Tsinghua-Berkeley Shenzhen Institute
LEARNING FROM DATA
Fall 2018

Homework 2

YOUR NAME: Guoqing Zhang October 27, 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.

  If you refer to other materials in your homework, please list here.

- **Collaborators:** I finish this homework by myself.But wiki (*Frobeniusinnerproduct*) and the Andrew Ng's pdf(*cs229 – notes2*) helped me a lot about solving the third question. I also imitate the format of Gaussian distribution($\sigma \neq 1$) in lecture4 to get the answer. But in fact I don't understand the third one very well.

2.1. We can always meet the situation that the dimension of $\mathbf{x}$ is bigger than the number of the sample.Or the rank of matrix$\mathbf{X}$ is less than its dimension.In this situation the X^TX is singular. In that case the equation is still right. The proof will be like follows:
In that case, we could find a lot θ to satisfy the equation: $\mathbf{y} = \mathbf{X}\theta$ (1) .
If we put (1) into the $\mathbf{X}^T\mathbf{y}$,we will get $\mathbf{X}^T\mathbf{X}\theta$, so of couse the old equation is right.
I don't know the other cases, so the proof may not be strict.

2.2. (a) Firstly, we unfold the function ℓ:

$$\begin{aligned}
\ell &= \sum_{i=1}^{m} \log P_{y|\mathbf{x}}(y_{(i)}|\mathbf{x}^{(i)}) \\
&= \sum_{i=1}^{m} \log \prod_{l=1}^{k} \left(\frac{exp(\theta_l^T\mathbf{x}^{(i)} + b_l)}{\sum_{j=1}^{k} exp(\theta_j^T\mathbf{x}^{(i)} + b_j)} \right)^{\mathbf{1}\{y^{(i)}=l\}} \\
&= \sum_{i=1}^{m} \sum_{l=1}^{k} \mathbf{1}\{y^{(i)} = l\} \log \left(\frac{exp(\theta_l^T\mathbf{x}^{(i)} + b_l)}{\sum_{j=1}^{k} exp(\theta_j^T\mathbf{x} + b_j)} \right)
\end{aligned}$$

so,if we want get the derivative of b_l,the ℓ will can also be wrote like this:

$$\begin{aligned}
\ell &= \sum_{i=1}^{m} \left(\mathbf{1}\{y^{(i)} = l\} \log \left(\frac{exp(\theta_l^T\mathbf{x}^{(i)} + b_l)}{\sum_{j=1}^{k} exp(\theta_j^T\mathbf{x}^{(i)} + b_j)} \right) + \mathbf{1}\{y^{(i)} \neq l\} \log \left(\frac{exp(\theta_{y^{(i)}}^T\mathbf{x}^{(i)} + b_{y^{(i)}})}{\sum_{j=1}^{k} exp(\theta_j^T\mathbf{x}^{(i)} + b_j)} \right) \right) \\
&= \sum_{i=1}^{m} (\mathbf{1}\{y^{(i)} = l\} f_1(b_l) + \mathbf{1}\{y^{(i)} \neq l\} f_2(b_l))
\end{aligned}$$

$$f_1(b_l) = \log \frac{exp(\theta_l^T \mathbf{x}^{(i)} + b_l)}{\sum_{j=1}^{k} exp(\theta_j^T \mathbf{x}^{(i)} + b_j)}, f_2(b_l) = \log \frac{exp(\theta_{y^{(i)}}^T \mathbf{x}^{(i)} + b_{y^{(i)}})}{\sum_{j=1}^{k} exp(\theta_j^T \mathbf{x}^{(i)} + b_j)}.$$

so we could get the derivative respectively.

$$\begin{aligned}
\nabla_{b_l} f_1 &= \frac{\sum_{j=1}^{k} exp(\theta_j^T \mathbf{x}^{(i)} + b_j)}{exp(\theta_l^T \mathbf{x}^{(i)} + b_l)} \cdot \frac{exp(\theta_l^T \mathbf{x}^{(i)} + b_l) \cdot \sum_{j=1}^{k} exp(\theta_j^T \mathbf{x}^{(i)} + b_j) - exp(2 \cdot (\theta_l^T \mathbf{x}^{(i)} + b_l))}{(\sum_{j=1}^{k} exp(\theta_j^T \mathbf{x}^{(i)} + b_j))^2} \\
&= 1 - \frac{exp(\theta_l^T \mathbf{x}^{(i)} + b_l)}{\sum_{j=1}^{k} exp(\theta_j^T \mathbf{x}^{(i)} + b_j)} \\
&= 1 - P_{y|\mathbf{x}}(l|\mathbf{x}^{(i)})
\end{aligned}$$

For the case "$y^{(i)} \neq l$",we assume that $y^{(i)} = v$.So the part of f_2 will be as follows:

$$\begin{aligned}
\nabla_{b_l} f_2 &= \frac{\sum_{j=1}^{k} exp(\theta_j^T \mathbf{x}^{(i)} + b_j)}{exp(\theta_v^T \mathbf{x}^{(i)} + b_v)} \cdot \frac{-exp(\theta_l^T \mathbf{x}^{(i)} + b_l) \cdot exp(\theta_v^T \mathbf{x}^{(i)} + b_v)}{(\sum_{j=1}^{k} exp(\theta_j^T \mathbf{x}^{(i)} + b_j))^2} \\
&= -\frac{exp(\theta_l^T \mathbf{x}^{(i)} + b_l)}{\sum_{j=1}^{k} exp(\theta_j^T \mathbf{x}^{(i)} + b_j)} \\
&= -P_{y|\mathbf{x}}(l|\mathbf{x}^{(i)})
\end{aligned}$$

