Tsinghua-Berkeley Shenzhen Institute
LEARNING FROM DATA
Fall 2018

Homework 2

Zi-Feng WANG October 28, 2018

- **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 my homework all by myself.

2.1. **Solution:** let $\nabla J(\theta) = 0$ and we can get the normal equation

$$X^{\mathrm{T}} X \boldsymbol{\theta} = X^{\mathrm{T}} \boldsymbol{y} \tag{1}$$

from $x \in \mathbb{R}^n$ we can get that the X is a $m \times n$ matrix, hence $X^{\mathrm{T}} X$ is a $n \times n$ matrix , $\boldsymbol{\theta}$ is $n \times 1$ vector and $X^{\mathrm{T}} \boldsymbol{y}$ is $n \times 1$ vector as well. the (1) can be represented as the following format

$$A\boldsymbol{\theta} = \boldsymbol{b} \tag{2}$$

in (2), the $A = X^{\mathrm{T}} X$, $b = X^{\mathrm{T}} \boldsymbol{y}$. if A is a singular and square matrix, the $rank(A) < n$. there are two possible cases:

(a) when $rank(A) = rank(A, b) < n$, there are infinite number of solutions of $\boldsymbol{\theta}$

(b) when $rank(A) < rank(A, b) < n$, there is none solution of $\boldsymbol{\theta}$

2.2. **Solution:**

(a) before getting $\nabla_{b_l} \ell$, unfold ℓ at first

$$\begin{aligned}
\ell &= \sum_{i=1}^{m} \log P_{\mathsf{y}|\mathsf{x}}(y^{(i)} | x^{(i)}) \\
&= \sum_{i=1}^{m} \log \frac{\exp(\theta_l^{\mathrm{T}} x^{(i)} + b_l)}{\sum_{j=1}^{k} \exp(\theta_j^{\mathrm{T}} x^{(i)} + b_j)} \\
&= \sum_{i=1}^{m} (\theta_l^{\mathrm{T}} x^{(i)} + b_l) - \sum_{i=1}^{m} \log \sum_{j=1}^{k} \exp(\theta_j^{\mathrm{T}} x^{(i)} + b_j)
\end{aligned} \tag{3}$$

from (3), the $\nabla_{b_l}\ell$ is

$$
\begin{aligned}
\nabla_{b_l}\ell &= \nabla_{b_l}\{\sum_{i=1}^{m}(\theta_l^{\mathrm{T}}x^{(i)}+b_l) - \sum_{i=1}^{m}\log\sum_{j=1}^{k}\exp(\theta_j^{\mathrm{T}}x^{(i)}+b_j)\} \\
&= \sum_{i=1}^{m}\mathbb{1}\{y^{(i)}=l\} - \sum_{i=1}^{m}\frac{1}{\sum_{j=1}^{k}\exp(\theta_j^{\mathrm{T}}x+b_j)}\frac{\partial(\sum_{j=1}^{k}\exp(\theta_j^{\mathrm{T}}x^{(i)}+b_j))}{\partial b_l} \\
&= \sum_{i=1}^{m}\mathbb{1}\{y^{(i)}=l\} - \sum_{i=1}^{m}\frac{\theta_l^{\mathrm{T}}x^{(i)}+b_l}{\sum_{j=1}^{k}\exp(\theta_j^{\mathrm{T}}x^{(i)}+b_j)} \\
&= \sum_{i=1}^{m}(\mathbb{1}\{y^{(i)}=l\} - P(y^{(i)}=l|x^{(i)};\theta,b))
\end{aligned}
\tag{4}
$$

(b) *Proof.* when $\nabla_{b_l}\ell = 0$, the (4) = 0, that is $\sum_{i=1}^{m}(\mathbb{1}\{y^{(i)}=l\} - P(y^{(i)}=l|x^{(i)};\theta,b)) = 0$.
so that

$$
\begin{aligned}
\sum_{i=1}^{m}\mathbb{1}\{y^{(i)}=l\} &= \sum_{i=1}^{m}P(y^{(i)}=l|x^{(i)};\theta,b) \\
m\frac{\sum_{i=1}^{m}\mathbb{1}\{y^{(i)}=l\}}{m} &= \sum_{i=1}^{m}\sum_{j=1}^{m}P(y^{(i)}=l|x^{(i)};\theta,b)\,\mathbb{1}\{x^{(i)}=x^{(j)}\} \\
\hat{P}_{\mathsf{y}}(l) &= \frac{1}{m}\sum_{i=1}^{m}P(y^{(i)}=l|x^{(i)};\theta,b)\sum_{j=1}^{m}\mathbb{1}\{x^{(i)}=x^{(j)}\} \\
\hat{P}_{\mathsf{y}}(l) &= \sum_{x\in\mathcal{X}}P_{\mathsf{y}|\mathsf{x}}(l|x)\hat{P}_{\mathsf{x}}(x)
\end{aligned}
$$

