your k-NN implementation on the images below.

i	y	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	$x_{3,1}$	$x_{3,2}$	$x_{3,3}$
1	$+$	0	0	1	0	1	1	0	0	0
2	$+$	1	0	1	0	1	0	1	0	0
3	$+$	0	1	0	1	1	1	1	0	1
4	$-$	0	0	1	0	0	0	0	0	0
5	$-$	1	0	1	1	1	0	0	1	1

Table 1: Training Data

i	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	$x_{3,1}$	$x_{3,2}$	$x_{3,3}$
6	1	0	1	0	0	0	1	0	1
7	0	0	1	0	0	1	1	1	0

Table 2: Test Data

(a) (1 point) What is the distance between $\mathbf{x}^{(2)}$ and $\mathbf{x}^{(6)}$?

Your answer:

(b) (1 point) **Select one:** What would a k-NN classifier with $k = 3$ predict as the label for test point $\mathbf{x}^{(7)}$?

○ $\hat{y} = +$

○ $\hat{y} = -$

○ $\hat{y} = ?$

(c) (1 point) **Select one:** What would a k-NN classifier with $k = 5$ predict as the label for test point $\mathbf{x}^{(7)}$?

○ $\hat{y} = +$

○ $\hat{y} = -$

○ $\hat{y} = ?$

(d) (2 points) **Short answer:** Your friend says that you should try using Euclidean distance because it might give better results. Do you agree that switching could lead to lower test error? Why or why not?

Your answer:

7. (3 points) **Numeric answer** Let's say you have a large labeled dataset and you want to train a k-NN classifier on it. You have decided you're going to perform hyperparameter optimization by performing a grid search. The specific hyperparameters you choose to vary are the value of k and the distance metric. You also decide you want to perform cross-validation when assessing these different classifiers during the course of your grid search.

   If you want to try 5 different values of k (3, 5, 7, 9 and 11), 2 different distance metrics (Euclidean distance and Hamming distance) and you choose to do 10-fold cross-validation. How many different classifiers will you end up learning during the hyperparameter optimization process?

   Your answer:

   100

4 Perceptron (17 points)

1. (1 point) **True or False:** Consider running the online perceptron algorithm on some sequence of examples S (an example is a data point and its label). Let S' be the same set of examples as S, but presented in a different order.

   The online perceptron algorithm is guaranteed to make the same number of mistakes on S as it does on S'.

   ○ True

   ● False

2. (3 points) **Select all that apply:** Suppose we have a perceptron whose inputs are 2-dimensional vectors and each feature vector component is either -1 or 1, i.e., $x_i \in \{-1, 1\}$. The prediction function is $y = \text{sign}(w_1 x_1 + w_2 x_2 + b)$, and

   $$\text{sign}(z) = \begin{cases} 1, & \text{if } z > 0 \\ -1, & \text{otherwise.} \end{cases}$$

   Which of the following functions can be implemented with the above perceptron? That is, for which of the following functions does there exist a set of parameters w, b that correctly define the function.

   ■ AND function, i.e., the function that evaluates to 1 if and only if all inputs are 1, and -1 otherwise.

   ■ OR function, i.e., the function that evaluates to 1 if and only if at least one of the inputs are 1, and -1 otherwise.

   □ XOR function, i.e., the function that evaluates to 1 if and only if the inputs are not all the same. For example

   $$\text{XOR}(1, -1) = 1, \text{ but } \text{XOR}(1, 1) = -1.$$

   □ None of the above.

3. (2 points) **Select one:** Suppose we have a dataset $\{(\mathbf{x}^{(1)}, y^{(1)}), \ldots, (\mathbf{x}^{(N)}, y^{(N)})\}$, where $\mathbf{x}^{(i)} \in \mathbb{R}^M$, $y^{(i)} \in \{+1, -1\}$. We would like to apply the perceptron algorithm on this dataset. Assume there is no intercept term. How many parameter values is the perceptron algorithm learning?

