Data Science & Social Inquiry HW3 Solution

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?

Solution:

- $(0,0)(3,0) \quad (0,4)$

$$0 + 1.5^2 + 1.5^2 = 4.5$$

- $(0,0)(0,4) \quad (3,0)$

$$0 + 2^2 + 2^2 = 8$$

- $(0,0) \quad (0,4)(3,0)$

$$0 + 2.5^2 + 2.5^2 = 12.5$$

The optimal clustering is thus $(0,0)(3,0) \quad (0,4)$.

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

Solution: There are two centroids $(0,2)$, $(3,0)$.

	(0,2)	(3,0)
(0,0)	4	9
(4,0)	4	25
(3,0)	13	0

After an iteration through the K-means algorithm, all points will remain in the same group.

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

> **Solution:** $\frac{1}{3}$

Question 2: OLS is Sample Mean

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

> **Solution:** $\hat{\alpha}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{-4000-4000}{4} = -2000$
>
> $\hat{\alpha}_0 = \bar{y} - \hat{\alpha}_1 \bar{x} = 18000 - (-2000)(2021) = 18000 + 4042000 = 4060000$

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

> **Solution:** $\hat{\beta}_1 = \frac{(18000+22000)}{2} = 20000$
>
> $\hat{\beta}_2 = \frac{(16000+20000)}{2} = 18000$
>
> $\hat{\beta}_3 = \frac{(14000+18000)}{2} = 16000$

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

> **Solution:** $\hat{\gamma}_0 = 20000$
>
> $\hat{\gamma}_1 = 18000 - 20000 = -2000$
>
> $\hat{\gamma}_2 = 16000 - 20000 = -4000$

(g) (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.

Solution: There are at most 4 values.

(h) (1 pt) Consider the regression:

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

If the coefficients δ_0, δ_1, δ_2, δ_3 are known, what would be the fitted value of new birth for Taipei in 2021?

Solution: $\delta_0 + \delta_2$

Question 3: Selection and shrinkage

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

Solution:

$$\begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} = \begin{pmatrix} 2 \\ 3 \\ 4 \\ 5 \end{pmatrix}$$

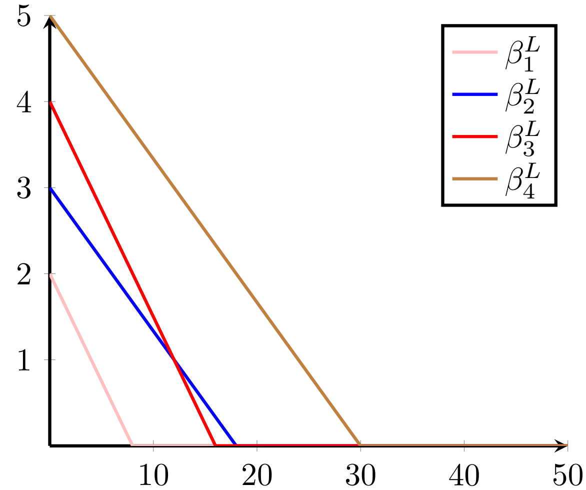
(j) (1 pt) What is the LASSO estimate with penalty term $\lambda = 12$?

Solution:

$$\begin{aligned} x_1 : \tfrac{12}{2\cdot 2} = 3 \Rightarrow &\quad \beta_1^L = 0 \\ x_2 : \tfrac{12}{2\cdot 3} = 2 \Rightarrow &\quad \beta_2^L = 1 \\ x_3 : \tfrac{12}{2\cdot 2} = 3 \Rightarrow &\quad \beta_3^L = 1 \\ x_4 : \tfrac{12}{2\cdot 3} = 2 \Rightarrow &\quad \beta_4^L = 3 \end{aligned}$$

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

Solution:



(l) (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?

Solution: Once λ becomes larger, LASSO shrinks small coefficients to 0.