Tsinghua-Berkeley Shenzhen Institute
LEARNING FROM DATA
Fall 2020

Writing Assignment 1

Issued: Wednesday 30th September, 2020 **Due:** Friday 16th October, 2020

1.1. (Multivariate Least Squares) A data set consists of m data pairs $(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}), i = 1, \ldots, m$, where $\boldsymbol{x} \in \mathbb{R}^n$ is the independent variable, and $\boldsymbol{y} \in \mathbb{R}^l$ is the dependent variable. Denote the design matrix by $\boldsymbol{X} \overset{\text{def}}{=} [\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}]^{\mathrm{T}}$, and let $\boldsymbol{Y} \overset{\text{def}}{=} [\boldsymbol{y}^{(1)}, \cdots, \boldsymbol{y}^{(m)}]^{\mathrm{T}}$. Please compute the optimal solution for $\boldsymbol{\Theta}$, where $\boldsymbol{\Theta} \in \mathbb{R}^{n \times l}$ is the parameter matrix you want to get, and $J(\boldsymbol{\Theta})$ is the square loss.

Hint: Hopefully you can write down the square loss without confusion. Just in case, we will write it as

$$J(\boldsymbol{\Theta}) = \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{l} \left((\boldsymbol{\Theta}^{\mathrm{T}} \boldsymbol{x}^{(i)})_j - \boldsymbol{y}_j^{(i)} \right)^2$$

Solution: (1 point)It can be written down as a trace

$$\begin{aligned} J(\boldsymbol{\Theta}) &= \frac{1}{2} \operatorname{Tr}\left((\boldsymbol{X}\boldsymbol{\Theta} - \boldsymbol{Y})^{\mathrm{T}} (\boldsymbol{X}\boldsymbol{\Theta} - \boldsymbol{Y}) \right) \\ &= \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{l} (\boldsymbol{X}\boldsymbol{\Theta} - \boldsymbol{Y})_{ij}^2 \end{aligned}$$

(1 point)Then do the derivative

$$\begin{aligned} \nabla_{\boldsymbol{\Theta}} J(\boldsymbol{\Theta}) &= \frac{1}{2} \nabla_{\boldsymbol{\Theta}} \left[\operatorname{Tr}\left(\boldsymbol{\Theta}^{\mathrm{T}} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{X} \boldsymbol{\Theta} \right) - \operatorname{Tr}\left(\boldsymbol{\Theta}^{\mathrm{T}} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{Y} \right) - \operatorname{Tr}\left(\boldsymbol{Y}^{\mathrm{T}} \boldsymbol{X} \boldsymbol{\Theta} \right) + \operatorname{Tr}\left(\boldsymbol{Y}^{\mathrm{T}} \boldsymbol{Y} \right) \right] \\ &= \frac{1}{2} \left[\boldsymbol{X}^{\mathrm{T}} \boldsymbol{X} \boldsymbol{\Theta} + \boldsymbol{X}^{\mathrm{T}} \boldsymbol{X} \boldsymbol{\Theta} - 2 \boldsymbol{X}^{\mathrm{T}} \boldsymbol{Y} \right] \\ &= \boldsymbol{X}^{\mathrm{T}} \boldsymbol{X} \boldsymbol{\Theta} - \boldsymbol{X}^{\mathrm{T}} \boldsymbol{Y} \\ &= 0 \end{aligned}$$

(0.5 point)Therefore, the solution is

$$\boldsymbol{\Theta} = \left(\boldsymbol{X}^{\mathrm{T}} \boldsymbol{X} \right)^{-1} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{Y}$$

1.2. (Softmax Regression) In multivariate classification problem, we use softmax function to derive the likelihood of each possible label y and predict the most probable one for data $\boldsymbol{x} \in \mathbb{R}^n$. To train parameter matrix $\boldsymbol{\Theta} \in \mathbb{R}^{n \times k}$ from the given samples $\left(\boldsymbol{x}^{(i)}, y^{(i)} \right), i = 1, \ldots, m$, we need to calculate the derivative of the softmax model's log-likelihood function

$$\ell(\boldsymbol{\Theta}) \overset{\text{def}}{=} \sum_{i=1}^{m} \log p(y^{(i)} | \boldsymbol{x}^{(i)}; \boldsymbol{\Theta}) = \sum_{i=1}^{m} \sum_{l=1}^{k} \mathbf{1}\left\{ y^{(i)} = l \right\} \log \frac{e^{\boldsymbol{\theta}_l^{\mathrm{T}} \boldsymbol{x}^{(i)}}}{\sum_{j=1}^{k} e^{\boldsymbol{\theta}_j^{\mathrm{T}} \boldsymbol{x}^{(i)}}}.$$

Calculate $\nabla_{\boldsymbol{\theta}_l}\ell(\boldsymbol{\Theta})$.

