Data Science and Social Inquiry: HW3

Yu-Chang Chen and Ming-Jen Lin

November 10, 2022

Question 1: K-means clustering by hand

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

Sol.

If $(0,4) \in C_1$ and $(0,0),(3,0) \in C_2$:
$\sum_{k=1}^{2} \sum_{i \in C_k} \sum_{j=1}^{2} (x_{ij} - \bar{x}_{kj})^2 = (0-1.5)^2 + (3-1.5)^2 = 4.5$

If $(3,0) \in C_1$ and $(0,0),(0,4) \in C_2$:
$\sum_{k=1}^{2} \sum_{i \in C_k} \sum_{j=1}^{2} (x_{ij} - \bar{x}_{kj})^2 = (0-2)^2 + (4-2)^2 = 8$

If $(0,0) \in C_1$ and $(3,0),(0,4) \in C_2$:
$\sum_{k=1}^{2} \sum_{i \in C_k} \sum j = 1^2 (x_{ij} - \bar{x}_{kj})^2 = \sqrt{(3-1.5)^2 + (0-2)^2} + \sqrt{(0-1.5)^2 + (4-2)^2} =$ 12.5

Hence, the optimal clustering is with the initial cluster assignments$(0,4) \in C_1$ and $(0,0),(3,0) \in C_2$

$\square$

(b) (1 pt) What would be the clustering the algorithm converges to? Is it the same as what you found in part (a)?

Sol.

$\bar{x_1} = (0,2)\bar{x_2} = (3,0)$

$d(x_1, \bar{x_1}) = 2, d(x_1, \bar{x_2}) = 3$
$d(x_2, \bar{x_1}) =\sqrt{13}$ d($\mathrm{x}_2, \bar{x_2}) = 0$

$d(x_3, \bar{x_1}) = 2, d(x_3, \bar{x_2}) = 5$

Hence, $(0,0), (0,4) \in C_1$ and $(3,0) \in C_2$ converge

It is not the same as what we 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?

Sol.

If we run the algorithm again with $(0,4) \in C_1$ and $(0,0), (3,0) \in C_2$, it converges to the optimal clustering. (part (a))

If we run the algorithm again with $(3,0) \in C_1$ and $(0,0), (0,4) \in C_2$, it converges to $(3,0) \in C_1$ and $(0,0), (0,4) \in C_2$. (part (b))

If we run the algorithm again with $(0,0) \in C_1$ and $(3,0), (0,4) \in C_2$, it converges to $(0,0) \in C_1$ and $(3,0), (0,4) \in C_2$.(below)

$\bar{x_1} = (0,0)\bar{x_2} = (1.5, 2)$

$d(x_1, \bar{x_1}) = 0, d(x_1, \bar{x_2}) = \frac{1}{2}\sqrt{5}$
$d(x_2, \bar{x_1}) = 3, d(x_2, \bar{x_2}) = 1.5$
$d(x_3, \bar{x_1}) = 4, d(x_3, \bar{x_2}) = 1.5$

Hence, the probability of converging to the global optimum is $\frac{1}{3}$if we run the algorithm again with random initial assignments.

□

Question 2: Selection and shrinkage

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

Sol.

$\hat{\beta}_1 = \frac{1+3}{1+1} = 2$

$\hat{\beta}_2 = \frac{2+2+5}{1+1+1} = 3$

$\hat{\beta}_3 = \frac{5+3}{1+1} = 4$

$\hat{\beta}_4 = \frac{4+6+5}{1+1+1} = 5$

Hence, the fitted value of y by OLS method is $\hat{y} = 2x_1 + 3x_2 + 4x_3 + 5x_4$

□

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

Sol.

$$\min \sum_{i=1}^{10} (y_i - \beta_1 x_1 - \beta_2 x_2 - \beta_3 x_3 - \beta_4 x_4)^2 + \lambda \sum_{j=1}^{10} |\beta_j|$$

$2 \sum_{i=1}^{10} (y_i x_i - \beta x_i^2) - \lambda = 0$

$\hat{\beta}_1 = \frac{2 \sum_{i=1}^{10} (x_i y_i) - \lambda}{2 \sum_{i=1}^{10} (x_i^2)}$

$2 < \sqrt{12}4 \Rightarrow \beta_1 = 0$

$\beta_2 = 3 - \sqrt{12}6 = 1$

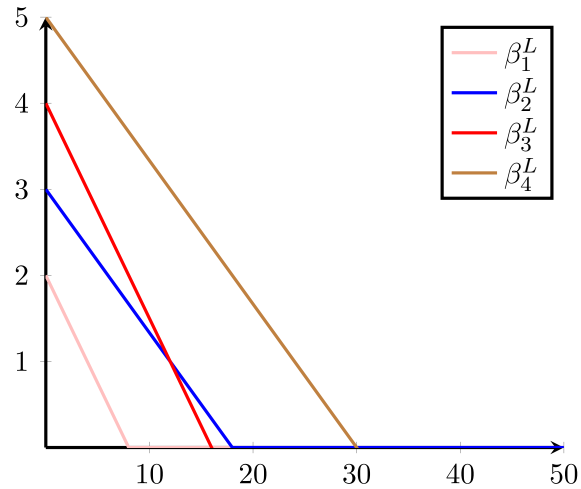
$\beta_3 = 4 - \sqrt{12}4 = 1$

$\beta_4 = 5 - \sqrt{12}6 = 3$

Hence, the LASSO estimator with penalty term $\lambda = 12$ is $y = x_2 + x_3 + 3x_4 + \epsilon$ □

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

Sol.



□

(*g*) (1 pt) Compare the three estimates you found. Can you see where the name "Least absolute and Shrinkage and Selection Operator" comes from?

Sol.

We can see that when λ gets larger, the more coefficients will be selected and shrink to 0. Only the ones who have larger average will be selected. In this case, when λ is larger or equal to 8,16,18,and 30 separately, β_1^L, β_3^L, β_2^L, and β_4^L will sequentially shrink to zero as well. □

Question 3: Predict stock returns with LASSO

(*h*) (1 pt) How many parameters (including the intercept) are we estimating?

Sol.

There are 3,067 companies,three periods (t = 3) and an intercept , which means there are 9,202 parameters. □

(*i*) (1 pt) Use the five-fold cross-validation to select the optimal penalty term λ. What is the optimal λ you find?

Sol.

The optimal λ we found was 0.000046. □

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

Sol.

With λ equals 0.000046, there are 23 non-zero coefficients.
The companies are :

BAC(t_1), BRK(t_1), IVR(t_1), LTRP(t_1), NGL(t_1), SMMC(t_1)

UONE(t_1), WPG(t_1), BAC(t_2), BRK(t_2), OPES(t_2), SMMC(t_2)

WFC(t_2), BAC(t_3), BRK(t_3), CLNY(t_3), GLOG(t_3), GLOP(t_3)

NYMT(t_3), PEI(t_3), QRTE(t_3), TWO(t_3), WFC(t_3).

□