This quiz has been regraded; your new score reflects 2 questions that were affected.

Homework 3: KNN, Perceptron, Linear Regression

Due Feb 14 at 11:59pm **Points** 100 **Questions** 31
Available Feb 7 at 11:35pm - Feb 18 at 11:59pm 11 days **Time Limit** None

Instructions

Due date: 2/14/18 11:59PM
TAs: Jennifer, Sienna, Yu, Qinghan

Read these instructions carefully!

Homework 3 covers topics on KNN, perceptron, and linear regression. The homework includes multiple choice, True/False, and short answer questions.

Important note on questions with numerical answers: When you enter numerical answers on Canvas, Canvas will automatically convert it to a decimal with **four** decimal places. So if you enter '7', Canvas will convert it into '7.0000'; if you enter '7.343', Canvas will convert into '7.3430'; if you enter '7.34388', Canvas will convert it into '7.3439'. Please just answer the numerical answers according to the questions, and don't worry about the extra 0s that get added at the end.

Important note on scoring on multiple answer questions: You will notice that there are questions like "Select all answers that are correct." in the assignment. The top right corner of the question shows the number of points the question is worth. Please keep in mind that the way Canvas grades this type of questions is as follows: Canvas divides the total points possible by the amount of correct answers for that question. This amount is awarded for every correct answer selected and deducted for every incorrect answer selected. For example, if the question has 2 options, is worth 2 points, and exactly one of the options is correct, then if you select both options, you would receive 2 - 2 = 0 points. However, the minimum score you could get on a problem is 0.

Important: Only **one submission** is allowed for this homework. Please make sure you're confident about your answers before you submit.

This quiz was locked Feb 18 at 11:59pm.

Attempt History

	Attempt	Time	Score	Regraded
LATEST	Attempt 1	4,303 minutes	86.33 out of 100	90.33 out of 100

Score for this quiz: **90.33** out of 100
Submitted Feb 12 at 9:18pm
This attempt took 4,303 minutes.

The following questions are on k-Nearest Neighbors.

Question 1

3 / 3 pts

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

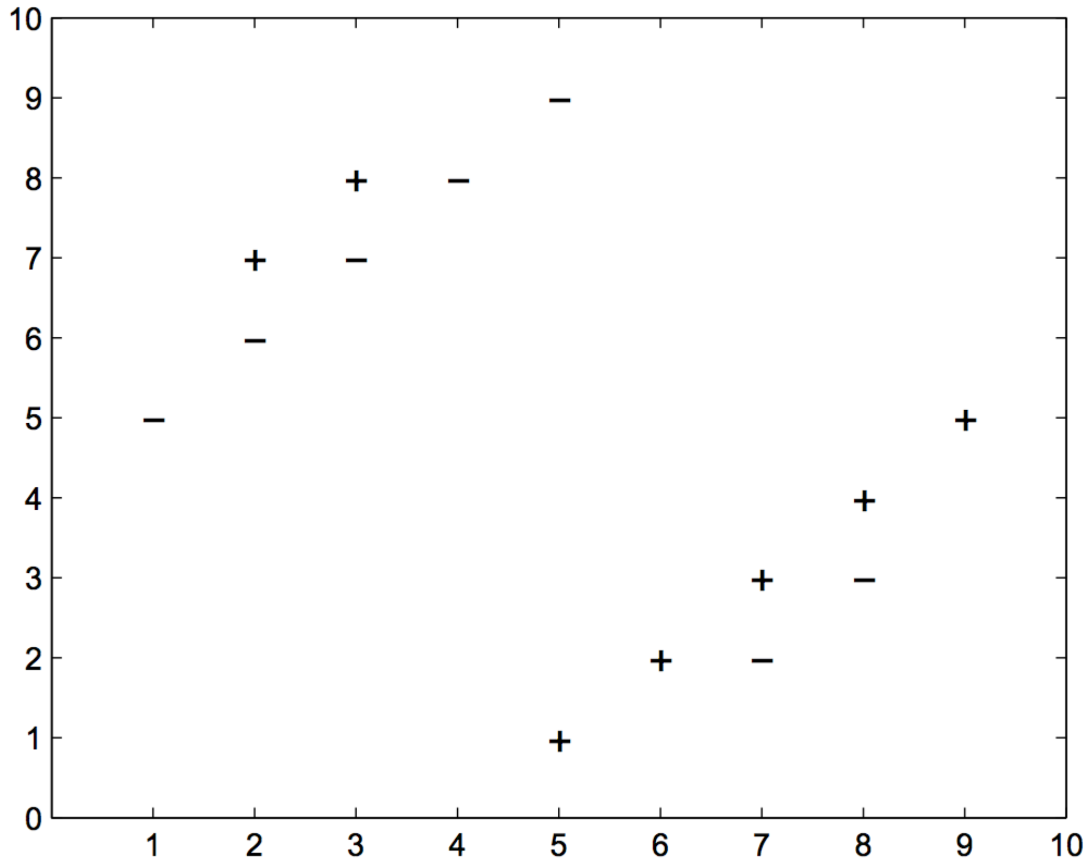
Correct!

