HOMEWORK 3
CLASSIFICATION AND REGRESSION[1]

10-301 / 10-601 INTRODUCTION TO MACHINE LEARNING (SPRING 2020)
https://mlcourse.org

OUT: Feb. 06, 2020
DUE: Feb. 12, 2020 11:59 PM
TAs: Ayushi Sood, Eu Jing Chua, Filipp Shelobolin, Mike Chen

START HERE: Instructions

> Homework 3 covers topics on decision trees, *k*-NN, perceptron, linear regression, and (stochastic) gradient descent. The homework includes multiple choice, True/False, and short answer questions. The total number of points is 100.

- **Collaboration policy:** Collaboration on solving the homework is allowed, after you have thought about the problems on your own. It is also OK to get clarification (but not solutions) from books or online resources, again after you have thought about the problems on your own. There are two requirements: first, cite your collaborators fully and completely (e.g., "Jane explained to me what is asked in Question 2.1"). Second, write your solution *independently*: close the book and all of your notes, and send collaborators out of the room, so that the solution comes from you only. See the Academic Integrity Section on the course site for more information: http://www.cs.cmu.edu/~mgormley/courses/10601/about.html#7-academic-integrity-policies

- **Late Submission Policy:** See the late submission policy here: http://www.cs.cmu.edu/~mgormley/courses/10601/about.html#6-general-policies

- **Submitting your work:**

  - **Gradescope:** For written problems such as short answer, multiple choice, derivations, proofs, or plots, we will be using Gradescope (https://gradescope.com/). 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. Regrade requests can be made, however this gives the TA the opportunity to regrade your entire paper, meaning if additional mistakes are found then points will be deducted. Each derivation/proof should be completed on a separate page. For short answer questions, you **should not** include your work in your solution. If you include your work in your solutions, your assignment may not be graded correctly by our AI assisted grader. In addition, please tag the problems to the corresponding pages when submitting your work.

For multiple choice or select all that apply questions, shade in the box or circle in the template document corresponding to the correct answer(s) for each of the questions. For LaTeX users, replace \choice with \CorrectChoice to obtain a shaded box/circle, and don't change anything else.

---

[1]Compiled on Sunday 9th February, 2020 at 00:28

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?

- ● Matt Gormley

- ○ 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
- ☐ None of the above

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 Tree (Revisited)

1. (3 points) Suppose you are the 10-601 instructor trying to predict students' final exam grade using only homework grades. You decide to use historic records on this course to build a predictive model. However, someone messed up CMU's academic records system and the only information you have on students from past semesters is (i) if a student has submitted all homework (ii) if a student has attained maximum score on any of the homework (iii) if a student has scored 0 on any of the homework and (iv) the students' letter grades on their final exams (this is the target, not attribute). You decide to use Decision Tree model, and since the information is so limited, you are satisfied as long as your model can predict the letter grade on the final exams (A+, A, A-, B+, B, B-, C+, C and C-). Suppose that you have designed the grading policy to be percentile-based so all letter grades will show up.

   Given enough past records and enough luck, will your model be able to do a perfect job (i.e., make no mistakes on students' letter grades of current semester)?
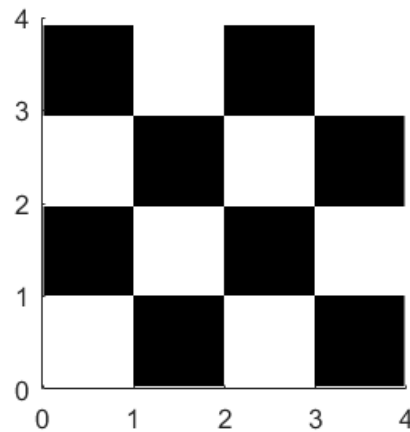
   ○ Yes

   ○ No

   Why or why not? Explain your reason briefly (you can use mathematical expressions).

   NOTE: Please do not change the size of the following text box, and keep your answer in it. Thank you!

   > Your answer.

2. (2 points) Consider the following 4×4 checkerboard pattern.