To sum up,we can get:

$$\nabla_{b_l} \ell = \sum_{i=1}^{m} (\mathbf{1}\{y_i = l\} - P_{y|\mathbf{x}}(l|\mathbf{x}^{(i)}))$$

(b) Because we get the optimal $(b_1, ..., b_k)$,so we could know that:

$$\nabla_{b_l} \ell = \sum_{i=1}^{m} (\mathbf{1}\{y_i = l\} - P_{y|\mathbf{x}}(l|\mathbf{x}^{(i)})) = 0$$

Because we know that the training dataset should not contain any duplicate samples. So we can get $\hat{P}_{\mathbf{x}}(\mathcal{X}) = \frac{1}{m}$.Then we get:

$$\begin{aligned}
&\sum_{i=1}^{m} \mathbf{1}\{y_i = l\} = \sum_{i=1}^{m} P_{y|\mathbf{x}}(l|\mathbf{x}^{(i)}) \\
&\frac{1}{m} \sum_{i=1}^{m} \mathbf{1}\{y_i = l\} = \frac{1}{m} \sum_{i=1}^{m} P_{y|\mathbf{x}}(l|\mathbf{x}^{(i)}) \\
&\hat{P}_y(l) = \sum_{i=1}^{m} \frac{1}{m} P_{y|\mathbf{x}}(l|\mathbf{x}^{(i)}) = \sum_{i=1}^{m} P_{y|\mathbf{x}}(l|\mathbf{x}^{(i)}) \cdot \hat{P}_{\mathbf{x}}(\mathbf{x}_i)
\end{aligned}$$

So:

$$\hat{P}_y(l) = \sum_{\mathbf{x} \in \mathcal{X}} P_{y|\mathbf{x}}(l|\mathbf{x}) \hat{P}_{\mathbf{x}}(\mathcal{X})$$

In some cases, the training dataset may have some duplicate samples. In that case, the set $\mathcal{X}$'s size will be less than m,but the duplicate samples' $\hat{P}_{\mathbf{x}}(\mathcal{X})$ will

be bigger than $\frac{1}{m}$, because of this offset, $\sum_{\mathbf{x}\in\mathcal{X}} P_{y|\mathbf{x}}(l|\mathbf{x})\hat{P}_{\mathbf{x}}(\mathcal{X})$ will still be equal to $\sum_{i=1}^{m} P_{y|\mathbf{x}}(l|\mathbf{x}^{(i)})$, so the conclusion will not be changed.

2.3. Like on the class, firstly I consider the situation that $\sum^{-1} = \mathbf{I}$, so the probability density function will be as follow:

$$\begin{aligned}
p_{\mathbf{y}}(\mathbf{y};\mu) &= \frac{1}{(2\pi)^{\frac{n}{2}}} exp(-\frac{1}{2}(\mathbf{y}-\mu)^T(\mathbf{y}-\mu)) \\
&= \frac{1}{(2\pi)^{\frac{n}{2}}} exp(-\frac{1}{2}\|\mathbf{y}-\mu\|^2) \\
&= \frac{1}{(2\pi)^{\frac{n}{2}}} exp(-\frac{1}{2}\|\mathbf{y}\|^2 + \mathbf{y}^T\mu - \frac{1}{2}\|\mu\|^2) \\
&= \frac{1}{(2\pi)^{\frac{n}{2}}} exp(-\frac{1}{2}\|\mathbf{y}\|^2) \cdot exp(\mathbf{y}^T\mu - \frac{1}{2}\|\mu\|^2)
\end{aligned}$$

So in this case :

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

In general:

$$\begin{aligned}
p_{\mathbf{y}}(\mathbf{y};\mu,\Sigma) &= \frac{1}{(2\pi)^{\frac{n}{2}}} exp(-\frac{1}{2}(\mathbf{y}-\mu)^T\Sigma^{-1}(\mathbf{y}-\mu)) \\
&= \frac{1}{(2\pi)^{\frac{n}{2}}} exp(-\frac{1}{2}\mathbf{y}^T\Sigma^{-1}\mathbf{y} + \mathbf{y}^T\Sigma^{-1}\mu - \frac{1}{2}\mu^T\Sigma^{-1}\mu) \\
&= \frac{1}{(2\pi)^{\frac{n}{2}}} exp(-\frac{1}{2}\mathbf{vec}^T(\Sigma^{-1}) \cdot \mathbf{vec}(\mathbf{y}\mathbf{y}^T) + \mu^T\Sigma^{-1}\mathbf{y} - \frac{1}{2}\mu^T\Sigma^{-1}\mu)
\end{aligned}$$

So in general cases :

$$\begin{aligned}
&\eta = \begin{bmatrix} u^T\Sigma^{-1} \\ -\frac{1}{2}\mathbf{vec}(\Sigma^{-1})\mu \end{bmatrix}_{(n+n^2)\times 1} \\
&b(\mathbf{y}) = \frac{1}{(2\pi)^{\frac{n}{2}}} \\
&T(\mathbf{y}) = \begin{bmatrix} \mathbf{y} \\ \mathbf{vec}(\mathbf{y}\mathbf{y}^T)\mu \end{bmatrix}_{(n+n^2)\times 1} \\
&a(\eta) = \frac{1}{2}\mu^T\Sigma^{-1}\mu
\end{aligned}$$

I am really not sure about the answer. I just watched the wiki about Frobenius inner product.And I saw the operator : **vec**. The answer is what I guessed to make sure they could dot to get a number.