UGBA 198 Homework 2

Due: February 27, 2018 at 6 PM on Gradescope

1 Justifying OLS with Maximum Likelihood Estimation

We started this course with an exploration of ordinary least squares. In the univariate case, we're solving for the value of w that minimizes our error function. Specifically:

$$w^* = arg\,min_w \sum_{i=1}^{n} (x_i * w - y_i)^2 \tag{1}$$

But where did this error function come from? That is, how do we know that this is a good error function to optimize? In this problem, we'll find that this ordinary least squares optimization comes directly from maximum likelihood estimation, when assuming unit normal noise on our data.

Let's start by statistically modeling the scenario. We want to model our measurements as having unit normal noise:

$$y_i = x_i * w + z_i \tag{2}$$

where:

$$z_i \sim N(0, 1)$$

(As a reminder, this means that $E[z_i] = 0$, and $\sigma_{z_i} = 1$.)

1. Notice that y_i is the output of a random variable, since it is a function of a random variable (z). In other words:

   $$Y_i = x_i * w + z_i$$

   What kind of distribution is Y_i? What is the mean of Y_i (i.e., $E[Y_i]$), and what is its standard deviation? (Hint: Remember that

$E[A+B] = E[A] + E[B]$, and that $E[k] = k$ for a constant k. $x_i * w$ is a constant.)

$Y_i \sim N(x_i * w, 1)$, since $E[x_i * w + z_i] = E[x_i * w] + E[z_i] = x_i * w$, and $Var(x_i * w + z) = Var(x_i * w) + Var(z) = Var(z) = 1$, since $x_i * w$ is a constant.

2. Recall the pdf of a normal distribution:

$$f(k|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(k-\mu)^2}{2\sigma^2}} \tag{3}$$

What is the probability density at the value that Y_i took on for a single data point? (i.e., **what is the pdf value for Y_i, for one value of y_i?**)

Hint: Look at your answer for part 1. What kind of distribution is Y_i? How can you apply the pdf above?

Hint: $k = y_i$.

$\frac{1}{\sqrt{2\pi}} e^{-\frac{(y_i - x_i * w)^2}{2}}$, by plugging $\sigma = 1, k = y_i, \mu = x_i * w$ into the PDF above.

3. Let's assume that the noise for each data point that we observe is independent (probably a bad assumption, which we might explore later in the course). In other words:

$$f(y_i, y_j) = f(y_i) * f(y_j) \tag{4}$$

for all i, j.

What is the probability density function for *all* of the values that the y_i took on? (i.e., what is the joint probability density? Remember that the realizations are independent.) Specifically, **we're looking for an expression that represents** $f(y_1, ..., y_n)$. It will be the product of multiple pdf values.

$\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{(y_i - x_i * w)^2}{2}}$, since we're assuming that each value of y_i is conditionally independent. (We can just multiply the PDF values.)

4. Consider getting the value of w that maximizes the probability density that you've solved for in part 3:

$$w^* = arg\,max_w f(y_1, ..., y_n)$$

As we described in class, we can take the logarithm of this expression, and we're still solving for the same maximum. **Is this the same as the least squares objective? Why?**

$$arg\,max_w \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{(y_i - x_i * w)^2}{2}}$$

$$= arg\,max_w log(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{(y_i - x_i * w)^2}{2}})$$

$$= arg\,max_w \sum_{i=1}^{n} log(\frac{1}{\sqrt{2\pi}} e^{-\frac{(y_i - x_i * w)^2}{2}})$$

$$= arg\,max_w \sum_{i=1}^{n} log(e^{-\frac{(y_i - x_i * w)^2}{2}})$$

$$= arg\,max_w \sum_{i=1}^{n} -(y_i - x_i * w)^2$$

$$= arg\,min_w \sum_{i=1}^{n} (y_i - x_i * w)^2$$

which is the same as the least squares objective.