Introduction to Machine Learing: Homework IV

Due on Dec 7th, 2022 at 11:59pm

Professor Ziping Zhao

Bingnan Li
2020533092

1. [*Clustering and Mixture Models*]

   (a) K-means algorithm.
   Solution:

      i. Initialize K cluster centers m_i by randomly selecting K input data points.
      ii. Repeat the following procedure until convergence:
         A. For all $x^{(l)} \in \mathcal{X}$, we obtain the estimated labels

$$b_i^{(l)} = \begin{cases} 1, \; if \;\; i = \arg\min_j ||x^{(l)} - m_j|| \\ 0, \; elsewhere \end{cases}$$

         B. For all m_i, we obtain

$$m_i = \frac{\sum_l b_i^{(l)} x^{(l)}}{\sum_l b_i^{(l)}}$$

   (b) Cluster the samples into 2 clusters.
   Solution:

   First, we select $m_1 = (0, 0)$ and $m_2 = (5, 0)$ as initialized cluster center. Then for the first iteration, we have the following result:

$$\begin{aligned} b_1^{(1)} &= 1 \quad b_2^{(1)} = 0 \\ b_1^{(2)} &= 1 \quad b_2^{(2)} = 0 \\ b_1^{(3)} &= 1 \quad b_2^{(3)} = 0 \\ b_1^{(4)} &= 0 \quad b_2^{(4)} = 1 \\ b_1^{(5)} &= 0 \quad b_2^{(5)} = 1 \\ m_1 &= \frac{(0,2) + (0,0) + (1,0)}{3} = (\frac{1}{3}, \frac{2}{3}) \\ m_2 &= \frac{(5,0) + (5,2)}{2} = (5, 1) \end{aligned}$$

   Next, for the second iteration, we find that

$$\begin{aligned} b_1^{(1)} &= 1 \quad b_2^{(1)} = 0 \\ b_1^{(2)} &= 1 \quad b_2^{(2)} = 0 \\ b_1^{(3)} &= 1 \quad b_2^{(3)} = 0 \\ b_1^{(4)} &= 0 \quad b_2^{(4)} = 1 \\ b_1^{(5)} &= 0 \quad b_2^{(5)} = 1 \\ m_1 &= \frac{(0,2) + (0,0) + (1,0)}{3} = (\frac{1}{3}, \frac{2}{3}) \\ m_2 &= \frac{(5,0) + (5,2)}{2} = (5, 1) \end{aligned}$$

   The result converged, so we terminated the algorithm and cluster centers are

$$m_1 = (\frac{1}{3}, \frac{2}{3}) \quad m_2 = (5, 1)$$

2. [*Clustering and Mixture Models*]

(a) Advantages of GMM and Why it can be used for clustering.
Solution:

Advantages: GMM is a kind of "soft-label" method, the projected data do not represent deterministic classification label but the probability of belonging to any classes.

Why it can be used for clustering: K-means is a special case of GMM. In practice, the higher the $h_i^{(l)}$ is, the more likely that $x^{(l)}$ is generated by component $\mathcal{G}_i$, which can be interpreted as $x^{(l)}$ belongs to cluster i.

(b) Estimate the parameters of the GMM.
Solution:

Define $\mathcal{Q}(\boldsymbol{\phi}|\boldsymbol{\phi^t})$ as following

$$\begin{aligned}
\mathcal{Q}(\boldsymbol{\phi}|\boldsymbol{\phi^t}) &= \mathbb{E}[\mathcal{L}_C(\boldsymbol{\phi}|\mathcal{X},\mathcal{Z})|\mathcal{X},\boldsymbol{\phi^t}] \\
&\mathit{where} \\
\mathcal{L}_C(\boldsymbol{\phi}) &= \log\prod_l p(x^{(l)}, z^{(l)}|\boldsymbol{\phi}) \\
&= \sum_l \left[\log P(\boldsymbol{z}^{(l)}|\boldsymbol{\phi}) + \log p(\boldsymbol{x}^{(l)}|\boldsymbol{z}^{(l)},\boldsymbol{\phi})\right] \\
&= \sum_l\sum_i z_i^{(l)}[\log\pi_i + \log p_i(\boldsymbol{x}^{(l)|\boldsymbol{\phi}})]
\end{aligned}$$