Hint: The index number of samples has nothing to do with $\boldsymbol{\theta}_l$, thus you just need to calculate $\nabla_{\boldsymbol{\theta}_l}\log p(y^{(i)}|\boldsymbol{x}^{(i)};\boldsymbol{\Theta})$ and sum them up. Indicator function $\mathbf{1}\left\{y^{(i)}=l\right\}=0$ when $y^{(i)}\neq l$, thus only one term in $\nabla_{\boldsymbol{\theta}_l}\log p(y^{(i)}|\boldsymbol{x}^{(i)};\boldsymbol{\Theta})$ will be left.

Solution:

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}_t}\ell(\boldsymbol{\Theta}) &= \nabla_{\boldsymbol{\theta}_t}\sum_{i=1}^{m}\sum_{l=1}^{k}\mathbf{1}\left\{y^{(i)}=t\right\}\log\frac{e^{\boldsymbol{\theta}_l^{\mathrm{T}}\boldsymbol{x}^{(i)}}}{\sum_{j=1}^{k}e^{\boldsymbol{\theta}_j^{\mathrm{T}}\boldsymbol{x}^{(i)}}}\\
&= \sum_{i=1}^{m}\mathbf{1}\left\{y^{(i)}=t\right\}\nabla_{\boldsymbol{\theta}_t}\log\frac{e^{\boldsymbol{\theta}_t^{\mathrm{T}}\boldsymbol{x}^{(i)}}}{\sum_{j=1}^{k}e^{\boldsymbol{\theta}_j^{\mathrm{T}}\boldsymbol{x}^{(i)}}}\\
&= \sum_{i=1}^{m}\mathbf{1}\left\{y^{(i)}=t\right\}\nabla_{\boldsymbol{\theta}_t}(\boldsymbol{\theta}_t^{\mathrm{T}}\boldsymbol{x}^{(i)}-\log\sum_{j=1}^{k}e^{\boldsymbol{\theta}_j^{\mathrm{T}}\boldsymbol{x}^{(i)}})\\
&= \sum_{i=1}^{m}\mathbf{1}\left\{y^{(i)}=t\right\}(\boldsymbol{x}^{(i)}-\frac{e^{\boldsymbol{\theta}_t^{\mathrm{T}}}\boldsymbol{x}^{(i)}}{\sum_{j=1}^{k}e^{\boldsymbol{\theta}_j^{\mathrm{T}}\boldsymbol{x}^{(i)}}}\boldsymbol{x}^{(i)})\\
&= \sum_{i=1}^{m}\left[\mathbf{1}\left\{y^{(i)}=t\right\}-p(y^{(i)}=t|\boldsymbol{x}^{(i)};\boldsymbol{\Theta})\right]\boldsymbol{x}^{(i)}.
\end{aligned}$$

1.3. (Ridge Regression) In PA1, a new method called *Ridge Regression* was introduced. By adding a regularization term in ordinary least square regression, the model can prevent the singularity when calculating matrix inverse. We can formulate ridge function as follows

$$J(\boldsymbol{\theta})\overset{\text{def}}{=}||\boldsymbol{y}-X\boldsymbol{\theta}||^2+\alpha||\boldsymbol{\theta}||^2,$$

where X is the design matrix, $\boldsymbol{y}$ is the corresponding label vector and $\boldsymbol{\theta}$ is the weight vector. For an appropriate α, calculate $\nabla_{\boldsymbol{\theta}}J(\boldsymbol{\theta})$ and give the optimal parameter $\boldsymbol{\theta}^*$.