(a) What is the minimum depth of decision tree that perfectly classifies the 4×4 colored regions, using x and y coordinates as separate features (how you use each of them is up to you)?

○ 1

○ 2

○ 4

○ 16

(b) What is the minimum depth of decision trees to perfectly classify the colored regions, using ANY features?

○ 1

○ 2

○ 4

○ 16

3. (3 points) **Ensemble of Decision Tree.** Say we have a data set shown below. In total, there are 12 data points, with 6 in label "-" and 6 in label "+". We would like to use Decision Tree to solve this binary classification problem. However, in our problem setting, each Decision Tree has access to only ONE line. That is to say, our Decision Tree would have access to only one attribute, and so has max-depth of 1.

By accessing this line, the Decision Tree could know (and only know) whether the data point is on the right side of this line or the left side. (Unofficial definition: let's assume the right side of a line shares the same direction with the **green** normal vector of that line.)

Finally, please use majority vote strategy to make classification decision at each leaf.
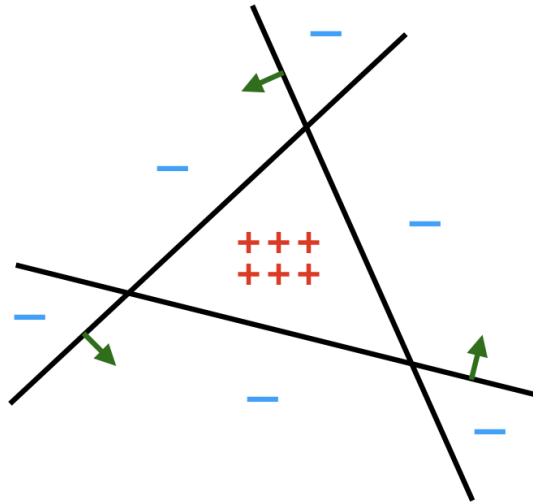
(a) If we train only one Decision Tree, what is the best/lowest error rate? Note that we have in total 12 data points. (Please round to 4 decimal.)

(b) If we could use two Decision Trees, what is the best/lowest error rate? Let's say, if we have two Decision Trees, then each would predict each data point with label like '+' or '-'. Then we would like to combine these predictions as the final result. If these two all predict '+', then the result is '+'. The same with '-'. However, if one predicts '+' while one predicts '-', then to break tie, we always choose '-' as the final result. (Please round to 4 decimal.)

(c) Now let's train three Decision Trees as a forest, what is the best/lowest error rate? The ensemble strategy is now unanimous voting. That is, if every Decision Tree agree, then the final result is positive. However, if one of them has a different answer from the other two, then we predict negative. That means, we train each DT individually, with each DT choose one unique line as its decision boundary. Each DT would try its best to achieve high accuracy. And, next, if all DTs agrees, then it will give positive label. (Please round to 4 decimal.)

4. (2 points) Consider a binary classification problem using 1-nearest neighbors. We have N 1-dimensional training points $x_1, x_2, \ldots x_N$ and corresponding labels $y_1, y_2, \ldots y_N$ with $x_i \in \mathbb{R}$ and $y_i \in \{0, 1\}$. Assume the points $x_1, x_2, \ldots x_N$ are in ascending order by value. If there are ties during the 1-NN algorithm, we break ties by choosing the label of the x_i with lower value. Assume we are using the Euclidean distance metric. Is it possible to build a decision tree where the decision at each node takes the form of "$x \leq t$ or $x > t$", where $t \in \mathbb{R}$ behaves exactly the same as the 1-nearest neighbor classifier?

○ Yes

○ No

If your answer is yes, please explain how you will construct the decision tree. If your answer is no, explain why it's not possible.

NOTE: Please do not change the size of the following text box, and keep your answer in it. Thank you!

Your answer.

2 *k*-Nearest Neighbors

1. (3 points) Consider the description of two objects below:

|	Object A	Object B
Feature 1	3	9.1
Feature 2	2.1	0.7
Feature 3	4.8	2.2
Feature 4	5.1	5.1
Feature 5	6.2	1.8

We can reason about these objects as points in high dimensional space.

Consider the two different distance functions below. Under which scheme are they closer in 5-D space?