Hence

$$\begin{aligned}
\mathcal{Q}(\boldsymbol{\phi}|\boldsymbol{\phi^t}) &= \mathbb{E}[\mathcal{L}_C(\boldsymbol{\phi}|\mathcal{X},\mathcal{Z})|\mathcal{X},\boldsymbol{\phi^t}] \\
&= \sum_l\sum_i \mathbb{E}[z_i^{(l)}|\mathcal{X},\boldsymbol{\phi^t}][\log\pi_i + \log p_i(\boldsymbol{x}^{(l)|\boldsymbol{\phi}})]
\end{aligned}$$

where

$$\begin{aligned}
\mathbb{E}[z_i^{(l)}|\mathcal{X},\boldsymbol{\phi^t}] &= \mathbb{E}[z_i^{(l)}|\boldsymbol{x}^{(l)},\boldsymbol{\phi}] \\
&= P(z_i^{(l)} = 1|\boldsymbol{x}^{(l)},\boldsymbol{\phi^t}) \\
&= \frac{p(\boldsymbol{x}^{(l)}|z_i^{(l)} = 1,\boldsymbol{\phi^t})P(z_i^{(l)} = 1|\boldsymbol{\phi^t})}{p(\boldsymbol{x}^{(l)}|\boldsymbol{\phi^t})} \\
&= \frac{p_i(\boldsymbol{x}^{(l)}|\boldsymbol{\phi^t})\pi_i}{\sum_j p_j(\boldsymbol{x}^{(l)}|\boldsymbol{\phi^t})\pi_j} \\
&= \frac{P(x^{(l)}|\mathcal{G}_i,\boldsymbol{\phi}^t)\pi_i}{\sum_j P(x^{(l)}|\mathcal{G}_j,\boldsymbol{\phi}^t)\pi_j} \\
&\equiv h_i^{(l)}
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
h_i^{(l)} &= \frac{P(x^{(l)}|\mathcal{G}_i,\boldsymbol{\phi}^t)\pi_i}{\sum_j P(x^{(l)}|\mathcal{G}_j,\boldsymbol{\phi}^t)\pi_j} \\
&= \frac{|\Sigma_i|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}(\boldsymbol{x_l}-\boldsymbol{\mu_i})^T(\Sigma)^{-1}(\boldsymbol{x_l}-\boldsymbol{\mu_i})\right]\pi_i}{\sum_{j=1}^K |\Sigma_j|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}(\boldsymbol{x_l}-\boldsymbol{\mu_j})^T(\Sigma)^{-1}(\boldsymbol{x_l}-\boldsymbol{\mu_j})\right]\pi_j} \\
&= \frac{\mathcal{N}(\boldsymbol{x}_l|\boldsymbol{\mu_i},\boldsymbol{\Sigma_i})\pi_i}{\sum_{j=1}^K\mathcal{N}(\boldsymbol{x}_l|\boldsymbol{\mu_j},\boldsymbol{\Sigma_j})\pi_j}
\end{aligned}$$

and

$$\mathcal{Q}(\boldsymbol{\phi}|\boldsymbol{\phi}^t) = \sum_l \sum_i h_i^{(l)} [\log \pi_i + \log \mathcal{N}(\boldsymbol{x}_l|\boldsymbol{\mu_i}, \boldsymbol{\Sigma_i})]$$

Then, maximization of $\mathcal{Q}(\boldsymbol{\phi}|\boldsymbol{\phi}^t)$ is equivalent to

$$\begin{array}{ll} \underset{\{\pi_i\},\{\boldsymbol{\mu}_i\},\{\Sigma_i\}}{\text{maximize}} & \mathcal{Q}(\boldsymbol{\phi}|\boldsymbol{\phi}^t) = \sum_l \sum_i h_i^{(l)} \log \pi_i + h_i^{(l)} \log \mathcal{N}(\boldsymbol{x}_l|\boldsymbol{\mu_i}, \boldsymbol{\Sigma_i}) \\ \text{subject to} & \sum_i \pi_i = 1 \end{array}$$