   ○ N

   ○ $N \times M$

   ● M

4. (2 points) **Select one:** Suppose we have been running the perceptron algorithm for 10 iterations. The following table shows a data set and the number of times each point has been misclassified so far. What is the current separating plane $\boldsymbol{\theta}$ found by the algorithm, i.e. $\boldsymbol{\theta} = [b, \theta_1, \theta_2, \theta_3]$? Assume that the initial weights and bias are all zero.

x_1	x_2	x_3	y	Times Misclassified
2	1	5	1	8
5	3	3	1	7
1	6	2	1	0
7	2	1	-1	5
3	2	6	-1	9

○ $[1, -11, 1, 2]$

○ $[1, -3, 0, 1]$

○ $[-1, 3, 0, -1]$

● $[-1, 11, -1, -2]$

5. (2 points) **Select one:** Suppose we have data whose examples are of the form $[x_1, x_2]$, where $x_1 - x_2 = 0$. We do not know the label for each element. Suppose the perceptron algorithm starts with $\boldsymbol{\theta} = [3, 5]$; which of the following values will $\boldsymbol{\theta}$ never take on in the process of running the perceptron algorithm on the data?

○ $[-1, 1]$

○ $[4, 6]$

● $[-3, 0]$

○ $[-6, -4]$

6. (2 points) **Select all that apply:** Consider the linear decision boundary below and the test dataset shown. Which of the following weight vectors $\boldsymbol{\theta}$ is paired with its corresponding test error on this dataset? (Note: Assume the decision boundary is fixed and does not change while evaluating error.)

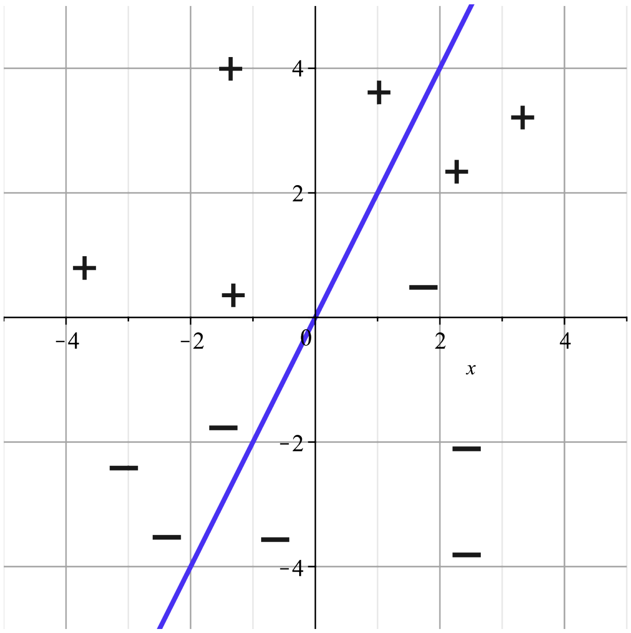
   □ $\boldsymbol{\theta} = [-2, 1]$, error = 5/13

   ■ $\boldsymbol{\theta} = [2, -1]$, error = 8/13

   □ $\boldsymbol{\theta} = [2, -1]$, error = 5/13

   □ $\boldsymbol{\theta} = [-2, 1]$, error = 8/13

   □ None of the above.



7. The following problem will walk you through an application of the Perceptron Mistake Bound. The following table shows a linearly separable dataset, and your task will be to determine the mistake bound for the dataset.

