Statistical Learning for Engineers (EN.530.641)
Homework 4

Jin Seob Kim, Ph.D.

Out: 09/30/2022
due: 10/07/2022 by midnight EST

1 Derive Eq. (2.16) in p.19 of Chapter 2. The equation is below:

$$\boldsymbol{\beta} = \left[E(\mathbf{X}\mathbf{X}^T)\right]^{-1} E(\mathbf{X}Y).$$

2 (related to Exercise problem 2.5(a))
Derive Eq. (2.27) in p.26 of ESL. The equation is below:

$$\begin{aligned} EPE(\mathbf{x}_0) &= E_{y_0|\mathbf{x}_0} E_{\mathcal{T}}(y_0 - \hat{y}_0)^2 \\ &= \mathrm{Var}(y_0 \mid \mathbf{x}_0) + E_{\mathcal{T}}[(\hat{y}_0 - E_{\mathcal{T}}\hat{y}_0)^2] + \left(E_{\mathcal{T}}\hat{y}_0 - \mathbf{x}_0^T\boldsymbol{\beta}\right)^2 \\ &= \mathrm{Var}(y_0 \mid \mathbf{x}_0) + \mathrm{Var}_{\mathcal{T}}(\hat{y}_0) + \mathrm{Bias}^2(\hat{y}_0) \\ &= \sigma^2 + E_{\mathcal{T}}[\mathbf{x}_0^T \left(X^T X\right)^{-1} \mathbf{x}_0 \sigma^2] + 0^2 \end{aligned}$$

Hint:

- As learned in the class,
$$E[Y \mid X = x] = \int_y yP(y \mid x)dy.$$
Likewise
$$E_{y|x}[A] = \int_y AP(y \mid x)dy.$$
- Given $\mathbf{x}_0$, y_0 is a random variable due to the random error term, as explained in the class.
- Since y_0 and $\mathcal{T}$ are independent, $E_{y_0|\mathbf{x}_0}$ and $E_{\mathcal{T}}$ commute.

3 Solve Exercise Problem 2.7 (a) and (b) in Chapter 2 (p. 40) of ESL. To repeat:

Suppose we have a sample of N pairs $\mathbf{x}_i, y_i$ drawn i.i.d. from the distribution characterized as follows:

$$\begin{aligned}
&\mathbf{x}_i \sim h(\mathbf{x}), \text{the design density} \\
&y_i = f(\mathbf{x}_i) + \epsilon_i,\ \ f:\ \text{regression function} \\
&\epsilon_i \sim \mathcal{N}(0, \sigma^2)
\end{aligned}$$

We construct an estimator for f *linear* in the y_i,

$$\hat{f}(\mathbf{x}_0) = \sum_{i=1}^{N} \ell_i(\mathbf{x}_0; \mathcal{X}) y_i,$$

where the weights ℓ_i do not depend on the y_i, but do depend on the entire training sequence of $\mathbf{x}_i$, denoted here as $\mathcal{X}$.

(a) Show that linear regression and k-nearest-neighbor regression are members of this class of estimators. Describe explicitly the weights $\ell_i(\mathbf{x}_0; \mathcal{X})$ in each of these cases.

(b) Decompose the conditional mean-squared error

$$E_{\mathcal{Y}|\mathcal{X}}\left[(f(\mathbf{x}_0) - \hat{f}(\mathbf{x}_0))^2\right]$$

into a conditional squared bias and a conditional variance component. Like $\mathcal{X}$, $\mathcal{Y}$ represents the entire training sequence of y_i.

4 Let us generate an artificial training data set as follows. First, we will define the output Y by the following additive error model

$$Y = 0.5X^2 + 0.5X + 1 + \epsilon \tag{1}$$

where $\epsilon \sim \mathcal{N}(0, 1)$. To complete a training data, let us generate the corresponding X as

$$X = 5\eta - 2$$

where η denotes uniform random number. Assume $N = 100$.

(a) Perform linear regression. What is the resulting line equation?

Now you will apply polynomial regression. Let p denote the highest order of polynomial of interest. Then the model in our case is

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \cdots \beta_p X^p.$$

In other words, the regression problem becomes a multi-dimensional input case where

$$\mathbf{X} = [X_1, X_2, \cdots, X_p]^T = [X, X^2, \cdots, X^p]^T.$$

Hence you can generate the corresponding training data from X and Y. Then as we discussed in the class, the same approach for linear regression can be applied.
Practically, when you use Scikit-learn, you will need `PolynomialFeatures` in `sklearn.preprocessing`. In `PolynomialFeatures`, you can set `degree` as polynomial oder p as in the above. Also use `include_bias = False` as another option. Then you will have to use `.fit_transform()` where the input is the random input X as above. Then finally perform `LinearRegression`.

(b) Perform polynomial regression when $p = 2$, 7, and 50. Then plot the original data with all the fitted models together (linear regression and three polynomial regression results). Try to run 3 times and generate the plot of each trial. From this, which models do you think are more sensitive to the changes of the training data?

(c) Which model do you think is the best for the given data and given true model in (1)? Also if you don't know (1), how do you select the best model among four models?

Submission Guideline

- For analytic parts (e.g., problems 1, 2 and 3, and the results of problem 4), submit your homework answers in a single pdf format, including plots, to "HW4_analytical" on the gradescope.
- No more than two (2) homework problems may be on the same page. In other words, for each problem your answers should be on a separate set of pages. Then when submitting, you should assign the pages to each problem on Gradescope.
- *Show your work.*
- Submit all your python codes in a single .zip file that contains codes for each problem (name them by including the problem number). Name your single zip file submission as "YourName_HW4.zip". For example, "JinSeobKim_HW4.zip" for a single zip file. Submission will be done through "HW4_computational" on the gradescope.
- Just in case you have related separate files, please make sure to include *all the necessary files.* If TAs try to run your function and it does not run, then your submission will have a significant points deduction.
- Make as much comments as possible so that the TAs can easily read your codes.