Problem 3 (Deriving Linear Regression, 10pts)

In class, we noted that the solution for the least squares linear regressions "looked" like a ratio of covariance and variance terms. In this problem, we will make that connection more explicit.

Let us assume that our data are tuples of scalars (x, y) that come from some distribution $p(x, y)$. We will consider the process of fitting these data with the best linear model possible, that is a linear model of the form $\hat{y} = wx$ that minimizes the expected squared loss $E_{x,y}[(y - \hat{y})^2]$.

Notes: The notation $E_{x,y}$ indicates an expectation taken over the joint distribution $p(x, y)$. Since x and y are scalars, w is also a scalar.

1. Derive an expression for the optimal w, that is, the w that minimizes the expected squared loss above. You should leave your answer in terms of moments of the data, e.g. terms like $E_x[x]$, $E_x[x^2]$, $E_y[y]$, $E_y[y^2]$, $E_{x,y}[xy]$ etc.

2. Provide unbiased and consistent formulas to estimate $E_{x,y}[yx]$ and $E_x[x^2]$ given observed data $\{(x_n, y_n)\}_{n=1}^N$.

3. In general, moment terms like $E_{x,y}[yx]$, $E_{x,y}[x^2]$, etc. can easily be estimated from the data (like you did above). If you substitute in these empirical moments, how does your expression for the optimal w^* in this problem compare with the optimal w^* that we derived in class/Section 2.6 of the cs181-textbook?

4. As discussed in lecture, many common probabilistic linear regression models assume that variables x and y are jointly Gaussian. Did any of your above derivations rely on the assumption that x and y are jointly Gaussian? Why or why not?

Solution

1. The expected squared loss is given as:
$$E_{x,y}[(y - \hat{y})^2]$$
The linear model has the form $\hat{y} = wx$, which can be substituted into the expected squared loss to be:
$$E_{x,y}[(y - wx)^2]$$
Expanding:
$$E_{x,y}[y^2 - 2wxy + w^2x^2]$$
By linearity of expectation we have:
$$E_y[y^2] - 2wE_{x,y}[xy] + w^2E_x[x^2]$$
To find the optimal w^*, we take the gradient of the above expression with respect to w and set it equal to 0 (that is solving the First Order Condition).
$$\frac{d}{dw}\left[E_y[y^2] - 2wE_{x,y}[xy] + w^2E_x[x^2]\right] = 0$$
$$-2E_{x,y}[xy] + 2w^*E_x[x^2] = 0$$
$$w^* = \frac{E_{x,y}[xy]}{E_x[x^2]}$$

2. An unbiased and consistent estimator for both of these quantities are their sample means:
$$E_{x,y}[yx] = \frac{1}{N}\sum_{n=1}^{N} y_n x_n$$
$$E_x[x^2] = \frac{1}{N}\sum_{n=1}^{N} x_n^2$$

3. Substituting the empirical moments above into our expression for w^* from 3.1, we have:
$$w^* = \frac{\frac{1}{N}\sum_{n=1}^{N} y_n x_n}{\frac{1}{N}\sum_{n=1}^{N} x_n^2}$$
$$w^* = \frac{\sum_{n=1}^{N} y_n x_n}{\sum_{n=1}^{N} x_n^2}$$
The optimal w^* derived in class/Section 2.6 of the cs181-textbook was:
$$w^* = (X^TX)^{-1}X^TY = \frac{X^TY}{X^TX}$$
Observe that both expressions for the optimal w^* are actually mathematically equivalent. The only difference is that the optimal w^* (derived in class/Section 2.6 of the cs181-textbook) is an analytical solution for data with higher order dimensionality (that is data with more features). In comparison, the optimal w^* we found in this problem is an analytical solution only for data that is 1-dimensional, that is scalar data like the one presented in this problem.
Assuming that X in the equation above are column vectors (Nx1 dimensions, where N is the number of observations such that each observations only has one feature), then we can directly relate the numerator and denominator of our two expressions for w^*:
If X is $Nx1$ (where there are N observations in the data), then: $\sum_{n=1}^{N} y_n x_n = X^TY$ and $\sum_{n=1}^{N} x_n^2 = X^TX$

4. None of the above derivations relied on making the assumption that x and y are jointly Gaussian. In fact, we make no assumptions about the distribution $p(x, y)$.
   In problem 3.1, to achieve the derivation we utilized laws of algebra, the law of linearity of expectation, and the rules of calculus to solve the First Order Condition. None of the derivation in problem 3.1 made any assumptions about the underlying distribution $p(x, y)$.
   Deriving the empirical moments in problem 3.2 only made use of the properties of expectation and sample means. The unbiasedness and consistency of the sample mean in estimating the expectation (empirical mean estimating theoretical mean) is simply a property of how expectation and the sample mean, and therefore we make no assumptions about the distribution $p(x, y)$. In particular, the sample mean is unbiased because the expectation of the sample mean is the theoretical population mean. Moreover, as the size of our data grows infinitely, the sample mean tends towards the true population mean (i.e. consistency). We can imagine that for if we had infinite data, we could directly find data for the entire population and therefore directly calculate the population mean using the sample mean formula.