Since the second term does not depend on π_i, the problem for $\{\pi_i\}$ is

$$\begin{array}{ll} \underset{\{\pi_i\}}{\text{maximize}} & \sum_l \sum_i h_i^{(l)} \log \pi_i \\ \text{subject to} & \sum_i \pi_i = 1 \end{array}$$

By using Lagrangian, we solve for

$$\frac{\partial}{\partial \pi_i} \left[\sum_l \sum_i h_i^{(l)} \log \pi_i - \lambda \left(\sum_i \pi_i - 1 \right) \right] = 0$$

And we get

$$\pi_i = \frac{\sum_l h_i^{(l)}}{N}$$

Then the first term of $\mathcal{Q}$ does not depend on $\boldsymbol{\mu_i}, \Sigma_i$. Hence, the problem for $\boldsymbol{\mu_i}, \Sigma_i$ is

$$\underset{\{\boldsymbol{\mu}_i\},\{\Sigma_i\}}{\text{maximize}} \quad \sum_l \sum_i h_i^{(l)} \log \mathcal{N}(\boldsymbol{x}_l|\boldsymbol{\mu_i}, \boldsymbol{\Sigma_i})$$

By solving

$$\frac{\partial}{\partial \mu_i} \left[\sum_l \sum_i h_i^{(l)} \log \mathcal{N}(\boldsymbol{x}_l|\boldsymbol{\mu_i}, \boldsymbol{\Sigma_i}) \right] = 0$$

and

$$\frac{\partial}{\partial \Sigma_i} \left[\sum_l \sum_i h_i^{(l)} \log \mathcal{N}(\boldsymbol{x}_l|\boldsymbol{\mu_i}, \boldsymbol{\Sigma_i}) \right] = 0$$

we get

$$\boldsymbol{\mu}_i^{t+1} = \frac{\sum_l h_i^{(l)} \boldsymbol{x}_l}{\sum_l h_i^{(l)}}$$

and

$$\boldsymbol{\Sigma}_i^{t+1} = \frac{\sum_l h_i^{(l)} (\boldsymbol{x}_l - \boldsymbol{\mu}_i^{t+1})(\boldsymbol{x}_l - \boldsymbol{\mu}_i^{t+1})^T}{\sum_l h_i^{(l)}}$$

3. [*Nonparametric Density Estimation*]

(a) Expression of $\hat{p}(x)$.

Proof:

By definition, the histogram estimator is defined as following:

$$\hat{p}(x) = \frac{\#\{x^{(l)} \textit{ in the same bin as } x\}}{nh} = \frac{\#\left\{x^{(l)} \in \left[\left\lfloor \frac{x}{h} \right\rfloor h, \left\lceil \frac{x}{h} \right\rceil h\right)\right\}}{nh}$$

(b) Expression of $L'(h)$ based on the histogram estimator $\hat{p}(x)$.
Proof:

First, for the first term of L', we can split the integral by bins:

$$\int_0^1 \hat{p}^2(x)dx = \sum_{j=0}^{\lfloor \frac{1}{h} \rfloor} \int_{jh}^{(j+1)h} \frac{\#^2\left\{x^{(l)} \in [jh,(j+1)h)\right\}}{n^2h^2}dx$$
$$= \frac{\sum_{j=0}^{\lfloor \frac{1}{h} \rfloor} \#^2\left\{x^{(l)} \in [jh,(j+1)h)\right\}}{n^2h}$$

For the second term of L', we can rewrite it as following:

$$\frac{2}{n}\sum_{i=1}^{n}\hat{p}(x_i) = \frac{2}{n}\sum_{i=1}^{n}\frac{\#\left\{x^{(l)} \in \left[\left\lfloor \frac{x}{h} \right\rfloor h, \left\lceil \frac{x}{h} \right\rceil h\right)\right\}}{nh}$$
$$= \frac{2\sum_{j=0}^{\lfloor \frac{1}{h} \rfloor} \#^2\left\{x^{(l)} \in [jh,(j+1)h)\right\}}{n^2h}$$

