DSC 40A - Homework 05

Due: Friday, February 14, 2020

Write your solutions to the following problems by either typing them up or handwriting them on another piece of paper. Unless otherwise noted by the problem's instructions, show your work or provide some justification for your answer. Homeworks are due via Gradescope on Friday afternoon at 5:00 p.m.

Note: We are collecting mid-quarter feedback! Please consider filling out the following Google Form. Your responses are totally anonymous.

Problem 1.

Let $\vec{y} \in \mathbb{R}^n$, $\mathbb{X}^{n \times (d+1)}$, and $\vec{w} \in \mathbb{R}^{d+1}$.

a) What should be the size of the gradient vector, $\frac{d}{d\vec{w}}[\vec{y}^T X \vec{w}]$? That is, how many entries does it have? You should be able to answer without actually computing the gradient.

Solution: It should have $d+1$ entries, since $\vec{w}$ has $d+1$ entries, and the gradient vector is the vector which contains an entry for each partial derivative with respect to an entry of $\vec{w}$.

b) Compute $\frac{d}{d\vec{w}}[\vec{y}^T X \vec{w}]$ by whatever means you'd like. Simplify your answer as much as possible – it should be some matrix times some vector.

Solution: Let $\vec{y} \in \mathbb{R}^n$, $\mathbb{X}^{n \times (d+1)}$, and $\vec{x} \in \mathbb{R}^{d+1}$. We'll start by unpacking $\vec{y}^T X \vec{w}$ until we see only dot products whose gradient is known. We have:

$$\begin{aligned}\vec{y}^T X \vec{w} &= \vec{y}^T \begin{pmatrix} \vec{X}_{1*} \cdot \vec{w} \\ \vec{X}_{2*} \cdot \vec{w} \\ \vdots \\ \vec{X}_{n*} \cdot \vec{w} \end{pmatrix} \\ &= y_1 \vec{X}_{1*} \cdot \vec{w} + y_2 \vec{X}_{2*} \cdot \vec{w} + \ldots + y_n \vec{X}_{n*} \cdot \vec{w}\end{aligned}$$

where $\vec{X}_{i*}$ denotes the ith row of X, as a vector. we know that if $\vec{a}$ is a vector of constants, then $\frac{d}{d\vec{w}} \vec{a} \cdot \vec{w} = \vec{a}$. Now, $\vec{X}_{i*} \cdot \vec{w}$ is a vector of constants when we are finding the gradient with respect to $\vec{w}$. So:

$$\begin{aligned}\frac{d}{d\vec{w}}[\vec{y}^T X \vec{w}] &= \frac{d}{d\vec{w}} \left[y_1 \vec{X}_{1*} \cdot \vec{w} + y_2 \vec{X}_{2*} \cdot \vec{w} + \ldots + y_n \vec{X}_{n*} \cdot \vec{w} \right] \\ &= \frac{d}{d\vec{w}} \left[y_1 \vec{X}_{1*} \cdot \vec{w} \right] + \frac{d}{d\vec{w}} \left[y_2 \vec{X}_{2*} \cdot \vec{w} \right] + \ldots + \frac{d}{d\vec{w}} \left[y_n \vec{X}_{n*} \cdot \vec{w} \right] \\ &= y_1 \frac{d}{d\vec{w}} \left[\vec{X}_{1*} \cdot \vec{w} \right] + y_2 \frac{d}{d\vec{w}} \left[\vec{X}_{2*} \cdot \vec{w} \right] + \ldots + y_n \frac{d}{d\vec{w}} \left[\vec{X}_{n*} \cdot \vec{w} \right] \\ &= y_1 \vec{X}_{1*} + y_2 \vec{X}_{2*} + \ldots + y_n \vec{X}_{n*}\end{aligned}$$

We are taking y_1 of the first row of X, y_2 of the second row, and so on. This is the same as taking y_1 of the first *column* of X^T, y_2 of the second column, and so on. That is just another way of viewing matrix multiplication of $X^T\vec{y}$. And so:

$$= X^T\vec{y}$$

Problem 2.

