Tsinghua-Berkeley Shenzhen Institute
Learning From Data
Fall 2021

Writing Assignment 1

Issued: Thursday 30th September, 2021 **Due:** Friday 22nd October, 2021

1.1. (Sigmoid Function) Show that the sigmoid function

$$\sigma(x) = \frac{1}{1+\exp(-x)}$$

satisfies the following properties.

(a) (1 point) $\sigma(-x) = 1 - \sigma(x)$.

(b) (1 point) $\dfrac{d\sigma(x)}{dx} = \sigma(x)(1-\sigma(x))$.

1.2. (Ridge Regression) In PA1, a new method called *Ridge Regression* was introduced. Ridge regression was developed as a possible solution to the imprecision of least square estimators when linear regression models have some multicollinear (highly correlated) independent variables.

We can formulate ridge regression loss function as the following

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

where X is the design matrix, $\boldsymbol{y}$ is the corresponding label vector and $\boldsymbol{\theta}$ is the weight vector. For an appropriate λ,

(a) (1 point) calculate $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$,

(b) (1 point) give the gradient descend iteration equation with learning rate α,

(c) (1 point) and derive the optimal parameter $\boldsymbol{\theta}^*$ for normal equation method. (Your solution should be consistent with Problem 2 in PA1.)

1.3. (Maximum Likelihood Estimation) In class, we have learnt maximum likelihood estimation for linear model assuming the error follows the Gaussian distribution. The maximization process results in an equivalent formulation as ordinary least square problem. But the maximum likelihood estimation is not always directing into the ℓ^2-norm measurement. It depends on the error distribution assumption.

As shown in Figure.1 and Figure.2, let's consider the linear regression problem with an error following Laplace distribution, also known as the least absolute deviation[1]: for the given m samples $(\boldsymbol{x}^{(1)}, y^{(1)}), \ldots, (\boldsymbol{x}^{(m)}, y^{(m)}), \boldsymbol{x}^{(i)} \in \mathbb{R}^n, y \in \mathbb{R}, i = 1, \cdots, m$, we need to determine the parameters $\boldsymbol{\theta} \in \mathbb{R}^n$ for the linear model:

$$y^{(i)} = \boldsymbol{\theta}^\top \boldsymbol{x}^{(i)} + \epsilon^{(i)},$$

$\epsilon^{(i)} \in \mathbb{R}$ are i.i.d. Laplacian random variables with density function:

$$P(z) = \frac{1}{2\tau} \exp(\frac{-|z-\mu|}{\tau})$$

where $\tau > 0$ and μ is the mean value.

(a) (1 point) Write down the expression of conditional distribution $P_{Y|X}(y|\boldsymbol{x};\boldsymbol{\theta})$.

(b) (1 point) Write down the log-likelihood function of this problem.

(c) (1 point) For data $((1,1)^\top, 1), ((1,2)^\top, -1)$ and $\tau = 1, \mu = 0$, derive the optimal parameter $\boldsymbol{\theta}^*$.

(d) (1 point) The ordinary least square uses ℓ^2-norm to measure the distances and wants to minimize overall distances of data points to a linear model. Try to give a geometric interpretation of the least absolute deviation.

[1] See `https://en.wikipedia.org/wiki/Least_absolute_deviations#Contrasting_ordinary_least_squares_with_least_absolute_deviations` for reference on least absolute deviation.

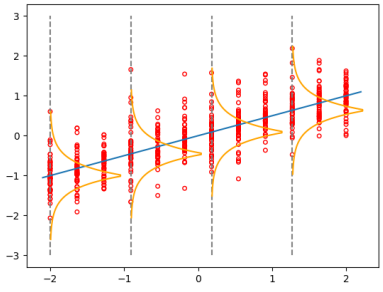

Figure 1: Linear regression with least absolute deviation

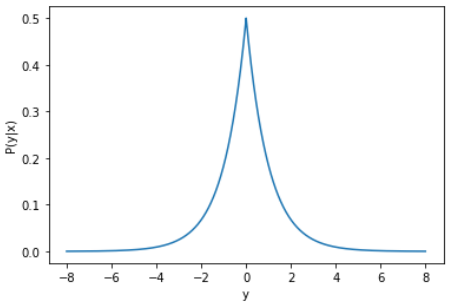

Figure 2: Error with Laplace distribution

1.4. (Softmax Regression)(3 points) 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 $\nabla_{\boldsymbol{\theta}_1} \ell(\boldsymbol{\Theta})$.

Hint: The index number of samples has nothing to do with $\boldsymbol{\theta}_1$, thus you just need to calculate $\nabla_{\boldsymbol{\theta}_1} \log p(y^{(i)} | \boldsymbol{x}^{(i)}; \boldsymbol{\Theta})$ and sum them up. Indicator function $\mathbf{1}\left\{y^{(i)} = 1\right\} = 0$ when $y^{(i)} \neq 1$, thus only one term in $\nabla_{\boldsymbol{\theta}_1} \log p(y^{(i)} | \boldsymbol{x}^{(i)}; \boldsymbol{\Theta})$ will be left.