Tsinghua-Berkeley Shenzhen Institute
LEARNING FROM DATA
Fall 2019

Homework 1

TIAN Chenyu — October 8, 2019

- **Acknowledgments:** This template takes some materials from course CSE 547/Stat 548 of Washington University: `https://courses.cs.washington.edu/courses/cse547/17sp/index.html`.
- **Collaborators:** I finish this template by myself.

1.1. The log-likelihood of the softmax regression model can be writen as

$$\begin{aligned} l &= \sum_{i=1}^{m} \log \frac{\exp(\boldsymbol{\theta}_{y^{(i)}}^T \boldsymbol{x}^{(i)} + b_{y^{(i)}})}{\sum_{j=1}^{k} \exp(\boldsymbol{\theta}_j^T \boldsymbol{x}^{(i)} + b_j)} \\ &= \sum_{i=1}^{m} [\boldsymbol{\theta}_{y^{(i)}}^T \boldsymbol{x}^{(i)} + b_{y^{(i)}} - \log(\sum_{j=1}^{k} \exp(\boldsymbol{\theta}_j^T \boldsymbol{x}^{(i)} + b_j))] \end{aligned}$$

(a) Evaluate the derivation of b_l:

$$\frac{\partial\ l}{\partial\ b_l} = \sum_{i}^{m} [\mathbb{1}(y^{(i)} = l) - \frac{\exp(\boldsymbol{\theta}_l^T\ \boldsymbol{x}^{(i)} + b_l)}{\sum_{j\ =\ 1}^{k}\ \exp(\boldsymbol{\theta}_j^T\ \boldsymbol{x}^{(i)} + b_j)}]$$

The $\mathbb{1}(y^{(i)} = l)$ function is defined as:

$$\mathbb{1}(x = l) = \begin{cases} 1, & \text{if } x = l, \\ 0, & \text{if } x \neq l. \end{cases}$$

(b) If we have set the biases to their optimal values, there exists $\frac{\partial\ l}{\partial\ b_l} = 0$. Based on (a):

$$\sum_{i}^{m}\ \mathbb{1}(y^{(i)} = l) = \frac{\exp(\boldsymbol{\theta}_l^T\ \boldsymbol{x}^{(i)} + b_l)}{\sum_{j\ =\ 1}^{k}\ \exp(\boldsymbol{\theta}_j^T\ \boldsymbol{x}^{(i)} + b_j)}$$

Based on the definition of $\hat{P}_y\ (l)$:

$$\begin{aligned} \hat{P}_y(l) &= \frac{1}{m} \sum_{i}^{m}\ \frac{\exp(\boldsymbol{\theta}_l^T\ \boldsymbol{x}^{(i)} + b_l)}{\sum_{j\ =\ 1}^{k}\ \exp(\boldsymbol{\theta}_j^T\ \boldsymbol{x}^{(i)} + b_j)}] \\ &= \frac{1}{m} \sum_{i}^{m} \sum_{\boldsymbol{x} \in \boldsymbol{X}} \frac{\exp(\boldsymbol{\theta}_l^T\ \boldsymbol{x} + b_l)}{\sum_{j\ =\ 1}^{k}\ \exp(\boldsymbol{\theta}_j^T\ \boldsymbol{x}^{(i)} + b_j)}]\, \mathbb{1}(\boldsymbol{x}^{(i)} = \boldsymbol{x}) \\ &= \frac{1}{m} \sum_{i}^{m} \sum_{\boldsymbol{x} \in \boldsymbol{X}} P_{(y|\boldsymbol{x})}(l \mid \boldsymbol{x})\, \mathbb{1}(\boldsymbol{x}^{(i)} = \boldsymbol{x}) \\ &= \sum_{\boldsymbol{x} \in \boldsymbol{X}} P_{(y|\boldsymbol{x})}(l \mid \boldsymbol{x}) \frac{1}{m} \sum_{i}^{m} \mathbb{1}(\boldsymbol{x}^{(i)} = \boldsymbol{x}) \\ &= \sum_{\boldsymbol{x} \in \boldsymbol{X}} P_{(y|\boldsymbol{x})} \hat{P}_{\boldsymbol{x}}(\boldsymbol{x}) \end{aligned}$$

1.2. (a) The MLE is to solve

$$\underset{\mu}{\text{maximize}} \prod_{i}^{n} P(x_i \mid \mu)$$

which is equal to:

$$\begin{aligned}
&\underset{\mu}{\text{maximize}} \sum_{i}^{n} \log P(x_i \mid \mu) \\
\Rightarrow\ &\underset{\mu}{\text{maximize}} \ \sum_{i}^{n} \ (\log \ \frac{1}{\sqrt{2 \ \pi \ \delta^2}} \ - \ \frac{(x \ - \ \mu)^2}{2 \ \delta^2}) \\
\Rightarrow\ &\underset{\mu}{\text{minimize}} \ \sum_{i}^{n} \ (x_i \ - \ \mu)^2
\end{aligned}$$

Here we define $f(\mu) = \sum_i^n \ (x_i \ - \ \mu)^2$ to find the μ^*. There exists $\frac{\partial \ f}{\partial \ \mu} = 2\sum_i^n(\mu - x_i) = 0$ when $\mu{=}\mu^*$.
Then $\mu^* = \frac{\sum_i^n x_i}{n}$.

