Data Science and Social Inquiry: HW3

Yu-Chang Chen

October 27, 2023

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: OLS is Sample Mean

Consider the data in Table 1 and the following OLS estimators:[2]

$$new\_births_{ct} = \alpha_0 + \alpha_1 year_t + \epsilon_{ct}$$
$$new\_births_{ct} = \beta_1 \mathbb{1}\{year_t = 2020\} + \beta_2 \mathbb{1}\{year_t = 2021\} + \beta_3 \mathbb{1}\{year_t = 2022\} + \epsilon_{ct}$$
$$new\_births_{ct} = \gamma_0 + \gamma_1 \mathbb{1}\{year_t = 2021\} + \gamma_2 \mathbb{1}\{year_t = 2022\} + \epsilon_{ct}$$

(a) (1 pt) What are $\hat{\alpha}_0$ and $\hat{\alpha}_1$?

[1] Ignore the case when all points are assigned to one clusters in the initial assignment.
[2] Subscript c stands for city and t for time.

city	year	new_births
Taipei	2020	18000
Taipei	2021	16000
Taipei	2022	14000
New Taipei	2020	22000
New Taipei	2021	20000
New Taipei	2022	18000

Table 1: Mock data

(*b*) (1 pt) What are $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$?

(*c*) (1 pt) What are $\hat{\gamma}_0$, $\hat{\gamma}_1$, and $\hat{\gamma}_2$?

(*d*) (1 pt) Define $X = \mathbb{1}\{year \geq 2022\}$ and $Z = \mathbb{1}\{city = Taipei\}$. Recall that the conditional expectation

$$E[\, new\_births \mid X, Z]$$

is a random variable. How many values (at most) would it possibly take? Hint: the variables $\mathbb{1}\{year \geq 2022\}, \mathbb{1}\{city = Taipei\}$ are both $0, 1$.

(*e*) (1 pt) Consider the regression:

$$new\_births_{ct} = \delta_0 + \delta_1 X_{ct} + \delta_2 Z_{ct} + \delta_3 X_{ct} Z_{ct} + \epsilon_{ct}.$$

If the coefficients $\delta_0, \delta_1, \delta_2, \delta_3$ are known, what would be the fitted value of new birth for Taipei in 2021?

Question 3: 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 2.

(*f*) (1 pt) What is the OLS estimate?

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

(*h*) (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 $\lambda \in [0, 50]$.

(*i*) (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?

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 2: The data set