1. Euclidean Distance: $d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$

2. Manhattan Distance: $d(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} |x_i - y_i|$

Select one:

- ⭘ Euclidean Distance

- ⭘ Manhattan Distance
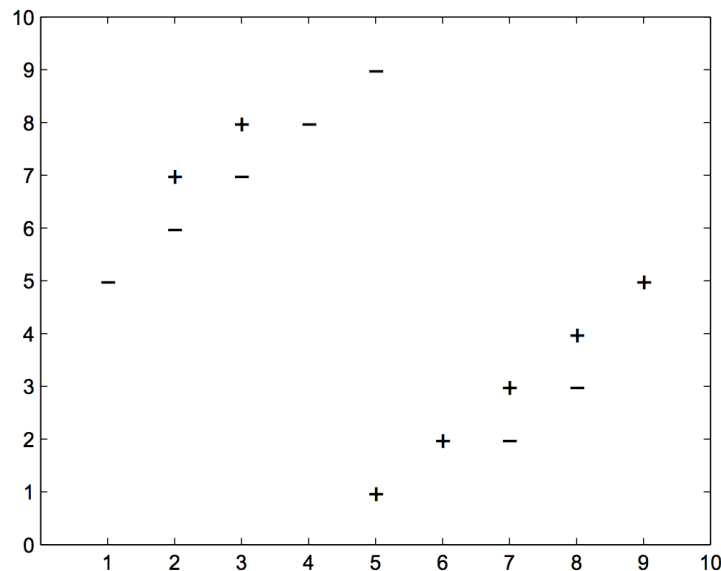
2. (3 points) Consider a k-nearest neighbors (k-NN) binary classifier which assigns the class of a test point to be the class of the majority of the k-nearest neighbors, according to a Euclidean distance metric. Using the data set shown above to train the classifier and choosing $k = 5$, which is the classification error on the training set? Assume that a point can be its own neighbor.

Answer as a decimal with precision 4, e.g. (6.051, 0.1230, 1.234e+7)

3. (3 points) In the data set shown above, what is the value of k that minimizes the training error? Note that a point can be its own neighbor. Let's assume we use random-picking as the tie-breaking algorithm.

> [ ]

4. (3 points) Assume we have a training set and a test set drawn from the same distribution, 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.

    Select one:

    ○ True

    ○ False

    (b) (2 points) Instead of choosing the hyper-parameters 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 hyper-parameters that lead to lower validation error. Is choosing hyper-parameters based on validation error better than choosing hyper-parameters based on training error? Justify your opinion with no more than 3 sentences.

    Select one:

    ○ Yes

    ○ No

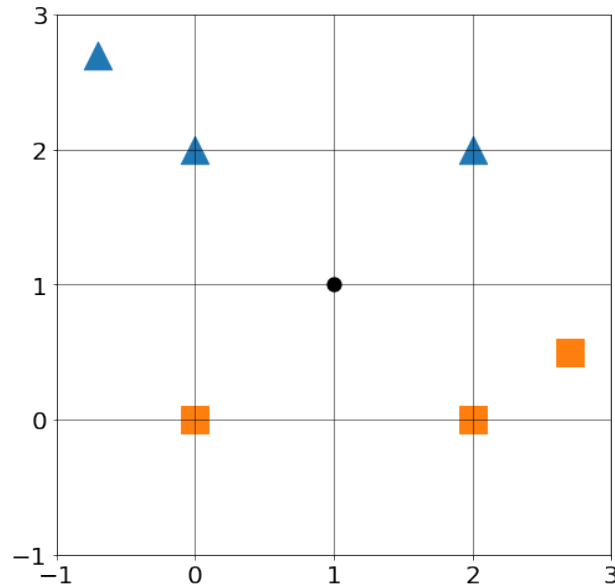   **NOTE: Please do not change the size of the following text box, and keep your answer in it. Thank you!**

    > Your answer.

5. (3 points) 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.

Which of the following methods can be used to break ties or avoid ties on this dataset?

