Course › Week 4... › Proble... › Proble...

Problem Set 4

Problems 1-4 correspond to "An introduction to linear regression"

Problem 1

6/6 points (graded)

Consider the following simple data set of four points (x, y):

$$(1,1),(1,3),(4,4),(4,6).$$

a) Suppose you had to predict y without knowledge of x. What value would you predict?

(1+3+4+6)/4 ✓ **Answer:** 3.5

$$\frac{1+3+4+6}{4}$$

Answer

Correct: This is simply the mean value of y: the average of the four observed values (1,3,4,6).

b) Continuing from part (a), what is the $\mathbf{mean\ squared\ error}$ (MSE) of your prediction, on the given four points?

((1-3.5)^2+(3-3.5)^2+(4-3 ✓ **Answer:** 3.25

$$\frac{(1-3.5)^2+(3-3.5)^2+(4-3.5)^2+(6-3.5)^2}{4}$$

Answer

Correct: This is the variance of the four observed values of y.

c) Now let's say you want to predict y based on x. Your initial choice of prediction rule is $y = x$. What is the MSE of the linear function $y = x$ on the four given points?

((1-3)^2+(4-6)^2)/4 ✓ **Answer:** 2

$\frac{(1-3)^2+(4-6)^2}{4}$

Answer
Correct:
This is not a good prediction function, as you can see if you plot it. The MSE is obtained by computing the (squared) error on each of the four points and averaging them.

d) Finally, you want to find the $\mathbf{best}$ prediction rule of the form $y = ax + b$. That is, you want to find the parameters $a, b \in \mathbb{R}$ such that this rule has the smallest possible mean squared error on the four training points. What are a and b?

$a =$

1 ✓ **Answer:** 1

1

$b =$

3.5-2.5 ✓ **Answer:** 1

$3.5 - 2.5$

e) Continuing from part (d), what is the MSE of this optimal linear predictor?

((1-(1+1))^2+(3-(1+1))^2+ ✓ **Answer:** 1

$\frac{(1-(1+1))^2+(3-(1+1))^2+(4-(4+1))^2+(6-(4+1))^2}{4}$

? Hint (1 of 1): It will be very helpful to actually plot the points on a piece of paper. The optimal line -- the one that cuts through the middle of the points -- is then pretty easy to see.

Next Hint

Submit

Answers are displayed within the problem

Problem 2

4/4 points (graded)

Suppose that we have data points $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})$, where $x^{(i)}, y^{(i)} \in \mathbb{R}$, and that we want to fit them with a line that passes through the origin. The general form of such a line is $y = ax$: that is, the sole parameter is $a \in \mathbb{R}$.

a) In this setting, what are the $\mathbf{predictor}$ and $\mathbf{response}$ variables?

predictor x and response y ▼ ✓ **Answer:** predictor x and response y

b) The goal is to find the value of a that minimizes the squared error on the data. We will do this by first writing down a $\mathbf{loss\ function}$ $L(\cdot)$. Which of the following statements is an accurate description of the loss function? Select all that apply.

☑ It takes a parameter a and returns a real number.

□ It takes a data set and returns a parameter a.

☑ It is based on the given data set.

□ It is the same regardless of the data set.

✓

c) Using calculus, find the optimal setting of a. The answer is of the form $a = N/D$ where the numerator N and the denominator D can be found in the following list.

$\sum_{i=1}^{n} (y^{(i)} - x^{(i)})\, x^{(i)}$

$\sum_{i=1}^{n} x^{(i)} y^{(i)}$

$\sum_{i=1}^{n} {y^{(i)}}^2$

$\sum_{i=1}^{n} {x^{(i)}}^2$

$\sum_{i=1}^{n}\left(y^{(i)}-x^{(i)}\right)^2$

Which of these is N and which is D?

