Database and Data Mining, Fall 2020
Homework 3

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

December 21, 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)

(b) The centers of the new clusters. (4 points)

(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)

(d) How many more iterations are needed to converge? Draw the result for each iteration. (8 points)

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)

(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)

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

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

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)

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

$$y_i = \theta x_i + \theta_0 + \epsilon_i, \quad \epsilon \sim \mathcal{N}(0, \sigma_i^2). \tag{3}$$

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

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),

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

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:

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

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)

(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:

$$\hat{\theta}_0^{ridge} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \tag{6}$$

$$\hat{\theta}^{ridge} = \underset{\theta}{\operatorname{argmin}} \left(\sum_{i=1}^{n} \left(\tilde{y}_i - \sum_{j=1}^{p} \tilde{x}_{ij}\theta_j \right)^2 + \lambda \sum_{j=1}^{p} \theta_j^2 \right). \tag{7}$$

(5 points)

(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)

(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 (7) 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)