Solution: (2 points)

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}}J(\boldsymbol{\theta}) &= \nabla_{\boldsymbol{\theta}}(\boldsymbol{y}-X\boldsymbol{\theta})^{\mathrm{T}}(\boldsymbol{y}-X\boldsymbol{\theta})+\alpha\nabla_{\boldsymbol{\theta}}(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\theta})\\
&= \nabla_{\boldsymbol{\theta}}\left(\boldsymbol{y}^{\mathrm{T}}\boldsymbol{y}-\boldsymbol{\theta}^{\mathrm{T}}X^{\mathrm{T}}\boldsymbol{y}-\boldsymbol{y}^{\mathrm{T}}X\boldsymbol{\theta}+\boldsymbol{\theta}^{\mathrm{T}}X^{\mathrm{T}}X\boldsymbol{\theta}\right)+2\alpha\boldsymbol{\theta}\\
&= -2X^{\mathrm{T}}\boldsymbol{y}+2X^{\mathrm{T}}X\boldsymbol{\theta}+2\alpha\boldsymbol{\theta}\\
&= 2(X^{\mathrm{T}}X+\alpha I)\boldsymbol{\theta}-2X^{\mathrm{T}}\boldsymbol{y}.
\end{aligned}$$

(0.5 point)For an appropriate α, let $\nabla_{\boldsymbol{\theta}}J(\boldsymbol{\theta})=0$, we have the optimal parameter

$$\boldsymbol{\theta}^*=(X^{\mathrm{T}}X+\alpha I)^{-1}X^{\mathrm{T}}\boldsymbol{y}.$$

1.4. (Newton's Method) Newton's method solves real functions $f(\boldsymbol{x}) = 0$ by iterative approximation. Thus, it can be used in logistic regression problem to calculate the optimal $\boldsymbol{\theta}^*$ when the derivative function is 0. When data $\boldsymbol{x}$ is multidimensional and label $y \in \{0, 1\}$, such iteration procedure is as follows:

$$\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t - (\boldsymbol{H}^{-1}(\boldsymbol{\theta})\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}))|_{\boldsymbol{\theta}_t} \tag{1}$$

where $J(\boldsymbol{\theta}) \overset{\text{def}}{=} \sum_{i=1}^{m} y^{(i)} \log(\frac{1}{1+e^{-\boldsymbol{\theta}^T \boldsymbol{x}^{(i)}}}) + (1-y^{(i)}) \log(1-\frac{1}{1+e^{-\boldsymbol{\theta}^T \boldsymbol{x}^{(i)}}})$, $\boldsymbol{H}(\boldsymbol{\theta})$ is the Hessian matrix of $J(\boldsymbol{\theta})$. Calculate $\boldsymbol{H}(\boldsymbol{\theta})$ and simplify iteration (1) without calculating the inverse of the Hessian matrix.

Hint: You may find PA1 question 1.2 very useful.

Solution:

$$\begin{aligned} J(\boldsymbol{\theta}) &= \sum_{i=1}^{m} y^{(i)} \log(\frac{1}{1+e^{-\boldsymbol{\theta}^T \boldsymbol{x}^{(i)}}}) + (1-y^{(i)}) \log(1-\frac{1}{1+e^{-\boldsymbol{\theta}^T \boldsymbol{x}^{(i)}}}) \\ &= \sum_{i=1}^{m} -\log(1+\exp(-\boldsymbol{\theta}^T \boldsymbol{x}^{(i)})) + (1-y^i)(-\boldsymbol{\theta}^T \boldsymbol{x}^{(i)}). \end{aligned}$$

(1 point)Thus, we have

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \sum_{i=1}^{m} (y^{(i)} - \frac{1}{1+e^{-\boldsymbol{\theta}^T \boldsymbol{x}^{(i)}}}) \boldsymbol{x}^{(i)}.$$

If we denote $\boldsymbol{\theta}_{(j)}$ and $\boldsymbol{x}^{(i)}_{(j)}$ as the j-th entry of vector $\boldsymbol{\theta}$ and $\boldsymbol{x}^{(i)}_{(j)}$ respectively. Then we have

$$\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_{(j)}} = \sum_{i=1}^{m} (y^{(i)} - \frac{1}{1+e^{-\boldsymbol{\theta}^T \boldsymbol{x}^{(i)}}}) \boldsymbol{x}^{(i)}_{(j)}.$$