$N =$

○ $\sum_{i=1}^{n}\left(y^{(i)}-x^{(i)}\right)x^{(i)}$

● $\sum_{i=1}^{n}x^{(i)}y^{(i)}$

○ $\sum_{i=1}^{n}{y^{(i)}}^2$

○ $\sum_{i=1}^{n}{x^{(i)}}^2$

○ $\sum_{i=1}^{n}\left(y^{(i)}-x^{(i)}\right)^2$

✓

$D =$

○ $\sum_{i=1}^{n}\left(y^{(i)}-x^{(i)}\right)x^{(i)}$

○ $\sum_{i=1}^{n}x^{(i)}y^{(i)}$

○ $\sum_{i=1}^{n}{y^{(i)}}^2$

● $\sum_{i=1}^{n}{x^{(i)}}^2$

○ $\sum_{i=1}^{n}\left(y^{(i)}-x^{(i)}\right)^2$

✓

Explanation

For a line $y = ax$, the total squared loss on the n data points is:

$L(a) = \sum_{i=1}^{n} \left(y^{(i)} - a x^{(i)}\right)^2.$

To minimize this, we take the derivative with respect to a:

$\frac{dL}{da} = -2 \sum_i \left(y^{(i)} - a x^{(i)}\right) x^{(i)}$

and then set this to zero, to get:

$a = \frac{\sum_i y^{(i)} x^{(i)}}{\sum_i {x^{(i)}}^2}.$

Submit

ℹ Answers are displayed within the problem

Problem 3

3/3 points (graded)

One fact that we used implicitly in the lecture is the following:

If we want to summarize a bunch of numbers $x_1, \ldots, x_n$ by a single number s, the best choice for s, the one that minimizes the average squared error, is the $\mathbf{mean}$ *of the x_i's.*

Let's see why this is true. We begin by defining a suitable loss function. Any value $s \in \mathbb{R}$ induces a mean squared loss (MSE) given by:

$$L(s) = \frac{1}{n} \sum_{i=1}^{n} (x_i - s)^2.$$

We want to find the s that minimizes this function.

a) Compute the derivative of $L(s)$. The answer is of the form

$$\frac{dL}{ds} = a x_1 + \cdots + a x_n + b s,$$

where a, b are some constants. What are a and b?

$a =$

-2 $/n$ ✓ **Answer:** -2

-2

$b =$

2 ✓ **Answer:** 2

2

b) Setting the derivative dL/ds to zero yields

$$s = -\frac{a}{b} \sum_{i=1}^{n} x_i .$$

With the right values of a, b, we have $-a/b = 1/n$, so this is exactly the mean of the x_i.

Were you able to do this problem quite easily? (Either answer will be graded as correct; we're just curious.)

● Yes

○ No ✓

✓

Explanation

The derivative, with respect to s, is: $\frac{dL}{ds} = -\frac{2}{n} \sum_{i=1}^{n} (x_i - s)$.

Submit

Answers are displayed within the problem

Problem 4

5/5 points (graded)

Let's continue the thought process of the previous problem. Again, we have a collection of numbers $x_1, \ldots, x_n$ that we wish to summarize by a single number s. But this time we want to minimize the average $\textbf{absolute error}$,

$$L\left(s\right)=\frac{1}{n}\sum_{i=1}^{n}\left|x_i-s\right|.$$

What value of s should we choose in this case?

a) Let's begin with an example. Suppose we have the following set of 9 numbers:

$$1,2,3,4,5,6,7,8,90.$$

What is their mean?

(1+2+3+4+5+6+7+8+90)/

✓ **Answer:** 14

$$\frac{1+2+3+4+5+6+7+8+90}{9}$$

b) Continuing with the previous example, what is the average absolute loss induced by setting s to the mean?

(abs(1-14)+abs(2-14)+ab:

✓ **Answer:** 152/9

$$\frac{\mathrm{abs}(1-14)+\mathrm{abs}(2-14)+\mathrm{abs}(3-14)+\mathrm{abs}(4-14)+\mathrm{abs}(5-14)+\mathrm{abs}(6-14)+\mathrm{abs}(7-14)+\mathrm{abs}(8-14)+\mathrm{abs}(90-14)}{9}$$

c) What is the average absolute loss induced by setting $s=5$?

