DSC 40A - Group Work Session 2

due Friday, July 14 at 11:59pm

Write your solutions to the following problems by either typing them up or handwriting them on another piece of paper. **One person** from each group should submit your solutions to Gradescope and **tag all group members** so everyone gets credit.

This worksheet won't be graded on correctness, but rather on good-faith effort. Even if you don't solve any of the problems, you should include some explanation of what you thought about and discussed, so that you can get credit for spending time on the assignment.

In order to receive full credit, you must work in a group of two to four students for at least 50 minutes in your assigned discussion section. You can also self-organize a group and meet outside of discussion section for 95 percent credit. You may not do the groupwork alone.

1 Error in a Prediction Rule

The problems in this section test your understanding of definitions only. You should be able to write down the answers to these questions without referring to any notes or resources.

Problem 1.

Consider the data set $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ and the linear prediction rule $y = 3x + 7$. Write down the expression for the mean squared error of this prediction rule on the data set.

Problem 2.

Consider the data set $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ and the quadratic prediction rule $y = 2x^2 - 4x + 1$. Write down the expression for the mean absolute error of this prediction rule on the data set.

2 Equivalent Formulas for Linear Regression

In class, we showed that the slope and intercept of the regression line $H^*(x) = w_0^* + w_1^* x$ are given by

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

$$w_0^* = \overline{y} - w_1^* \overline{x},$$

where $\bar{x}$ and $\bar{y}$ represent the mean of the x's and y's, respectively.

We also showed an equivalent form of the slope:

$$w_1^* = r\frac{\sigma_y}{\sigma_x},$$

where σ_x and σ_y represent the standard deviations of the x's and y's, respectively.

Now, you will show the equivalence of another common form for the slope. It can be useful to have multiple equivalent formulas because some properties can be easier to prove when we start with a certain form. After doing this problem, feel free to start at any of these equivalent forms when solving other problems in this class.

Problem 3.

Show that

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

Substituting this into the numerator of w_1^* gives an equivalent formulation of the slope of the regression line:

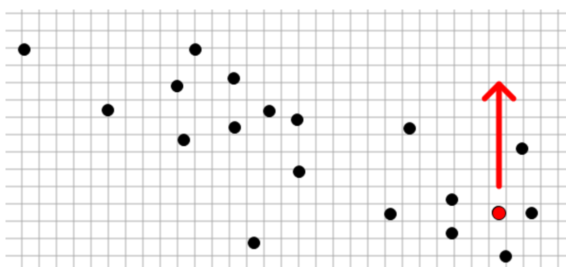
$$w_1^* = \frac{\displaystyle\sum_{i=1}^{n}(x_i - \overline{x})y_i}{\displaystyle\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

3 Visualizing Changes in the Data

The problems in this section will help you visualize how changes in the data affect the regression line. Assume all data is in the first quadrant (positive x and y coordinates).

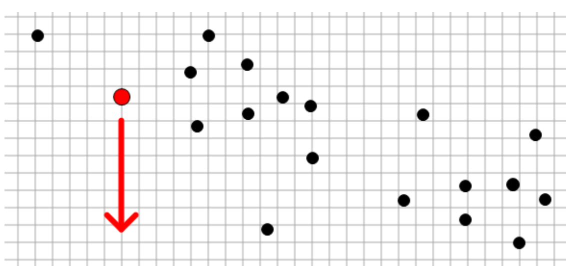
Problem 4.

For the data set shown below, how will the slope and intercept of the regression line change if we move the red point in the direction of the arrow?



Problem 5.

For the data set shown below, how will the slope and intercept of the regression line change if we move the red point in the direction of the arrow?



Problem 6.

Suppose we transform a data set of $\{(x_i, y_i)\}$ pairs by doubling each y-value, creating a transformed data set $\{(x_i, 2y_i)\}$. How does the slope of the regression line fit to the transformed data compare to the slope of the regression line fit to the original data? Can you prove your answer from the formula for the slope of the regression line?

Problem 7.