Beginning with the normal equations, $\vec{w} = (X^\intercal X)^{-1}X^\intercal\vec{y}$, and assuming that $\vec{y}$ is $n \times 1$ and

$$X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix},$$

derive the familiar formula

$$w_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Hint: the inverse of a 2×2 matrix $A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$ is given by $A^{-1} = \frac{1}{\det(A)}\begin{pmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{pmatrix}$, where $\det(A) = a_{11}a_{22} - a_{12}a_{21}$. This is the only time you'll need to know how to invert a matrix in this class.

Solution: We start by multiplying $X^\intercal X$:

$$X^\intercal X = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{pmatrix}\begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$$
$$= \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix},$$

where all of the sums range from $i = 1$ to $i = n$. Now we take the inverse of this matrix. To use the formula provided for the inverse of a 2×2 matrix, we'll first need to calculate the determinant:

$$\det\begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix} = n\sum x_i^2 - \left(\sum x_i\right)^2$$

Now we apply the formula:

$$(X^\intercal X)^{-1} = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}^{-1}$$
$$= \frac{1}{n\sum x_i^2 - \left(\sum x_i\right)^2}\begin{pmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{pmatrix}.$$

Now we compute $X^\intercal \vec{y}$:

$$X^\intercal \vec{y} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$
$$= \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix}$$

Now we can calculate $\vec{w}$:

$$\begin{aligned} \vec{w} &= (X^\intercal X)^{-1} X^\intercal \vec{y} \\ &= \frac{1}{n\sum x_i^2 - \left(\sum x_i\right)^2} \begin{pmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{pmatrix} \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix} \\ &= \frac{1}{n\sum x_i^2 - \left(\sum x_i\right)^2} \begin{pmatrix} \sum x_i^2 y_i - \sum x_i \sum x_i y_i \\ -\sum x_i \sum y_i + n\sum x_i y_i \end{pmatrix} \end{aligned}$$

Recalling that we have defined $\vec{w} = (w_0, w_1)^\intercal$, w_1 is the second entry of this vector. That is:

$$w_1 = \frac{-\sum x_i \sum y_i + n\sum x_i y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2}$$

Now we'll work backwards from the familiar formula for w_1 in order to show that it equals the above. We have:

$$\begin{aligned} w_1 &= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \\ &= \frac{\sum \left(x_i y_i - x_i \bar{y} - \bar{x} y_i + \bar{x}\bar{y}\right)}{\sum \left(x_i^2 + \bar{x}^2 - 2x_i\bar{x}\right)} \\ &= \frac{\sum x_i y_i - \bar{y}\sum x_i - \bar{x}\sum y_i + n\bar{x}\bar{y}}{\sum x_i^2 + n\bar{x}^2 - 2\bar{x}\sum x_i} \end{aligned}$$

Writing $\bar{x} = \frac{1}{n}\sum x_i$, and similarly for $\bar{y}$:

$$\begin{aligned} &= \frac{\sum x_i y_i - \frac{1}{n}\left(\sum y_i\right)\left(\sum x_i\right) - \frac{1}{n}\left(\sum x_i\right)\left(\sum y_i\right) + n\left(\frac{1}{n}\right)^2 \left(\sum x_i\right)\left(\sum y_i\right)}{\sum x_i^2 + n\left(\frac{1}{n}\right)^2 \left(\sum x_i\right)^2 - 2\frac{1}{n}\left(\sum x_i\right)^2} \\ &= \frac{\sum x_i y_i - \frac{1}{n}\left(\sum x_i\right)\left(\sum y_i\right)}{\sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2} \end{aligned}$$

Multiplying by n/n:

$$= \frac{n\sum x_i y_i - \left(\sum x_i\right)\left(\sum y_i\right)}{n\sum x_i^2 - \left(\sum x_i\right)^2}$$

This is the expression we found above, so this proves the claim.

