Machine Learning - HW4

Kai Liao

September 2020

1 K-means and K-medians for Cute Raccoon

1.1 K-medians

1.1.1 e

Firstly, I provide steps for the algorithm.
1. Input number of clusters, randomly initialize centers
2. Assign each point to the its nearest median, with distance given by l_1.
3. Change each cluster center to be the median of point in the group in each single dimension.

Proof:
The minimization problem

$$\min_{c_1,\dots,c_K} \sum_{i,j=1}^{N} \min_{k\in\{1,\dots,K\}} |x_{ij} - c_k| \tag{1}$$

can be rewritten to

$$\min_{c_1,\dots,c_K} \min_{k\in\{1,\dots,K\}} \sum_{k} \sum_{ij:x_{ij} \text{ is in cluster } _k} |x_{ij} - c_k| \tag{2}$$

For each iteration, we can assign data point to the closest cluster center first, i.e., assign it the cluster where distance $|x_{ij} - c_k|$ is the smallest. Then, we can fix this cluster assignment. Therefore, the minimization problem for this step is given by

$$\begin{aligned} &\min_{c_1',\cdots,c_k'} \textstyle\sum_{i=1}^{n} |x_{ij} - c_k'| \\ &= \min_{c_1',\cdots,c_k'} \left(\textstyle\sum_{x_{ij}\in c_1} |x_{ij} - c_1'| + \cdots + \sum_{x_{ij}\in c_k} |x_{ij} - c_k'| \right) \end{aligned} \tag{3}$$

The FOC is given by

$$\sum_{k} \sum_{ij\in c_k} -\operatorname{sign}(x_{ij} - c_k') = 0 \tag{4}$$

Clearly,when c_k' is the median of $x_{ij} \in c_k$, the FOC is satisfied and the cost function reaches its minimization point.

1.1.2 d

It is not always the same. Consider a 'outlier' x_i which is significantly larger than the mean and the median of the points belong to the same cluster. When x_i gets greater and greater, the K-means will increase the center for this cluster. However the K-medians will not change the center because the median remains unchanged. Therefore, they will not always give the same result.

1.1.3 e

The image is NOT exactly the same, but they do look very similar(qualitatively the same).

2 Ridge Regression

2.1 a

Since $y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i$ where $\epsilon_1, \cdots, \epsilon_n$ are independent and identically distributed from a $N(0,1)$ distribution, $y_i \sim N(\beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j, 1)$. Therefore, the likelihood function is given by

$$P(\mathbf{Y} \mid \boldsymbol{\beta}, \mathbf{X}) = \prod_{i=1}^{n} p\left(y_i \mid x_i; \beta\right) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{\left(y_i - \left(\beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j\right)\right)^2}{2}} \tag{5}$$

Using the notation of course note,

$$P(\mathbf{Y} \mid \boldsymbol{\beta}, \mathbf{X}) = \frac{1}{(2\pi)^{p/2}} \exp\left(-\frac{1}{2}\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2\right) \tag{6}$$

2.2 b

$$\begin{aligned}
P(\boldsymbol{\beta} \mid \mathbf{Y}, \mathbf{X}) &= P(\mathbf{Y} \mid \boldsymbol{\beta}, \mathbf{X}) P(\boldsymbol{\beta}) \frac{1}{Z} \\
&= \frac{1}{Z} \frac{1}{(2\pi)^{p/2}} \exp\left(-\frac{1}{2}\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2\right) \frac{1}{(2\pi)^{p/2}\lambda^{-p/2}} \exp\left(-\frac{1}{2\lambda^{-1}}\|\boldsymbol{\beta}\|_2^2\right) \\
&= \frac{1}{Z} \frac{1}{(2\pi)^{p}\lambda^{-p/2}} \exp\left(-\frac{1}{2}\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2\right) \exp\left(-\frac{1}{2\lambda^{-1}}\|\boldsymbol{\beta}\|_2^2\right) \\
&= \frac{\lambda^{p/2}}{(2\pi)^{p} \int P(Y \mid \beta X) p(\beta) d\beta} \exp\left(-\frac{1}{2}\|Y - X\beta\|_2^2\right) \cdot \exp\left(-\frac{\lambda}{2}\|\beta\|_2^2\right)
\end{aligned} \tag{7}$$

2.3 c

Without normalization of the raw data, the regularization of each variable could be very imbalance. The model may weight regularization of some parameters too heavy or too light. Suppose the dataset of features of vehicles. One variable is the height of the car, ranging from 1 m to 3 m. The other is the mileage, ranging for 0 m to 100,000,000 m. The effect of regularization on these two variables will be very different if without normalization.

