General Machine Learning Questions

Denis Ovchinnikov

5/8/2021

Problem 1. *Consider a linear model $Y = X\beta + \epsilon$ where Y is an $N \times 1$ vector, X is an $N \times K$ matrix, β is a $K \times 1$ vector of coefficients and ϵ is an $N \times 1$ vector.*

1. *What assumptions are required to estimate the coefficients vector β with Ordinary Least Squares (OLS)? Explain.*

   You don't need any assumptions to just estimate β using OLS. However, you do need assumptions if you want to guarantee that this estimate is good.

   Usually the following assumptions are made to ensure that OLS gives the best (smallest variance) unbiased estimate:

   (a) As mentioned above, true distribution can be modeled by a linear model with additive error: $Y = \beta_0 + \sum X_j\beta_j + \epsilon = X\beta + \epsilon$. This is quite essential, and if your data doesn't follow this pattern, you should consider using different techniques. You can sometimes "add" new features to make regression "not linear", e.g. adding X_iX_j for all i, j as new features will be a way to model Y's that are quadratic in X's.

   (b) $E(\epsilon \mid X) = 0$ for all X (i.e. error is "symmetric" around 0).

   (c) $Var(\epsilon \mid X) = \sigma^2 I$, where σ is a constant. That means that variation of the error doesn't change between observations and there is no autocorrelation between errors of different observations.

   (d) It is also common to make some assumptions on X:

      i. $N > K$ (i.e. number of samlpes is greater than the number of features). This avoids having models that fit the set perfectly and will fail to generalize.

      ii. Points in X are drawn at random from a population (or fixed). That means you can't just sample convenient points (e.g. if I live in NYC and want to study US population, I can not only consider NYC residents for my dataset)

      iii. There is no perfect collinearity between different features. In practice, even if there is no perfect collinearity, but there is a strong correlations between features, some modifications of OLS are recommended (e.g. Ridge, Lasso, etc.). This usually reduces variance and even though the estimate is no longer unbiased, the estimated error is smaller).

Often for convenience, instead of items (b),(c) , one assumes that $\epsilon \sim N(0, \sigma^2)$ is a Gaussian random variable independent on X. This allows to get strong precise formulas for various parameters of the model (e.g. confidence intervals of $\hat{\beta}$). In this case OLS is also the Maximum Likelihood model.

2. *What assumptions are required to provide an asymptotically valid confidence interval for β? What about a confidence interval which is valid for any value of N?*

   The most common assumption is that $\epsilon \sim N(0, \sigma^2)$ (for some unknown but fixed σ^2). Then we can construct a valid confidence interval for β_j: $(\hat{\beta}_j - z^{(1-\alpha)}\sqrt{\nu_j}\hat{\sigma}, \hat{\beta}_j + z^{(1-\alpha)}\sqrt{\nu_j}\hat{\sigma})$. Here $\hat{\beta}_j$ is the OLS estimate for β_j, $\hat{\sigma}$ is the (unbiased) sample variance, ν_j is the j's diagonal element in $(X^T X)^{-1}$ and $z^{(1-\alpha)}$ is $1-\alpha$ percentile of t-distribution. In practice, for large sample sizes, we just take percentiles of normal distribution instead. This expression can be simplified if σ is known.

   If the assumption that $\epsilon \sim N(0, \sigma^2)$ doesn't hold, the above interval is asymptotically valid (by the Central Limit Theorem, when we average over large number of independent equally distributed variables, we approach a normal distribution). We just need that ϵ are independent, identically distributed for all X, and $E(\epsilon \mid X) = 0$, (and $E(\epsilon^2 < \infty)$). There are generalized versions of CLT that do not require full independence too.

3. *Assume you estimate the model by OLS. What is the relationship between:*

   (a) Estimated ϵ and X.

   (b) Estimated ϵ and estimated Y (equal to $X\hat{\beta}$)

   (c) The mean of the "true" ϵ and the mean of the estimated $\hat{\epsilon}$.

   I'm not sure if I understand the question right, but here is what I think what it means. By the definition,

$$\hat{\epsilon} = Y - \hat{Y} = Y - X\hat{\beta} = Y - X(X^T X)^{-1} X^T Y = (I - X(X^T X)^{-1} X)Y.$$

   (a),(b) The formula above provides a direct relationship between $\hat{\epsilon}, Y$ and X. For a fixed X, $\hat{\epsilon}$ is linear in Y.