(b) The MAP problem can be writen as:

$$\underset{\mu}{\text{maximize}} \, P(\mu \mid x_1, ..., x_n)$$

which is equal to:

$$\begin{aligned}
&\underset{\mu}{\text{maximize}} \, P(\mu \mid x_1, ..., x_n) \\
\Rightarrow\ &\underset{\mu}{\text{maximize}} \ \ P(x_1, ..., x_n \mid \mu)P(\mu) \\
\Rightarrow\ &\underset{\mu}{\text{maximize}} \ \prod_{i}^{n} P(x_i \mid \mu)P(\mu) \\
\Rightarrow\ &\underset{\mu}{\text{maximize}} \ \sum_{i}^{n} \log P(x_i \mid \mu) + \log P(\mu) \\
\Rightarrow\ &\underset{\mu}{\text{maximize}} \ \sum_{i}^{n} \ -\frac{(x_i - \mu)^2}{2\delta^2} - \log \ \sqrt{2\pi\theta^2} \ - \ \frac{(\mu - \nu)^2}{2 \ \delta^2} \\
\Rightarrow\ &\underset{\mu}{\text{minimize}} \ \sum_{i}^{n} \ \frac{(x_i - \mu)^2}{2\delta^2} + \frac{(\mu - \nu)^2}{2 \ \theta^2}
\end{aligned}$$

Here we define $g(\mu) = \sum_i^n \ \frac{(x_i - \mu)^2}{2\delta^2} + \frac{(\mu-\nu)^2}{2 \ \theta^2}$ to find the μ^*. There exists $\frac{\partial \ g}{\partial \ \mu} = \sum_i^n \ \frac{\mu(\mu - x_i)}{\delta^2} + \frac{\mu(\mu-\nu)}{\theta^2} = 0$ when $\mu{=}\mu^*$.
Then $\mu^* = \frac{\sum_i^n x_i + \delta^2\nu}{\theta^2 n + \delta^2}$.
When $n \to \infty$, μ^* for MLE and MAP is equal.

1.3. First, the square error can be written as

$$\begin{aligned} J(\boldsymbol{\Theta}) &= \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{l}((\boldsymbol{\Theta}^T\boldsymbol{x}^{(i)})_j - \boldsymbol{y}_j^{(i)})^2 \\ &= \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{l}(\sum_{k=1}^{n}\boldsymbol{\Theta}_{kj}\boldsymbol{x}_k^{(i)} - \boldsymbol{y}_j^{(i)})^2 \end{aligned}$$

In order to compute the solution, it needs to find the minimum $J(\Theta)$ where $\frac{\partial\ J}{\partial\ \Theta} = 0$. To find the solution, the derivative is:

$$\begin{aligned} \frac{\partial\ J}{\partial\ \boldsymbol{\Theta}_{\alpha\beta}} &= \sum_{i=1}^{m}(\boldsymbol{x}_\alpha^{(i)}(\sum_{k}^{n}\boldsymbol{\Theta}_{k\beta}\boldsymbol{x}_k^{(i)} - \boldsymbol{y}_\beta^{(i)})) \\ &= \sum_{i=1}^{m}(\boldsymbol{x}_\alpha^{(i)}(\boldsymbol{\Theta}_\beta^T\boldsymbol{x}^{(i)} - \boldsymbol{y}_\beta^{(i)})) \end{aligned}$$

Then we have $\frac{\partial\ J}{\partial\ \Theta} = 0$, which means:

$$\begin{aligned} \sum_{i=1}^{m}\boldsymbol{x}_\alpha^{(i)}(\boldsymbol{\Theta}_\beta^T\boldsymbol{x}^{(i)}) &= \sum_{i=1}^{m}\boldsymbol{x}_\alpha^{(i)}\boldsymbol{y}_\beta^{(i)}) \\ \Rightarrow\ \boldsymbol{X}_\alpha^T\boldsymbol{X}\boldsymbol{\Theta}_\beta &= \boldsymbol{X}_\alpha^T\boldsymbol{Y}_\beta \\ \Rightarrow\ \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\Theta}_\beta &= \boldsymbol{X}^T\boldsymbol{Y}_\beta \end{aligned}$$

So, for $\beta \in (1, 2, ..., l)$ we have

$$\boldsymbol{\Theta}_\beta = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}_\beta$$

Also,

$$\boldsymbol{\Theta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

1.4. Based on Σ is symmetrical as well as the properties of matrix trace, the multivariate normal distribution can be written as

$$\begin{aligned} P_{\boldsymbol{y}}(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Sigma}) &= \frac{1}{(2\pi)^{\frac{n}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}}\exp(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})) \\ &= (2\pi)^{-\frac{n}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp(-\frac{1}{2}(\boldsymbol{y}^T\boldsymbol{\Sigma}^{-1} - \boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1})(\boldsymbol{y}-\boldsymbol{\mu})) \\ &= (2\pi)^{-\frac{n}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp(-\frac{1}{2}(\boldsymbol{y}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{y} - \boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}) \\ &= (2\pi)^{-\frac{n}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp(-\frac{1}{2}tr(\boldsymbol{y}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{y} - \boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}) \\ &= (2\pi)^{-\frac{n}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp(-\frac{1}{2}(tr(\boldsymbol{y}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{y}) - tr(\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{y}) - tr(\boldsymbol{y}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}) + tr(\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})) \\ &= (2\pi)^{-\frac{n}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp(tr(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})\boldsymbol{y} - \frac{1}{2}tr(\boldsymbol{\Sigma}^{-1}\boldsymbol{y}\boldsymbol{y}^T) - \frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}) \end{aligned}$$

Thus, we can see that multivariate normal distribution is an exponential family with:

$$
\begin{aligned}
\eta &= \begin{pmatrix} \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \\ -\frac{1}{2}\boldsymbol{\Sigma}^{-1} \end{pmatrix} \\
b(\boldsymbol{y}) &= (2\pi)^{-\frac{n}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}} \\
T(\boldsymbol{y}) &= \begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{y}\boldsymbol{y}^T \end{pmatrix} \\
a(\eta) &= \frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}
\end{aligned}
$$