DSC 40A - Extra Practice for Final Part 1

Spring 2023

Problem 1. Sloped Mean

Suppose you have a data set $y_1, y_2, \ldots, y_n$ with at least three values, $n \geq 3$, and the values are arranged such that $y_1 \leq y_2 \leq \cdots \leq y_n$.

We know from class that the mean of the data minimizes mean squared error,

$$R_{sq}(h) = \sum_{i=1}^{n}(h - y_i)^2.$$

Define a new function that weights larger data points less heavily:

$$S(h) = \left(\sum_{i=1}^{n-2}(h - y_i)^2\right) + 0.5 \cdot (h - y_{n-1})^2 + 0.1 \cdot (h - y_n)^2.$$

a) What value of h minimizes $S(h)$? We'll call the value of h that minimizes $S(h)$ the **sloped mean**, since the coefficients of the data values decrease for larger data.

Solution:

$$h = \frac{\left(\sum_{i=1}^{n-2} y_i\right) + 0.5 \cdot y_{n-1} + 0.1 \cdot y_n}{n - 1.4}$$

We can prove this using calculus.

$$\begin{aligned}
S'(h) &= \left(\sum_{i=1}^{n-2} 2 \cdot (h - y_i)\right) + 2 \cdot 0.5 \cdot (h - y_{n-1}) + 2 \cdot 0.1 \cdot (h - y_n) \\
0 &= 2 \cdot \left(\left(\sum_{i=1}^{n-2}(h - y_i)\right) + 0.5 \cdot (h - y_{n-1}) + 0.1 \cdot (h - y_n)\right) \\
0 &= \left(\sum_{i=1}^{n-2}(h - y_i)\right) + 0.5 \cdot (h - y_{n-1}) + 0.1 \cdot (h - y_n) \\
0 &= (n-2) \cdot h - \left(\sum_{i=1}^{n-2} y_i\right) + 0.5 \cdot h - 0.5 \cdot y_{n-1} + 0.1 \cdot h - 0.1 \cdot y_n \\
(n-2) \cdot h + 0.5 \cdot h + 0.1 \cdot h &= \left(\sum_{i=1}^{n-2} y_i\right) + 0.5 \cdot y_{n-1} + 0.1 \cdot y_n \\
h \cdot (n - 1.4) &= \left(\sum_{i=1}^{n-2} y_i\right) + 0.5 \cdot y_{n-1} + 0.1 \cdot y_n \\
h &= \frac{\left(\sum_{i=1}^{n-2} y_i\right) + 0.5 \cdot y_{n-1} + 0.1 \cdot y_n}{n - 1.4}
\end{aligned}$$

We know that this critical point corresponds to a minimum because $S(h)$ is a quadratic with a positive leading coefficient, so it's an upward-facing parabola.

b) Which do you think is a better hypothesis, the mean or the sloped mean? Is your answer always the same, or does it depend on some property of the data set? Give an example of when you might prefer to use the sloped mean, and when you might prefer the (regular) mean.

Solution: Our answer depends on the shape and size of our dataset and whether we think extreme values are important. For example, for very small data sets, say data sets of size $n = 3$, it's almost always better to use the mean over the sloped mean since the sloped mean weights the smallest data point as way more important than the other two, meaning your prediction is based primarily on a single value.

The sloped mean might be a better hypothesis if the maximum value for the dataset is a large outlier. In this case, by using the sloped mean, we reduce the effect of this outlier in our predictions.

For other data sets, however, we may not want to devalue outliers. If there are a lot of outliers, rather than just one or two, or if outliers exist in our data set for good reason, then we should probably still incorporate them in the determination of our hypothesis as we would any other value. For example, if our data set describes the daily rainfall in San Diego, in inches, we would expect that most of the values in our data set would be zero, but it would not be a good idea to ignore the highest value in the data set, since that day contributed a lot to the annual rainfall, and ignoring it when calculating the average daily rainfall would paint a picture of even less rainfall than in reality.

Problem 2. Which is bigger? By how much?

