Database and Data Mining, Fall 2020
Homework 3

(Due Friday, Dec. 25 at 11:59pm (CST))

December 24, 2020

Note that: solutions with the correct answer but without adequate explanation will not earn marks.

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

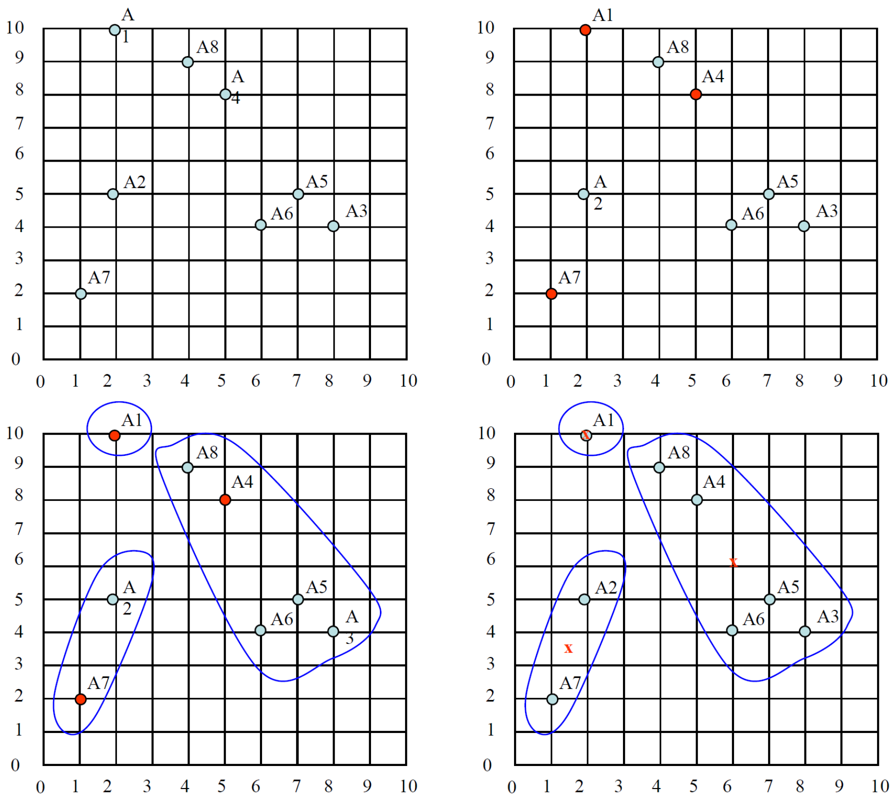
$$x_1 = (2, 10),\ x_2 = (2, 5),\ x_3 = (8, 4),\ x_4 = (5, 8),$$
$$x_5 = (7, 5),\ x_6 = (6, 4),\ x_7 = (1, 2),\ x_8 = (4, 9).$$

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:

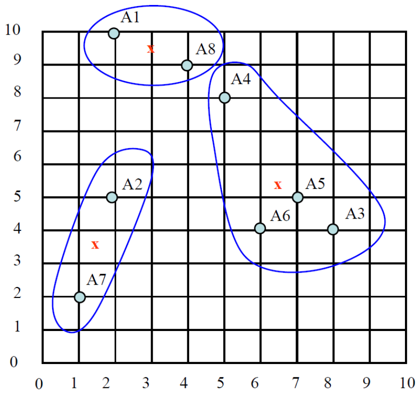
(a) The new clusters, i.e., the example assignment. (4 points)
Solution:
Cluster 1: $\{x_1\}$; Cluster 2: $\{x_3, x_4, x_5, x_6, x_8\}$; Cluster 3: $\{x_2, x_7\}$.

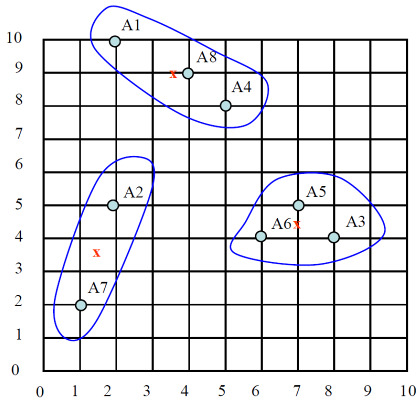
(b) The centers of the new clusters. (4 points)
Solution:
$c_1 = (2, 10)$, $c_2 = (6, 6)$, $c_3 = (1.5, 3.5)$.

(c) Draw a 10 by 10 space with all the 8 points, and show the clusters after the first iteration and the new centroids. (4 points)
Solution:



(d) How many more iterations are needed to converge? Draw the result for each iteration. (8 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 $c_1 = (3.66, 9)$, $c_2 = (7, 4.33)$, $c_3 = (1.5, 3.5)$.





2. 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 $(\bar{x}, \bar{y})$.

