Lab Solutions 02

1. To start, download and open the class02.Rmd file in RStudio. Follow the script until you get to the section that asks you to return to these notes.

I didn't include answers to the Rmarkdown file questions today because I thought they were straightforward. The point was more to get used to using R for computations. If you have any questions, just let me know!

2. Last time we started with the basic idea of statistical learning. We observe pairs (x_i, y_i) and want to construct a function $\widehat{f}(x)$ from this training data that does a good job of predicting future values of y_i given new values of x_i. One of the simplest such models for predicting a continuous response y is simple linear regression. Visually this corresponds to fitting a linear function f to the data such that:

$$\widehat{f}(x_i) = a + b \cdot x_i. \tag{2.1}$$

Where the parameters a (the intercept) and b (the intercept) are *learned* from the data. Write down, symbolically, what the mean squared loss function is of using the above f to predict the values y_i.

The squared loss function is given by:

$$\mathcal{L}(a, b) = \sum_i \left(y_i - (a + b \cdot x_i)\right)^2. \tag{2.2}$$

3. We are going to simplify things further by removing the intercept term a from the model and assuming that we have only:

$$\widehat{f}(x_i) = b \cdot x_i.$$

Taking the equation you had from the previous question, write down the loss function for the new value of $\widehat{f}$. Take the derivative with respect to b and set it equal to zero. Can you find a formula for b that minimizes the loss function?

The new loss function is given by:

$$\mathcal{L}(b) = \sum_i \left(y_i - b \cdot x_i\right)^2.$$

The derivative with respect to b is given by:

$$\begin{aligned}
\frac{d}{db}\mathcal{L}(b) &= \frac{d}{db}\sum_i (y_i - b \cdot x_i)^2 \\
&= \sum_i \frac{d}{db}(y_i - b \cdot x_i)^2 \\
&= \sum_i 2 \cdot (y_i - b \cdot x_i) \cdot \frac{d}{db}(y_i - b \cdot x_i) \\
&= \sum_i 2 \cdot (y_i - b \cdot x_i) \cdot (-x_i) \\
&= \sum_i 2 \cdot \left(b \cdot x_i^2 - y_i x_i\right).
\end{aligned}$$

Here I used the chain rule, but you can also expand the quadratic term and take the derivative of each term directly.

Setting the loss equal to zero we see:

$$\begin{aligned}
\sum_i 2 \cdot \left(\widehat{b} \cdot x_i^2 - y_i x_i\right) = 0 \\
\sum_i \widehat{b} \cdot x_i^2 &= \sum_i y_i x_i \\
\widehat{b} \times \sum_i x_i^2 &= \sum_i y_i x_i \\
\widehat{b} &= \frac{\sum_i y_i x_i}{\sum_i x_i^2}
\end{aligned}$$

This gives us an explicit way of going from the data (x_i, y_i) to an estimate of the slope parameter in our model.

4. Taking the second derivative of the loss function, prove that you found a global minimizer in the previous question rather than a saddle point or maximum.

Taking the second derivative of the loss function yields:

$$\begin{aligned}
\frac{d^2}{db^2}\mathcal{L}(b) &= \frac{d}{db}\sum_i 2 \cdot \left(b \cdot x_i^2 - y_i x_i\right) \\
&= 2 \cdot \sum_i x_i^2.
\end{aligned}$$

Unless every data point $\{x_i\}_i$ is equal to zero, the sum $\sum_i x_i^2$ will be positive and therefore the second derivative will be positive. The second derivative test then tells us that the value of $\widehat{b}$ is a local minimum. Since this is a function with a continuous first derivative and only one local minimum it must be a global minimum.

5. We typically write the learned parameters in a model with a 'hat'. So the slope you computed above becomes $\widehat{b}$. Can you re-write $\widehat{b}$ such that the estimator is written a weighted sum of the values y_i?

This equation just requires being comfortable with the summation notation. I will go through this slowly is at seemed to cause some trouble. Start by noticing that we can change the index variable used in a summation because it is a dummy variable:

$$\sum_i x_i^2 = \sum_j x_j^2.$$

Now, with a different index, we can put the denominator *inside* the other summation sign:

$$\begin{aligned}\widehat{b} &= \frac{\sum_i y_i x_i}{\sum_j x_j^2} \\ &= \sum_i \left(y_i \cdot \frac{x_i}{\sum_j x_j^2} \right).\end{aligned}$$

Defining weights given by:

$$w_i = \frac{x_i}{\sum_j x_j^2}$$

We can then write:

$$\widehat{b} = \sum_i y_i \cdot w_i.$$

While we won't be able to get into a lot of the details for a lack of probability theory, the fact that $\widehat{b}$ is a linear combination of the y_i's is an important theoretical property.

6. So far, we have made no assumptions about the 'true' nature of the relationship between x and y. Assume that we can write:

$$y_i = b \cdot x_i + \epsilon_i \tag{2.3}$$

For some term ϵ_i known as the *error term*. Plugging this into your equation for $\widehat{b}$,

can you argue that $\widehat{b}$ will be close to b if the error terms are small?

Plugging this value into the equation for $\widehat{b}$, we have:

$$\begin{aligned}
\widehat{b} &= \frac{\sum_i y_i x_i}{\sum_i x_i^2} \\
&= \frac{\sum_i (b \cdot x_i + \epsilon_i) \cdot x_i}{\sum_i x_i^2} \\
&= \frac{\sum_i b \cdot x_i^2}{\sum_i x_i^2} + \frac{\sum_i \epsilon_i \cdot x_i}{\sum_i x_i^2} \\
&= b \cdot \frac{\sum_i x_i^2}{\sum_i x_i^2} + \frac{\sum_i \epsilon_i \cdot x_i}{\sum_i x_i^2} \\
&= b + \sum_i \left(\epsilon_i \cdot \frac{x_i}{\sum_j x_j^2} \right).
\end{aligned}$$

So $\widehat{b}$ is equal to the 'true' slope b plus some weighted sum of the errors. If the errors are small, we would expect that $\widehat{b}$ is therefore close to b.

7. Return to the R code to complete today's lab.

Again, please ask if you have any questions with the lab for today. I will supply solutions when the questions in the R code are more involved.