FBA | QUANTITATIVE FINANCE RESEARCH GROUP

Financial Machine Learning

Homework 2

Due at 07:00 pm (Korea Standard Time) on Sunday, August 21.

Submit one file: written solutions with executable Python code

Problem 1. Text book: Hands-on Machine Learning. Submit .ipynb file.

(a) Practice all the codes in the Text book Chapter 4. And show that they work well.
(b) Implement Batch Gradient Descent with early stopping for Softmax Regression (without using Scikit-Learn)
(c) Why would you want to use:
- Ridge Regression instead of plain Linear Regression (i.e., without any regularization)?
- Lasso instead of Ridge Regression?
- Elastic Net instead of Lasso?

Problem 2. What happens to linear regression in high dimensions?

Simulate the following setting: To generate each data point, generate independent variables $X_1, \dots X_p \sim N(0,1)$, and let $Y = 4X_1 + \varepsilon$, where $\varepsilon \sim N(0,1)$. This means that the linear model assumption is true for this data for any number of variables p, though there are many useless additional variables when p is large.

Holding n fixed at 100 observations, vary the number of variables p from 2 to 80. For each setting, run 100 simulations of:

(a) Draw n = 100 data points.
(b) Fit a linear regression of Y on $X_1, \dots X_p$ using these data points. Call the resulting coefficient vector $\hat{\beta}$.
(c) Draw a separate test set of m=100 data points, compute predictions at these points using your estimated $\hat{\beta}$, and compute mean squared prediction error of these points:

$$\frac{1}{m}\sum_{i=1}^{m}\left(y_i - x_i^T\hat{\beta}\right)^2$$

(d) Plot average (over simulations) mean squared prediction error vs. p.

Problem 3. Solve the following.

(a) Prove that the estimates (β) given polynomial regression are solved by the following normal equation. (The least squares method is used for model estimation, and k=3)

$$Y_i = b_1 + b_2 X_{2i} + \cdots + b_k X_{ki}$$

$$b_1 = \bar{Y} - b_2 \bar{X}_2 - b_3 \bar{X}_3$$

$$b_2 = \frac{S_{2y} S_{33} - S_{3y} S_{23}}{S_{22} S_{33} - S_{23}^2}$$

$$b_3 = \frac{S_{3y^2} S_{22} - S_{2y} S_{23}}{S_{22} S_{33} - S_{23}^2}$$

(b) Show that the F statistic for dropping a single coefficient from a model is equal to the square of the corresponding z-score

$$F = \frac{(\mathrm{RSS}_0 - \mathrm{RSS}_1)/(p_1 - p_0)}{\mathrm{RSS}_1/(N - p_1 - 1)}, \quad z_j = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{v_j}}$$