(1 point)Further more, we have

$$\begin{aligned} \frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_{(j)} \partial \boldsymbol{\theta}_{(k)}} &= \frac{\partial}{\partial \boldsymbol{\theta}_{(k)}} \sum_{i=1}^{m} (y^{(i)} - \frac{1}{1+e^{-\boldsymbol{\theta}^T \boldsymbol{x}^{(i)}}}) \boldsymbol{x}^{(i)}_{(j)} \\ &= -\frac{\partial}{\partial \boldsymbol{\theta}_{(k)}} \sum_{i=1}^{m} \frac{\boldsymbol{x}^{(i)}_{(j)}}{1+e^{-\boldsymbol{\theta}^T \boldsymbol{x}^{(i)}}} \\ &= -\sum_{i=1}^{m} \boldsymbol{x}^{(i)}_{(j)} \frac{\partial}{\partial \boldsymbol{\theta}_{(k)}} (\frac{1}{1+e^{-\boldsymbol{\theta}^T \boldsymbol{x}^{(i)}}}) \\ &= -\sum_{i=1}^{m} \frac{\boldsymbol{x}^{(i)}_{(j)} \boldsymbol{x}^{(i)}_{(k)} e^{-\boldsymbol{\theta}^T \boldsymbol{x}^{(i)}}}{(1+e^{-\boldsymbol{\theta}^T \boldsymbol{x}^{(i)}})^2} \end{aligned}$$

If you want to rewrite it into matrix form, the following process will help:

$$\begin{aligned}\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_{(j)}\partial \boldsymbol{\theta}_{(k)}} &= -\sum_{i=1}^{m} \boldsymbol{x}_{(j)}^{(i)} \frac{e^{-\boldsymbol{\theta}^T \boldsymbol{x}^{(i)}}}{(1+e^{-\boldsymbol{\theta}^T \boldsymbol{x}^{(i)}})^2} \boldsymbol{x}_{(k)}^{(i)} \\ &= -\sum_{i=1}^{m} \boldsymbol{x}_{(j)}^{(i)} h_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)})(1-h_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}))\boldsymbol{x}_{(k)}^{(i)},\end{aligned}$$

where $h_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}) \overset{\text{def}}{=} (1+e^{-\boldsymbol{\theta}^T \boldsymbol{x}^{(i)}})^{-1}$.
(0.5 point)Thus $\boldsymbol{H}(\boldsymbol{\theta}) = \boldsymbol{X}^T \boldsymbol{R} \boldsymbol{X}$, where $\boldsymbol{X} \in \mathbb{R}^{m\times n}$ is the matrix composed by data $\boldsymbol{x}^{(i)}, i = 1, \ldots, m$, $\boldsymbol{R} \in \mathbb{R}^{m\times m}$ is a diagnal matrix whose entry $R_{ii} = h_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)})(1-h_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)})), i = 1, \ldots, m$. Finally, the iteration (1) can be simplified as

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + (\boldsymbol{X}^T \boldsymbol{R} \boldsymbol{X})^{-1} \boldsymbol{X}^T (\boldsymbol{y} - \boldsymbol{\mu}),$$

where $\boldsymbol{\mu}_i = h_{\boldsymbol{\theta_t}}(\boldsymbol{x}^{(i)})$.

1.5. (Multivariate Gaussian) The multivariate normal distribution can be written as

$$P_{\mathbf{y}}(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{n}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} (\boldsymbol{y} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{y} - \boldsymbol{\mu}) \right),$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the parameters. Show that the family of multivariate normal distributions is an exponential family, and derive the generalized linear model for Multivariate Gaussian with known Σ.

Hint: The parameters η and $T(\boldsymbol{y})$ are not limited to be vectors, but can also be matrices. In this case, the Frobenius inner product can be used to define the inner product between two matrices, which is represented as the trace of their products

$$\langle \boldsymbol{A}, \boldsymbol{B} \rangle_F = \mathrm{trace}(\boldsymbol{A}^{\mathrm{T}} \boldsymbol{B}).$$

The properties of matrix trace might also be useful.

