Tsinghua-Berkeley Shenzhen Institute
Learning From Data
Autumn 2024

Written Assignment 1

Issued: Friday 27th September, 2024 **Due:** Friday 11th October, 2024

COMMENTS

- Mention *collaborators* in your assignments. See the policies for details.
- Provide sufficient arguments in your proof.

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:

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

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

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

Solution:
(a)(0.5 point)

$$\begin{aligned}\frac{d\sigma(x)}{dx} &= \frac{\exp(-x)}{(1+\exp(-x))^2}\\ &= \frac{1+\exp(-x)-1}{(1+\exp(-x))^2}\\ &= \frac{1}{1+\exp(-x)}\left(1-\frac{1}{1+\exp(-x)}\right)\\ &= \sigma(x)(1-\sigma(x))\end{aligned}$$

(b)(1 point)

$$\begin{aligned}\frac{\partial I(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_j} &= \frac{\partial}{\partial \boldsymbol{\theta}_j}\sum_{i=1}^{m} y^{(i)}\log\sigma(\boldsymbol{\theta}^\top \boldsymbol{x}^{(i)}) + (1-y^{(i)})\log\left(1-\sigma(\boldsymbol{\theta}^\top \boldsymbol{x}^{(i)})\right)\\ &= \sum_{i=1}^{m} y^{(i)}\cdot\frac{1}{\sigma(\boldsymbol{\theta}^\top \boldsymbol{x}^{(i)})}\cdot\sigma(\boldsymbol{\theta}^\top \boldsymbol{x}^{(i)})(1-\sigma(\boldsymbol{\theta}^\top \boldsymbol{x}^{(i)}))\cdot\boldsymbol{x}_j^{(i)}\\ &+ \sum_{i=1}^{m}(1-y^{(i)})\cdot\frac{-1}{1-\sigma(\boldsymbol{\theta}^\top \boldsymbol{x}^{(i)})}\cdot\sigma(\boldsymbol{\theta}^\top \boldsymbol{x}^{(i)})(1-\sigma(\boldsymbol{\theta}^\top \boldsymbol{x}^{(i)}))\cdot\boldsymbol{x}_j^{(i)}\\ &= \sum_{i=1}^{m}\left(y^{(i)}-\sigma(\boldsymbol{\theta}^\top \boldsymbol{x}^{(i)})\right)\boldsymbol{x}_j^{(i)}\end{aligned}$$

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

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

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.

Solution:

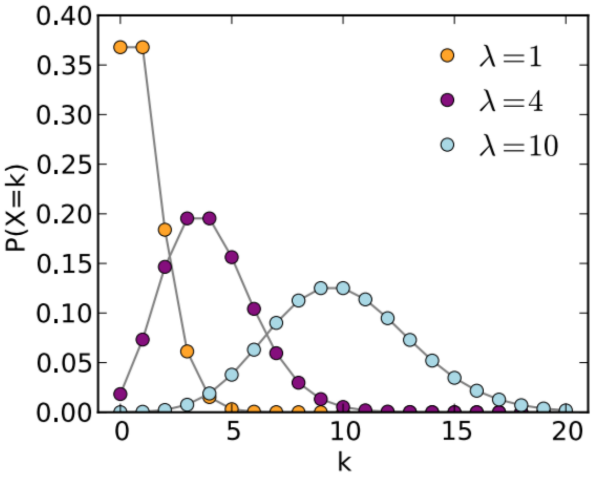

Figure 1: Poisson distribution

(a)(1 point)

