DSC 40A - Homework 03

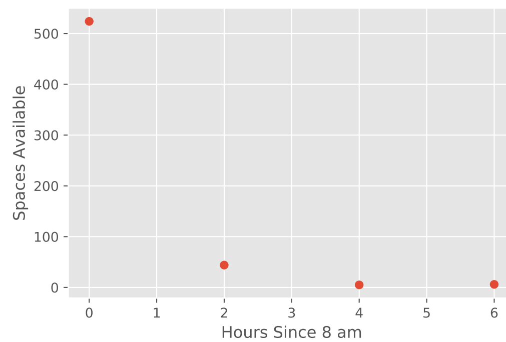
Due: Friday, January 31, 2020

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.

Problem 1.

The table below and the accompanying plot show the total number of "S" parking spaces available on the UCSD campus at various times on a typical Tuesday:[1]

Hours Since 8 am	Spaces Available
0	524
2	44
4	5
6	6



a) Use least squares regression to find a prediction rule of the form $H(x) = w_1 x + w_0$ for the number of spaces available. The variable x represents the number of hours since 8 am. *Hint*: you can check whether your answer is reasonable by plotting.

b) Use your prediction rule from above to predict the number of parking spots available at 9 am. Use it again to predict the number of parking spaces available at 4 pm. Do you believe that your prediction rule makes good predictions? Why or why not?

c) Use least squares regression to find a prediction rule of the form $H(x) = \frac{w_1}{x+1} + w_0$.

d) It looks like the number of parking spaces decreases exponentially as the day goes on. A better prediction rule might be $H_{\text{exp}}(x) = 524 \cdot e^{-wx}$, where w is a parameter that we want to learn from data. Write down the general formula for computing the mean squared error of this prediction rule as a function of w, using x_i for the hours since 8 a.m. and y_i for the number of parking spaces, and n for the number of data points.

Hint: the formula for the mean squared error of a linear prediction rule, $H(x) = w_1 x + w_0$, is: $R(w_1, w_0) = \frac{1}{n} \sum_{i=1}^{n} \left((w_1 x_i + w_0) - y_i\right)^2$.

Bonus (+3 points) Unlike the case of linear prediction rules, there's no formula for the minimizer of R_{exp}. Instead, we have to minimize the mean squared error numerically by using gradient descent or some other method. The Python package `scipy` has a function called `scipy.optimize.minimize` which numerically minimizes a function. Use it to find the value of w which minimizes the mean squared error of the exponential prediction rule given the data in the table above.

Hint: There is a short demo notebook on using SciPy for Numerical Optimization Demo. It will show you how to use `scipy.optimize.minimize`, and you can perform your analysis in the notebook itself.

[1] Parking availability for January 14, 2020 was scraped from the UCSD transportation office website. Lot P701 was excluded.

Problem 2.

In lecture, we derived the least squares solutions for linear prediction rules $H(x) = w_1 x + w_0$. They were:

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

$$w_0 = \bar{y} - b_1 \bar{x}$$

Where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

You may see these solutions written in various equivalent forms. In this problem, we'll derive another form that you may find useful in solving other problems.

a) Show that $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$.

b) Use the result of the previous part to show that

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

is equivalent to the formula for w_1 that was given in lecture.

Problem 3.

A *Boolean feature* is one that is either true or false. For example, when predicting the price of a car, a useful feature might be whether or not the car has an automatic transmission. We can perform least squares regression with Boolean features by "encoding" true and false as numbers: a common choice is to encode true as 1 and false as 0.

In this problem, suppose we have a data set $(x_1, y_1), \ldots, (x_n, y_n)$ of n cars, where the feature x_i is either 1 or 0 (has automatic transmission, or does not) and where y_i is the price of the car. Furthermore, suppose that n_1 of the cars have automatic transmissions, while n_0 do not. Assume for simplicity that the data are sorted so that the first n_0 cars do not have automatic transmissions while the rest do, so that $x_1, \ldots, x_{n_0} = 0$ and $x_{n_0+1}, \ldots, x_n = 1$.

