Tsinghua-Berkeley Shenzhen Institute
LEARNING FROM DATA
Fall 2019

Written Assignment 1

Issued: Sunday 29th September, 2019 **Due:** Sunday 13th October, 2019

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

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

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

The log-likelihood of the softmax regression model is

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

(a) Evaluate $\dfrac{\partial \ell}{\partial b_l}$.

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

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

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

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

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

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

where $\mathcal{X} = \{\boldsymbol{x}^{(i)} : i = 1, \ldots, m\}$ is the set of all samples of $\boldsymbol{x}$.

Hint: The optimality implies $\dfrac{\partial \ell}{\partial b_1} = \dfrac{\partial \ell}{\partial b_2} = \cdots = \dfrac{\partial \ell}{\partial b_k} = 0$.

Solution: *This problem shows that, with the help of biases, the conditional probability estimated by softmax regression behaves like the empirical conditional probability. Specifically, the result of (b) would become the formula of total probability if we substitute* $P_{\mathsf{y}|\mathsf{x}}$ *with* $\hat{P}_{\mathsf{y}|\mathsf{x}}$. *It turns out that this formula also holds for our estimated* $P_{\mathsf{y}|\mathsf{x}}$.

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

(a) We have

$$\begin{aligned}\frac{\partial \ell}{\partial b_l} &= \frac{\partial}{\partial b_l}\sum_{i=1}^{m}\left[\boldsymbol{\theta}_{y^{(i)}}^{\mathrm{T}}\boldsymbol{x}^{(i)} + b_{y^{(i)}} - \log\left(\sum_{j=1}^{k}\exp(\boldsymbol{\theta}_j^{\mathrm{T}}\boldsymbol{x}^{(i)} + b_j)\right)\right]\\ &= \sum_{i=1}^{m}\left[\mathbb{1}\left\{y^{(i)} = l\right\} - \frac{\exp(\boldsymbol{\theta}_l^{\mathrm{T}}\boldsymbol{x}^{(i)} + b_l)}{\sum_{j=1}^{k}\exp(\boldsymbol{\theta}_j^{\mathrm{T}}\boldsymbol{x}^{(i)} + b_j)}\right]\\ &= \sum_{i=1}^{m}\left[\mathbb{1}\left\{y^{(i)} = l\right\} - P_{\mathsf{y}|\mathsf{x}}(l|\boldsymbol{x}^{(i)})\right], \quad l = 1, \ldots, k.\end{aligned}$$

(b) The optimality implies $\frac{\partial \ell}{\partial b_l} = 0$ for all l's, i.e.,

$$\sum_{i=1}^{m}\left[\mathbb{1}\left\{y^{(i)} = l\right\} - P_{\mathsf{y}|\mathsf{x}}(l|\boldsymbol{x}^{(i)})\right] = 0, \quad l = 1, \ldots, k.$$

As a result, $\forall\, l = 1, \ldots, k$,

$$\begin{aligned}\hat{P}_{\mathsf{y}}(l) = \frac{1}{m}\sum_{i=1}^{m}\mathbb{1}\left\{y^{(i)} = l\right\} &= \frac{1}{m}\sum_{i=1}^{m}P_{\mathsf{y}|\mathsf{x}}(l|\boldsymbol{x}^{(i)})\\ &= \frac{1}{m}\sum_{i=1}^{m}P_{\mathsf{y}|\mathsf{x}}(l|\boldsymbol{x}^{(i)})\sum_{\boldsymbol{x}\in\mathcal{X}}\mathbb{1}\left\{\boldsymbol{x}^{(i)} = \boldsymbol{x}\right\}\\ &= \frac{1}{m}\sum_{i=1}^{m}\sum_{\boldsymbol{x}\in\mathcal{X}}P_{\mathsf{y}|\mathsf{x}}(l|\boldsymbol{x}^{(i)})\cdot\mathbb{1}\left\{\boldsymbol{x}^{(i)} = \boldsymbol{x}\right\}\\ &= \frac{1}{m}\sum_{i=1}^{m}\sum_{\boldsymbol{x}\in\mathcal{X}}P_{\mathsf{y}|\mathsf{x}}(l|\boldsymbol{x})\cdot\mathbb{1}\left\{\boldsymbol{x}^{(i)} = \boldsymbol{x}\right\}\\ &= \sum_{\boldsymbol{x}\in\mathcal{X}}P_{\mathsf{y}|\mathsf{x}}(l|\boldsymbol{x})\cdot\left(\frac{1}{m}\sum_{i=1}^{m}\mathbb{1}\left\{\boldsymbol{x}^{(i)} = \boldsymbol{x}\right\}\right)\\ &= \sum_{\boldsymbol{x}\in\mathcal{X}}P_{\mathsf{y}|\mathsf{x}}(l|\boldsymbol{x})\hat{P}_{\mathsf{x}}(\boldsymbol{x}),\end{aligned}$$

1.2. We have mentioned maximun likelyhood estimation (MLE) in the class.

(a) Now let's do it on a very simple case.