Given a data set $y_1 \leq y_2 \leq \cdots \leq y_n$, define the following empirical risk functions:

$$R_{\text{abs}}(h) = \frac{1}{n}\sum_{i=1}^{n} |y_i - h| \qquad\qquad R_{\text{sq}}(h) = \frac{1}{n}\sum_{i=1}^{n}(y_i - h)^2$$

Parts (a), (b), and (c) below concern R_{abs}. Parts (d) and (e) concern R_{sq}.

a) For an arbitrary c with $c < c + 1 < y_1$, how does $R_{\text{abs}}(c)$ compare to $R_{\text{abs}}(c + 1)$? Can you determine which is bigger, and by how much?

Solution: $R_{\text{abs}}(c) > R_{\text{abs}}(c + 1)$.

Given $y_1 \leq y_2 \leq \cdots \leq y_n$, and $c < c + 1 < y_1$, we can say c and $c + 1$ are smaller than every point in the data set. Therefore, $|y_i - c| = y_i - c$ and $|y_i - (c+1)| = y_i - (c+1)$.

$$\begin{aligned}
R_{\text{abs}}(c) - R_{\text{abs}}(c+1) &= \frac{1}{n}\sum_{i=1}^{n}|y_i - c| - \frac{1}{n}\sum_{i=1}^{n}|y_i - (c+1)| \\
&= \frac{1}{n}\left(\sum_{i=1}^{n} y_i - c - \sum_{i=1}^{n} y_i - (c+1)\right) \\
&= \frac{1}{n}\left(\sum_{i=1}^{n}(y_i - c) - (y_i - (c+1))\right) \\
&= \frac{1}{n}\left(\sum_{i=1}^{n} 1\right) \\
&= \frac{1}{n}\cdot n \\
&= 1
\end{aligned}$$

This says that $R_{\text{abs}}(c)$ is exactly 1 larger than $R_{\text{abs}}(c+1)$.

Another way to do this problem is to notice that when we start with a c that is sufficiently to the left of all data points and move it to the right by one unit, its distance from each data point decreases by one unit. So the total distance to all data points decreases by n units, or the average distance to all data points decreases by 1 unit.

b) For an arbitrary c with $y_n < c < c+2$, how does $R_{\text{abs}}(c)$ compare to $R_{\text{abs}}(c+2)$? Can you determine which is bigger, and by how much?

Solution: $R_{\text{abs}}(c) < R_{\text{abs}}(c+2)$.

Given $y_1 \leq y_2 \leq \cdots \leq y_n$, and $y_n < c < c+2$, we can say c and $c+2$ are larger than every point in the data set. Therefore, $|y_i - c| = c - y_i$ and $|y_i - (c+2)| = c + 2 - y_i$.

$$\begin{aligned}
R_{\text{abs}}(c+2) - R_{\text{abs}}(c) &= \frac{1}{n}\sum_{i=1}^{n}|y_i - (c+2)| - \frac{1}{n}\sum_{i=1}^{n}|y_i - c| \\
&= \frac{1}{n}\left(\sum_{i=1}^{n} c + 2 - y_i - \sum_{i=1}^{n} c - y_i\right) \\
&= \frac{1}{n}\left(\sum_{i=1}^{n}(c + 2 - y_i) - (c - y_i)\right) \\
&= \frac{1}{n}\left(\sum_{i=1}^{n} 2\right) \\
&= \frac{1}{n}\cdot 2n \\
&= 2
\end{aligned}$$

This says that $R_{\text{abs}}(c+1)$ is exactly 2 larger than $R_{\text{abs}}(c)$.

Another way to do this problem is to notice that when we start with a c that is greater than all data points and move it over to the right by two units, its distance from each data point increases by two units. So the total distance to all data points increases by $2n$ units, or the average distance to all data points increases by 2 units.

c) Suppose $n = 10$. For an arbitrary c with $y_3 < c < c+1 < y_4$, how does $R_{\text{abs}}(c)$ compare to $R_{\text{abs}}(c+1)$? Can you determine which is bigger, and by how much?

Solution: $R_{\text{abs}}(c) > R_{\text{abs}}(c+1)$.