◉ Euclidean Distance

○ Manhattan Distance

Question 2

4 / 4 pts



Consider a k-nearest neighbors 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 $k = 5$, what 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. (1.234e+5, 6.501 or 0.4310).

Correct!

0.2857

orrect Answers

Between 0.285 and 0.286

Question 3

3 / 3 pts

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.

Correct!

1.0000

orrect Answers

1.0 (with margin: 0.0)

Question 4

2 / 2 pts

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 kNN classifier. 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

Correct!

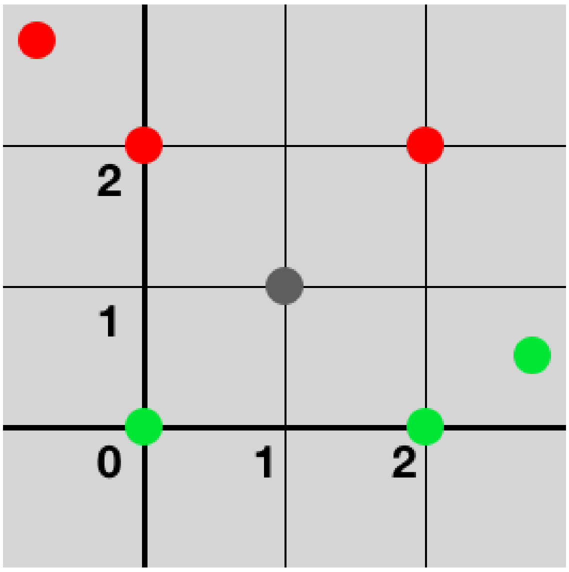
● False

Question 5

3 / 3 pts

Consider a binary k-NN classifier where $k = 4$ and the two labels are "red" and "green".

Consider classifying a new point $x = (1, 1)$, where two of x's nearest neighbors are labeled "red" and two are labeled "green" 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 x (from the labels of its 4 closest points)

3) Use $k = 3$ instead

4) Use $k = 5$ instead

○ 1 only

○ 2 only

○ 2, 3, 4

Correct! ● 2, 4

○ 4 only

○ 1, 2, 3, 4

Question 6

3 / 3 pts

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.

Correct!

3.0000

orrect Answers

3.0 (with margin: 0.0)

Question 7

0 / 4 pts

In the data set shown above, our task is to predict the salary Student 9 earns after graduation. We apply kNN 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.

'ou Answered

117.0000

orrect Answers

132.0 (with margin: 0.0)

Question 8

2 / 2 pts

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 KNN 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.

Correct!

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

Correct!

☑ 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

Question 9

4 / 4 pts

In this question, we would like to compare the differences among KNN, the perceptron algorithm, and linear regression. Please select all that apply in the following options.

Correct!

☑ For classification tasks, both KNN and the perceptron algorithm can have linear decision boundaries.