$$\begin{aligned}\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) &= \nabla_{\boldsymbol{\theta}}(\boldsymbol{y}-X\boldsymbol{\theta})^\top(\boldsymbol{y}-X\boldsymbol{\theta})+\gamma\nabla_{\boldsymbol{\theta}}(\boldsymbol{\theta}^\top\boldsymbol{\theta})\\ &= \nabla_{\boldsymbol{\theta}}\left(\boldsymbol{y}^\top\boldsymbol{y}-\boldsymbol{\theta}^\top X^\top\boldsymbol{y}-\boldsymbol{y}^\top X\boldsymbol{\theta}+\boldsymbol{\theta}^\top X^\top X\boldsymbol{\theta}\right)+2\gamma\boldsymbol{\theta}\\ &= -2X^\top\boldsymbol{y}+2X^\top X\boldsymbol{\theta}+2\gamma\boldsymbol{\theta}\\ &= 2(X^\top X+\gamma I)\boldsymbol{\theta}-2X^\top\boldsymbol{y}.\end{aligned}$$

(b)(1 point) In i-th iteration,

$$\boldsymbol{\theta}_i=\boldsymbol{\theta}_{i-1}-2\alpha(X^\top X+\gamma I)\boldsymbol{\theta}_{i-1}+2\alpha X^\top\boldsymbol{y}$$

(c)(1 point) For an appropriate γ, let $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})=0$, we have the optimal parameter

$$\boldsymbol{\theta}^*=(X^\top X+\gamma I)^{-1}X^\top\boldsymbol{y}.$$

1.3. (Poisson Distribution and Generalized Linear Model)(2 points)

As shown in Figure. 1, the Poisson distribution is used to model count data, where the probability of observing $y\in\mathbb{Z}_{\geq 0}$ given a rate parameter $\lambda>0$ is:

$$p(y\mid\lambda)=\frac{\lambda^y e^{-\lambda}}{y!}.$$

In this problem, you will express the Poisson distribution as a member of the exponential family and identify the relevant components.

Exponential Family: A probability distribution is said to belong to the exponential family if it can be written in the following canonical form:

$$p(y\mid\eta)=b(y)\exp\left(\eta^T T(y)-a(\eta)\right),$$

where:

- y: the observed random variable (in this case, the count data).
- η: the **natural (canonical) parameter**, which is a function of the distribution's parameters. It serves as the primary variable that links the data to the distribution.
- $T(y)$: the **sufficient statistic** of the distribution. This is a function of y that summarizes all the information needed from the data.
- $b(y)$: a function of the data y that typically ensures normalization and varies depending on the distribution.
- $a(\eta)$: the **log partition function** (or cumulant function), which ensures the distribution sums (or integrates) to one and plays a key role in controlling the variability of the distribution.

For the Poisson distribution, derive the following components within the exponential family framework:

(a) (1.5 points) Derive the exponential form of Poisson distribution.

(b) (2 points) Now we derive the GLM for Poisson distribution, also known as Poisson regression. Write the hypothesis function $h_\theta(x)$. What is the canonical link function in this case? (Hint: the canonical link function g^{-1} maps distribution parameter λ to the natural parameter η)

Solution:

(a)(1.5 points)

$$P(y|\lambda) = \frac{\lambda^y \cdot e^{-\lambda}}{y!} = \frac{1}{y!} \cdot e^{y \ln \lambda - \lambda}$$

- $b(y) = \frac{1}{y!}$, $T(y) = y$, $\eta = \ln \lambda$, $a(\eta) = e^\eta$

(b)(2 points)

- The hypothesis function is:

$$h_\theta(x) = E[T(y)|x;\theta] = E[y|x;\theta] = e^\eta = e^{\theta^T x}$$

- The canonical response function is:

$$\eta = g^{-1}(E[T(y)|x;\theta]) = g^{-1}(E[y|x;\theta])$$

$$g(\eta) = E[y|x;\theta] = \lambda = e^\eta$$

- The canonical link function is:

$$\eta = g^{-1}(\lambda) = \ln \lambda$$

1.4. (Softmax Regression)(2 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

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

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

Solution:

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}_t}\ell(\boldsymbol{\Theta}) &= \nabla_{\boldsymbol{\theta}_t} \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)}}} \\
&= \sum_{i=1}^{m} \nabla_{\boldsymbol{\theta}_t} \sum_{l=1}^{k} \mathbf{1}\left\{y^{(i)} = l\right\} \left(\boldsymbol{\theta}_l^\top \boldsymbol{x}^{(i)} - \log \sum_{j=1}^{k} e^{\boldsymbol{\theta}_j^\top \boldsymbol{x}^{(i)}}\right) \\
&= \sum_{i=1}^{m} \nabla_{\boldsymbol{\theta}_t} \left(\sum_{l=1}^{k} \mathbf{1}\left\{y^{(i)} = l\right\} \boldsymbol{\theta}_l^\top \boldsymbol{x}^{(i)} - \sum_{l=1}^{k} \mathbf{1}\left\{y^{(i)} = l\right\} \log \sum_{j=1}^{k} e^{\boldsymbol{\theta}_j^\top \boldsymbol{x}^{(i)}}\right) \\
&= \sum_{i=1}^{m} \nabla_{\boldsymbol{\theta}_t} \left[\sum_{l=1}^{k} \mathbf{1}\left\{y^{(i)} = l\right\} \boldsymbol{\theta}_l^\top \boldsymbol{x}^{(i)} - \left(\sum_{l=1}^{k} \mathbf{1}\left\{y^{(i)} = l\right\}\right) \log \sum_{j=1}^{k} e^{\boldsymbol{\theta}_j^\top \boldsymbol{x}^{(i)}}\right] \\
&= \sum_{i=1}^{m} \nabla_{\boldsymbol{\theta}_t} \left(\mathbf{1}\left\{y^{(i)} = 1\right\} \boldsymbol{\theta}_1^\top \boldsymbol{x}^{(i)} + ... + \mathbf{1}\left\{y^{(i)} = k\right\} \boldsymbol{\theta}_k^\top \boldsymbol{x}^{(i)} - \log \sum_{j=1}^{k} e^{\boldsymbol{\theta}_j^\top \boldsymbol{x}^{(i)}}\right) \\
&= \sum_{i=1}^{m} \left(\mathbf{1}\left\{y^{(i)} = t\right\} \boldsymbol{x}^{(i)} - \frac{\nabla_{\boldsymbol{\theta}_t} \sum_{j=1}^{k} e^{\boldsymbol{\theta}_j^\top \boldsymbol{x}^{(i)}}}{\sum_{j=1}^{k} e^{\boldsymbol{\theta}_j^\top \boldsymbol{x}^{(i)}}}\right) \\
&= \sum_{i=1}^{m} \left(\mathbf{1}\left\{y^{(i)} = t\right\} \boldsymbol{x}^{(i)} - \frac{\nabla_{\boldsymbol{\theta}_t} e^{\boldsymbol{\theta}_t^\top \boldsymbol{x}^{(i)}}}{\sum_{j=1}^{k} e^{\boldsymbol{\theta}_j^\top \boldsymbol{x}^{(i)}}}\right) \\
&= \sum_{i=1}^{m} \left(\mathbf{1}\left\{y^{(i)} = t\right\} \boldsymbol{x}^{(i)} - \frac{e^{\boldsymbol{\theta}_t^\top \boldsymbol{x}^{(i)}} \boldsymbol{x}^{(i)}}{\sum_{j=1}^{k} e^{\boldsymbol{\theta}_j^\top \boldsymbol{x}^{(i)}}}\right) \\
&= \sum_{i=1}^{m} \left[\mathbf{1}\left\{y^{(i)} = t\right\} - p(y^{(i)} = t|\boldsymbol{x}^{(i)})\right] \boldsymbol{x}^{(i)}
\end{aligned}$$

Thus,

$$\nabla_{\boldsymbol{\theta}_1}\ell(\boldsymbol{\Theta}) = \sum_{i=1}^{m} \left[\mathbf{1}\left\{y^{(i)} = 1\right\} - p(y^{(i)} = 1|\boldsymbol{x}^{(i)})\right] \boldsymbol{x}^{(i)}.$$

