Tsinghua-Berkeley Shenzhen Institute
Learning From Data
Spring 2024

Written Assignment 1

Issued: Friday 15th March, 2024 **Due:** Friday 29th March, 2024

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.
- **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. (Logistic Regression) Given random vectors $\boldsymbol{x} \in \mathbb{R}^n$, logistic regression models the conditional distribution of class y given $\boldsymbol{x}$ with a Bernoulli distribution parameterized by the Sigmoid function of $\boldsymbol{\theta}^\top \boldsymbol{x}$, i.e.

$$P(y|\boldsymbol{x};\boldsymbol{\theta}) = (\sigma(\boldsymbol{\theta}^\top \boldsymbol{x}))^y (1-\sigma(\boldsymbol{\theta}^\top \boldsymbol{x}))^{1-y},$$

where $\boldsymbol{\theta} \in \mathbb{R}$ is the weighting parameter for $\boldsymbol{x}$ and $\sigma(\cdot)$ denotes the Sigmoid function.

(a) (0.5 points) Show that the sigmoid function

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

satisfies the property $\dfrac{d\sigma(z)}{dz} = \sigma(z)(1-\sigma(z))$.

(b) (1 point) Suppose we have m independently generated training examples $\left(\boldsymbol{x}^{(1)}, y^{(1)}\right), \ldots, \left(\boldsymbol{x}^{(m)}, y^{(m)}\right), \boldsymbol{x}^{(i)} \in \mathbb{R}^n, y^{(i)} \in \mathbb{R}, i = 1, \ldots, m$, the log-likelihood function can be written as:

$$I(\boldsymbol{\theta}) = \sum_{i=1}^{m} y^{(i)} \log \sigma(\boldsymbol{\theta}^\top \boldsymbol{x}^{(i)}) + (1 - y^{(i)}) \log \left(1 - \sigma(\boldsymbol{\theta}^\top \boldsymbol{x}^{(i)})\right).$$

Prove that for $\boldsymbol{\theta}_j, \forall j \in \{1, \ldots, n\}$,

$$\frac{\partial I(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_j} = \sum_{i=1}^{m} \left(y^{(i)} - \sigma(\boldsymbol{\theta}^\top \boldsymbol{x}^{(i)})\right) \boldsymbol{x}_j^{(i)}.$$

(c) (0.5+1 points) Based on former results, give the pseudo code for solving $arg\max_{\boldsymbol{\theta}} I(\boldsymbol{\theta})$ using: 1) stochastic gradient ascent; 2) batch gradient ascent.

1.2. (Ridge Regression) 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 the 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) derive the optimal parameter $\boldsymbol{\theta}^*$ for the normal equation method.

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 $\left(\boldsymbol{x}^{(1)}, y^{(1)}\right), \ldots, \left(\boldsymbol{x}^{(m)}, y^{(m)}\right), \boldsymbol{x}^{(i)} \in \mathbb{R}^n, y^{(i)} \in \mathbb{R}, i = 1, \ldots, m$, 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:

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

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

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

(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) (Bonus 2 points) 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.

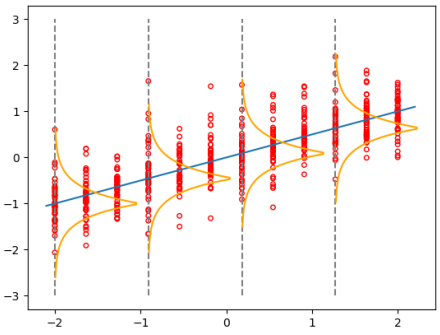

Figure 1: Linear Regression with Least Absolute Deviation

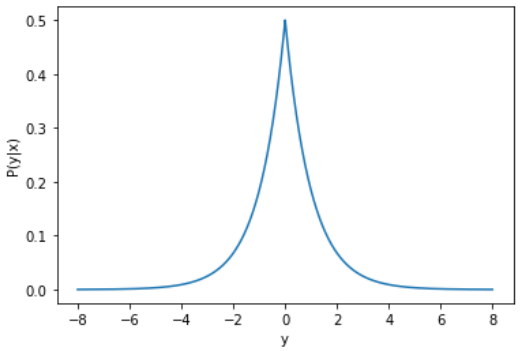

Figure 2: Error with Laplace Distribution

1.4. (MAP) Suppose we have m samples $x_1, x_2, ..., x_m$ independently drawn from a normal distribution with known variance σ^2 and unknown mean θ, i.e.

$$P(x_i|\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x_i-\theta)^2}{2\sigma^2}).$$

Maximum likelihood estimation (MLE) assumes that the optimal parameter θ is the one that generates the observed data with the highest probability, i.e. $\theta_{MLE} \overset{\text{def}}{=} \arg\max_\theta P(x_1, x_2, ..., x_m|\theta)$. However, what if we know some additional prior information about the distribution of θ? e.g. Let θ be a random variable following a Gaussian distribution, i.e. $\theta \sim \mathcal{N}(\nu, \mu^2)$. We can calculate the posterior distribution of θ using Bayes' theorem and derive the MAP estimator θ_{MAP}, i.e.

$$\theta_{MAP} \overset{\text{def}}{=} argmax_\theta P(\theta|x_1, \ldots, x_m) = \frac{P(x_1, \ldots, x_m|\theta)P(\theta)}{P(x_1, \ldots, x_m)}.$$

(a) (1 point) Find the MLE estimator for θ;

(b) (1 point) Find the MAP estimator for θ;

(c) (1 point) Compare the estimators of MLE and MAP when n is very large.

1.5. (Softmax Regression)(3 points) In multivariate classification problems, 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

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

Calculate $\nabla_{\boldsymbol{\theta}_1} \ell(\mathbf{\Theta})$.