   NOTE: The proof of the perceptron mistake bound requires that the optimal linear separator passes through the origin. To make the linear separator pass through the origin, we fold the bias into the weights and prepend a 1 to each training example's input. The original data is on the left, and the result of this prepending is shown on the right. **Be sure to use the modified dataset on the right in your calculations.**

x_1	x_2	y
-4	3	1
-2	5	-1
-1	4	-1
1	1	1
2	-1	1
4	3	1

$\boldsymbol{x_0}$	x_1	x_2	y
1	-4	3	-1
1	-2	5	-1
1	-1	4	-1
1	1	1	1
1	2	-1	1
1	4	3	1

   (a) (2 points) Compute the radius R of the "circle" centered at the origin that bounds the data points. **Round to 4 decimal places after the decimal point.**

   Radius:

   5.4772

   (b) (2 points) Assume that the linear separator with the largest margin is given by

$$\boldsymbol{\theta}^{*T}\begin{bmatrix}1\\x_1\\x_2\end{bmatrix}=0,\text{ where }\boldsymbol{\theta}^*=\begin{bmatrix}1.1538\\0.3077\\-0.4615\end{bmatrix}$$

   Now, compute the margin of the dataset.
   Round to 4 decimal places after the decimal point.

   Margin:

   0.7811

   (c) (1 point) Based on the above values, what is the theoretical perceptron mistake bound for this dataset, given this linear separator? **Give the tightest bound possible, as an integer**

   Mistake Bound:

   49

5 Linear Regression (17 points)

1. Consider the following dataset:

x	9.0	2.0	6.0	1.0	8.0
y	1.0	0.0	3.0	0.0	1.0

Let $\boldsymbol{x}$ be the vector of datapoints and $\boldsymbol{y}$ be the label vector. Here, we are fitting the data using gradient descent, and our objective function is $J(w, b) = \frac{1}{N}\sum_{i=1}^{N}(wx_i + b - y_i)^2$ where N is the number of data points, w is the weight, and b is the intercept.

Note: Showing your work in these questions is optional, but it is recommended to help us understand where any misconceptions may occur. We may give partial credit for correct work if your answer is incorrect.

(a) (2 points) If we initialize the weight as 3.0 and intercept as 0.0, what is the gradient of the loss function with respect to the weight w, calculated over all the data points, in the first step of the gradient descent update?
Round to 4 decimal places after the decimal point.

Gradient:

209.2

Work

(b) (2 points) What is the gradient of the loss function with respect to the intercept b, calculated over all the data points, in the first step of the gradient descent update?

Gradient:

29.2

Work

(c) (2 points) Let the learning rate be 0.01. Perform one step of gradient descent on the data. Fill in the following blanks with the value of the weight and the value of the intercept after this step. **Round to 4 decimal places after the decimal point.**

Weight:

0.908

Intercept:

-0.292

2. Consider a dataset $\mathcal{D}_1 = \{(x^{(1)}, y^{(1)}), \ldots, (x^{(N)}, y^{(N)})\}$. Assume the linear regression model that minimizes the mean-squared error on $\mathcal{D}_1$ is $y = w_1 x + b_1$.

   (a) (2 points) **Select one:** Now, suppose we have the dataset $\mathcal{D}_2 = \{(x^{(1)} + \alpha,\ y^{(1)} + \beta), \ldots, (x^{(N)} + \alpha,\ y^{(N)} + \beta)\}$ where $\alpha > 0, \beta > 0$ and $w_1\alpha \neq \beta$. Assume the linear regression model that minimizes the mean-squared error on $\mathcal{D}_2$ is $y = w_2 x + b_2$. Select the correct statement about w_1, w_2, b_1, b_2 below. Note that the statement should hold no matter what values α, β take on within the specified constraints.

   ○ $w_1 = w_2, b_1 = b_2$

   ○ $w_1 \neq w_2, b_1 = b_2$

   ● $w_1 = w_2, b_1 \neq b_2$

   ○ $w_1 \neq w_2, b_1 \neq b_2$

   (b) (2 points) We decide to ask a friend to analyze $\mathcal{D}_1$; however, he makes a mistake by duplicating a subset of the rows in $\mathcal{D}_1$. Explain why the linear regression parameters that minimize mean-squared error on the duplicated data *may* differ from the parameters learned on $\mathcal{D}_1$, i.e. w_1 and b_1.

