Tsinghua-Berkeley Shenzhen Institute
LEARNING FROM DATA
Fall 2020

Writing Assignment 1

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

POLICIES

- **Acknowledgments:** We expect you to make an honest effort to solve the problems individually. As we sometimes reuse problem set questions from previous years, covered by papers and web pages, we expect the students **NOT** to copy, refer to, or look at the solutions in preparing their answers (relating to an unauthorized material is considered a violation of the honor principle). Similarly, we expect to not to google directly for answers (though you are free to google for knowledge about the topic). If you do happen to use other material, it must be acknowledged here, with a citation on the submitted solution.

- **Required homework submission format:** You can submit homework either as one single PDF document or as handwritten papers. Written homework needs to be provided during the class in the due date, and PDF document needs to be submitted through Tsinghua's Web Learning (`http://learn.tsinghua.edu.cn/`) before the end of due date.

  It is encouraged you LaTeX all your work, and we would provide a LaTeX template for your homework.

- **Collaborators:** In a separate section (before your answers), list the names of all people you collaborated with and for which question(s). If you did the HW entirely on your own, **PLEASE STATE THIS**. Each student must understand, write, and hand in answers of their own.

1.1. (Multivariate Least Squares) A data set consists of m data pairs $(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}), i = 1, \dots, m$, 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 $\boldsymbol{X} \overset{\text{def}}{=} [\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(m)}]^{\mathrm{T}}$, 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 sqaure loss.

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

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

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

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

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.

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

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

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 $\boldsymbol{\theta}^*$.

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 $J(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \sum\limits_{i=1}^{m} y^{(i)} \log(\frac{1}{1+e^{-\boldsymbol{\theta}^T \boldsymbol{x}^{(i)}}}) + (1 - y^{(i)}) \log(1 - \frac{1}{1+e^{-\boldsymbol{\theta}^T \boldsymbol{x}^{(i)}}})$, $\boldsymbol{H}(\boldsymbol{\theta})$ 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.

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 response function of 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.