(c) Suppose the observed label value y_i $(i = 1, 2, ..., n)$ is generated according to the non-deterministic linear model:

$$y_i = \theta x_i + \theta_0 + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2), \tag{4}$$

where $\mathcal{N}(0, \sigma^2)$ denotes a Gaussian distribution with mean 0 and variance σ^2. Calculate the expectation and variance of y_i $(i = 1, 2, ..., n)$, and use Maximum Likelihood Estimation (MLE) to estimate the model parameters θ and θ_0. (5 points)
Solution:

$$\mathbb{E}(y_i) = \theta x_i + \theta_0, \quad \operatorname{Var}(y_i) = \operatorname{Var}(\epsilon) = \sigma^2, \quad y_i \sim \mathcal{N}(\theta x_i + \theta_0, \sigma^2). \tag{5}$$

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ denote the dataset. According to MLE, we have

$$\begin{aligned}\max_{\theta,\theta_0} L(\mathcal{D}|\theta,\theta_0) &= \max_{\theta,\theta_0} \mathrm{P}(\mathcal{D}|\theta,\theta_0) && (6)\\ &= \max_{\theta,\theta_0} \prod_{i=1}^n \mathrm{P}(x_i, y_i|\theta,\theta_0) && (7)\\ &= \max_{\theta,\theta_0} \prod_{i=1}^n \mathrm{P}(y_i|x_i,\theta,\theta_0)\mathrm{P}(x_i). && (8)\end{aligned}$$

Since $\mathrm{P}(x_i)$ is irrelevant with θ and θ_0, the above problem is equivalent to

$$\begin{aligned}\max_{\theta,\theta_0} \ell(\mathcal{D}|\theta,\theta_0) &= \max_{\theta,\theta_0} \ln L(\mathcal{D}|\theta,\theta_0) && (9)\\ &= \max_{\theta,\theta_0} \sum_{i=1}^n \ln \mathrm{P}(y_i|x_i,\theta,\theta_0) + C && (10)\\ &= \max_{\theta,\theta_0} \sum_{i=1}^n \ln\left(\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(y_i-\theta x_i-\theta_0)^2}{2\sigma^2}\right)\right) + C && (11)\\ &= \min_{\theta,\theta_0} \frac{1}{2\sigma^2}\sum_{i=1}^n (y_i-\theta x_i-\theta_0)^2 + \tilde{C}, && (12)\end{aligned}$$

where C and $\tilde{C}$ denote the constant terms. The solutions are the same with (a).

(d) Suppose the observed label value y_i $(i = 1, 2, ..., n)$ is generated according to the non-deterministic linear model:

$$y_i = \theta x_i + \theta_0 + \epsilon_i, \quad \epsilon \sim \mathcal{N}(0, \sigma_i^2). \tag{13}$$

Use MLE to estimate the model parameters θ and θ_0, and discuss the difference with the results in (c). (5 points)

Solution:

According to

$$\mathbb{E}(y_i) = \theta x_i + \theta_0, \quad \mathrm{Var}(y_i) = \mathrm{Var}(\epsilon) = \sigma^2, \quad y_i \sim \mathcal{N}(\theta x_i + \theta_0, \sigma_i^2), \tag{14}$$

we have

$$\max_{\theta,\theta_0} L(\mathcal{D}|\theta,\theta_0) \Leftrightarrow \min_{\theta,\theta_0} \sum_{i=1}^n \frac{1}{2\sigma_i^2}(y_i-\theta x_i-\theta_0)^2, \tag{15}$$

which is weighted linear regression with weights $w_i = \frac{1}{2\sigma_i^2}$, $i = 1, 2, ..., n$.

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

$$\hat{\theta}^{ridge},\ \hat{\theta}_0^{ridge} = \underset{\theta,\ \theta_0}{\mathrm{argmin}} \left(\sum_{i=1}^n \left(y_i - \theta_0 - \sum_{j=1}^p x_{ij}\theta_j\right)^2 + \lambda \sum_{j=1}^p \theta_j^2\right). \tag{16}$$

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

(a) Show that the ridge regression problem in (16) is equivalent to the problem:

$$\hat{\theta}^c,\ \hat{\theta}_0 = \underset{\theta^c,\ \theta_0}{\mathrm{argmin}} \left(\sum_{i=1}^n \left(y_i - \theta_0^c - \sum_{j=1}^p (x_{ij}-\bar{x}_j)\theta_j^c\right)^2 + \lambda \sum_{j=1}^p \theta_j^{c2}\right), \tag{17}$$

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 (16). Characterize the solution to this modified criterion. (5 points)

Solution:

Rewrite above objective function as

$$Q(\theta^c, \theta_0^c) = \left(\sum_{i=1}^n \left(y_i - \left(\theta_0^c - \sum_{j=1}^p \bar{x}_j\theta_j^c\right) - \sum_{j=1}^p x_{ij}\theta_j^c\right)^2 + \lambda \sum_{j=1}^p \theta_j^{c2}\right). \tag{18}$$

Compared with (16), we get the following correspondence:

$$\theta_0 = \theta_0^c - \sum_{j=1}^{p} \bar{x}_j \theta_j^c, \tag{19}$$

$$\theta_j = \theta_j^c, \quad j = 1, 2, ..., p. \tag{20}$$

(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 (16) can be separated into following two parts:

$$\hat{\theta}_0^{ridge} = \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i, \tag{21}$$

$$\hat{\theta}^{ridge} = \operatorname*{argmin}_{\theta} \left(\sum_{i=1}^{n} \left(\tilde{y}_i - \sum_{j=1}^{p} \tilde{x}_{ij} \theta_j \right)^2 + \lambda \sum_{j=1}^{p} \theta_j^2 \right). \tag{22}$$