(abs(1-5)+abs(2-5)+abs(3

✓ **Answer:** 101/9

$$\frac{\mathrm{abs}(1-5)+\mathrm{abs}(2-5)+\mathrm{abs}(3-5)+\mathrm{abs}(4-5)+\mathrm{abs}(5-5)+\mathrm{abs}(6-5)+\mathrm{abs}(7-5)+\mathrm{abs}(8-5)+\mathrm{abs}(90-5)}{9}$$

d) From parts (b) and (c), we see that the value of s that minimizes absolute loss is $\mathbf{not}$ the mean. In fact, it is the $\mathbf{median}$: if you arrange the set of numbers in order, the median is the number right in the middle (if the set has odd size) or any number between the two middle numbers (if the set has even size).

What is the median in the example above?

5

✓ **Answer:** 5

5

e) To see why the median is the solution in general (not just for the specific numbers in the example, but always, for any numbers), we could try to use calculus, as we did in the case of squared loss. But this is tricky, because the absolute value function $|x|$ is not differentiable (at $x = 0$). A related approach is to reason that if s is less than the median, then the loss function gets lower when you increase s; and if s is more than the median, then the loss function gets lower when you decrease s.

Work through this reasoning on your own, and then select one of the following. (You'll be marked correct either way.)

○ I was able to work through the entire argument. ✓

● I get the general idea, but got stuck on some details.

○ I don't know how to start. ✓

✓

Explanation

The mean, the average of the numbers, is 14. This leads to error:
$L(14) = \frac{13+12+11+10+9+8+7+6+76}{9} = \frac{152}{9}$.
On the other hand, the median of the numbers, 5, leads to a lower error value
$L(5) = 101/9$.

Submit

Answers are displayed within the problem

Problems 5-7 correspond to "Linear regression"

Problem 5

1/1 point (graded)

Let's say we have a regression problem with predictor variables $x_1, \ldots, x_d$ and response variable y. Sometimes, we are interested in solutions that do not necessarily use all the predictor variables. For instance, the linear function

$$f(x) = 3x_2 - 7x_9$$

uses just two of the features. We call this a $sparse$ solution.

Here is a question: does adding more features always help? That is, let S be some subset of the features and define $\mathrm{LOSS}(S)$ to be the loss obtained by doing least-squares regression using just these features. Now let's say we add another feature to S, to get S'. How does $\mathrm{LOSS}(S')$ compare to $\mathrm{LOSS}(S)$? Select all that apply.

☑ $\mathrm{LOSS}(S') \leq \mathrm{LOSS}(S)$ always

□ $\mathrm{LOSS}(S') \leq \mathrm{LOSS}(S)$ sometimes, but not always

□ $\mathrm{LOSS}(S') < \mathrm{LOSS}(S)$ always

☑ $\mathrm{LOSS}(S') < \mathrm{LOSS}(S)$ sometimes, but not always.

✓

Explanation

If $S \subset S'$ then any linear function expressible using S is also expressible using S'. This means that the optimal linear function using S' has loss at most that of the optimal linear function using just S. That is, $\mathrm{LOSS}(S') \leq \mathrm{LOSS}(S)$. Sometimes this will be a strict inequality, if S' contains some feature not in S that allows the data to be better modeled.

Submit

ℹ Answers are displayed within the problem

Problem 6

1/1 point (graded)

We have a data set $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})$, where $x^{(i)} \in \mathbb{R}^d$ and $y^{(i)} \in \mathbb{R}$. We want to express y as a linear function of x, of the form $w \cdot x + b$, but the error penalty we have in mind is not the usual squared loss: if we predict $\hat{y}$ and the true value is y, then the penalty should be the absolute difference, $|y - \hat{y}|$. Write down the loss function that corresponds to the total penalty on the training set.

● Click here where you think you have the correct loss function.

✓

Answer

Correct:

Your loss function should be $L(w, b) = \sum_{i=1}^{n} |y^{(i)} - (w \cdot x^{(i)} + b)|$. Check to ensure that your loss function is correct, and let us know on the forums what you came up with.