□ For classification tasks, both KNN and the perceptron algorithm always have linear decision boundaries.

Correct!

☑ 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.

Question 10

4 / 4 pts

Please select all that apply about kNN in the following options.

Correct!

☑ Large k gives a smoother decision boundary.

Correct!

☑ 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.

Correct!

☑ 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.

The following questions are on perceptron.

Question 11

2 / 2 pts

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

○ True

Correct!

● False

Question 12

4 / 4 pts

Suppose we have a perceptron whose inputs are two dimensional vectors and the vector component is either 0 or 1, i.e., $x_i \in \{0,\ 1\}$. The prediction function $y\ =\ \text{sign}\left(w_1x_1 + w_2x_2 + b\right)$, and $\text{sign}\left(z\right)\ = \begin{cases} 1, & \text{if } z \geq 0 \\ 0, & \text{otherwise} \end{cases}$. 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.

Correct!

☑ AND function

Correct!

☑ OR function

□ XOR function

□ None of the above

Question 13

3 / 3 pts

Suppose we have a dataset $\{\left(x^{(1)},\, y^{(1)}\right),\, \ldots,\, \left(x^{(N)},\, y^{(N)}\right)\}$, where $x^{(i)} \in \mathbb{R}^M,\;\; y^{(i)} \in \{+1, -1\}$. We would like to apply the perceptron algorithm on this dataset. Assume there's no bias term. How many parameter values is the perceptron algorithm learning?

○ N

○ N * M

Correct!

● M

Question 14

4 / 4 pts

Which of the following are true about the perceptron algorithm? Select all that apply.

□ The number of mistakes the perceptron makes is proportional to the size of the dataset.

□ The perceptron algorithm converges on any dataset.

Correct!

☑ Perceptron algorithm can be used in the context of online learning.

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

Question 15

4 / 4 pts

Suppose we have the following data:

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

$y^{(1)} = 1,\ y^{(2)} = -1,\ y^{(3)} = -1,\ y^{(4)} = 1.$

Starting from $w = [0,\ 0]$, what is the vector θ after running the perceptron algorithm with exactly one pass over the data? Assume we are running the perceptron algorithm without a bias term. If the value of the dot product of a data point and the weight vector is 0, the algorithm makes the prediction 1.

Correct!

● $[1,\ -2]$

○ $[2,\ 0]$

○ $[-1,\ 1]$

○ $[1,\ -3]$

Question 16

Original Score: 3 / 3 pts **Regraded Score: 3 / 3 pts**

This question has been regraded.

Please refer to previous question for the data. Assume we're running perceptron in the batch setting, how many passes will the perceptron algorithm make before termination?

Correct!

● 2

○ 3

○ 5

○ Infinitely many (the algorithm does not converge)

In this question, we give full credits for students who chose either 2 or 3 as the answer. The algorithm will arrive at the parameter settings that make no mistakes on the entire dataset after 2 passes, however, it will take another pass through the data to check that the learned parameters make no mistakes on the entire dataset.

Question 17

4 / 4 pts

We can view the perceptron algorithm as trying to minimize which of the following objective functions with stochastic gradient descent? Assume that we apply the notation where $x_0 = 1$, θ_0 is the bias term, and N is the number of data points. Note the function

$$f(x) = (x)_+ = \begin{cases} x, & \text{if } x \geq 0 \\ 0, & \text{otherwise} \end{cases}.$$

○ $J(\theta) = \sum_{i=1}^{N} -y^{(i)} \left(\theta \cdot x^{(i)}\right)$

○ $J(\theta) = \sum_{i=1}^{N} y^{(i)} \left(\theta \cdot x^{(i)}\right)$

Correct!