(5 points)
Solution:
Due to the equivalence between (16) and (17), we consider to solve (17) instead. Let $Q(\theta^c, \theta_0^c)$ denote the objective function of (17), we have

$$\frac{\partial Q}{\partial \theta_0^c} = -2 \sum_{i=1}^{n} \left(y_i - \theta_0^c - \sum_{j=1}^{p} (x_{ij} - \bar{x}_j) \theta_j^c \right) = 0, \tag{23}$$

leading to

$$\begin{aligned}
\theta_0^c &= \frac{1}{n} \left(\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \sum_{j=1}^{p} (x_{ij} - \bar{x}_j) \theta_j^c \right) \\
&= \frac{1}{n} \sum_{i=1}^{n} y_i - \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{p} x_{ij} \theta_j^c + \sum_{j=1}^{p} \bar{x}_j \theta_j^c \\
&= \frac{1}{n} \sum_{i=1}^{n} y_i - \sum_{j=1}^{p} \left(\frac{1}{n} \sum_{i=1}^{n} x_{ij} \right) \theta_j^c + \sum_{j=1}^{p} \bar{x}_j \theta_j^c \\
&= \bar{y}.
\end{aligned} \tag{24}$$

Substituting the above equation into (17), we have

$$\begin{aligned}
\hat{\theta}^c &= \operatorname*{argmin}_{\theta^c} \left(\sum_{i=1}^{n} \left(y_i - \bar{y} - \sum_{j=1}^{p} (x_{ij} - \bar{x}_j) \theta_j^c \right)^2 + \lambda \sum_{j=1}^{p} \theta_j^{c2} \right) \\
&= \operatorname*{argmin}_{\theta^c} \left(\sum_{i=1}^{n} \left(\tilde{y}_i - \sum_{j=1}^{p} \tilde{x}_{ij} \theta_j^c \right)^2 + \lambda \sum_{j=1}^{p} \theta_j^{c2} \right).
\end{aligned} \tag{25}$$

(c) Based on the ridge regression model learned in (b), show its prediction $\hat{y}_0$ on an arbitrary testing point $\mathbf{x}_0 = [x_{01}, x_{02}, ..., x_{0p}]^\top \in \mathbb{R}^p$. (4 points)
Solution:
Given the model $(\hat{\theta}^{ridge}, \hat{\theta}_0^{ridge})$ learned in (b), the prediction $haty_0$ on $\mathbf{x}_0$ is made by

$$\begin{aligned}
\hat{y}_0 &= \sum_{j=1}^{p} (x_{0j} - \bar{x}_j) \hat{\theta}_j^{ridge} + \hat{\theta}_0^{ridge} \\
&= \sum_{j=1}^{p} (x_{0j} - \bar{x}_j) \hat{\theta}_j^{ridge} + \bar{y},
\end{aligned} \tag{26}$$

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.

(d) Given $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n]^\top \in \mathbb{R}^{n\times p}$ ($\mathbf{x}_i \in \mathbb{R}^p$ is the i-th example, $i = 1, 2, ..., n$), $\mathbf{y} = [y_1, y_2, ..., y_n]^\top \in \mathbb{R}^n$, and $\boldsymbol{\theta} = [\theta_1, \theta_2, ..., \theta_p]^\top \in \mathbb{R}^p$. Show the optimization problem (22) and its closed-form solution in the matrix form. (Suppose $\mathbf{X}$ and $\mathbf{y}$ have been removed the sample means in column-wise.) (6 points)

Solution:

The optimization problem (22) in matrix form:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 + \lambda \|\boldsymbol{\theta}\|_2^2 . \tag{27}$$

Its objective function $Q(\boldsymbol{\theta})$ can be rewritten as followings:

$$\begin{aligned} Q(\boldsymbol{\theta}) &= (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) + \lambda \boldsymbol{\theta}^\top \boldsymbol{\theta} \\ &= \mathbf{y}^\top \mathbf{y} - \mathbf{y}^\top \mathbf{X}\boldsymbol{\theta} - \boldsymbol{\theta}\mathbf{X}^\top \mathbf{y} + \boldsymbol{\theta}^\top \mathbf{X}^\top \mathbf{X}\boldsymbol{\theta} + \lambda \boldsymbol{\theta}^\top \boldsymbol{\theta}. \end{aligned} \tag{28}$$

Let the derivative of $Q(\boldsymbol{\theta})$ w.r.t. θ equal to 0, we have

$$\begin{aligned} \left(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}\right) \boldsymbol{\theta} &= \mathbf{X}^\top \mathbf{y} \\ \boldsymbol{\theta} &= \left(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}\right)^{-1} \mathbf{X}^\top \mathbf{y}. \end{aligned} \tag{29}$$

The matrix $\left(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}\right)$ is always invertible once $\lambda > 0$.