1. Assign x the label of its nearest neighbor

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

3. Use $k = 3$ instead

4. Use $k = 5$ instead

Select one:

○ 1 only

○ 2 only

○ 2, 3, 4

○ 2, 4

○ 4 only

○ 1, 2, 3, 4

○ None of the above

6. (2 points) Consider the following data concerning the relationship between academic performance and salary after graduation. High school GPA and university GPA are two numerical variables (predictors) 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	2.2	3.4	45
2	3.9	2.9	55
3	3.7	3.6	91
4	4.0	4.0	142
5	2.8	3.5	88
6	3.5	1.0	2600
7	3.8	4.0	163
8	3.1	2.5	67
9	3.5	3.6	unknown

Among Students 1 to 8, who is the nearest neighbor to Student 9, using Euclidean distance?

Answer the Student ID only.

[ ]

7. (3 points) In the data set shown above, our task is to predict the salary Student 9 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 9's salary?

Round your answer to the nearest integer. Be sure to use the same unit of measure (thousands of dollars per year) as the table above.

[ ]

8. (3 points) Suppose that the first 8 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?

   Select all that apply:

   ☐ 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", so that each is a percentage of 4.0 (scale by the same percentage).

   ☐ None of the above.

9. (2 points) In this question, we would like to compare the differences among k-NN, the perceptron algorithm, and linear regression.

   Select all that apply:

   ☐ For classification tasks, both k-NN and the perceptron algorithm can have linear decision boundaries.

   ☐ For classification tasks, both k-NN and the perceptron algorithm always have linear decision boundaries.

   ☐ All three models can be susceptible to overfitting.

   ☐ In all three models, after the training is completed, we must store the training data to make predictions on the test data.

   ☐ None of the above.

10. (3 points) Please select all that apply about k-NN in the following options.
    Select all that apply:

    ☐ A larger k gives 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.

3   Perceptron

1. (1 point) 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.

   True or False: the online perceptron algorithm is guaranteed to make the same number of mistakes on S as it does on S'.

   Select one:

   ○ True

   ○ False

2. (3 points) Suppose we have a perceptron whose inputs are 2-dimensional vectors and each feature vector component is either 0 or 1, i.e., $x_i \in \{0, 1\}$. The prediction function $y = \text{sign}(w_1 x_1 + w_2 x_2 + b)$, and

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

   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. **Select all that apply:**

   □ AND function, i.e., the function that evaluates to 1 if and only if all inputs are 1, and 0 otherwise.

   □ OR function, i.e., the function that evaluates to 1 if and only if at least one of the inputs are 1, and 0 otherwise.

   □ XOR function, i.e., the function that evaluates to 1 if and only if the inputs are not all the same. For example
   $$\text{XOR}(1, 0) = 1, \text{ but } \text{XOR}(1, 1) = 0.$$

   □ None of the above.

3. (2 points) Suppose we have a dataset $\{(\mathbf{x}^{(1)}, y^{(1)}), \dots, (\mathbf{x}^{(N)}, y^{(N)})\}$, 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?

   Select one:

   ○ N

   ○ $N \times M$

   ○ M

4. (2 points) Which of the following are true about the perceptron algorithm?

   Select all that apply:

   ☐ The number of mistakes the perceptron algorithm makes is proportional to the number of points in the dataset.

   ☐ The perceptron algorithm converges on any dataset.

   ☐ The perceptron algorithm can be used in the context of online learning.

   ☐ For linearly separable data, the perceptron algorithm always finds the separating hyperplane with the largest margin.

   ☐ None of the above.

5. (3 points) Suppose we have the following data:

   $$\mathbf{x}^{(1)} = [1, 2] \qquad \mathbf{x}^{(2)} = [-1, 2] \qquad \mathbf{x}^{(3)} = [-2, 3] \qquad \mathbf{x}^{(4)} = [1, -1]$$
   $$y^{(1)} = 1 \qquad\quad y^{(2)} = -1 \qquad\quad y^{(3)} = -1 \qquad\quad y^{(4)} = 1$$

   Starting from $\mathbf{w} = [0, 0]$, what is the vector $\mathbf{w}$ after running the perceptron algorithm in the online setting with exactly one pass over the data (one pass is one iteration of the algorithm, going through all datapoints)?