   (c) Since $\hat{\beta}$ is unbiased, $E[\hat{\beta}] = \beta$, so (for fixed X),

$$E(\hat{\epsilon}) = E(Y) - XE(\hat{(}\beta)) = E(Y) - X\beta = E(Y - X\beta) = E(\epsilon).$$

   This changes for biased

Problem 2. *Suppose you have a regression model of Y $(N \times 1)$ on X $(N \times 1)$. Suppose that you estimate the model with Ridge regression picking a given value for the penalty multiplier (the scalar which multiplies the L^2 norm of the betas). Suppose now that you add a perfect copy of X to the model. You now therefore have a multiple regression of Y $(N \times 1)$ onto $X^* = [X, X]$ a $N \times 2$ matrix where the two columns are exactly equal to each other.*

1. *What happens to the coefficients if you run a Ridge regression with the same value of the penalty multiplier of Y onto X^*?*

   After expanding the definitions and doing some computations, if the original problem can be formulated as

   $$\text{argmin}\left[\sum(\beta_0+\beta_1 x_i-y_i)^2+\lambda(\beta_0^2+\beta_1^2)\right]$$

   , after adding a copy of the X column, we need to find

   $$\text{argmin}\left[\sum(\beta_0'+(\beta_1'+\beta_2')x_i-y_i i)^2+\lambda((\beta_0')^2+(\beta_1')^2+(\beta_2')^2)\right]$$

   . Now given $\beta_1'+\beta_2'=\gamma_1$ (i.e. fixing the residual sum of squares), the minimum of $(\beta_1')^2+(\beta_2')^2$ is $\frac{\gamma_1^2}{2}$ (achieved when $\beta_1'=\beta_2'$). Applying ridge regression is equivalent to ridge regression

   $$\text{argmin}\left[\sum(\gamma_0+\gamma_1 x_i-y_i)^2+\lambda(\gamma_0^2+1/2\gamma_1^2))\right].$$

   That is also equivalent to applying the ridge regression to Y on $\sqrt{2}\cdot X$ with the same weight λ.

2. *What happens if you instead run a Lasso regression of Y onto X^*?*

   The computations are similar to the previous item. The main difference is that for fixed $\beta_1'+\beta_2'$, the minimal value of $|\beta_1'|+|\beta_2'|$ is just $|\beta_1'+\beta_2'|$ and is achieved for any β_1', β_2' that have the same sign. It follows that:

   (a) The prediction of Y using Lasso for X would be the same as the prediction of Y using Lasso for X^*.

   (b) in the second case, minimizing coefficients would not be unique, which might deteriorate performance.

   (c) if $\hat{\beta}_0,\hat{\beta}_1$ were the original coefficients that we got from Lasso for Y,X, then $\hat{\beta}_0'=\hat{\beta}_0$ and any $\hat{\beta}_1',\hat{\beta}_2'$ that have the same sign and $\hat{\beta}_1'+\hat{\beta}_2'=\hat{\beta}_1$ would be the coefficients we get from the X^* case.

3. *What happens if you use Ordinary Least Squares (OLS) to regress Y onto X^*? Consider both coefficients and standard errors.*

   Similarly to the last two items, in the first case we are looking for

   $$\text{argmin}\left[\sum(\beta_0+\beta_1 x_i-y_i)^2\right],$$

   in the second case we are looking for

   $$\text{argmin}\left[\sum(\beta_0'+(\beta_1'+\beta_2')x_i-y_i i)^2\right]$$

   It is clear that for coefficients $\hat{\beta}_0,\hat{\beta}_1$ that give solutions of the original system, $\hat{\beta}_0'=\hat{\beta}_0$ and any $\hat{\beta}_1',\hat{\beta}_2'$ such that $\hat{\beta}_1'+\hat{\beta}_2'=\hat{\beta}_1$ would be solutions of the second system. So:

- The resulting estimate $\hat{Y^*}$ is the same as the original estimate
- Due to perfect correlation between X and X, the solutions $\hat{\beta}_1'$, $\hat{\beta}_2'$ are not unique (and are not bounded), so we can't provide any estimate on them or their standard errors. (See problem 1.d.iii). Moreover, standard formula for OLS wouldn't even work (since matrix $X^T X$ is not invertible).
- If we are guaranteed that the second column is just a copy of the first column, we would drop it in the OLS and get the same regression model as before.

4. *Suppose now that you instead stack the data. You now have Y^{**} a $2N \times 1$ vector and $X^{**} = [[X, X]]$ a $2N \times 1$ vector and run an OLS regression of Y^{**} onto X^{**}. What happens to the coefficient and its standard error?*

   In this case, the new RSS (residual sum of squares) that we need to minimize is

$$2\left[\sum(\beta_0 + \beta_1 x_i - y_i)^2\right],$$

   simply the multiple of the original RSS by 2. Then the coefficients β_0, β_1 that minimize the RSS are exactly the same as before.

   Let's look at the variance: the original estimated variance is $\hat{\sigma}^2 = \frac{1}{N-2}\sum(y_i - \hat{y_i})^2$, while the new one would be

$$\hat{(\sigma^{**})}^2 = \frac{2}{2N-2}\sum(y_i - \hat{y_i})^2.$$

   If N is big, $\frac{1}{N-1} \approx \frac{2}{2N-2}$, so the sample variance would be almost the same.

   It is worth noting that due to having "extra" sample points, even though $\hat{\beta}$ and $\hat{\sigma^2}$ are (almost) the same, the standard error

$$s_{\hat{\beta}} = \frac{\hat{\sigma}}{\sum(y_i - \hat{y^i})^2}$$

   would decrease by a factor of $\sqrt{2}$ (approximately). This is because $\hat{\sigma}$ stays about the same, while new $RSS(\hat{Y^{**}}, Y^{**}) = 2RSS(\hat{Y}, Y)$ is doubled. This shows that simply duplicating data points might create a false sense of the model being better.