Submit

Answers are displayed within the problem

Problem 7

4/4 points (graded)

Let $x^{(1)}, \ldots, x^{(n)}$ be a set of n data points in $\mathbb{R}^d$, and let $y^{(1)}, \ldots, y^{(n)} \in \mathbb{R}$ be corresponding response values. In this problem, we will see how to rewrite several basic functions of the data using matrix-vector calculations. To this end, define:

X, the $n \times d$ matrix whose rows are the $x^{(i)}$

y, the n-dimensional vector with entries $y^{(i)}$

$\mathbf{1}$, the n-dimensional vector whose entries are all 1

and consider the matrix-vector expressions:

i) XX^T

ii) $(1/n)\,\mathbf{1}^T y$

iii) $(1/n)\,X^T X$

iv) $(1/n)\,X^T\mathbf{1}$

Each of the following quantities is equivalent to one of the expressions (i)-(iv) above. In each case, choose the correct match (i,ii,iii,iv) from the list.

a) The average of the $y^{(i)}$ values, that is, $\left(y^{(1)}+\cdots+y^{(n)}\right)/n$.

ii ▼ ✓ **Answer:** ii

b) The $n \times n$ matrix whose (i,j) entry is the dot product $x^{(i)} \cdot x^{(j)}$.

i ▼ ✓ **Answer:** i

c) The average of the $x^{(i)}$ vectors, that is, $\left(x^{(1)}+\cdots+x^{(n)}\right)/n$.

iv ▼ ✓ **Answer:** iv

d) The empirical covariance matrix, assuming the points $x^{(i)}$ are centered (that is, assuming the average of the $x^{(i)}$ vectors is zero). This is the $d \times d$ matrix whose (i,j) entry is

$$\frac{1}{n}\sum_{k=1}^{n} x_i^{(k)} x_j^{(k)}.$$

iii ▼ ✓ **Answer:** iii

Submit

🛈 Answers are displayed within the problem

Problems 8-11 correspond to "Regularized linear regression"

Problem 8

1/1 point (graded)

Suppose we want to predict a response value y using d predictor variables. The way we will do this is to use training data, consisting of n points $(x^{(i)}, y^{(i)})$, to fit a linear function. In general, how would the dimension of the data (d) influence the number of training points (n) that we need to fit a good model?

● For larger d, we need more data points.

○ For larger d, we need fewer data points.

○ The number of data points needed is unrelated to d.

✓

Explanation
Most machine learning methods require more data points as more features are added, even if the number of *relevant features* remains fixed.

Submit

ℹ Answers are displayed within the problem

Problem 9

2/2 points (graded)

In lecture, we asserted that in d-dimensional space, it is possible to perfectly fit (almost) any set of $d+1$ points $(x^{(0)}, y^{(0)}), (x^{(1)}, y^{(1)}), \ldots, (x^{(d)}, y^{(d)})$. Let's see how this works in the specific case where:

- $x^{(0)} = 0$

- $x^{(i)}$ is the ith coordinate vector (the vector that has a 1 in position i, and zeros everywhere else), for $i = 1, \ldots, d$

- $y^{(i)} = c_i$, where $c_0, c_1, \ldots, c_d$ are arbitrary constants.

Find $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$ such that $w \cdot x^{(i)} + b = y^{(i)}$ for all i. You should express your answer in terms of $c_0, c_1, \ldots, c_d$.

a) What is b?

○ $b = c_0 + c_1 + \ldots + c_d$

● $b = c_0$

○ $b = c_d$

○ $b = (1/d)\,(c_0 + c_1 + \ldots + c_d)$

✓

b) Write $w = (w_1, \ldots, w_d)$. What is w_i?

○ $w_i = c_i$

○ $w_i = c_0$

● $w_i = c_i - c_0$

○ $w_i = c_i = c_0$

✓

Explanation

We need to fit the given $d + 1$ points: that is, we want w and b such that $c_i = w \cdot x^{(i)} + b$. For $x^{(0)} = 0$, we get $c_o = b$. For $x^{(i)} = (\text{ith coordinate vector})$, we get $c_i = w_i + b$. Thus $b = c_o$ and each $w_i = c_i - c_o$.