   Assume we are running the perceptron algorithm without an intercept term. If the value of the dot product of a data point and the weight vector is 0, the algorithm makes the prediction 1. Also process in the given order $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \mathbf{x}^{(3)}$, then $\mathbf{x}^{(4)}$.

   Select one:

   ○ $[1, -2]$

   ○ $[2, 0]$

   ○ $[-1, 1]$

   ○ $[1, -3]$

6. (2 points) Please refer to previous question for the data. Assume we are running perceptron in the batch setting. How many passes will the perceptron algorithm make before converging to a perfect classifier, i.e., one that does not make any false predictions on this dataset?

   Select one:

   ○ 2

   ○ 3

   ○ 5

   ○ Infinitely many (the algorithm does not converge)

7. (3 points) Please select the correct statement(s) about the mistake bound of the perceptron algorithm.

   Select all that apply:

   □ If the minimum distance from any data point to the separating hyperplane is increased, without any other change to the data points, the mistake bound will also increase.

   □ If the whole dataset is shifted away from origin, then the mistake bound will also increase.

   □ If the size of the data set (i.e., the maximum pair-wise distance between data points) is increased, then the mistake bound will also increase.

   □ The mistake bound is linearly inverse-proportional to the minimum distance of any data point to the separating hyperplane of the data.

   □ None of the above.

8. (2 points) Given a zero-centered 3-dimensional dataset, the coordinate of the point with the highest L2 norm is $(2, 2, 2)$. Assuming that the dataset is linearly separable with margin 2, what is the greatest number of mistakes that Perceptron could make?

   Select one:

   ○ 1

   ○ 2

   ○ 3

   ○ 4

9. (2 points) Suppose we have data whose elements 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?
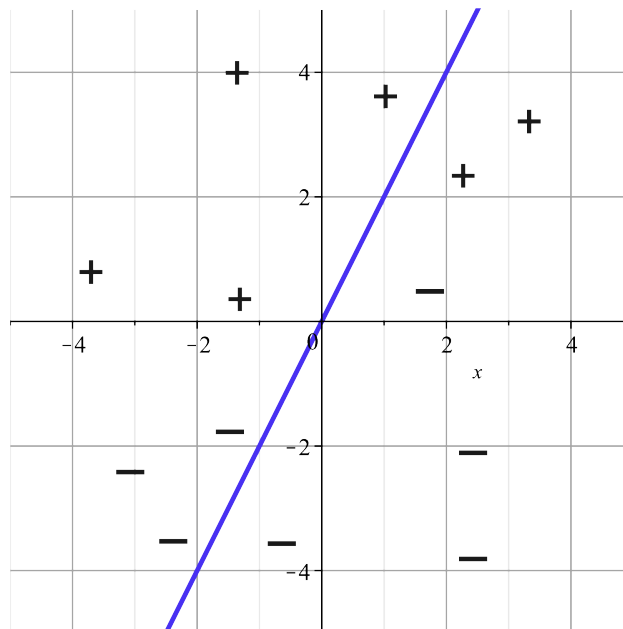
Select one:

○ $[-1, 1]$

○ $[4, 6]$

○ $[-3, -1]$

○ $[5, 5]$

10. (2 points) Consider the linear decision boundary below and the training dataset shown. Which of the following are valid weights $\boldsymbol{\theta}$ and its corresponding training error? (Note: Assume the decision boundary is fixed and does not change while evaluating training error.)

Select all that apply:

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

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

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

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

4 Linear Regression

1. (4 points) Suppose you have data $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})$ and the solution to linear regression on this data is $y = w_1 x + b_1$. Now suppose we have the dataset $(x^{(1)} + \alpha, y^{(1)} + \beta), \ldots, (x^{(n)} + \alpha, y^{(n)} + \beta)$ where $\alpha > 0, \beta > 0$ and $w_1 \alpha \neq \beta$. The solution to the linear regression on this dataset is $y = w_2 x + b_2$. Please 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.