3 Neural Networks

$$\begin{aligned}
&\mathbb{E}_{w_1,\cdots,w_m,s_1,\cdots,s_m} \left\langle \phi\left(x_1\right), \phi\left(x_2\right) \right\rangle \\
=&\mathbb{E}\left[\sum_{i=1}^{m} x_1 x_2^\top \frac{1}{m} s_i^2 \mathbb{I}\left[w_i^\top x_1 \geq 0\right] \mathbb{I}\left[w_i^\top x_2 \geq 0\right]\right] \\
=&\mathbb{E}\left[x_1 x_2^\top \frac{1}{m} \sum_{i=1}^{m} \left\{\mathbb{I}\left[w_i^\top x_1 \geq 0\right] \mathbb{I}\left[w_i^\top x_2 \geq 0\right]\right\}\right] \\
=&x_1 x_2^\top \mathbb{P}_w\left(w^\top x_1 \geq 0 \quad \text{and} \quad w^\top x_2 \geq 0\right) \\
=&x_1 x_2^\top \mathbb{P}_w\left(\frac{w^\top x_1}{\|w\|_2 \|x_1\|_2} \geq 0 \quad \text{and} \quad \frac{w^\top x_2}{\|w\|_2 \|x_2\|_2} \geq 0\right) \\
=&x_1 x_2^\top \mathbb{P}_w\left(\arccos \frac{w^\top x_1}{\|w\|_2 \|x_1\|_2} \leq \frac{\pi}{2} \quad \text{and} \quad \arccos \frac{w^\top x_2}{\|w\|_2 \|x_2\|_2} \leq \frac{\pi}{2}\right) \\
=&x_1 x_2^\top \frac{\pi - \arccos \frac{x_1^\top x_2}{\|x_1\|_2 \|x_2\|_2}}{2\pi}
\end{aligned} \tag{8}$$

4 Expectation Maximization (EM) for a Mixture of Two Different Distributions

4.1 a

The data generating process is given by

$$p\left(X_i = x_i \mid \lambda\right) = \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} \tag{9}$$

Therefore, the likelihood function is

$$L\left(\lambda; x_1, \ldots, x_n\right) = \prod_{j=1}^{n} \exp(-\lambda)\frac{1}{x_j!}\lambda^{x_j} \tag{10}$$

Take *log* to acquire the log-likelihood:

$$\begin{aligned} l\left(\lambda; x_1, \ldots, x_n\right) &= \ln\left(\prod_{j=1}^{n} \exp(-\lambda)\frac{1}{x_j!}\lambda^{x_j}\right) \\ &= \sum_{j=1}^{n} \ln\left(\exp(-\lambda)\frac{1}{x_j!}\lambda^{x_j}\right) \\ &= -n\lambda - \sum_{j=1}^{n} \ln\left(x_j!\right) + \ln(\lambda)\sum_{j=1}^{n} x_j \end{aligned} \tag{11}$$

Therefore, we can define the minimization problem by

$$\widehat{\lambda} = \arg\max_{\lambda} l\left(\lambda; x_1, \ldots, x_n\right) \tag{12}$$

Then the FOC of the minimization problem of log-likelihood is given by

$$\frac{d}{d\lambda} l\left(\lambda; x_1, \ldots, x_n\right) = 0 \tag{13}$$

Solve for λ,

$$\frac{dl}{d\lambda} = -n + \frac{1}{\lambda}\sum_{j=1}^{n} x_j = 0 \tag{14}$$

$$\widehat{\lambda}_n = \frac{1}{n}\sum_{j=1}^{n} x_j \tag{15}$$

4.2 b

$$p\left(X_i = 0 \mid Z_i = 1, \lambda\right) = 1 \tag{16}$$

$$p\left(X_i \neq 0 \mid Z_i = 1, \lambda\right) = 0 \tag{17}$$

$$p\left(X_i = x_i \mid Z_i = 2, \lambda\right) = \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} \tag{18}$$

4.3 c

$$p\left(X_i = 0 \mid \lambda\right) = p\left(Z_i = 1\right) + p\left(Z_i = 2\right)e^{-\lambda} = w + (1-w)e^{-\lambda} \tag{19}$$

For $x_i \neq 0$

$$p\left(X_i = x_i \mid \lambda\right) = p\left(Z_i = 2\right)\frac{e^{-\lambda}\lambda^{x_i}}{x_i!} = (1-w)\frac{e^{-\lambda}\lambda^{x_i}}{x_i!} \tag{20}$$