For any choice of h between y_3 and y_4, there are 3 data points that are smaller than h and 7 data points that are larger than h. Therefore, using the formula for the slope of $R_{\text{abs}}(h)$ that we derived in class,

$$\frac{1}{n}\left(\#y_i < h - \#y_i > h\right),$$

the slope of $R_{\text{abs}}(h)$ at any h between y_3 and y_4 is $\frac{1}{10}(3-7) = -0.4$. Since $y_3 < c < c+1 < y_4$, the slope of -0.4 for this portion of the graph means that $R_{\text{abs}}(c)$ is exactly 0.4 units larger than $R_{\text{abs}}(c+1)$.

d) For an arbitrary c with $c < y_1$, how does $R_{\text{sq}}(c)$ compare to $R_{\text{sq}}(c-1)$? Can you determine which is bigger, and by how much?

Solution: $R_{\text{sq}}(c-1) > R_{\text{sq}}(c)$.

Notice that $c-1$ and c are both smaller than y_1 so they are smaller than the mean of $y_1, y_2, \ldots, y_n$. This means both c and $c+1$ are to the left of the vertex on an upward-facing parabola, which means the graph is decreasing here. Therefore, $R_{\text{sq}}(c-1) > R_{\text{sq}}(c)$. As before, we cannot determine how much they differ by because it depends on the data set and the value of c.

e) For an arbitrary c with $c > y_n$, how does $R_{\text{sq}}(c)$ compare to $R_{\text{sq}}(c+1)$? Can you determine which is bigger, and by how much?

Solution: $R_{\text{sq}}(c) < R_{\text{sq}}(c+1)$.

Notice that c and $c+1$ are both larger than y_n so they are larger than the mean of $y_1, y_2, \ldots, y_n$. This means both c and $c+1$ are to the right of the vertex on an upward-facing parabola, which means the graph is increasing here. Therefore, $R_{\text{sq}}(c) < R_{\text{sq}}(c+1)$. We cannot determine how much $R_{\text{sq}}(c)$ is less than $R_{\text{sq}}(c+1)$ because it depends on the data set and the value of c. In other words, if we look at pairs of points on a parabola that are equally spaced horizontally, their vertical spacing will not all be the same. It depends on where on the parabola they fall.

Problem 3. Matrix, Vector, Scalar, or Nonsense?

Suppose M is an $m \times n$ matrix, v is a vector in $\mathbb{R}^n$, and s is a scalar. Determine whether each of the following quantities is a matrix, vector, scalar, or nonsense (undefined).

a) Mv

Solution: This is a vector in $\mathbb{R}^m$.

b) vM

Solution: This is nonsense. v has dimensions $n \times 1$ and M has dimensions $m \times n$. Since the inner dimensions don't match, they can't be multiplied like this.

c) v^2

Solution: This is nonsense. We can't square a vector because the dimensions don't work out to multiply a vector by itself. The closest thing we have to this is taking the dot product of a vector with itself, which squares each component and sums them.

d) $M^T M$

Solution: This is a matrix, with dimensions $n \times n$.

e) MM^T

Solution: This is a matrix, with dimensions $m \times m$.

f) $v^T Mv$

Solution: This is nonsense because v^T has dimensions $1 \times n$ and M has dimensions $m \times n$ so the inner dimensions don't match.

g) $(sMv) \cdot (sMv)$

Solution: This is a scalar. Mv is a vector, as we saw in part (a), so sMv is also a vector. This is the dot product of a vector with itself, which is a scalar.

h) $(sv^T M^T)^T$

Solution: This is a vector in $(R)^m$. v^T has dimensions $1 \times n$ and M^T has dimensions $n \times m$ so $sv^T M^T$ has dimensions $1 \times m$ and its transpose has dimensions $m \times 1$.

Another way to think about this is to use the property of transposes that says the transpose of a product is the product of the transposes in the opposite order. Since a scalar is its own transpose, $(sv^T M^T)^T = sMv$. We know Mv is a vector in $\mathbb{R}^m$ from part (a), so a scalar multiple of it is also a vector in $\mathbb{R}^m$.

i) $v^T M^T Mv$

Solution: This is a scalar, which we can figure out by looking at the dimensions of each of the components from left to right: $\times n$, $n \times m$, $m \times n$, and $n \times 1$.

