Homework 4

()

Exercise 1.
The goal of this problem is to apply least-squares regression to analyze data sets.

First, download the dataset from the course website. You may also find additional information about the data set `prostate` at `https://web.stanford.edu/~hastie/ElemStatLearn/`

(a) Pre-processing Dataset. Divide the dataset into `training` and `testing`. For each dataset, there is an outcome vector `lpsa`. Let's denote this vector as $\boldsymbol{y} \in \mathbb{R}^n$. There are 8 features for this dataset, namely: `lcavol, lweight, age, lbph, svi, lcp, gleason, pgg45`. Denote these feature vectors as $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_8$, and use a matrix $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_8] \in \mathbb{R}^{n \times 8}$ to represent these feature vectors. In this pre-processing step, we like to make sure

- The average of $\boldsymbol{y}$ is 0, i.e., $(1/n)\sum_{i=1}^n y_i = 0$. You can do this by

$$\boldsymbol{y} \leftarrow \boldsymbol{y} - \frac{1}{n}\sum_{i=1}^n y_i$$

- The average of each column of $\boldsymbol{X}$ is 0, i.e., $(1/n)\sum_{i=1}^n x_{ij} = 0$ for every j. Implement this step by yourself.
- After taking off the average of each column, we also want that the sum square of each column of $\boldsymbol{X}$ is 1, i.e., $\sum_{i=1}^n x_{ij}^2 = 1$ for every j. Implement this step by yourself.
- Explain why these steps are useful.

(b) Least-Squares Fitting. This part uses the pre-processed data from the `training` dataset. First, pick a positive constant λ. Then, solve the following least squares fitting problem:

$$\widehat{\boldsymbol{\beta}}_\lambda = (\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}.$$

Note that $\widehat{\boldsymbol{\beta}}_\lambda$ is a function of λ.

Now, compute $\widehat{\boldsymbol{\beta}}_\lambda$ for λ chosen from the set `numpy.logspace(-10,10,1000)`. This will give you $\widehat{\boldsymbol{\beta}}_1, \ldots, \widehat{\boldsymbol{\beta}}_{1000}$. Plot the trajectory of $\widehat{\boldsymbol{\beta}}_\lambda$ as λ changes. To better visualize your result, use `plt.semilogx`. Mark the x-axis and y-axis of your plot clearly.

(c) Cross-Validation of λ. This part uses the pre-processed data from the `testing` dataset. For every λ you use in part (b), compute the mean square error of the predicted value with respect to the true value in the testing dataset. That is, if we denote the observation of the testing dataset as $\boldsymbol{y}_0$, and the corresponding matrix as $\boldsymbol{X}_0$, then, the mean squared error is

$$\text{MSE}(\lambda) = \frac{1}{m}\|\boldsymbol{y}_0 - \boldsymbol{X}_0\widehat{\boldsymbol{\beta}}_\lambda\|^2, \tag{1}$$

where m is the number of samples in the testing dataset. Plot MSE_λ as a function of λ. Plot using `plt.semilogx`. Mark your axes clearly. What λ is the best value? At this optimal λ, how do you interpret the corresponding regression coefficients $\widehat{\boldsymbol{\beta}}_\lambda$?

Exercise 2.
The goal of this problem is to apply least squares fitting to predict stock market price.

We will use the publicly available datasets on Yahoo: `https://finance.yahoo.com/`. On the course website, you can see two stocks: GOOGLE, and NOVARTIS. There are 200 days of closing price of each stock.

To make our exercise simple, we will use an auto-regressive model. This model assumes that

$$y_n = \sum_{i=1}^{k} \beta_i y_{n-i}. \tag{2}$$

In words, it says that the current value y_n is a linear combination of the its previous k sample values. Putting the auto-regressive model into matrix-vector notation, we have

$$\begin{bmatrix} y_{k+1} \\ y_{k+2} \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} y_k & y_{k-1} & \cdots & y_1 \\ y_{k-1} & y_{k-2} & \cdots & y_2 \\ & & \ddots & \\ y_{n-1} & y_{n-2} & \cdots & y_{n-k} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \tag{3}$$

which is in the form of $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}$. The least-squares fitting therefore gives

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}.$$

(Note: It is okay to add an additional term $\lambda\boldsymbol{I}$ but not necessary for this exercise.)

(a) In this exercise, let's set $k = 25$. Estimate the regression coefficients $\widehat{\boldsymbol{\beta}}$ for `google` and `nvs`.

(b) Predict the stock value for next 30 days. To do so, you need to start with the last k days of known price, and then predict the value of the $(k+1)$-th price. Once the $(k+1)$-th price is estimated, treat it as the true value and proceed to estimate the $(k+2)$-th price.

(c) Plot the known stock prices using one color, and then append your predicted prices using another color.