4.4 d

$$\begin{aligned}\gamma_{i,k} &:= p\left(Z_i = k \mid X_i = x_i, \lambda\right) \\ &= \frac{p\left(X_i = x_i \mid Z_i = k, \lambda\right) p\left(Z_i = k \mid \lambda\right)}{p\left(X_i = x_i \mid \lambda\right)}\end{aligned} \tag{21}$$

$$p\left(Z_i = 1 \mid X_i = 0, \lambda\right) = \frac{w}{w + (1-w)e^{-\lambda}} \tag{22}$$

$$p\left(Z_i = 1 \mid X_i = x_i, \lambda\right) = 0,\ x_i \neq 0 \tag{23}$$

$$p\left(Z_i = 2 \mid X_i = 0, \lambda\right) = \frac{e^{-\lambda}(1-w)}{w + (1-w)e^{-\lambda}} \tag{24}$$

$$p\left(Z_i = 2 \mid X_i = x_i, \lambda\right) = 1,\ x_i \neq 0 \tag{25}$$

4.5 e

$$p\left(X_i = x_i, Z_i = k \mid \lambda\right) = \begin{cases} w, & \text{if } x_i = 0, z_i = 1. \\ 0, & \text{if } x_i \neq 0, z_i = 1. \\ (1-w)\frac{e^{-\lambda}\lambda^{x_i}}{x_i!}, & \text{if } z_i = 2. \end{cases} \tag{26}$$

$$\begin{aligned}\arg\max_{\lambda} A(w, \lambda) &= \arg\max_{\lambda} \sum_{i=1}^{N}\sum_{k=1}^{K} \gamma_{i,k} \log \frac{p\left(X_i = x_i, Z_i = k \mid \lambda\right)}{\gamma_{i,k}} \\ &= \arg\max_{\lambda} \sum_{i=1}^{N}\sum_{k=1}^{K} \gamma_{i,k} \log p\left(X_i = x_i, Z_i = k \mid \lambda\right) \\ &= \arg\max_{\lambda} \sum_{i=1}^{N} \gamma_{i,2} \log p\left(X_i = x_i, Z_i = 2 \mid \lambda\right) \\ &= \arg\max_{\lambda} \sum_{i=1}^{N} \gamma_{i,2} \log (1-w)\frac{e^{-\lambda}\lambda^{x_i}}{x_i!} \\ &= \arg\max_{\lambda} \sum_{i=1}^{N} \gamma_{i,2} \log \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} \\ &= \arg\max_{\lambda} \sum_{i=1}^{N} \gamma_{i,2} \left(-\lambda - \ln\left(x_i!\right) + \ln(\lambda) x_i\right)\end{aligned} \tag{27}$$

FOC condition is given by

$$\sum_{i=1}^{N}(-\gamma_{i,2} + \gamma_{i,2} x_i \frac{1}{\lambda}) = 0 \tag{28}$$

Solve for λ:

$$\lambda^* = \frac{\sum_{i=1}^{N} \gamma_{i,2} x_i}{\sum_{i=1}^{N} \gamma_{i,2}} \tag{29}$$

It is different from the result of (a). It is weighted by $\gamma_{i,2}$

4.6 f

$$
\begin{aligned}
\arg\max_w A(w,\lambda) &= \arg\max_w \sum_{i=1}^{N}\sum_{k=1}^{K} \gamma_{i,k} \log \frac{p\left(X_i = x_i, Z_i = k \mid \lambda\right)}{\gamma_{i,k}} \\
&= \arg\max_w \sum_{i=1}^{N}\sum_{k=1}^{K} \gamma_{i,k} \log p\left(X_i = x_i, Z_i = k \mid \lambda\right) \\
&= \arg\max_w \left(\sum_{i=1}^{N} \gamma_{i,1} \ln w + \sum_{i=1}^{N} \gamma_{i,2} \ln(1-w) \ln \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} \right) \\
&= \arg\max_w \left(\sum_{i=1}^{N} \gamma_{i,1} \ln w + \sum_{i=1}^{N} \left(\gamma_{i,2} \ln(1-w) - \lambda - \ln\left(x_i!\right) + \ln(\lambda) x_i\right) \right)
\end{aligned}
\tag{30}
$$

Note that when $k = 1$, if the data generating process is valid, $\mathbb{I}_{[x_i=1]} = 1$ always true.

FOC condition is given by

$$
\frac{1}{w}\sum_{i=1}^{N} \gamma_{i,1} - \frac{1}{1-w}\sum_{i=1}^{N} \gamma_{i,2} = 0 \tag{31}
$$

Solve for w:

$$
w^* = \frac{\sum_{i=1}^{n} \gamma_{i,2}}{\sum_{i=1}^{n} \left(\gamma_{i,1} + \gamma_{i,2}\right)} = \frac{\sum_{i=1}^{N} \gamma_{i,1}}{n} \tag{32}
$$

It is the same as the mixture of Gaussian distributions case.