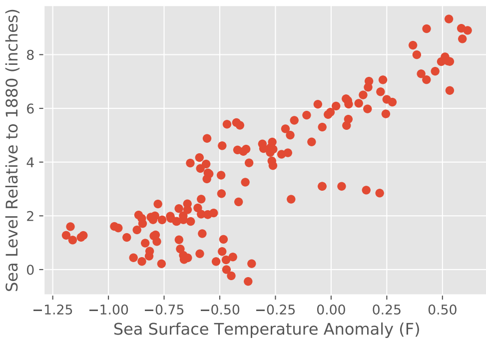
a) Show that $\bar{x} = \frac{n_1}{n}$.

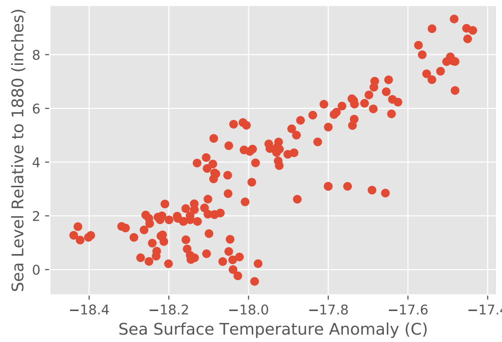
b) Show that $\displaystyle\sum_{i=1}^{n}(x_i - \bar{x})y_i = \frac{n_0}{n}\sum_{i=n_0+1}^{n} y_i - \frac{n_1}{n}\sum_{i=1}^{n_0} y_i$

c) Suppose least squares regression is used to fit a linear prediction rule $H(x) = w_1 x + w_0$ to this data. Show that the prediction $H(0)$ is the mean price of cars without automatic transmissions ($\frac{1}{n_0}\sum_{i=1}^{n_0} y_i$) and the prediction $H(1)$ is the mean price of cars with automatic transmissions ($\frac{1}{n_1}\sum_{i=n_0+1}^{n} y_i$).

Problem 4.

The figures below show the difference in sea level between today and 1880 plotted against the anomaly in sea surface temperature (relative to the 1971-2000 average).[2] Both plots are exactly the same except for the units used to measure temperature; the plot on the left uses Fahrenheit, while the plot on the right uses Celsius.





Suppose a linear prediction $H_1(x) = w_1 x + w_0$ rule is fit to the data on the left using least squares. And suppose that another prediction rule $H_2(x) = b_1 x + b_0$ is fit to the data on the right.

a) Are the slopes of the two lines the same? I.e., is $b_1 = w_1$?

b) Suppose H_1 is used to make a prediction for some temperature anomaly t_F measured in Fahrenheit, and H_2 is used to make a prediction for an anomaly t_C, where t_C is the temperature t_F, but converted to Celsius. Are the two predictions the same? Prove your answer.

Problem 5.

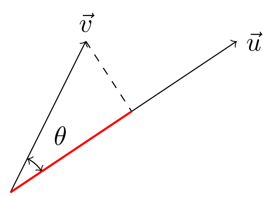
Looking Ahead. We will soon need to remember the key properties of the dot product. This question is meant to help you remember them.

Recall from your class on vector algebra that one way to define the dot product of two vectors, $\vec{u}$ and $\vec{v}$, is:

$$\vec{u} \cdot \vec{v} = \|\vec{u}\| \|\vec{v}\| \cos\theta,$$

where $\|\vec{u}\|$ is the length of the vector $\vec{u}$, $\|\vec{v}\|$ is the length of $\vec{v}$, and θ is the angle between the two vectors.

Two vectors $\vec{u}$ and $\vec{v}$ are shown below.



Argue that the length of the red segment is $(\vec{u} \cdot \vec{v})/\|\vec{u}\|$.

[2] The data were scraped from the EPA website.