   Select one:

   - ○ $w_1 = w_2, b_1 = b_2$

   - ○ $w_1 \neq w_2, b_1 = b_2$

   - ○ $w_1 = w_2, b_1 \neq b_2$

   - ○ $w_1 \neq w_2, b_1 \neq b_2$

2. (4 points) We would like to fit a linear regression estimate to the dataset

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

   with $\mathbf{x}^{(i)} \in \mathbb{R}^M$ by minimizing the ordinary least square (OLS) objective function:

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

   Specifically, we solve for each coefficient w_k ($1 \leq k \leq M$) by deriving an expression of w_k from the critical point $\frac{\partial J(\mathbf{w})}{\partial w_k} = 0$. What is the expression for each w_k in terms of the dataset $(\mathbf{x}^{(1)}, y^{(1)}), \cdots, (\mathbf{x}^{(N)}, y^{(N)})$ and $w_1, \cdots, w_{k-1}, w_{k+1}, \cdots, w_M$?

   Select one:

   - ○ $w_k = \dfrac{\sum_{i=1}^{N} x_k^{(i)} (y^{(i)} - \sum_{j=1, j \neq k}^{M} w_j x_j^{(i)})}{\sum_{i=1}^{N} (x_k^{(i)})^2}$

   - ○ $w_k = \dfrac{\sum_{i=1}^{N} x_k^{(i)} (y^{(i)} - \sum_{j=1, j \neq k}^{M} w_j x_j^{(i)})}{\sum_{i=1}^{N} (y^{(i)})^2}$

   - ○ $w_k = \sum_{i=1}^{N} x_k^{(i)} (y^{(i)} - \sum_{j=1}^{M} w_j x_j^{(i)})$

   - ○ $w_k = \dfrac{\sum_{i=1}^{N} x_k^{(i)} (y^{(i)} - \sum_{j=1, j \neq k}^{M} w_j x_j^{(i)})}{\sum_{i=1}^{N} (x_k^{(i)} y^{(i)})^2}$

3. (3 points) Continuing from the above question, how many coefficients do you need to estimate? When solving for these coefficients, how many equations do you have?

   Select one:

   ○ N coefficients, M equations

   ○ M coefficients, N equations

   ○ M coefficients, M equations

   ○ N coefficients, N equations

4. (3 points) Consider the following 3 data points for linear regression: $\mathbf{x}^{(1)} = [0, 1, 2]^T$, $\mathbf{x}^{(2)} = [1, 0, 2]^T$ and $\mathbf{x}^{(3)} = [2, 1, 0]^T$. The corresponding y values are $y^{(1)} = 3$, $y^{(2)} = 6$, $y^{(3)} = 9$.

   Assume the intercept to be 0. Find the weights $\boldsymbol{\theta} = [\theta_1, \theta_2, \theta_3]^T \in \mathbb{R}^3$ such that the mean squared error $J(\boldsymbol{\theta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})$ is minimized on this training set. $\mathbf{X}$ is the design matrix where $\mathbf{X}_{ij} = \mathbf{x}_j^{(i)}$.

   θ_1: [   ]

   θ_2: [   ]

   θ_3: [   ]

5. (2 points) Assume that a data set has M data points and N variables, where $M > N$. Different loss functions would return the same sets of solutions as long as they are convex.

   Select one:

   ○ True

   ○ False

6. (2 points) Suppose we are working with datasets where the number of features is 3. The optimal solution for linear regression is always unique regardless of the number of data points that are in this dataset.

   Select one:

   ○ True

   ○ False

7. (1 point) Consider the following dataset:

x_1	1.0	2.0	3.0	4.0	5.0
x_2	2.0	4.0	6.0	8.0	10.0
y	4.0	7.0	8.0	11.0	17.0

We want to carry out a multiple-linear regression between y (dependent variable) and x_1 and x_2 (independent variables). The closed-form solution given by $\mathbf{w} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}$ will return the unique solution.

Note: The ith row of $\mathbf{X}$ contains the ith data point $(x_1^{(i)}, x_2^{(i)})$ while the ith row of $\mathbf{Y}$ contains the ith data point $y^{(i)}$.