● $J(\theta) = \sum_{i=1}^{N} \left(-y^{(i)} \left(\theta \cdot x^{(i)}\right)\right)_+$

○ $J(\theta) = \sum_{i=1}^{N} \left(y^{(i)} \left(\theta \cdot x^{(i)}\right)\right)_+$

Question 18

0 / 3 pts

Continuing with the above question, what is the gradient of the correct cost function when the current data we're seeing is $\left(x^{(i)},\ y^{(i)}\right)$?

orrect Answer

○ $\begin{cases} -y^{(i)}x^{(i)}, & \text{if } -y^{(i)}\left(\theta \cdot x^{(i)}\right) \geq 0 \\ 0, & \text{otherwise} \end{cases}$

○ $-y^{(i)}x^{(i)}$

○ $y^{(i)}x^{(i)}$

'ou Answered

● $\begin{cases} y^{(i)}x^{(i)}, & \text{if } -y^{(i)}\left(\theta \cdot x^{(i)}\right) \geq 0 \\ 0, & \text{otherwise} \end{cases}$

Question 19

Original Score: 0 / 4 pts Regraded Score: 4 / 4 pts

This question has been regraded.

Please select the correct statement(s) about the mistake bound of the perceptron algorithm. Select all that apply.

□ If the minimum distance of any data point to the separating hyperplane of the data is increased, the mistake bound will also increase.

Correct!

☑ If the maximum distance of any data point to the origin is increased, then the mistake bound will also increase.

□ If the maximum distance of any data point to the mean of the all data points is increased, then the mistake bound will also increase.

'ou Answered

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

Question 20

3 / 3 pts

Suppose we have data whose elements are of the form $\begin{pmatrix}x_1\\x_2\end{pmatrix}$, where $x_1 + x_2 = 0$. We don't know the label for each element. Suppose the perceptron algorithm starts with $\theta = \begin{pmatrix}1\\2\end{pmatrix}$, which of the following values will θ never take on in the process of running the perceptron algorithm on the data?

○ $\begin{pmatrix}3\\0\end{pmatrix}$

○ $\begin{pmatrix}-2\\5\end{pmatrix}$

○ $\begin{pmatrix}1\\2\end{pmatrix}$

Correct!

● $\begin{pmatrix}2\\3\end{pmatrix}$

The following questions are on linear regression.

Question 21

4 / 4 pts

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 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 the statement should hold no matter what values $\alpha,\ \beta$ take on within the specified constraints.

○ $w_1 = w_2, b_1 = b_2$

○ $w_1 \neq w_2, b_1 = b_2$

Correct!

● $w_1 = w_2, b_1 \neq b_2$

○ $w_1 \neq w_2, b_1 \neq b_2$

Question 22

4 / 4 pts

We are trying to derive the closed form solution for linear regression:

In the following, each row in $\mathbf{X}$ denotes one data point and Y is column vector.

First we take the derivative of the objective function $L = \frac{1}{2}(\mathbf{X}\mathbf{w} - Y)^T(\mathbf{X}\mathbf{w} - Y)$ with respect to w and set it to zero, arriving at equation (a).

Then after some algebraic manipulation, we get the solution $\mathbf{w} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^TY$. What should equation (a) be?

Correct!

● $(\mathbf{X}\mathbf{w} - Y)^T\mathbf{X} = 0$

○ $(\mathbf{X}\mathbf{w} + Y)^T\mathbf{X} = 0$

○ $\mathbf{X}^T\mathbf{X}\mathbf{w} + Y^T\mathbf{X} = 0$

○ $Y\mathbf{X}^T\mathbf{X}\mathbf{w} + \mathbf{X} = 0$

Question 23

2 / 2 pts

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 in the dataset.

○ True

Correct!

● False

Question 24

1.33 / 4 pts

Identifying whether a function is a convex function is useful because a convex function's local minimum has to be its global minimum. Please select all functions below that are convex functions. Note $dom\left(f\right)$ denotes the domain of the function f.

orrect Answer

□ $f\left(x\right)\ =\ x,\ dom\left(f\right)\ =\ \mathbb{R}$

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

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

Correct!

