# CSE 151A - Homework 04

Due: Wednesday, April 29, 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 Wednesday at 11:59 p.m.

**Essential Problem 1.**

Consider two $d \times 1$ vectors, $x$ and $y$. The dot product (also called the inner product) of $x$, $y$ is defined as:

$$\vec{x} \cdot \vec{y} = \vec{x}^\intercal \vec{y}$$

Notice that the inner product of two vectors is a scalar.
Outer product of $\vec{x}$, $\vec{y}$ is defined as:

$$\vec{x} \circ \vec{y} = \vec{x}\vec{y}^\intercal$$

Notice that if $\vec{x}$ and $\vec{y}$ are $d \times 1$ vectors, then $x \circ y$ is a $d \times d$ matrix.
Using basic properties of matrix multiplication, determine whether the following statement is true or false:

$$x^\top (yy^\top)x = y^\top (xx^\top)y$$

Justify your answer.

**Solution:** Yes

**Explanation**:
Recall that matrix multiplication is *associative*, meaning we can move those parenthesis around. A rather simpler way of looking at this equality is

$$(x^\intercal y)(y^\intercal x) \stackrel{?}{=} (y^\intercal x)(x^\intercal y)$$

This is much easier to process. Remember that vector inner-products are commutative, and equal a scalar constant, (say $a$):

$$x^\intercal y = y^\intercal x = \sum_{i=1}^{d} y_i x_i = a$$

and the above equality can easily be verified

$$(x^\intercal y)(y^\intercal x) \stackrel{?}{=} (y^\intercal x)(x^\intercal y)$$
$$(a)(a) \stackrel{?}{=} (a)(a)$$
$$a^2 \stackrel{\checkmark}{=} a^2$$

This is the solution.

**Essential Problem 2.**

This problem will check that we're all on the same page when it comes to the notation used in lecture.

The table below shows data we've collected on the salaries of three data scientists (don't worry, the Salary column is in thousands of dollars).

| Person | GPA | Experience | Salary |
|---|---|---|---|
| 1 | 3.3 | 4 | 95 |
| 2 | 3.9 | 10 | 120 |
| 3 | 3.2 | 3 | 80 |

Suppose we have decided on a prediction rule:

$$H(\vec{x}) = 50 + 10 \times x_1 + 2 \times x_2,$$

where the first component of $\vec{x}$, $x_1$, represents GPA and the second component, $x_2$, represents Experience.

**a)** Write down the parameter vector, $\vec{w}$. Assume that it includes $w_0$. Your answer should be a vector with three elements.

**Solution:**

$$\vec{w} = \begin{pmatrix} 50 \\ 10 \\ 2 \end{pmatrix}$$

**b)** Write down the data vectors $\vec{x}^{(1)}$, $\vec{x}^{(2)}$, and $\vec{x}^{(3)}$ for the first, second, and third person in data set, respectively.

**Solution:**

$$\vec{x}^{(1)} = \begin{pmatrix} 3.3 \\ 4 \end{pmatrix} \qquad \vec{x}^{(2)} = \begin{pmatrix} 3.9 \\ 10 \end{pmatrix} \qquad \vec{x}^{(3)} = \begin{pmatrix} 3.2 \\ 3 \end{pmatrix}$$

**c)** Compute the predicted salaries $H(\vec{x}^{(1)})$, $H(\vec{x}^{(2)})$, $H(\vec{x}^{(3)})$ for each of the three people in the data set.

**Solution:**

$$\begin{aligned}
H(\vec{x}^{(1)}) &= \vec{w} \cdot \text{Augmented}(\vec{x}^{(1)}) \\
&= \begin{pmatrix} 50 \\ 10 \\ 2 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 3.3 \\ 4 \end{pmatrix} \\
&= 50 + 33 + 8 \\
&= 91
\end{aligned}$$

$$\begin{aligned}
H(\vec{x}^{(2)}) &= \vec{w} \cdot \text{Augmented}(\vec{x}^{(2)}) \\
&= \begin{pmatrix} 50 \\ 10 \\ 2 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 3.9 \\ 10 \end{pmatrix} \\
&= 50 + 39 + 20 \\
&= 109
\end{aligned}$$

$$\begin{aligned}
H(\vec{x}^{(3)}) &= \vec{w} \cdot \text{Augmented}(\vec{x}^{(3)}) \\
&= \begin{pmatrix} 5010 \\ 2 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 3.2 \\ 3 \end{pmatrix} \\
&= 50 + 32 + 6 \\
&= 88
\end{aligned}$$

**d)** Compute the mean squared error of this prediction rule (with this particular choice of parameters).

**Solution:** The squared error of each prediction is:

$$\begin{aligned}
(H(\vec{x}^{(1)}) - y_1)^2 &= (91 - 95)^2 = 16 \\
(H(\vec{x}^{(2)}) - y_2)^2 &= (109 - 120)^2 = 121 \\
(H(\vec{x}^{(3)}) - y_3)^2 &= (88 - 80)^2 = 64
\end{aligned}$$

That makes the mean squared error:

$$(16 + 121 + 64)/3 = 201 = 67.$$

**e)** Write down the *design matrix*, $X$.

**Solution:**

$$X = \begin{pmatrix} 1 & 3.3 & 4 \\ 1 & 3.9 & 10 \\ 1 & 3.2 & 3 \end{pmatrix}$$

**f)** Check that the entries of $X\vec{w}$ are the predicted salaries you found above.

**Solution:**

$$
\begin{aligned}
X\vec{w} &= \begin{pmatrix} 1 & 3.3 & 4 \\ 1 & 3.9 & 10 \\ 1 & 3.2 & 3 \end{pmatrix} \begin{pmatrix} 50 \\ 10 \\ 2 \end{pmatrix} \\
&= \begin{pmatrix} 50 + 33 + 8 \\ 50 + 39 + 20 \\ 50 + 32 + 6 \end{pmatrix} \\
&= \begin{pmatrix} 91 \\ 109 \\ 88 \end{pmatrix}
\end{aligned}
$$

**g)** Calculate the norm of the vector $X\vec{w} - \vec{y}$, where $\vec{y} = (95, 120, 80)^\intercal$ is the vector of observations.