Problem 3.

Compute the gradient $\frac{d}{d\vec{x}}\|\vec{x}\|$.

Solution: There are two main ways of computing this: the easy way and the hard way.

The "hard" way is to expand the norm until we see indices. By definition of the vector norm:

$$\|\vec{x}\| = \sqrt{x_1^2 + \ldots + x_d^2}$$

Now we take partial derivatives. Starting with $\partial/\partial x_1$:

$$\frac{\partial}{\partial x_1}\|\vec{x}\| = \frac{\partial}{\partial x_1}\sqrt{x_1^2 + \ldots + x_d^2}$$

Using the standard chain rule for partial derivatives:

$$= \frac{1}{2}\left(x_1^2 + \ldots + x_d^2\right)^{-1/2} \cdot 2x_1 \qquad = \left(x_1^2 + \ldots + x_d^2\right)^{-1/2} \cdot x_1$$

Next we compute $\partial/\partial x_2$. The only difference is that the trailing x_1 will be replaced by x_2:

$$\frac{\partial}{\partial x_2}\|\vec{x}\| = \left(x_1^2 + \ldots + x_d^2\right)^{-1/2} \cdot x_2$$

Seeing the pattern, we write down the gradient vector:

$$\begin{aligned}
\frac{d}{d\vec{x}}\|x\| &= \begin{pmatrix} \left(x_1^2 + \ldots + x_d^2\right)^{-1/2} x_1 \\ \left(x_1^2 + \ldots + x_d^2\right)^{-1/2} x_2 \\ \vdots \\ \left(x_1^2 + \ldots + x_d^2\right)^{-1/2} x_d \end{pmatrix} \\
&= (x_1^1 + \ldots + x_d^2)^{-1/2} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix} \\
&= (x_1^1 + \ldots + x_d^2)^{-1/2}\vec{x} \\
&= \frac{\vec{x}}{(x_1^1 + \ldots + x_d^2)^{-1/2}} \\
&= \frac{\vec{x}}{\|\vec{x}\|}
\end{aligned}$$

Here's a slightly easier way of getting this answer. We know that $\frac{d}{d\vec{x}}\|\vec{x}\|^2 = 2\|\vec{x}\|\frac{d}{d\vec{x}}\|\vec{x}\|$. We also know that $\frac{d}{d\vec{x}}\|\vec{x}\|^2 = \frac{d}{d\vec{x}}\vec{x}^T\vec{x} = 2\vec{x}$. Equating these two, we have: $2\vec{x} = 2\|\vec{x}\|\frac{d}{d\vec{x}}\|\vec{x}\|$. Solving, we find $\frac{d}{d\vec{x}}\|\vec{x}\| = \vec{x}/\|\vec{x}\|$.

Problem 4.

Suppose you have collected the following data in a survey of data scientists:

Experience	GPA	Salary
5	3.2	85
7	3.7	110
3	3.1	87
9	3.5	105
2	3.2	80

Using least squares, fit a prediction rule of the form $H(\text{experience}, \text{GPA}) = w_0 + w_1 \times \text{experience} + w_2 \times \text{GPA}$. You'll need to solve a system of three equations with three unknowns. You can do this by hand, or you can use `np.linalg.solve`.

Solution: The following code will compute the least squares solution:

```
import numpy as np

X = np.array([
    [1, 5, 3.2],
    [1, 7, 3.7],
    [1, 3, 3.1],
    [1, 9, 3.5],
    [1, 2, 3.2]
])

y = np.array([85, 110, 87, 105, 80])

np.linalg.solve(X.T @ X, X.T @ y)
```

The result is `array([-29.30968968, 1.69284357, 34.1038632 ])`. That is, $w_0 = -29.3$, $w_1 = 1.69$, and $w_2 = 34.1$.

Problem 5.

