CS150A Database, Fall 2021
Homework 3 Solutions

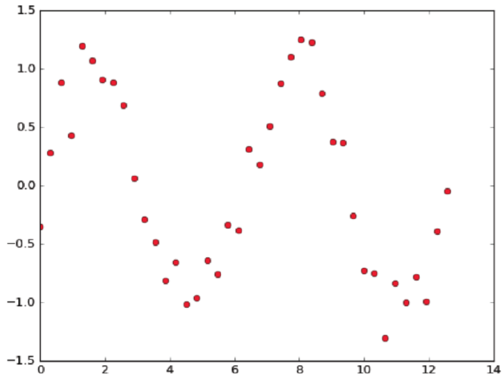
(Due Sunday, Jan. 2 at 11:59pm (CST))

January 3, 2022

Note that: For Q2, Q3 and Q4, solutions with the correct answer but without adequate explanation will not earn marks.

1. (1) Suppose we've already run k-means with $k = 3$, and have the following cluster centers: Red=(1, 6), Green=(5, 3), Blue=(2, 2). We then receive a new point (4, 5), which cluster do you predict it belongs to? (5 points)
   Solution:
   We predict based on the cluster center which has the shortest distance to the point.
   Distance((4,5), (1,6)) = 3.16
   Distance((4,5), (5,3)) = 2.2
   Distance((4,5), (2,2)) = 3.6
   So the closest cluster is Green.

   (2) Suppose we wanted to apply linear regression to this data. Which of the following features would you include to better fit the data? Check all that apply. (5 points)



   A. 1
   B. x
   C. x^2
   D. x^3
   E. Not possible to apply linear regression to this data
   Solution:
   Higher degree polynomials can help better fit the nonlinear data. Choose 1, x, x^2 and x^3.

   (3) Increasing the number of features will guarantee your model to perform better. Mark only one answer. (5 points)
   A. True, because you have more information, and more is always better
   B. False, because your computational performance will slow drastically
   C. False because your model may overfit on the training data
   Solution:
   False, because your model may overfit on the training data.

2. Use the k-means algorithm and Euclidean distance to cluster the following 8 data points:

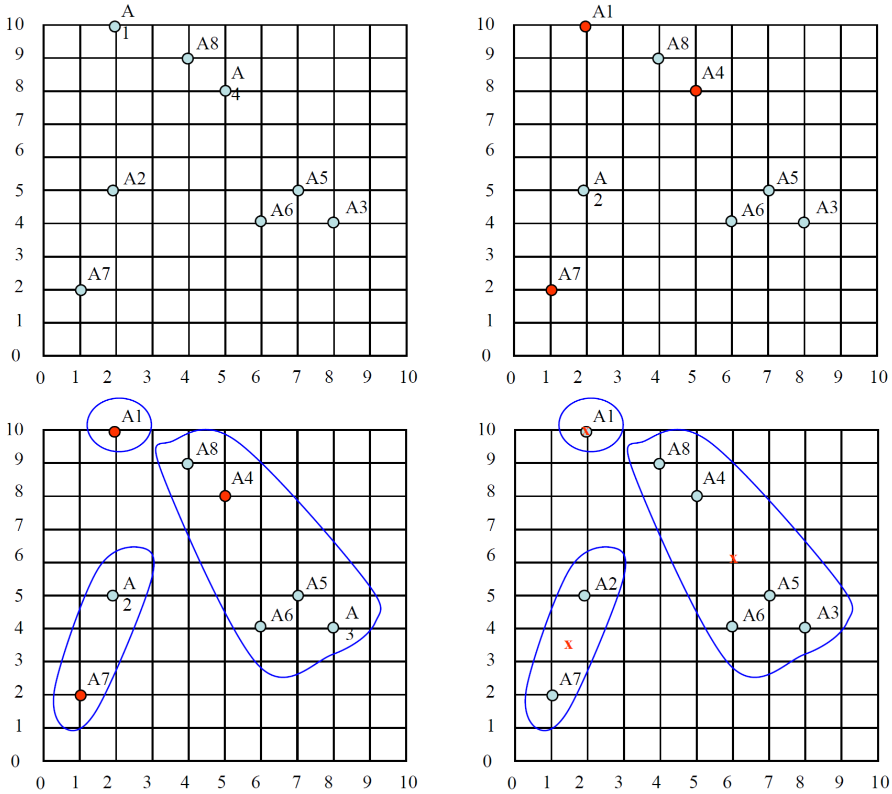
$$x_1 = (2, 10),\ x_2 = (2, 5),\ x_3 = (8, 4),\ x_4 = (5, 8),$$
$$x_5 = (7, 5),\ x_6 = (6, 4),\ x_7 = (1, 2),\ x_8 = (4, 9).$$

Suppose the number of clusters is 3, and the Lloyd's algorithm is applied with the initial cluster centers x_1, x_4 and x_7. At the end of the first iteration show:

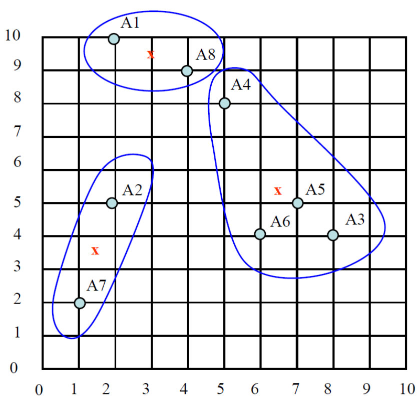
(a) The new clusters, i.e., the example assignment. (5 points)
Solution:
Cluster 1: $\{x_1\}$; Cluster 2: $\{x_3, x_4, x_5, x_6, x_8\}$; Cluster 3: $\{x_2, x_7\}$.

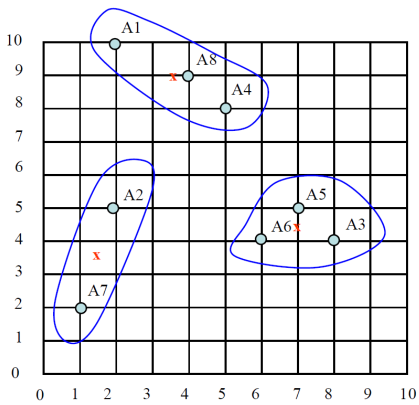
(b) The centers of the new clusters. (5 points)
Solution:
$c_1 = (2, 10)$, $c_2 = (6, 6)$, $c_3 = (1.5, 3.5)$.

(c) Draw a 10 by 10 space with all the 8 points, and show the clusters after the first iteration and the new centroids. (5 points)
Solution:



(d) How many more iterations are needed to converge? Draw the result for each iteration. (5 points)
Solution:
Two more iterations are needed.
After the 2nd iteration the results would be
Cluster 1: $\{x_1, x_8\}$; Cluster 2: $\{x_3, x_4, x_5, x_6\}$; Cluster 3: $\{x_2, x_7\}$.
With centers $c_1 = (3, 9.5)$, $c_2 = (6.5, 5.25)$, $c_3 = (1.5, 3.5)$.
After the 3rd iteration the results would be
Cluster 1: $\{x_1, x_4, x_8\}$; Cluster 2: $\{x_3, x_5, x_6\}$; Cluster 3: $\{x_2, x_7\}$.
With centers $c_1 = (3.66, 9)$, $c_2 = (7, 4.33)$, $c_3 = (1.5, 3.5)$.


A1
A8
A4
A2
A5
A6
A3
A7


A1
A8
A4
A2
A5
A6
A3
A7

3. Given a set of i.i.d. observation pairs $(x_1, y_1) \cdots (x_n, y_n)$, where $x_i, y_i \in \mathbb{R}$, $i = 1, 2, ..., n$.

(a) By assuming the linear model is a reasonable approximation, we consider fitting the model via least squares approaches, in which we choose coefficients θ and θ_0 to minimize the Residual Sum of Squares (RSS),

$$\hat{\theta},\ \hat{\theta}_0 = \underset{\theta,\ \theta_0}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \theta x_i - \theta_0)^2. \tag{1}$$

Estimate the model parameters θ and θ_0. (5 points)
Solution:

$$\hat{\theta} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}, \tag{2}$$

$$\hat{\theta}_0 = \bar{y} - \hat{\beta}\bar{x}, \tag{3}$$

(b) Using (1), argue that in the case of simple linear regression, the least squares line always passes through the point $(\bar{x}, \bar{y})$, where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. (5 points)
Solution:
We can plug $(\bar{x}, \bar{y})$ into the equation $\hat{y} = \hat{\theta}x_i + \theta_0$, and we find $\bar{y} = \hat{\theta}\bar{x} + (\bar{y} - \hat{\theta}\bar{x}) = \bar{y}$ satisfies. So the least squares line always passes through the point $(\bar{x}, \bar{y})$.

4. Ridge regression shrinks the regression coefficients by imposing a penalty on their size. The ridge coefficients minimize a penalized Residual Sum of Squares (RSS),

$$\hat{\theta}^{ridge},\ \hat{\theta}_0^{ridge} = \underset{\theta,\ \theta_0}{\text{argmin}} \left(\sum_{i=1}^{n} \left(y_i - \theta_0 - \sum_{j=1}^{p} x_{ij}\theta_j \right)^2 + \lambda \sum_{j=1}^{p} \theta_j^2 \right). \tag{4}$$

Here $\lambda \geq 0$ is a complexity parameter that controls the amount of shrinkage.