Submit

Answers are displayed within the problem

Problem 10

4/4 points (graded)

Continuing from the previous problem, let's keep the same set of $d+1$ points $(x^{(0)}, y^{(0)}), (x^{(1)}, y^{(1)}), \ldots, (x^{(d)}, y^{(d)})$. As we saw, we can find w, b that perfectly fit these points; hence least-squares regression would find this perfect solution and have zero loss on the training set.

Now, let us instead use ridge regression, with parameter $\lambda \geq 0$, to obtain a solution. We can denote this solution by w_λ, b_λ. Also define the squared training loss associated with this solution,

$$L(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \left(y^{(i)} - (w_\lambda \cdot x^{(i)} + b_\lambda)\right)^2 .$$

a) What is $L(0)$?

0 ✓ **Answer:** 0

0

Answer
Correct: This is just the least-squares solution

b) As λ increases, how does $\|w_\lambda\|$ behave?

○ It increases

● It decreases

○ It initially decreases, then increases

○ It doesn't necessarily follow any of these trends

✓

c) As λ increases, how does $L(\lambda)$ behave?

● It increases

○ It decreases

○ It initially increases, then decreases

○ It doesn't necessarily follow any of these trends

✓

d) As λ goes to infinity, what value does $L(\lambda)$ approach?

○ Zero

○ Infinity

● The variance of the y-values

○ The average of the y-values

✓

Explanation

When $\lambda = 0$, we are in the least-squares situation. As λ grows, the loss function places a higher penalty on $\|w\|$, as a result of which smaller $\|w\|$ is preferred. Ultimately, as λ reaches ∞, the penalty on $\|w\|$ is so high that the solution is $w = 0$. This essentially predicts the mean value of y without looking at x, and thus the loss is just the variance of y.

? **Hint (1 of 2):** To figure out $L(0)$: when $\lambda = 0$, we are back in what situation?

Hint (2 of 2): As λ increases: the loss function assigns a higher penalty to $\|w\|$.

Next Hint

Submit

Answers are displayed within the problem

Problem 11

1/1 point (graded)

We saw in lecture that the Lasso tends to produce $\textbf{sparse}$ solutions to regression problems. We will now look at an alternative strategy for doing this.

Suppose we have d predictor variables. For any subset of features $S \subset \{1, 2, \ldots, d\}$ define $\text{LOSS}\left(S\right)$ to be the loss obtained by doing least-squares regression using just the features S. Given a sparsity level k, we would ideally like to find the subset S of size k such that $\text{LOSS}\left(S\right)$ is as small as possible. Unfortunately, this problem is NP-hard, which means that there is unlikely to be an efficient algorithm for it (the brute-force strategy, trying all possible subsets of k features, is inefficient when k is large).

Instead, an approximate solution can be obtained using a $\textbf{greedy}$ algorithm called $\textbf{forward stepwise regression}$. It chooses one feature at a time, as follows:

- Let S be the set of features chosen so far. Initially S is empty.

- Repeat while $|S| < k$: Find the feature x_i such that $\text{LOSS}\left(S \cup \{i\}\right)$ is as small as possible, and add i to S

Does this make sense? (Just give an honest answer; you'll get marked correct either way.)

○ This makes perfect sense to me. ✓

● I am somewhat hazy on what is going on here.

✓

Submit

ℹ Answers are displayed within the problem

Problems 12-13 correspond to "Linear models for conditional probability estimation"

Problem 12

1/1 point (graded)

We identified *inherent uncertainty* as one reason why it might be difficult to get perfect classifiers, even with a lot of training data. In which of the following situations is there likely to be a significant amount of inherent uncertainty? Select two of the four below.

☐ x is a picture of an animal and y is the name of the animal

☑ x consists of the dating profiles of two people and y is whether they will be interested in each other

☐ x is a speech recording and y is the transcription of the speech into words

☑ x is the recording of a new song and y is whether it will be a big hit

✓

Submit

ℹ Answers are displayed within the problem

Problem 13

3/3 points (graded)

A logistic regression model given by parameters $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$ is fit to a data set of points $x \in \mathbb{R}^d$ with binary labels $y \in \{-1, 1\}$. For any constant c, the model assigns the same conditional probability $\Pr(y = 1|x)$ to all points x that satisfy $w \cdot x + b = c$.