☑ $f\left(x\right)\ =\ \left|x\right|,\ dom\left(f\right)\ =\ \mathbb{R}$

orrect Answer

□ $f(x) = ||x||_2, dom(f) = \mathbb{R}$

Question 25

3 / 3 pts

Typically, we can solve linear regression problems in two ways. One is through direct methods, e.g. closed form solution, and the other is through iterative methods, e.g. stochastic or batch gradient descent methods. Consider performing linear regression on data $\left(\mathbf{X},\mathbf{y}\right)$. We assume each row in $\mathbf{X}$ denotes one input in the dataset. Please select all options that are correct about the two methods.

□ If the matrix $\mathbf{X}^T\mathbf{X}$ is invertible, 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 that N is the number of examples and M is the number of features. The computational complexity of N iterations of batch gradient descent is $O\left(MN\right)$.

Correct!

☑

When the dataset is large, stochastic gradient descent is often the preferred method because it gets us reasonably close to the solution faster than both the direct method and batch gradient descent.

Question 26

3 / 3 pts

A data scientist is working on a regression problem on a large data set. After trying stochastic gradient descent (gradient is evaluated on a portion of the data set in each step) and batch gradient descent (gradient is evaluated on the entire data set in each step), the scientist obtained the values of the loss function for the two methods with respect to training time. Note that the same learning rate is used in both cases.

Time in hours	Stochastic GD	Batch GD
1	102.34	120.12
2	80.45	92.37
3	65.23	73.64
4	58.77	58.23
5	52.33	49.21
6	50.74	45.98
7	49.88	43.64

Select all the choices consistent with this table.

□

This table shows, in practice, that stochastic gradient descent can compute more accurate gradient direction compared to batch gradient descent.

□

Within the first 3 hours, the table suggests that batch gradient descent makes more progress in finding the optimum of the objective function than the stochastic gradient descent.

Correct!

☑

In general, stochastic gradient descent does not necessarily take a descent step in each step. However, stochastic gradient descent takes much less time to evaluate per step. In this table, during the first hour, stochastic gradient descent makes more update steps to the weights while batch gradient descent makes less updates. Hence it could be reasonable that using batch gradient descent results in a higher value for the loss function than that of the stochastic gradient descent at the 1 hour time point

Question 27

3 / 3 pts

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

If we initialize the weight as 2.0 and bias term 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 gradient descent update? Note that we do not introduce any regularization in this problem 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 bias term.

Fill in the blank with the gradient you computed. Round to 2 decimal places after decimal point.

Correct!

-23.6000

orrect Answers

-23.6 (with margin: 0.1)

Question 28

4 / 4 pts

Based on the data of previous questions, please compute the direct solution of the weight and the bias for the objective function defined in the previous question. Please fill in the blank with the weight you computed. Round to 2 decimal places after decimal point.

Correct!

2.8000

orrect Answers

2.8 (with margin: 0.1)

Question 29

3 / 3 pts

Please use the dataset and model given in question 27. Perform two steps of batch gradient descent on the data. Please 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 decimal point.

Correct!

2.4200

orrect Answers

2.416 (with margin: 0.1)

Question 30

4 / 4 pts

Using the dataset and model given in question 27, which of the following learning rates leads to the most optimal weight and bias 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.)

○ 1

○ 0.1

Correct!

● 0.01

○ 0.001

Question 31

0 / 0 pts

Please answer the following questions:

1. Did you receive any help whatsoever from anyone in solving this assignment? Yes / No.
   - If you answered ‘yes’, give full details: ____________
   - (e.g. “Jane Doe explained to me what is asked in Question 3.4”)
2. Did you give any help whatsoever to anyone in solving this assignment? Yes / No.
   - If you answered ‘yes’, give full details: _____________
   - (e.g. “I pointed Joe Smith to section 2.3 since he didn’t know how to proceed with Question 2”)
3. Did you find or come across code that implements any part of this assignment ? Yes / No. (See below policy on “found code”)
   - If you answered ‘yes’, give full details: _____________
   - (book & page, URL & location within the page, etc.).

Your Answer:

1. No

2. No

3. No

Quiz Score: **90.33** out of 100