Solution: (1 point) *Using the Frobenius inner product* $\langle \cdot, \cdot \rangle_{\mathrm{F}}$ *to deal with the inner products between matrices. Or, equivalently, use the vectorization* $\mathrm{vec}(\cdot)$ *to convert matrices into vectors.*

$$\begin{aligned}-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{y} - \boldsymbol{\mu}) &= \boldsymbol{y}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} - \frac{1}{2} \boldsymbol{y}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{y} - \frac{1}{2} \boldsymbol{\mu}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} \\ &= \langle \boldsymbol{y}, \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} \rangle + \left\langle \boldsymbol{y}\boldsymbol{y}^{\mathrm{T}}, -\frac{1}{2}\boldsymbol{\Sigma}^{-1} \right\rangle_{\mathrm{F}} - \frac{1}{2} \boldsymbol{\mu}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}\end{aligned}$$

(1 point)Let $\boldsymbol{\eta}_1 \overset{\text{def}}{=} \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$, $\boldsymbol{\eta}_2 \overset{\text{def}}{=} -\frac{1}{2}\boldsymbol{\Sigma}^{-1}$, $\boldsymbol{T}_1(\boldsymbol{y}) = \boldsymbol{y}$, $\boldsymbol{T}_2(\boldsymbol{y}) = \boldsymbol{y}\boldsymbol{y}^{\mathrm{T}}$. Then we have

$$-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{y} - \boldsymbol{\mu}) = \langle \boldsymbol{T}(\boldsymbol{y}), \boldsymbol{\eta} \rangle + \frac{1}{4} \boldsymbol{\eta}_1^{\mathrm{T}} \boldsymbol{\eta}_2^{-1} \boldsymbol{\eta}_1,$$

where we have defined $\boldsymbol{T}(\boldsymbol{y}) \overset{\text{def}}{=} (\boldsymbol{T}_1(\boldsymbol{y}), \boldsymbol{T}_2(\boldsymbol{y})), \boldsymbol{\eta} \overset{\text{def}}{=} (\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)$, and the inner product $\langle \boldsymbol{T}(\boldsymbol{y}), \boldsymbol{\eta} \rangle \overset{\text{def}}{=} \langle \boldsymbol{T}_1(\boldsymbol{y}), \boldsymbol{\eta}_1 \rangle + \langle \boldsymbol{T}_2(\boldsymbol{y}), \boldsymbol{\eta}_2 \rangle_{\mathrm{F}}$. As a result,

$$\begin{aligned} p_{\mathsf{y}}(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \frac{1}{(2\pi)^{\frac{n}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} (\boldsymbol{y} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{y} - \boldsymbol{\mu}) \right) \\ &= \frac{1}{(2\pi)^{\frac{n}{2}}} \exp\left(\langle \boldsymbol{T}(\boldsymbol{y}), \boldsymbol{\eta} \rangle + \frac{1}{4} \boldsymbol{\eta}_1^{\mathrm{T}} \boldsymbol{\eta}_2^{-1} \boldsymbol{\eta}_1 + \frac{1}{2} \log |-2\boldsymbol{\eta}_2| \right). \end{aligned}$$

(0.5 point)Further, with $b(\boldsymbol{y}) \overset{\text{def}}{=} (2\pi)^{-\frac{n}{2}}$ and

$$a(\boldsymbol{\eta}) \overset{\text{def}}{=} -\frac{1}{4} \boldsymbol{\eta}_1^{\mathrm{T}} \boldsymbol{\eta}_2^{-1} \boldsymbol{\eta}_1 - \frac{1}{2} \log |-2\boldsymbol{\eta}_2|,$$

we can obtain

$$p_{\mathsf{y}}(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = b(\boldsymbol{y}) \exp\left(\langle \boldsymbol{T}(\boldsymbol{y}), \boldsymbol{\eta} \rangle - a(\boldsymbol{\eta}) \right).$$

The response function are $\mu = -\frac{1}{2} \boldsymbol{\eta}_2^{-1} \boldsymbol{\eta}_1$ and $\boldsymbol{\Sigma} = -\frac{1}{2} \boldsymbol{\eta}_2^{-1}$. For further reading and understanding, you may refer to Generalized Linear Model.