1.5. (Maximum Likelihood Estimation) [Bonus Question] In class, we have learnt maxi-

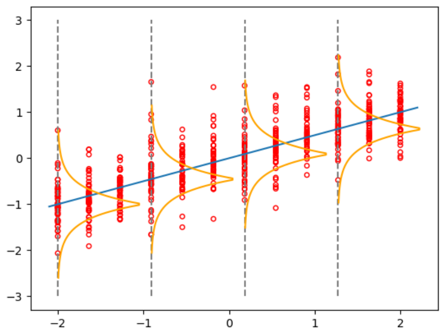

Figure 2: Linear Regression with Least Absolute Deviation

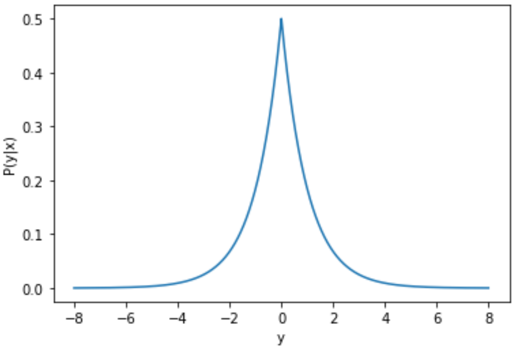

Figure 3: Error with Laplace Distribution

mum likelihood estimation for linear model assuming the error follows the Gaussian distribution. The maximization process results in an equivalent formulation as an ordinary least square problem. However, the maximum likelihood estimation is not always directed into the ℓ^2-norm measurement. It depends on the error distribution assumption.

As shown in Figure. 2 and Figure. 3, 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:

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

where $\tau > 0$.

(a) (0.5 points) Write down the expression of conditional distribution $P_{Y|X}(y|\boldsymbol{x};\boldsymbol{\theta})$.

(b) (0.5 points) Write down the log-likelihood function of this problem.

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

Solution:

(a)(1 point) The expression of conditional distribution is as the following.

$$P_{Y|X}(y|\boldsymbol{x};\boldsymbol{\theta}) = \frac{1}{2\tau} \exp(\frac{-|y - \boldsymbol{\theta}^\top \boldsymbol{x} - \mu|}{\tau})$$

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

(b)(1 point) The log-likelihood function of this problem is as the following.

$$\begin{aligned}\log \boldsymbol{L}(\boldsymbol{\theta}) &= \log \prod_{i=1}^{m} P_{Y|X}(y^{(i)}|\boldsymbol{x}^{(i)}) \\ &= \log \prod_{i=1}^{m} \frac{1}{2\tau} \exp\left(\frac{-|y^{(i)} - \boldsymbol{\theta}^\top \boldsymbol{x}^{(i)} - \mu|}{\tau}\right) \\ &= -m \log(2\tau) - \frac{1}{\tau} \sum_{i=1}^{m} |y^{(i)} - \boldsymbol{\theta}^\top \boldsymbol{x}^{(i)} - \mu|\end{aligned}$$

(c)(1 point) Ordinary Least Squares (OLS) uses the ℓ^2 norm to measure the distance from data points to the linear model, aiming to minimize the sum of the squares of these distances. In contrast, Least Absolute Deviations (LAD) uses the ℓ^1 norm, that is, the absolute value, to measure distances.

Geometric interpretation:

- **OLS**: By minimizing the sum of squared errors, OLS is more sensitive to outliers because squaring amplifies larger errors. Geometrically, this means we are looking for a line such that the sum of the squared vertical distances from all points to this line is minimized.
- **LAD**: By minimizing the sum of absolute errors, LAD is more robust to outliers. Geometrically, this means we are looking for a line such that the sum of the absolute vertical distances from all points to this line is minimized.