(a) Show that the ridge regression problem in (4) is equivalent to the problem:

$$\hat{\theta}^{c},\ \hat{\theta}_0 = \underset{\theta^c,\ \theta_0}{\text{argmin}} \left(\sum_{i=1}^{n} \left(y_i - \theta_0^c - \sum_{j=1}^{p} (x_{ij} - \bar{x}_j)\theta_j^c \right)^2 + \lambda \sum_{j=1}^{p} \theta_j^{c2} \right), \tag{5}$$

where $\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$, $j = 1, 2, ..., p$. Given the correspondence between θ^c and the original θ in (4). Characterize the solution to this modified criterion. (5 points)

Solution:

Rewrite above objective function as

$$Q(\theta^c, \theta_0^c) = \left(\sum_{i=1}^{n} \left(y_i - \left(\theta_0^c - \sum_{j=1}^{p} \bar{x}_j \theta_j^c \right) - \sum_{j=1}^{p} x_{ij}\theta_j^c \right)^2 + \lambda \sum_{j=1}^{p} \theta_j^{c2} \right). \tag{6}$$

Compared with (4), we get the following correspondence:

$$\theta_0 = \theta_0^c - \sum_{j=1}^{p} \bar{x}_j \theta_j^c, \tag{7}$$

$$\theta_j = \theta_j^c, \quad j = 1, 2, ..., p. \tag{8}$$

(b) After reparameterization using centered inputs ($\tilde{x}_{ij} \leftarrow x_{ij} - \bar{x}_j$, $\tilde{y}_i \leftarrow y_i - \bar{y}$, $\forall i, j$), show that the solution to (4) can be separated into following two parts:

$$\hat{\theta}_0^{ridge} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \tag{9}$$

$$\hat{\theta}^{ridge} = \underset{\theta}{\text{argmin}} \left(\sum_{i=1}^{n} \left(\tilde{y}_i - \sum_{j=1}^{p} \tilde{x}_{ij}\theta_j \right)^2 + \lambda \sum_{j=1}^{p} \theta_j^2 \right). \tag{10}$$

(5 points)

Solution:

Note: This question isn't clearly described, and we will be awarding full points for any attempt.

Due to the equivalence between (4) and (5), we consider to solve (5) instead. Let $Q(\theta^c, \theta_0^c)$ denote the objective function of (5), we have

$$\frac{\partial Q}{\partial \theta_0^c} = -2\sum_{i=1}^{n} \left(y_i - \theta_0^c - \sum_{j=1}^{p} (x_{ij} - \bar{x}_j)\theta_j^c \right) = 0, \tag{11}$$

leading to

$$\begin{aligned}
\theta_0^c &= \frac{1}{n} \left(\sum_{i=1}^{n} y_i - \sum_{i=1}^{n}\sum_{j=1}^{p} (x_{ij} - \bar{x}_j)\theta_j^c \right) \\
&= \frac{1}{n}\sum_{i=1}^{n} y_i - \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p} x_{ij}\theta_j^c + \sum_{j=1}^{p} \bar{x}_j \theta_j^c \\
&= \frac{1}{n}\sum_{i=1}^{n} y_i - \sum_{j=1}^{p} \left(\frac{1}{n}\sum_{i=1}^{n} x_{ij} \right) \theta_j^c + \sum_{j=1}^{p} \bar{x}_j \theta_j^c \\
&= \bar{y}.
\end{aligned} \tag{12}$$

Substituting the above equation into (5), we have

$$\begin{aligned}
\hat{\theta}^c &= \underset{\theta^c}{\text{argmin}} \left(\sum_{i=1}^{n} \left(y_i - \bar{y} - \sum_{j=1}^{p} (x_{ij} - \bar{x}_j)\theta_j^c \right)^2 + \lambda \sum_{j=1}^{p} \theta_j^{c2} \right) \\
&= \underset{\theta^c}{\text{argmin}} \left(\sum_{i=1}^{n} \left(\tilde{y}_i - \sum_{j=1}^{p} \tilde{x}_{ij}\theta_j^c \right)^2 + \lambda \sum_{j=1}^{p} \theta_j^{c2} \right). \qquad (13)
\end{aligned}$$

(c) Based on the ridge regression model learned in (b), show its prediction $\hat{y}_0$ on an arbitrary testing point $\mathbf{x}_0 = [x_{01}, x_{02}, ..., x_{0p}]^\top \in \mathbb{R}^p$. (5 points)
Solution:
Note: This question isn't clearly described, and we will be awarding full points for any attempt.
Given the model $(\hat{\theta}^{ridge}, \hat{\theta}_0^{ridge})$ learned in (b), the prediction $haty_0$ on $\mathbf{x}_0$ is made by

$$\begin{aligned}
\hat{y}_0 &= \sum_{j=1}^{p} (x_{0j} - \bar{x}_j)\hat{\theta}_j^{ridge} + \hat{\theta}_0^{ridge} \\
&= \sum_{j=1}^{p} (x_{0j} - \bar{x}_j)\hat{\theta}_j^{ridge} + \bar{y}, \qquad (14)
\end{aligned}$$

where $\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$ $(\forall j)$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ are calculated based on the training data.