Data Science and Social Inquiry: HW3

Yu-Chang Chen and Ming-Jen Lin

November 2, 2022

Question 1: K-means clustering by hand

Suppose that we would like group three data points $(x_{i1}, x_{i2}) = (0,0), (3,0), (0,4)$ into $K = 2$ clusters by K-means clustering. Answer the following questions.

(a) (1 pt) What is the optimal clustering that minimizes the total within-cluster sum of squared Euclidean distance?

$$\sum_{k=1}^{2} \sum_{i \in C_k} \sum_{j=1}^{2} (x_{ij} - \bar{x}_{kj})^2$$

(b) (1 pt) Suppose that we run the iterative K-means clustering algorithm (see p. 519 in the textbook) with the initial cluster assignments $(0,0), (0,4) \in C_1$ and $(3,0) \in C_2$. What would be the clustering the algorithm converges to? Is it the same as what you found in part (a)?

(c) (1 pt) What is the probability of converging to the global optimum if we run the algorithm again with random initial assignments? [1]

Question 2: Selection and shrinkage

Suppose that we want to fit the following linear model:

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon$$

with the data set given in Table 1.

(d) (1 pt) What is the OLS estimate?

(e) (1 pt) What is the LASSO estimate with penalty term $\lambda = 12$?

[1] Ignore the case when all points are assigned to one clusters in the initial assignment.

y	x_1	x_2	x_3	x_4
2	0	1	0	0
1	1	0	0	0
5	0	0	1	0
2	0	1	0	0
4	0	0	0	1
6	0	0	0	1
3	1	0	0	0
3	0	0	1	0
5	0	1	0	0
5	0	0	0	1

Table 1: The data set

(f) (1 pt) How does the size of the penalty term affect our LASSO estimates? Plot the LASSO estimates $(\hat{\beta}_1^L, \hat{\beta}_2^L, \hat{\beta}_3^L, \hat{\beta}_4^L)$ as functions of $\lambda \in [0, 50]$.

(g) (1 pt) What happens when the penalty term gets larger? Can you see where the name "Least absolute and Shrinkage and Selection Operator" comes from?

Question 3: Predict stock returns with LASSO

In this question, you are given a data set of per-minute stock prices for all stocks listed in the New York Stock Exchange (NYSE).[2] We are going to predict the one-minute-ahead return of Pfizer (*stock symbol PFE*) using all the past stock returns (including Pfizer's own past return). Specifically, let

$$r_{i,t} = \frac{p_{i,t} - p_{i,t-1}}{p_{i,t-1}}, \quad t = 1, 2, ..., 33$$

be the one-minute return of stock i at minute t. We are interested in the linear model:

$$r_{\text{Pfizer},t} = \alpha_{\text{Pfizer}} + \sum_{j \in \text{NYSE}} \left[\beta_{j,1} r_{j,t-1} + \beta_{j,2} r_{j,t-2} + \beta_{j,3} r_{j,t-3}\right] + \epsilon_{\text{Pfizer},t}.$$

3 x 3067 +1

(h) (1 pt) How many parameters (including the intercept) are we trying to estimate?

Since the our data is high-dimensional, we need to use LASSO to estiamte our model, and we need to select an appropriate penalty term λ. In below, we outline how to select λ in a data-driven fashion with *cross-validation*.

[2] The data is stored in the *parquet* format. Parquet is a *column-oriented* data storage format designed for efficient storage and process of big data. You can load parquet files in Python with `pandas.read_parquet`.

First, calculate the one-minute return for $t = 1, 2, ..., 33$ and drop observation $t = 0$ since $t = 0$ has no base period. Then, split the data into five subsamples in which subsample $\mathcal{D}_k$ contains observation

$$t = 6 \cdot (k - 1) + 1, 6 \cdot (k - 1) + 2, ..., 6 \cdot (k - 1) + 9.$$

Notice that $\mathcal{D}_k$'s are partly overlapping since we need $r_{t-1}, r_{t-2}, r_{t-3}$ to predict r_t.[3]

The idea of cross-validation is that, for a given candidate λ, we can train the LASSO using one part of the data set (the *training* set) then test the model's performance (with penalty λ) on the other part of the data set (the *test* set). Specifically, with five-fold cross-validation, we split our data into five subsamples and use the four of the five subsamples together as training set, while keeping one of subsample out for testing. Below is an outline of the algorithm.

1. for $\lambda \in [a, b]$:[4]
   - (i) Leave the first subsample $\mathcal{D}_1$ out and use the other four subsamples together to train the model of interest $\hat{f}_\lambda$.
   - (ii) Evaluate the model $\hat{f}_\lambda$ on the testing set $\mathcal{D}_1$ by calculating:

     $$\mathcal{R}_\lambda(\mathcal{D}_1) = \sum_{t \in \mathcal{D}_1} (r_{\text{Pfizer},t} - \hat{f}(x))^2,$$

     where x refers to the past returns $(r_{j,t-1}, r_{j,t-2}, r_{j,t-3}), j \in$ NYSE.
   - (iii) Repeat (i) and (ii) with $\mathcal{D}_2$, $\mathcal{D}_3$, $\mathcal{D}_4$, $\mathcal{D}_5$ as training set. Calculate $\mathcal{R}_\lambda = \sum_{k=1}^{5} \mathcal{R}_\lambda(D_k)$.
2. Choose the λ that minimizes $\mathcal{R}_\lambda$.

(i) (1 pt) What is the optimal λ you find using five-fold cross-validation? When estimating LASSO, please standardize every regressor so that each regressor is mean zero with variance one.

(j) (1 pt) Use the λ you found to estimate the coefficients. How many coefficients are non-zero? What are the stocks with non-zero coefficients?

[3] The standard way to split data is to split it randomly. Here, we split our data into subsamples of consecutive minutes to account for time dependence and for the ease of grading.

[4] The interval $[a, b]$ is the search grid in which you think the optimal λ belongs to.