What is the value of $w \cdot x + b$ for which the following conditional probabilities are assigned? In each case, just give a single real number, to two decimal places.

a) $\Pr(y = 1|x) = 1/2$

0 ✓ **Answer:** 0

0

b) $\Pr(y = 1|x) = 3/4$

1.1

✓ **Answer:** 1.10

1.1

c) $\Pr(y = 1|x) = 1/4$

-1.1

✓ **Answer:** -1.10

-1.1

Explanation

Recall that

$\Pr(y = 1|x) = \frac{1}{1+e^{-(w \cdot x+b)}}$.

Part (a): this probability is $1/2$ when $w \cdot x + b = 0$, that is, on the decision boundary.

Part (b): the probability in the formula above is $3/4$ when $w \cdot x + b = \ln 3$.

Part (c): the probability in the formula above is $1/4$ when $w \cdot x + b = -\ln 3$: the symmetric situation to part (b).

? **Hint (1 of 3):** For part (a): probability $1/2$ is assigned to points that lie exactly on the decision boundary.

Next Hint

Hint (2 of 3): For part (b): look at the specific formula for the logistic regression conditional probability model. For what value of $w \cdot x + b$ will this formula give a probability of $3/4$ for $y = 1$?

Hint (3 of 3): For part (c): this is symmetric to part (b), on the other side of the boundary. No calculation should be necessary.

Submit

Answers are displayed within the problem

Problems 14-16 correspond to "Logistic regression"

Problem 14

1/1 point (graded)

When learning a logistic regression model from training data $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})$, which of the following do we try to do? Select all that apply.

☐ Maximize the probabilities of the $x^{(i)}$

☑ Maximize the conditional probabilities of the $y^{(i)}$ given $x^{(i)}$

☐ Maximize the joint probabilities of $x^{(i)}$ and $y^{(i)}$

✓

Submit

ℹ Answers are displayed within the problem

Problem 15

1/1 point (graded)

Which of the following best describes the process of learning a logistic regression model? Select all that apply.

☐ There is a simple closed-form formula for the solution.

☑ The solution is the minimum point of a loss function that is convex.

☐ The solution is the minimum point of a loss function that might have multiple local optima of varying quality.

☑ The solution is found by local search.

✓

Submit

ℹ Answers are displayed within the problem

Problem 16

1/1 point (graded)

In many situations, it would be helpful to have a conditional probability function that is easy to interpret, for instance, one that depends on just a few features. Which of the following are reasonable strategies for learning a *sparse* logistic regression model (that is, one in which w has only a few nonzero entries)? Select all that apply.

☐ Modify the loss function for logistic regression by adding a regularization term of the form $\lambda \|w\|_2^2$.

☑ Modify the loss function for logistic regression by adding a regularization term of the form $\lambda \|w\|_1$.

☑ Use a forward stepwise strategy to select one feature at a time: the feature that (when added to those already chosen) most reduces the loss function.

☐ First fit a logistic regression model, and then retain the features with the highest-valued coefficients.

✓

Explanation

The first option corresponds to ridge regression while the second is the Lasso. As we have seen, it is the Lasso that tends to produce sparse solutions.
The third option is reasonable.
The fourth is dangerous: if the data is very high-dimensional and all the coordinates of the logistic regression solution have significant weights, then keeping just a small number of them could lead to a very poor classifier.

Submit

ℹ Answers are displayed within the problem

Problem 17 corresponds to "Logistic regression in use"

Problem 17

1/1 point (graded)

Suppose that in a bag-of-words representation, we decide to use the following vocabulary of five words: **`is`**, **`flower`**, **`rose`**, **`a`**, **`an`**. What is the vector form of the following sentence: *A rose is a rose is a rose*?

○ (2,3,3)

● (2,0,3,3,0)

○ (1,0,1,1,0)

○ (3,3,2,0,0)

✓

Submit

Answers are displayed within the problem