Data Science and Social Inquiry - Midterm Exam

No explanation is needed for multiple choice questions.

1. (2 points) Select all options with correct statements.

   A. One can choose the optimal penalty for LASSO by gradient descent.

   B. LASSO is a clustering algorithm.

   C. LASSO applies only when p, the number of regressors, is greater than n, the number of observations.

   D. LASSO with penalty $\lambda = 0$ is ordinary least squares regression.

2. (1 point) The class of penalized regression takes the form

$$\min_{\alpha,\beta} \sum_{i=1}^{n} (y_i - \alpha - \boldsymbol{x_i'}\boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} |\beta_j|^s.$$

   Which value of s corresponds to LASSO?

   A. $s = \infty$ B. $s = 2$ C. $s = 0$ **D.** $s = 1$

3. (1 point) Suppose that the covariance matrix $\boldsymbol{\Sigma} = \begin{bmatrix} 10 & -1 \\ -1 & 10 \end{bmatrix}$. Let $PC_1 = \begin{bmatrix} a_{11} \\ a_{12} \end{bmatrix}$ be the coefficient of the first principal component. If $a_{11} > 0$, then

   A. $a_{12} > 0$ B. Can't determine. C. $a_{12} = 0$ **D.** $a_{12} < 0$

4. (1 point) Following the previous question. Let $PC_2 = \begin{bmatrix} a_{21} \\ a_{22} \end{bmatrix}$ be the coefficient of the second principal component. If $a_{21} > 0$, then

   A. Can't determine. B. $a_{22} = 0$ C. $a_{22} < 0$ **D.** $a_{22} > 0$

5. (2 points) Suppose that our data has 5 observations:

$$(X_i, Y_i) = (-2, -3),\ (-1, 2),\ (3, 2),\ (-1, 5),\ (3, -2).$$

   Define the distance between two points as the square of their euclidean distance, i.e.,

$$(x_i - x_j)^2 + (y_i - y_j)^2.$$

Perform the hierarchical clustering algorithm with the average linkage. Draw a dendrogram to indicate which observations are pooled in each step. Make sure you mark the distances between clusters on the y-axis.

Solution:

1.

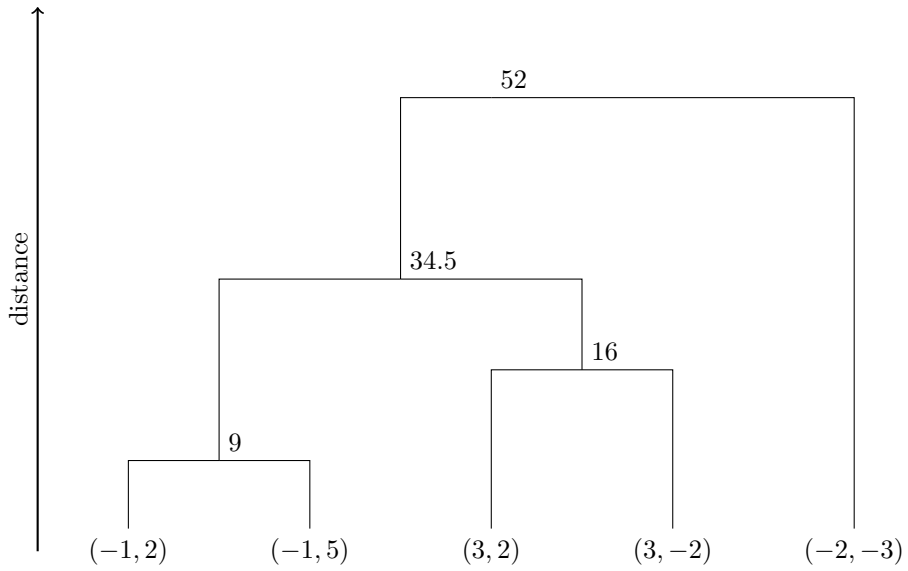
	(-2,-3)	(-1,2)	(3,2)	(-1,5)	(3,-2)
(-2,-3)	0	26	50	65	26
(-1,2)		0	16	9	32
(3,2)			0	25	16
(-1,5)				0	65
(3,-2)					0

2.

	(-2,-3)	{(-1,2),(-1,5)}	(3,2)	(3,-2)
(-2,-3)	0	45.5	50	26
{(-1,2),(-1,5)}		0	20.5	48.5
(3,2)			0	16
(3,-2)				0

3.

	(-2,-3)	{(-1,2),(-1,5)}	{(3,2), (3,-2)}
(-2,-3)	0	45.5	38
{(-1,2),(-1,5)}		0	34.5
{(3,2),(3,-2)}			0



6. (1 point) Bandiera et al. (2020) uses dimension reduction method to identify two types of CEO. What are the two types?

   A. directors & entrepreneurs.

   B. managers & leaders.

   C. leaders & directors.

   D. entrepreneurs & managers.

   E. entrepreneurs & leaders.

   F. directors & managers.

7. (1 point) Which of the following matrices can possibly be a covariance matrix? Select all that apply.

   A. $\begin{bmatrix} 3 & -2 \\ -2 & 4 \end{bmatrix}$ B. $\begin{bmatrix} 3 & 4 \\ 4 & 5 \end{bmatrix}$ C. $\begin{bmatrix} -3 & 1 \\ 1 & 2 \end{bmatrix}$ D. $\begin{bmatrix} 3 & -2 \\ -1 & 5 \end{bmatrix}$

8. (1 point) "Functional data" refers to data set that contains continuous functions $f_1, f_2, ..., f_n$ as observations. For continuous functions on $[-1, 1]$, we may consider the *uniform norm*:

$$d(f_1, f_2) = \max_{x \in [-1,1]} |f_1(x) - f_2(x)|$$

   to define distances between functions. Suppose that our data contains three functions

$$f_1(x) = x,\ f_2(x) = |x|,\ f_3(x) = x^2$$

   for $x \in [-1, 1]$. What is the optimal K-means clustering with $K = 2$? Show your work.

Solution: Three possibilities:

- $\{f_1, f_2\}$, $\{f_3\}$,

$$\frac{1}{2}2 + 0 = 1$$

- $\{f_1, f_3\}$, $\{f_2\}$,

$$\frac{1}{2}2 + 0 = 1$$

- $\{f_2, f_3\}$, $\{f_1\}$,

$$\frac{1}{2}\frac{1}{2} + 0 = \frac{1}{4}.$$

The optimal K-means is therefore $\{|x|,\ x^2\}$, $\{x\}$.