Hence, we have

$$L'(h) = \int_0^1 \hat{p}^2(x)dx - \frac{2}{n}\sum_{i=1}^{n}\hat{p}(x_i) = -\frac{\sum_{j=0}^{\lfloor \frac{1}{h} \rfloor} \#^2\left\{x^{(l)} \in [jh,(j+1)h)\right\}}{n^2h}$$

(c) h that minimizes $L'(h)$.
Proof:

Since as h goes to 0, as long as there are not coincide sample points, the numerator of $L'(h)$ will go to n. However, the denominator of $L'(h)$ is n^2h, which goes to 0 as $h \to 0$, thus $\lim_{h\to 0} L'(h) = -\infty$.

$$h = \arg minimize(L'(h)) = 0$$

4. [Nonparametric Regression]

(a) Estimated output $\hat{y}$ and is linear regression a linear smoother?
Solution:

Given that the least squares estimate for $\boldsymbol{w}$ is

$$\boldsymbol{w}^* = \left(H^TH\right)^{-1}H^TY$$

we have the following estimated output $\hat{y}$

$$\begin{aligned}
\hat{y} &= (\boldsymbol{w}^*)^T \cdot \boldsymbol{h}(\boldsymbol{x}) \\
&= \left[\left(H^TH\right)^{-1}H^TY\right]^T \boldsymbol{h}(\boldsymbol{x}) \\
&= Y^TH\left(H^TH\right)^{-1}\boldsymbol{h}(\boldsymbol{x}) \\
&= \left[H\left(H^TH\right)^{-1}\boldsymbol{h}(\boldsymbol{x})\right]^T Y \\
&= \boldsymbol{h}(\boldsymbol{x})^T\left(H^TH\right)^{-1}H^TY \\
&\Rightarrow
\end{aligned}$$
$$\boldsymbol{l}(\boldsymbol{x}) = H\left(H^TH\right)^{-1}\boldsymbol{h}(\boldsymbol{x})$$

Hence, linear regression is a linear smoother.

(b) In kernel regression, if we use kernel $K(x_i, x) = exp\left\{\frac{-||x_i - x||^2}{2\sigma^2}\right\}$, given an input x, please derive the estimated output $\hat{y}$. Furthermore, is this kernel regression a linear smoother?

Solution:

By the definition of Kernel regression, we have the estimated output $\hat{y}$ as following:

$$\begin{aligned}
\hat{y} &= \frac{\sum_{i=1}^{n} K(x_i, x) y_i}{\sum_{i=1}^{n} K(x_i, x)} \\
&= \frac{\sum_{i=1}^{n} \exp\left\{-\frac{||x_i - x||^2}{2\sigma^2}\right\} y_i}{\sum_{i=1}^{n} \exp\left\{-\frac{||x_i - x||^2}{2\sigma^2}\right\}} \\
&= \sum_{i=1}^{n} softmax\left(\frac{||x_i - x||^2}{2\sigma^2}\right) y_i
\end{aligned}$$

Then define that

$$S = \begin{bmatrix} softmax\left(\frac{||x_1 - x||^2}{2\sigma^2}\right) \\ softmax\left(\frac{||x_2 - x||^2}{2\sigma^2}\right) \\ \vdots \\ softmax\left(\frac{||x_n - x||^2}{2\sigma^2}\right) \end{bmatrix}$$

we have

$$\begin{aligned}
\hat{y} &= \sum_{i=1}^{n} softmax\left(\frac{||x_i - x||^2}{2\sigma^2}\right) y_i \\
&= S^T Y \\
&\Rightarrow
\end{aligned}$$

$$\boldsymbol{l(x)} = S = \begin{bmatrix} softmax\left(\frac{||x_1 - x||^2}{2\sigma^2}\right) \\ softmax\left(\frac{||x_2 - x||^2}{2\sigma^2}\right) \\ \vdots \\ softmax\left(\frac{||x_n - x||^2}{2\sigma^2}\right) \end{bmatrix}$$

Hence, this kernel regression is a linear smoother.