Alternatively, we can rewrite $v^T M^T = (Mv)^T$, so that $v^T M^T Mv = (Mv)^T (Mv)$, which is the dot product of the vector Mv with itself. Therefore, the result is a scalar.

j) $vv^T + M^T M$

Solution: This is a matrix, with dimensions $n \times n$. Since v has dimensions $n \times 1$ and v^T has dimensions $1 \times n$, the product vv^T has dimensions $n \times n$. We already saw in part (d) that $M^T M$ is a matrix with dimensions $n \times n$, so we can add two matrices with the same dimensions.

Problem 4. Orthogonality

a) Is it possible for a vector to be orthogonal to itself?

Solution: It is possible, but only if the vector is the zero vector. For a vector $\vec{v} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}$ to be orthogonal to itself, that would require $\vec{v} \cdot \vec{v} = \sum_{i=1}^{n} v_i^2 = 0$. Since squares are non-negative, the only way for this sum to total 0 is for each term of the sum to be 0, which means $\vec{v}$ is the zero vector.

b) Show that if $\vec{u}$ is orthogonal to both $\vec{v}$ and $\vec{w}$, then $\vec{u}$ is also orthogonal to any linear combination of $\vec{v}$ and $\vec{w}$, $\alpha\vec{v} + \beta\vec{w}$.

Solution: Suppose $\vec{u}$ is orthogonal to both $\vec{v}$ and $\vec{w}$. By the definition of orthogonality, this means $\vec{u}^T\vec{v} = 0$ and $\vec{u}^T\vec{w} = 0$. Then,

$$\begin{aligned} \vec{u}^T(\alpha\vec{v} + \beta\vec{w}) &= \vec{u}^T\alpha\vec{v} + \vec{u}^T\beta\vec{w}) \\ &= \alpha\vec{u}^T\vec{v} + \beta\vec{u}^T\vec{w}) \\ &= \alpha * 0 + \beta * 0 \\ &= 0, \end{aligned}$$

which means $\vec{u}$ is orthogonal to any linear combination of $\vec{v}$ and $\vec{w}$.

c) Show that if $A^T\vec{b} = 0$, then $\vec{b}$ is orthogonal to the **column space** of A, which is the space of all linear combinations of the columns of A.

Solution: Notice that the columns of A are the rows of A^T. Since $A^T\vec{b} = 0$, this means the dot product of $\vec{b}$ with each row of A^T equals 0, as matrix-vector multiplication involves taking the dot product of the vector with each row of the matrix. This says $\vec{b}$ is orthogonal to each row of A^T, or equivalently, to each column of A. Using the result of the previous part, if $\vec{b}$ is orthogonal to the columns of A, then it is also orthogonal to linear combinations of the columns of A.

Problem 5. Regression

Suppose you have a dataset

$$\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$$

where the standard deviation of the x-values, $SD(x)$, is twice the standard deviation of the y-values, $SD(y)$. Let

$$y = a + bx$$

be the regression line with x as the predictor variable and y as the response variable. Let

$$x = c + dy$$

be the regression line with y as the predictor variable and x as the response variable. **Express b in terms of d.**

Solution: Using the formula for the slope of the regression line we have

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

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

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

Interchanging the roles of x and y gives us a formula for d as well:

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

Since the expression for b and the expression for d have the same numerator, we can solve each expression for that numerator and set them equal to get a relationship between b and d, which we can then solve to get b in terms of d as desired.

$$b * n * (SD(x))^2 = d * n * (SD(y))^2$$
$$b = d * \frac{(SD(y))^2}{(SD(x))^2}$$

Since we are told that $SD(x) = 2 * SD(y)$, this means $(SD(x))^2 = 4 * (SD(y))^2$ and so the relationship simplifies to

$$b = \frac{d}{4}.$$

Problem 6. Farmfluencer

Billy the avocado farmer heard about the success of 72 year-old Gerald Stratford's viral gardening videos on Twitter and Instagram. After witnessing Gerald turn into the so-called King of Big Veg overnight, Billy is feeling inspired to up his social media game (he's also feeling a little bit jealous).