Select one:

○ True

○ False

8. (3 points) Order the following different formulations of the regression cost function according to sensitivity to outliers from the most sensitive to the least sensitive.

1. $J(\mathbf{w}) = \sum_i |(\mathbf{x}^{(i)})^T\mathbf{w} - y^{(i)}|^2$

2. $J(\mathbf{w}) = \sum_i |(\mathbf{x}^{(i)})^T\mathbf{w} - y^{(i)}|^4$

3. $J(\mathbf{w}) = \sum_i |(\mathbf{x}^{(i)})^T\mathbf{w} - y^{(i)}|$

Order the cost functions here:

1: ☐

2: ☐

3: ☐

9. (3 points) Identifying whether a function is a convex function is useful because a convex function's local minimum has the nice property that it has to be the global minimum. Please select all functions below that are convex functions. Note $dom(f)$ denotes the domain of the function f.
   Select all that apply:

   ☐ $f(x) = x, dom(f) = \mathbb{R}$

   ☐ $f(x) = x^3 + 2x + 3, dom(f) = \mathbb{R}$

   ☐ $f(x) = \log x, dom(f) = \mathbb{R}_{++}$ (the set of positive real numbers)

   ☐ $f(x) = |x|, dom(f) = \mathbb{R}$

   ☐ $f(x) = ||\mathbf{x}||_2, dom(f) = \mathbb{R}^n$

   ☐ None of the above.

10. (2 points) Typically we can solve linear regression problems in two ways. One is through direct methods, e.g. solving the closed form solution, and the other is through iterative methods (gradient descent). Consider a linear regression on data $(\mathbf{X}, \mathbf{y})$. We assume each row in $\mathbf{X}$ denotes one input in the dataset. Please select all correct options.

   Select all that apply:

   ☐ If the matrix $\mathbf{X}^T\mathbf{X}$ is invertible, the exact solution is always preferred for solving the solution to linear regression as computing matrix inversions and multiplications are fast regardless of the size of the dataset.

   ☐ Assume N is the number of examples and M is the number of features. The computational complexity of N iterations of batch gradient descent is $\mathcal{O}(MN)$.

   ☐ The computational complexity of the closed form solution is linear in number of parameters/features.

   ☐ None of the above.

11. (2 points) Consider the following dataset:

x	1.0	2.0	3.0	4.0	5.0
y	3.0	8.0	9.0	12.0	15.0

Let x be the vector of datapoints and y be the label vector. Here, we are fitting the data using gradient descent. If we initialize the weight as 2.0 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? Note that we do not introduce any regularization in this problem and and our objective function looks like $\frac{1}{N}\sum_{i=1}^{N}(wx_i + b - y_i)^2$, where N is the number of data points, w is the weight, and b is the intercept.

Fill in the blank with the gradient on the weight you computed, rounded to 2 decimal places after the decimal point.

12. (4 points) Based on the data of the previous question, please compute the direct solution of the weight and the intercept for the objective function defined in the previous question, rounded to 2 decimal places after the decimal point.

Weight:

Intercept:

13. (2 points) Using the dataset and model given in question 11, perform two steps of batch gradient descent on the data. Fill in the blank with the value of the weight after two steps of batch gradient descent. Let the learning rate be 0.01. Round to 2 decimal places after the decimal point.

<div style="border:1px solid black; width:250px; height:70px;"></div>

14. (2 points) Using the dataset and model given in question 11, which of the following learning rates leads to the most optimal weight and intercept after performing two steps of batch gradient descent? (Hint: The most optimal learned parameters are the parameters that lead to the lowest value of the objective function.)

    Select one:

    ○ 1

    ○ 0.1

    ○ 0.01

    ○ 0.001

Collaboration Questions Please answer the following:

After you have completed all other components of this assignment, report your answers to the collaboration policy questions detailed in the Academic Integrity Policies found here.

1. Did you receive any help whatsoever from anyone in solving this assignment? Is so, include full details.

2. Did you give any help whatsoever to anyone in solving this assignment? Is so, include full details.

3. Did you find or come across code that implements any part of this assignment ? If so, include full details.

Solution: