CS534 Written Homework 1

Yongsung CHO

1. (Weighted linear regression) (15 pts) In class when discussing linear regression, we assume that the Gaussian noise is independently identically distributed. Now we assume the noises $\epsilon_1, \epsilon_2, \cdots \epsilon_n$ are independent but each $\epsilon_i \sim N(0, \sigma_i^2)$, i.e., it has its own distinct variance.

   (a) (3pts) Write down the log likelihood function of $\mathbf{w}$. $f(x) = \frac{e^{-x^2/2\sigma_i^2}}{\sigma_i\sqrt{2\pi}}$

$$y = w_0x_0 + w_1x_1 + w_2x_2 \cdots w_nx_n + \epsilon = w \cdot x + \epsilon$$

$$E(y|x,w) = w^Tx$$

$$P(y|x,w) = w^Tx + N(0,\sigma^2)$$

$$\prod_{i=1}^{n} p(x_i,y_i;w) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(y_i - w^Tx_i)^2}{2\sigma_i^2}\right)$$

$$= \left(\frac{1}{2\pi}\right)^{\frac{N}{2}} \sum_{i=1}^{n} \frac{1}{\sigma_i} \cdot \exp\left(-\frac{1}{2\sigma_i^2} \sum_{i=1}^{N} (y_i - w^Tx_i)^2\right)$$

$$\log \prod_{i=1}^{n} p(x_i,y_i;w) = -\frac{N}{2}\log(2\pi) + \log \sum_{i=1}^{n} \frac{1}{\sigma_i} - \sum_{i=1}^{N} \frac{1}{2\sigma_i^2}(y_i - w^Tx_i)^2$$

   (b) (4pts) Show that maximizing the log likelihood is equivalent to minimizing a weighted least square loss function $J(\mathbf{W}) = \frac{1}{2}\sum_{i=1}^{n} a_i(\mathbf{w}^T\mathbf{x}_i - y_i)^2$, and express each a_i in terms of σ_i.

log likelihood is

$$\log \prod_{i=1}^{n} p(x_i,y_i;w) = \underbrace{-\frac{N}{2}\log(2\pi) + \log \sum_{i=1}^{n} \frac{1}{\sigma_i}}_{\text{constant}} - \sum_{i=1}^{N} \frac{1}{2\sigma_i^2}(y_i - w^Tx_i)^2$$

To maximize the log likelihood,

$$\frac{\partial}{\partial w} L(w) = \left(-\sum_{i=1}^{N} \frac{1}{2\sigma_i^2}(y_i - w^Tx_i)^2\right)\frac{\partial}{\partial w}$$

To minimize a weighted least square loss func

$$\frac{\partial}{\partial w} J(w) = \left(\frac{1}{2}\sum_{i=1}^{N} a_i(w^Tx_i - y_i)^2\right)\frac{\partial}{\partial w}$$

$$\therefore a_i = \frac{1}{\sigma_i^2}$$

(c) (4 pts) Derive a batch gradient descent update rule for optimizing this objective.

$$\frac{d}{dw}J(w) = \left(\frac{1}{2}\sum_{i=1}^{N} a_i (w^T x_i - y_i)^2 \right)\frac{\partial}{\partial w}$$

$$= \frac{1}{2}\sum_{i=1}^{N} \frac{2}{\sigma_i^2}(w^T x_i - y_i)\, x_i$$

$$\Delta J(w) = \sum_{i=1}^{N} \frac{1}{\sigma_i^2}(w^T x_i - y_i)\, x_i$$

To optimize this object update w by using $\Delta J(w)$

Weight vector(w) needs to update values with opposite way of gradient

in addition, you need to set learning rate to update w

as a result,

$$w \leftarrow w - \underset{\text{(Learning rate)}}{\lambda} \cdot \Delta J(w) \quad \text{until } |\Delta J(w)| < \varepsilon$$

(until $\Delta J(w)$ is enough small to do not affect w)

(d) (4 pts) Derive a closed form solution to this optimization problem.

To derive a closed form solution, w doesn't need to update anymore. Because, the value of w is minimum point of loss function. The feature of the point that makes minimum of loss function is that $\Delta J(w) = 0$

As a result,

$$\sum_{i=1}^{N} \frac{1}{\sigma_i^2}(\hat{y}_i - y_i)\, x_i = 0 \quad \text{is}$$

a closed form solution

2. (14 pts) Consider the maximum likelihood estimation problem for multi-class logistic regression using the soft-max function defined below:

$$p(y=k|\mathbf{x}) = \frac{\exp(\mathbf{w}_k^T\mathbf{x})}{\sum_{j=1}^{K}\exp(\mathbf{w}_j^T\mathbf{x})}$$

(x_1, y_1) (x_2, y_2) $x_3\ y_3$ $\quad y = 1, 2, 3, 4, \ldots, K$

We can write out the likelihood function as:

$$L(\mathbf{w}) = \prod_{i=1}^{N}\prod_{k=1}^{K} p(y=k|\mathbf{x}_i)^{I(y_i=k)}$$

where $I(y_i = k)$ is the indicator function, taking value 1 if y_i is k.

(a) (2 pts) Compute the log-likelihood function.

$$\log L(w) = \log\left(\sum_{i=1}^{N}\sum_{k=1}^{K} P(y=k|x_i)^{I(y_i=k)}\right)$$

$$= \sum_{i=1}^{N}\sum_{k=1}^{K} I(y_i=k)\log(P(y=k|x_i))$$

$$= \sum_{i=1}^{N}\sum_{k=1}^{K} I(y_i=k)\log\left(\frac{\exp(w_k^T x)}{\sum_{j=1}^{K}\exp(w_j^T x)}\right)$$

$$= \sum_{i=1}^{N}\sum_{k=1}^{K} t_{ik}\log p_i \qquad T = \begin{bmatrix} t_{11} & t_{12} & \cdots & t_{1K} \\ \vdots & \vdots & \vdots & \vdots \\ t_{i1} & t_{i2} & \cdots & t_{iK} \end{bmatrix}$$

$$= \sum_{i=1}^{N} y_i \log(p_i)$$

(b) (12 pts) Compute the gradient of the log-likelihood function w.r.t the weight vector $\mathbf{w}_c$ of class c. (Precursor to this question, which terms are relevant for $\mathbf{w}_c$ in the loglikelihood function? Also hint: Logistic regression slide provides the solution to this problem, just need to fill in what is missing in between.)

let $a_j = w_k^T x_j + b \qquad p_j = \frac{\exp(w_j^T x)}{\sum_{k=1}^{K}\exp(w_k^T x)}$

$$\frac{\partial p_j}{\partial a_j} = \frac{\exp(w_j^T x)\left(\sum_{k=1}^{K}\exp(w_k^T x) - \exp(w_j^T x)\right)}{\left(\sum_{k=1}^{K}\exp(w_k^T x)\right)^2}$$

$$= \frac{\exp(w_j^T x)}{\sum_{k=1}^{K}\exp(w_k^T x)} \times \frac{\sum_{k=1}^{K}\exp(w_k^T x) - \exp(w_j^T x)}{\sum_{k=1}^{K}\exp(w_k^T x)}$$

$$= p_i(1 - p_j)$$

log likelihood $L(w) = \sum_{i}^{N} y_i \log(p_i)$

$$\frac{\partial L}{\partial a_i} = \sum_{k=1}^{N} y_k \frac{\partial \log(p_k)}{\partial a_i} = \sum_{k=1}^{N} y_k \frac{\partial \log(p_k)}{\partial p_k} \times \frac{\partial p_k}{\partial a_i}$$

$$= \sum_{k=1}^{N} y_k \frac{1}{p_k} \times \frac{\partial p_k}{\partial a_i}$$

$$p_k(1-p_i)$$

$$= y_k(1-p_i)$$

W.r.t weight vector w_c and C

$$\delta L(w) = y_c(1-p_c)$$

$$= I(y_i = k)\left(1 - \frac{\exp(w_c^T x)}{\sum_{j=1}^{k} \exp(w_j^T x)}\right)$$

3. (11 pts) (Maximum A Posterior Estimation.) Suppose we observe the values of n IID random variables $X_1, \ldots, X_n$ drawn from a single Bernoulli distribution with parameter θ. In other words, for each X_i, we know that $P(X_i = 1) = \theta$ and $P(X_i = 0) = 1 - \theta$. In the Bayesian framework, we treat θ as a random variable, and use a prior probability distribution over θ to express our prior knowledge/preference about θ. In this framework, $X_1, \ldots, X_n$ can be viewed as generated by:

   - First, the value of θ is drawn from a given prior probability distribution
   - Second, $X_1, \ldots, X_n$ are drawn independently from a Bernoulli distribution with this θ value.

   In this setting, Maximum A Posterior (MAP) estimation is a natural way to estimate the value of θ by choosing the most probable value given both its prior distribution and the observed data $X_1, \ldots, X_n$. Specifically, the MAP estimation of θ is given by

$$\begin{aligned}\hat{\theta}_{MAP} &= \underset{\hat{\theta}}{\operatorname{argmax}}\, P(\theta = \hat{\theta} | X_1, \ldots, X_n) \\ &= \underset{\hat{\theta}}{\operatorname{argmax}}\, P(X_1, \ldots, X_n | \theta = \hat{\theta}) P(\theta = \hat{\theta}) \\ &= \underset{\hat{\theta}}{\operatorname{argmax}}\, L(\hat{\theta}) p(\hat{\theta})\end{aligned}$$

where $L(\hat{\theta})$ is the data likelihood function and $p(\hat{\theta})$ is the density function of the prior. Now consider using a beta distribution for prior: θ $Beta(\alpha, \beta)$, whose PDF function is

$$p(\hat{\theta}) = \frac{\hat{\theta}^{(\alpha-1)}(1-\hat{\theta})^{(\beta-1)}}{B(\alpha, \beta)}$$

where $B(\alpha, \beta)$ is a normalizing constant to make it a proper probability density function.

(a) (5 pts) Derive the posterior distribution $p(\hat{\theta} | X_1, \ldots, X_n, \alpha, \beta)$ and show that it is also a Beta distribution.

According to Bayes rule,

$$P(\hat{\theta} | X_1, \cdots X_n, \alpha, \beta) = \frac{P(X_1 X_2 X_3 \cdots X_n \alpha, \beta | \hat{\theta})\, P(\hat{\theta})}{P(X_1, X_2, X_3 \cdots X_n, \alpha, \beta)}$$

$P(X_1, X_2 \cdots X_n, \alpha\beta) \Rightarrow$ Constant ($\because$ independent of θ)

$P(\hat{\theta}) = \dfrac{\hat{\theta}^{(\alpha-1)}(1-\hat{\theta})^{(\beta-1)}}{B(\alpha, \beta)} \Rightarrow$ prior distribution of $\hat{\theta}$

$P(X_1 X_2 X_3 \cdots X_n, \alpha, \beta | \hat{\theta}) \Rightarrow$ likelihood function of $\hat{\theta}$

Bc $X_1, X_2 \cdots X_n$ are drawn from a Bernoulli distribution,

$$L(\hat{\theta}) = \prod \theta^{X_i} ((1-\theta)^{1-X_i}$$

$L(\hat{\theta})$ is some specific value.

As a result $P(\hat{\theta} | X_1, \cdots X_n, \alpha, \beta) =$ $\dfrac{P(X_1 X_2 X_3 \cdots X_n \alpha, \beta | \hat{\theta})\, P(\hat{\theta})}{P(X_1, X_2, X_3 \cdots X_n, \alpha, \beta)}$

Beta distribution

$L(\hat{\theta})$ follow Beta distribution

constant

(b) (6 pts) Suppose we use $Beta(2,2)$ as the prior, What is the posterior distribution of θ after we observe 5 coin tosses and 2 of them are head? What is the posterior distribution of θ after we observe 50 coin tosses and 20 of them are head? Plot the pdf function of these two posterior distributions. Assume that $\theta = 0.4$ is the true probability, as we observe more and more coin tosses from this coin, what do you expect to happen to the posterior?

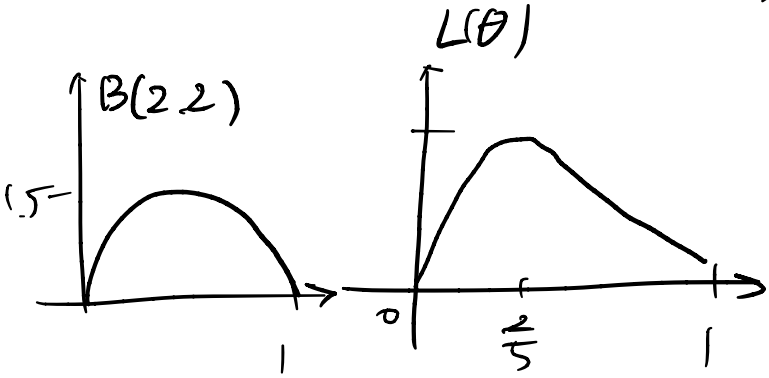
we can observe 5 cointosses and 2 of them are head, according to the result, we can know that the value of θ is $\frac{2}{5} = 0.4$

posterior distribution of $\theta = \frac{P(X_1 X_2 X_3 \cdots \alpha\beta|\theta) \cdot P(\theta)}{P(X_1 X_2 X_3 \cdots \alpha\beta)}$

let $\frac{1}{P(X_1 X_2 \cdots \alpha\beta)} = \mu$ = constant

$$= \prod_{i=1} \theta^{x_i}(1-\theta)^{1-x_i} \cdot \frac{\theta \cdot (1-\theta)}{B(2,2)} \cdot \mu$$

$$= \frac{\frac{2}{5} \times \frac{3}{5}}{B(2,2)} \times \prod_{i=1} \left(\frac{2}{5}\right)^{x_i} \left(\frac{3}{5}\right)^{1-x_i} \cdot \mu$$



if try more time,