Suppose we collect salaries as well as experience, GPA, and number of internships. To begin, we fit a linear prediction rule $w_0 + w_1 x_1 + w_2 x_2$ to the data using only the experience and GPA as features – we find that our rule has a mean squared error of E_1. Next, we fit a linear prediction rule $w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3$ to the same data, but using all three features: experience, GPA, and number of internships. We find that our new line has a mean squared error of E_2. Is it possible for the new mean squared error to be larger than the old? That is, can $E_2 > E_1$? Justify your answer.

Solution: The mean squared error cannot be increased by adding a new feature. Here's why:

Suppose when we use two features, the best parameters (in the least squares sense) are a_0, a_1, a_2. Then suppose that when we use three features, the best parameters (in the least squares sense), are b_0, b_1, b_2, b_3. There's no reason that $b_0 = a_0$, $b_1 = a_1$, and $b_2 = a_2$; in fact, they're most likely different.

Because b_0, b_1, b_2, b_3 are the *least squares solutions*, then have the smallest mean squared error out of any choice of parameters. That is, the MSE E_2 of the prediction rule $b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3$ is as small as it can possibly be. If we were to pick other coefficients – *whatever* they may be – the MSE cannot be smaller. So what if we picked coefficients $a_0, a_1, a_2, 0$? Our prediction rule would be $a_0 + a_1 x_1 + a_2 x_2 + 0 x_3 = a_0 + a_1 x_1 + a_2 x_2$; in other words, it would make the same exact predictions as the least squares solution with only two features. As a result, its MSE would be E_1. And since we said that E_2 is the smallest error possible with three features, $E_2 < E_1$.

Problem 6.

A categorical variable is one which can take only a small number of distinct values. For example, the college that a UCSD student belongs to is a categorical variable: it takes values in the set "Warren", "Muir", "Roosevelt", "Revelle", "Marshall", "Sixth".

We can use categorical variables in regression, but we must first encode the value of a categorical variable as a number. There are two main ways of encoding a categorical variable. The first is to map each possible value to a number; for instance, "Warren" might become 1, "Muir" might become 2, "Roosevelt" becomes

3, and so on. The other approach is called "one-hot encoding". In this situation, we introduce a new feature for every possible value of the categorical variable. To encode someone's college, for instance, we create 6 new features, $x_1, x_2, \ldots, x_6$, one for each college. The feature is one if the student is in that college, and zero otherwise. For instance, supposing that x_1 represents "Warren", x_2 represents "Muir", x_3 represents "Roosevelt", and so on, a student in Muir would be represented as the vector $(0, 1, 0, 0, 0, 0)$.

Suppose we are using least squares regression to find a linear prediction rule for predicting the salary of a UCSD graduate, and we will use their college as a feature. Which method of encoding their college should be used – mapping to a number, or one-hot encoding? Why?

Solution: We should use one-hot encoding so that each college is associated with its own feature, as opposed to using one feature to encode each college.

Suppose we use a single feature to represent the college by mapping each college to a number. We solve the least squares problem and find that this feature is given a weight of w. Since Warren is encoded as 1, someone in Warren will have a predicted salary of w. Since Muir is encoded as 2, someone in that college will have a predicted salary of $2w$. And since Roosevelt is encoded as 3, someone in that college has a predicted salary of $3w$.

As we see, the predicted salary depends strongly on the order in which the colleges are enumerated when choosing their (arbitrary) labels. The fact that Muir was second and Warren was first was totally arbitrary, but our prediction rule will always say that people in Muir make twice as much as those in Warren simply because Muir was 2 and Warren was 1. This is strange and not at all desirable. The prediction will probably be pretty poor, too.

On the other hand, if we use one-hot encoding, each college has its own feature. When we fit a prediction rule using least squares regression, we get six weights $w_1, \ldots, w_6$; one for each college. Since a person will only ever belong to one college, when evaluating the prediction rule only one of these weights will be "active"; the rest are multiplied by zeros. As such, we can interpret the weight w_i as the adjustment to someone's salary that comes specifically from their membership in the ith college.