□

2.3. the multivariate normal distribution can be written as

$$
\begin{aligned}
p_{\mathsf{y}}(y;\mu,\Sigma) &= \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}}\exp(-\frac{1}{2}(y-\mu)^{\mathrm{T}}\Sigma^{-1}(y-\mu)) \\
&= \frac{1}{(2\pi)^{\frac{n}{2}}}\exp\{-\frac{1}{2}\log|\Sigma| - \frac{1}{2}(y-\mu)^{\mathrm{T}}\Sigma^{-1}(y-\mu)\}
\end{aligned}
\tag{5}
$$

the part in $\exp(\cdot)$ of (5) is

$$
\begin{aligned}
&-\frac{1}{2}\log|\Sigma| - \frac{1}{2}(y-\mu)^{\mathrm{T}}\Sigma^{-1}(y-\mu) \\
&= -\frac{1}{2}\log|\Sigma| - \frac{1}{2}(y^{\mathrm{T}}\Sigma^{-1}y - y^{\mathrm{T}}\Sigma^{-1}\mu - \mu^{\mathrm{T}}\Sigma^{-1}y + \mu^{\mathrm{T}}\Sigma^{-1}\mu) \\
&= -\frac{1}{2}(tr(y^{\mathrm{T}}\Sigma^{-1}y) - tr(y^{\mathrm{T}}\Sigma^{-1}\mu) - tr(\mu^{\mathrm{T}}\Sigma^{-1}y)) - \frac{1}{2}(\log|\Sigma| + \mu^{\mathrm{T}}\Sigma^{-1}\mu) \\
&= -\frac{1}{2}(tr(\Sigma^{-1}yy^{\mathrm{T}}) - tr(\Sigma^{-1}\mu y^{\mathrm{T}}) - tr(\mu^{\mathrm{T}}\Sigma^{-1}y)) - \frac{1}{2}(\log|\Sigma| + \mu^{\mathrm{T}}\Sigma^{-1}\mu) \\
&= -\frac{1}{2}tr(\Sigma^{-1}yy^{\mathrm{T}} - 2\mu^{\mathrm{T}}\Sigma^{-1}y) - \frac{1}{2}(\log|\Sigma| + \mu^{\mathrm{T}}\Sigma^{-1}\mu)
\end{aligned}
\tag{6}
$$

let

$$\begin{aligned}\boldsymbol{\eta} &= (\mathbf{vec}^{\mathrm{T}}(-\frac{1}{2}\Sigma^{-1}), \mu^{\mathrm{T}}\Sigma^{-1})^{\mathrm{T}} \\ T(y) &= (\mathbf{vec}(yy^{\mathrm{T}}), y)^{\mathrm{T}}\end{aligned} \tag{7}$$

therefore the (6) can be written as

$$tr(\boldsymbol{\eta}^{\mathrm{T}}T(y)) - \frac{1}{2}(\log|\Sigma| + \mu^{\mathrm{T}}\Sigma^{-1}\mu) = \langle\boldsymbol{\eta}, T(y)\rangle_F{}^1 - a(\boldsymbol{\eta}) \tag{8}$$

combine the (5), (6), (7) and (8), we can show that the multivariate normal distribution is an exponential family

$$P_{\mathsf{y}}(y;\boldsymbol{\eta}) = b(y)\exp(\langle\boldsymbol{\eta}, T(y)\rangle_F - a(\boldsymbol{\eta})) \tag{9}$$

where

$$\begin{aligned}b(y) &= \frac{1}{(2\pi)^2} \\ \boldsymbol{\eta} &= (\mathbf{vec}^{\mathrm{T}}(-\frac{1}{2}\Sigma^{-1}), \mu^{\mathrm{T}}\Sigma^{-1})^{\mathrm{T}} \\ T(y) &= (\mathbf{vec}(yy^{\mathrm{T}}), y)^{\mathrm{T}} \\ a(\boldsymbol{\eta}) &= \frac{1}{2}(\log|\Sigma| + \mu^{\mathrm{T}}\Sigma^{-1}\mu)\end{aligned}$$

[1] the $\langle A, B\rangle_F$ is the *Frobenius inner product* used to define the inner product between two matrices A and B, which is represented as the trace of their products i.e. $tr(A^{\mathrm{T}}B)$.