Tsinghua-Berkeley Shenzhen Institute
Learning from Data
Fall 2018

---

**Problem Set 2**

**Issued:** Monday 22nd October, 2018 **Due:** Monday 29th October, 2018

---

2.1. A data set consists of $m$ data pairs $(\boldsymbol{x}^{(1)}, y^{(1)}), \dots, (\boldsymbol{x}^{(m)}, y^{(m)})$, where $\boldsymbol{x} \in \mathbb{R}^n$ is the independent variable, and $y \in \mathbb{R}$ is the dependent variable. Denote the design matrix by $\boldsymbol{X} \triangleq [\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(m)}]^{\mathrm{T}}$, and let $\boldsymbol{y} \triangleq [y^{(1)}, \cdots, y^{(m)}]^{\mathrm{T}}$. The least-squares method then minimizes the square loss $J(\boldsymbol{\theta})$ defined as

$$J(\boldsymbol{\theta}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2,$$

where $\boldsymbol{\theta} \in \mathbb{R}^n$ is the parameter to be estimated. To find the optimal $\boldsymbol{\theta}$, let $\nabla J(\boldsymbol{\theta}) = 0$, and we can get the normal equation:

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

When $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ is invertible, we have $\boldsymbol{\theta} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}\boldsymbol{y}$.

Now, suppose $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ is singular. Does the solution of (1) still exist? Prove your result, and explain its meaning in plain words.

2.2. A data set consists of $m$ data pairs $(\boldsymbol{x}^{(1)}, y^{(1)}), \dots, (\boldsymbol{x}^{(m)}, y^{(m)})$, where $\boldsymbol{x} \in \mathbb{R}^n$ is the independent variable, and $y \in \{1, \dots, k\}$ is the dependent variable. The conditional probability $P_{\mathsf{y}|\mathbf{x}}(y|\boldsymbol{x})$[1] estimated by the softmax regression is

$$P_{\mathsf{y}|\mathbf{x}}(l|\boldsymbol{x}) = \frac{\exp(\boldsymbol{\theta}_l^{\mathrm{T}}\boldsymbol{x} + b_l)}{\sum_{j=1}^{k}\exp(\boldsymbol{\theta}_j^{\mathrm{T}}\boldsymbol{x} + b_j)}, \quad l = 1, \dots, k,$$

where $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_k \in \mathbb{R}^n$ and $b_1, \dots, b_k \in \mathbb{R}$ are the parameters of softmax regression. The term $b_i$ is called a bias term.

The log-likelihood of the softmax regression model is

$$\ell = \sum_{i=1}^{m} \log P_{\mathsf{y}|\mathbf{x}}(y^{(i)}|\boldsymbol{x}^{(i)}).$$

(a) Evaluate $\nabla_{b_l}\ell$.

The data set can be described by its empirical distribution $\hat{P}_{\mathbf{x},\mathsf{y}}(\boldsymbol{x}, y)$ defined as

$$\hat{P}_{\mathbf{x},\mathsf{y}}(\boldsymbol{x}, y) = \frac{1}{m}\sum_{i=1}^{m}\mathbb{1}\{\boldsymbol{x} = \boldsymbol{x}^{(i)}, y = y^{(i)}\},$$

where $\mathbb{1}\{\cdot\}$ is the indicator function. Similarly, the empirical marginal distributions of this data set are

$$\hat{P}_{\mathbf{x}}(\boldsymbol{x}) = \frac{1}{m}\sum_{i=1}^{m}\mathbb{1}\{\boldsymbol{x} = \boldsymbol{x}^{(i)}\}, \quad \hat{P}_{\mathsf{y}}(y) = \frac{1}{m}\sum_{i=1}^{m}\mathbb{1}\{y = y^{(i)}\}.$$

[1] The notation $P_{\mathsf{y}|\mathbf{x}}(y|\boldsymbol{x})$ stands for $\mathbb{P}(\mathsf{y} = y|\mathbf{x} = \boldsymbol{x})$, i.e., the conditional probability of $\mathsf{y} = y$ given $\mathbf{x} = \boldsymbol{x}$.

(b) Suppose we have set the biases $(b_1, \ldots, b_k)$ to their optimal values, prove that

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

where $\mathcal{X} = \{\boldsymbol{x}^{(i)} : i = 1, \ldots, m\}$ is the set of all samples of $\boldsymbol{x}$.
*Hint: The optimality implies* $\nabla_{b_1} \ell = \nabla_{b_2} \ell = \cdots = \nabla_{b_k} \ell = 0$.

2.3. The multivariate normal distribution can be written as

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

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the parameters. Show that the family of multivariate normal distributions is an exponential family, and find the corresponding $\eta$, $b(\boldsymbol{y})$, $T(\boldsymbol{y})$, and $a(\eta)$.

*Hints: The parameters $\eta$ 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. The properties of matrix trace might be useful.*