Billy is new to Instagram and is trying to understand how people gain followers. In particular, he wants to be able to predict the number of followers, y, based on these features:

- number of people they follow, $x^{(1)}$
- number of years since first post, $x^{(2)}$
- average number of posts per day, $x^{(3)}$

a) Suppose Billy has access to a large data set of Instagram accounts, and he uses multiple regression on this data to fit a linear prediction rule of the form

$$H(\vec{x}) = w_0 + w_1 x^{(1)} + w_2 x^{(2)} + w_3 x^{(3)}.$$

What does w_2 represent in terms of Instagram followers?

Solution: The coefficient w_2 is like the slope when $x^{(2)}$ is our variable. It represents the change in y for each unit of change in $x^{(2)}$. That is, it represents a typical increase in Instagram followers for each year of having an Instagram account, or how many followers someone might gain in a year.

b) What if instead of the number of years since the first post, $x^{(2)}$, Billy instead uses the number of days since the first post, $x^{(4)}$. Now he uses multiple regression to fit a prediction rule of the form

$$H'(\vec{x}) = w'_0 + w'_1 x^{(1)} + w'_3 x^{(3)} + w'_4 x^{(4)}.$$

How do the parameters of this prediction rule (w'_0, w'_1, w'_3, w'_4) compare to the parameters of original prediction rule (w_0, w_1, w_2, w_3)?

Solution: By the same logic as above, w'_4 represents a typical increase in followers for each additional day of having an account. Since there are 365 days in a year, we should expect that $w'_4 * 365 = w_2$. Since the other variables are unchanged, we'd have $w'_0 = w_0, w'_1 = w_1, w'_3 = w_3$.

Problem 7. Changing the Prediction Rule

Suppose we have a dataset consisting of variables $x^{(1)}, x^{(2)}$, and y. We use multiple regression to fit a prediction rule of the form

$$H(x^{(1)}, x^{(2)}) = w_0 + w_1(x^{(1)} + x^{(2)}) + w_2 x^{(1)} x^{(2)} + w_3(x^{(1)} + 1)(x^{(2)} + 1) \tag{1}$$

and then again use multiple regression to fit a different prediction rule of the form

$$H'(x^{(1)}, x^{(2)}) = w'_0 + w'_1 x^{(1)} + w'_2 x^{(2)} + w'_3 x^{(1)} x^{(2)}. \tag{2}$$

Which form, (1) or (2), will yield a prediction rule with lower mean squared error? Justify your answer.

Solution: The prediction rule of form (2) will have a lower MSE. We can see this by rewriting form (1) as

$$\begin{aligned} H(x^{(1)}, x^{(2)}) &= w_0 + w_1(x^{(1)} + x^{(2)}) + w_2 x^{(1)} x^{(2)} + w_3(x^{(1)} + 1)(x^{(2)} + 1) \\ &= w_0 + w_1 x^{(1)} + w_1 x^{(2)} + w_2 x^{(1)} x^{(2)} + w_3 x^{(1)} x^{(2)} + w_3 x^{(1)} + w_3 x^{(2)} + w_3 \\ &= (w_0 + w_3) + (w_1 + w_3) x^{(1)} + (w_1 + w_3) x^{(2)} + (w_2 + w_3) x^{(1)} x^{(2)} \end{aligned}$$

If we define new variables $c_0 = w_0 + w_3$, $c_1 = w_1 + w_3$, $c_2 = w_2 + w_3$, we can write this as

$$= c_0 + c_1 x^{(1)} + c_1 x^{(2)} + c_2 x^{(1)} x^{(2)}$$

This is a very similar form to (2), except the weights of $x^{(1)}$ and $x^{(2)}$ are required to be the same, which constrains the solution. In other words, with a prediction rule of form (2), multiple regression picks the best parameters w'_0, w'_1, w'_2, w'_3 to minimize the mean squared error. The weights w'_1 and w'_2 which are the coefficients of $x^{(1)}$ and $x^{(2)}$ may not turn out to be the same. This says that the MSE is lower when we allow for these coefficients to be different instead of forcing them to be the same.