**Solution:**

$$
\begin{aligned}
\|X\vec{w} - \vec{y}\| &= \left\| \begin{pmatrix} 91 \\ 109 \\ 88 \end{pmatrix} - \begin{pmatrix} 95 \\ 120 \\ 80 \end{pmatrix} \right\| \\
&= \left\| \begin{pmatrix} 4 \\ 11 \\ 8 \end{pmatrix} \right\| \\
&= \sqrt{4^2 + 11^2 + 8^2} \\
&= \sqrt{16 + 121 + 64} \\
&= \sqrt{201}
\end{aligned}
$$

**h)** Check that 1/3 of the *squared norm* of $X\vec{w} - \vec{y}$ is the mean squared error you found above.

**Solution:**

$$
\frac{1}{3} \cdot \sqrt{201}^2 = \frac{201}{3} = 67
$$

**Essential Problem 3.**

Suppose you have collected the following data in a survey of machine learning engineers (don't worry, the salary is in thousands of dollars).

| 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 a library function, like `np.linalg.solve`.

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

```python
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 $w_2 = 34.1$.

**Essential Problem 4.**

In this problem, recall that the aim of logistic regression is to predict the probability that an input vector belongs to a positive class – for example, the probability that a particular patient has heart disease.

**a)** Suppose that a patient is represented by three features: their age ($x_1$), cholesterol level ($x_2$), and exercise frequency in days per week ($x_3$). Assume that the weight associated to the age feature is $w_1 = .005$, the weight associated to cholesterol level is $w_2 = .02$, and the weight given to exercise frequency is $w_3 = -.2$. Also assume that the "bias" weight is $w_0 = -.1$.

Consider a new patient who is 62 years old, has a cholesterol level of 242, and exercises 1 time a week. Under the logistic regression model, what is the predicted probability that this patient has heart disease? Show your work.

**Solution:**

$$\begin{aligned}\sigma(\vec{w} \cdot \operatorname{Aug}(\vec{x})) &= \sigma(-0.1 + 0.005(62) + 0.02(242) - 0.2(1)) \\ &= \frac{1}{1 + e^{-4.85}} \\ &\approx 0.992\end{aligned}$$

**b)** Suppose we now wish to predict the probability that someone exercises given their age, cholesterol level, and lung efficiency. Suppose $w_1, w_2$, and $w_3$ are the weights assigned to these three features, respectively, in a logistic regression model for this problem. If the model is to make good predictions, what do you expect the sign of each weight to be? Provide some justification for your answers.

**Solution:** We may expect that: older people and people with high cholesterol are less likely to exercise, and people with high lung efficiency are more likely to exercise. Thus, the signs of weights $w_1, w_2$, and $w_3$ may be $-, -, +$ respectively.

**Plus Problem 1.** (6 plus points)

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

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

Hint: the inverse of a $2 \times 2$ 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 $X^\intercal X$:

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

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 $2 \times 2$ 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:

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

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

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

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:

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

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.

**Plus Problem 2.** (5 plus points)

Recall that in logistic regression, the predicted probability that an input vector $\vec{x}$ belongs to the positive class is given by

$$H(\vec{x}) = \sigma(\vec{w} \cdot \vec{x}),$$

where $\vec{w}$ is a vector of parameters that is learned from data and $\sigma$ is the logistic function.

Although the output of $H(\vec{x})$ is a probability, we can turn it into a binary classification by thresholding. For instance, we might say that if $H(\vec{x}) \geq 0.5$, return "Yes", otherwise return "No".

Suppose we train a logistic regression model using two features, $x_1$ and $x_2$, and find a parameter vector $\vec{w} = (-.5, -1, 2)^\intercal$. The input space is 2-dimensional here; draw the decision boundary which partitions the input space into a region where the prediction is "Yes" and a region where the prediction is "No", assuming a threshold of 0.5. Mark where the decision boundary crosses each axis.

**Solution:** We have $\vec{x} = (1, x_1, x_2)^\intercal, \vec{w} = (-0.5, -1, 2)^\intercal$, and we want to find where $H(\vec{x}) = \sigma(\vec{w} \cdot \vec{x}) = 0.5$. Solving for $x_2$ in terms of $x_1$ will allow us to plot a decision boundary in our feature-space.

$$\begin{aligned} 0.5 &= \frac{1}{1 + e^{-(\vec{w} \cdot \vec{x})}} \\ \implies \quad 1 &= 0.5(1 + e^{-(\vec{w} \cdot \vec{x})}) \\ \implies \quad 2 &= 1 + e^{-(\vec{w} \cdot \vec{x})} \\ \implies \quad 0 &= -(\vec{w} \cdot \vec{x}) \\ \implies \quad x_2 &= \frac{0.5 + x_1}{2} \end{aligned}$$

We will have an $x_1$-intercept at -0.5 and an $x_2$-intercept at 0.25.