Suppose we have N samples (x_1,x_2,...,x_n) independently drawn from a normal distribution with **KNOWN** variance σ^2 and **UNKNOWN** mean μ

$$P\left(x_i|\mu\right) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right),$$

please find the MLE for μ

(b) It is always a task to find a satisfying loss function.

However MLE is not always the best. To explain MLE simply, it is trying to maximize the possiblity of the events under the parameter you want to compute. Think about it, you have known the N samples. Why should we consider them as the events that should be estimated instead of the conditions we have known. Based on Bayes Rule, we can cultivate another method called Maximum A Posteriori (MAP).

$$P\left(\mu|x_1, x_2, ..., x_n\right) = \frac{P\left(x_1, x_2, ..., x_n|\mu\right) P(\mu)}{P\left(x_1, x_2, ..., x_n\right)}$$

There exist some assumptions on $P\left(\mu\right)$. If μ is purely at random, it will make no difference between MLE and MAP. Here suppose

$$P\left(\mu\right) = \frac{1}{\sqrt{2\pi\theta^2}} \exp\left(-\frac{(\mu-\nu)^2}{2\theta^2}\right),$$

do the calculation to find the estimator for μ. Also, compare the estimators of MLE and MAP when n is very large.

Solution:

(a) Just follow the way,

$$\log P\left(x_1, x_2, ..., x_n|\mu\right) = \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{(x_i-\mu)^2}{2\sigma^2},$$

then take the derivatives with respect to μ,

$$\frac{\partial \log P\left(x_1, x_2, ..., x_n|\mu\right)}{\partial\mu} = \sum_{i=1}^{n} \frac{(x_i-\mu)}{\sigma^2} = 0$$

Therefore

$$\hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n}$$

(b) Find the posteriori we want to maximize

$$P\left(\mu|x_1, x_2, ..., x_n\right) = \frac{\left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right) \frac{1}{\sqrt{2\pi\theta^2}} e^{-\frac{(\mu-\nu)^2}{2\theta^2}}}{C},$$

then take the derivative

$$\frac{\partial \log P\left(\mu|x_1, x_2, ..., x_n\right)}{\partial\mu} = \left(\sum_{i=1}^{n} \frac{x_i-\mu}{\sigma^2}\right) - \frac{\mu-\nu}{\theta^2} = 0$$

$$\frac{\mu-\nu}{\theta^2} = \frac{\sum_{i=1}^{n} x_i}{\sigma^2} - \frac{n\mu}{\sigma^2}$$

Therefore,

$$\hat{\mu} = \frac{\sigma^2\nu + \theta^2 \sum_{i=1}^{n} x_i}{\sigma^2 + n\theta^2}$$

When $n \to \infty$,

$$\hat{\mu}_{MAP} \to \hat{\mu}_{MLE}$$

1.3. We want to talk about Multivariate Least squares in this problem.
A data set consists of m data pairs $(\boldsymbol{x}^{(1)}, \boldsymbol{y}^{(1)}), \ldots, (\boldsymbol{x}^{(m)}, \boldsymbol{y}^{(m)})$, 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 $\boldsymbol{X} \triangleq [\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}]^{\mathrm{T}}$, and let $\boldsymbol{Y} \triangleq [\boldsymbol{y}^{(1)}, \cdots, \boldsymbol{y}^{(m)}]^{\mathrm{T}}$. Please write down the square loss $J(\boldsymbol{\Theta})$, where $\boldsymbol{\Theta} \in \mathbb{R}^{n \times l}$ is the parameter matrix you want to get, and then compute the solution.

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

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

You can do the following things.

Solution: It can be written down as a trace

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

Then do the derivative

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

Therefore, the solution is

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

1.4. 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 find the corresponding η, $b(\boldsymbol{y})$, $T(\boldsymbol{y})$, and $a(\eta)$.

Hints: 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. The properties of matrix trace might be useful.

Solution: *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.*

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

Let $\boldsymbol{\eta}_1 \triangleq \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$, $\boldsymbol{\eta}_2 \triangleq -\frac{1}{2}\boldsymbol{\Sigma}^{-1}$, $\boldsymbol{T}_1(\boldsymbol{y}) = \boldsymbol{y}$, $\boldsymbol{T}_2(\boldsymbol{y}) = \boldsymbol{y}\boldsymbol{y}^{\mathrm{T}}$. 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 $\boldsymbol{T}(\boldsymbol{y}) \triangleq (\boldsymbol{T}_1(\boldsymbol{y}), \boldsymbol{T}_2(\boldsymbol{y}))$, $\boldsymbol{\eta} \triangleq (\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)$, and the inner product $\langle \boldsymbol{T}(\boldsymbol{y}), \boldsymbol{\eta}\rangle \triangleq \langle \boldsymbol{T}_1(\boldsymbol{y}), \boldsymbol{\eta}_1\rangle + \langle \boldsymbol{T}_2(\boldsymbol{y}), \boldsymbol{\eta}_2\rangle_{\mathrm{F}}$. As a result,

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

Further, with $b(\boldsymbol{y}) \triangleq (2\pi)^{-\frac{n}{2}}$ and

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

we can obtain

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