Your answer:

Duplicating a subset of rows in the dataset can change the linear regression parameters if the duplicated points are not representative of the overall dataset's trend. The mean-squared error (MSE) loss function gives equal weight to all data points. When a subset of points is duplicated, they are effectively given more weight in the calculation, which can pull the best-fit line towards them.

3. We wish to learn a linear regression model on the dataset $\mathcal{D} = \{(\boldsymbol{x}^{(1)}, y^{(1)}), \ldots, (\boldsymbol{x}^{(N)}, y^{(N)})\}$ where $\boldsymbol{x} \in \mathbb{R}^k$. The loss function that we are going to use is called the Cauchy loss and is defined as follows:

$$\ell(\hat{y}, y) = \frac{c^2}{2} \log\left(1 + \left(\frac{y - \hat{y}}{c}\right)^2\right)$$

Where c is the Cauchy loss function constant. In this question, we will use $c = 1$, the base of the $\log$ function is **e** and we do not include an intercept term. Therefore, for a given point $\boldsymbol{x}^{(i)}$, the Cauchy loss of a model with parameters $\boldsymbol{\theta}$ is

$$J^{(i)}(\boldsymbol{\theta}) = \frac{1}{2} \log\left(1 + \left(y^{(i)} - \boldsymbol{\theta}^T \boldsymbol{x}^{(i)}\right)^2\right)$$

We are interested in minimizing the loss over our training data, so we minimize the average Cauchy loss over all points in $\mathcal{D}$. Note that the bias here is 0.

(a) (3 points) What is the partial derivative of $J^{(i)}(\boldsymbol{\theta})$ with respect to the j^{th} parameter, θ_j? You should not include an intercept term.

Your answer:

$$-\frac{\left(y^{(i)} - \boldsymbol{\theta}^T \boldsymbol{x}^{(i)}\right) x_j^{(i)}}{1 + \left(y^{(i)} - \boldsymbol{\theta}^T \boldsymbol{x}^{(i)}\right)^2}$$

(b) (2 points) What is the gradient of $J^{(i)}(\boldsymbol{\theta})$ with respect to the entire parameter vector $\boldsymbol{\theta}$? (Hint: Your result from (a) should be helpful)

Your answer:

$$-\frac{y^{(i)} - \boldsymbol{\theta}^T \boldsymbol{x}^{(i)}}{1 + \left(y^{(i)} - \boldsymbol{\theta}^T \boldsymbol{x}^{(i)}\right)^2} \boldsymbol{x}^{(i)}$$

(c) (2 points) Using your answer from (b), what happens to the gradient of $J^{(i)}(\boldsymbol{\theta})$ when the error of this regression model on the *ith* data point $(\boldsymbol{x}^{(i)}, y^{(i)})$ increases drastically. Explain your answer.

Your Answer

As the magnitude of the error $|y^{(i)} - \boldsymbol{\theta}^T \boldsymbol{x}^{(i)}|$ increases drastically, the magnitude of the gradient $\nabla_{\boldsymbol{\theta}} J^{(i)}(\boldsymbol{\theta})$ approaches zero. This occurs because the scalar factor $\frac{y^{(i)} - \boldsymbol{\theta}^T \boldsymbol{x}^{(i)}}{1 + \left(y^{(i)} - \boldsymbol{\theta}^T \boldsymbol{x}^{(i)}\right)^2}$ in the gradient expression tends toward zero for large errors, making the overall gradient vector small regardless of the input $\boldsymbol{x}^{(i)}$. This behavior reflects the robustness of the Cauchy loss to outliers, as large errors contribute minimally to parameter updates during optimization.

6 Collaboration Questions

After you have completed all other components of this assignment, report your answers to these questions regarding the collaboration policy. Details of the policy can be found here.

1. Did you receive any help whatsoever from anyone in solving this assignment? If so, include full details.
2. Did you give any help whatsoever to anyone in solving this assignment? If so, include full details.
3. Did you find or come across code that implements any part of this assignment? If so, include full details.

Your Answer