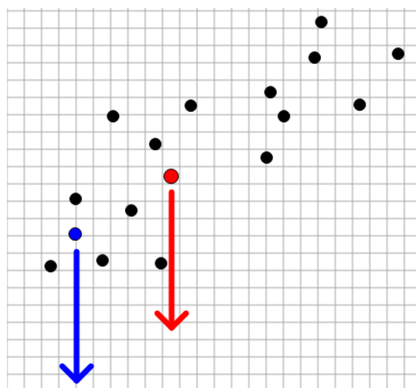
Suppose we transform a data set of $\{(x_i, y_i)\}$ pairs by doubling each x-value, creating a transformed data set $\{(2x_i, y_i)\}$. How does the slope of the regression line fit to the transformed data compare to the slope of the regression line fit to the original data? Can you prove your answer from the formula for the slope of the regression line?

Problem 8.

Compare two different possible changes to the data set shown below.

- Move the red point down c units.
- Move the blue point down c units.

Which move will change the slope of the regression line more? Why?



4 Gradient with Respect to a Vector

The derivative of a scalar-valued function $f(\vec{w})$ with respect to a vector input $\vec{w} \in \mathbb{R}^n$ is called the gradient. The gradient of $f(\vec{w})$ with respect to $\vec{w}$, written $\nabla_{\vec{w}} f$ or $\frac{df}{d\vec{w}}$, is defined to be the vector of partial derivatives:

$$\frac{df}{d\vec{w}} = \begin{bmatrix} \frac{\partial f}{\partial w_1} \\ \frac{\partial f}{\partial w_2} \\ \vdots \\ \frac{\partial f}{\partial w_n}, \end{bmatrix}$$

where $w_1, \ldots, w_n$ are the components of the vector $\vec{w}$. In other words, the gradient of $f(\vec{w})$ with respect to $\vec{w}$ is the same as the gradient of $f(w_1, \ldots, w_n)$, a multivariable function of the components of $\vec{w}$.

Problem 9.

If $\vec{w} \in \mathbb{R}^n$, show that the gradient of $\vec{w}^T \vec{w}$ with respect to $\vec{w}$ is given by

$$\frac{d}{d\vec{w}}(\vec{w}^T \vec{w}) = 2\vec{w}$$

This should remind you of the familiar rule from single-variable calculus that says $\frac{d}{dx}(x^2) = 2x$.

Hints:

- First, start by writing $\vec{w}^T \vec{w}$ as a sum. What is the partial derivative of that sum with respect to w_1? With respect to w_2?
- If you're still confused, look at the example from Lecture 11 called "Example gradient calculation."

5 Multiple Regression

This problem will check that we're all on the same page when it comes to the notation and basic concepts of regression with multiple features.

Problem 10.

The table below shows the softness and color of several different avocados, which we want to use to predict their ripeness. Each variable is measured on a scale of 1 to 5. For softness, 5 is softest, for color, 5 is darkest, and for ripeness, 5 is ripest.

Avocado	Softness	Color	Ripeness
1	3	4	2.5
2	1	2	2
3	4	5	5

Suppose we have decided on the following prediction rule: given an avocado's softness and color, we average these numbers to produce a predicted ripeness.

a) Is this prediction rule a linear prediction rule or not?

b) Write down the prediction rule as a function $H(\vec{x})$, where

$$\vec{x} = \begin{bmatrix} x^{(1)} \\ x^{(2)} \end{bmatrix},$$

with $x^{(1)}$, representing softness and $x^{(2)}$ representing color.

c) Write down the feature vectors $\vec{x}_1$, $\vec{x}_2$, and $\vec{x}_3$ for the first, second, and third avocados in the data set, respectively.

d) Compute the predicted ripeness $H(\vec{x}_1)$, $H(\vec{x}_2)$, $H(\vec{x}_3)$ for each of the three avocados in the data set.

e) Compute the mean squared error of this prediction rule on our data set.

f) Write down the *design matrix*, X.

g) Write down the *parameter vector*, $\vec{w}$ that corresponds to this particular choice of prediction rule. The parameter vector should have three components, one for the bias, and one for each of the features.

h) Check that the entries of $X\vec{w}$ are the predicted ripenesses you found above.

i) Write down the *observation vector* $\vec{y}$.

j) Calculate the length of the vector $X\vec{w} - \vec{y}$.

k) What is the relationship between the length of the vector $X\vec{w} - \vec{y}$ and the mean squared error you found above?

l) Is the prediction rule we've used so far, the average of softness and color, the best prediction rule, or can you find a better one? By better, we mean having lesser mean squared error. Is there a single best prediction rule or multiple?