Haizhuan Wang
UCID: 12213470
Cnetid: haichuan
Collaborator: Faradawn (Zeyuan) Yang

$$w^* = \underset{w}{\arg\min} \left\{ \sum_{i=1}^{N} L(w, x_i, y_i) + \lambda \sum_{j=1}^{d} |w_j|^p \right\} \quad [1]$$

Problem 1: show that the objective in (1) is equivalent to

$$w^* = \underset{w}{\arg\min} \sum_{i=1}^{N} L(w, x_i, y_i) \quad \text{subject to} \quad \sum_{j=1}^{d} |w_j|^p \le r \quad [2]$$

Proof: Both [1] and [2] are convex and smooth

The solution for the first problem is for $j \in [d]$:

$$\sum_{i=1}^{N} \frac{\partial L(\hat{w}, x_i, y_i)}{\partial w_j} + \lambda p |\hat{w}_j|^{p-1} \cdot \operatorname{sgn}(\hat{w}_j) = 0$$

The KKT condition of the second problem is

$$\sum_{i=1}^{N} L(w, x_i, y_i) + u\left(r - \sum_{j=1}^{d} |w_j|^p\right) \quad (3)$$

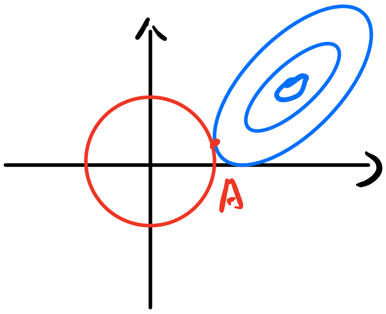
for $j \in [d]$ $\displaystyle\sum_{i=1}^{N} \frac{\partial L(\hat{w}, x_i, y_i)}{\partial w_j} - u p |\hat{w}_j|^{p-1} \cdot \operatorname{sgn}(\hat{w}_j) = 0$

and

$$u\left(r - \sum_{j=1}^{d} |\hat{w}_j|^p\right) = 0$$

First of all. By KKT, only when of u and $r - \sum_{j=1}^{d} |w_j|^p$ equals to zero. Now observe that when $u = -\lambda$, then essentially (1) and (3) are doing the same minimization problem as λr is a constant term. Hence, we can substitute $u = -\lambda$

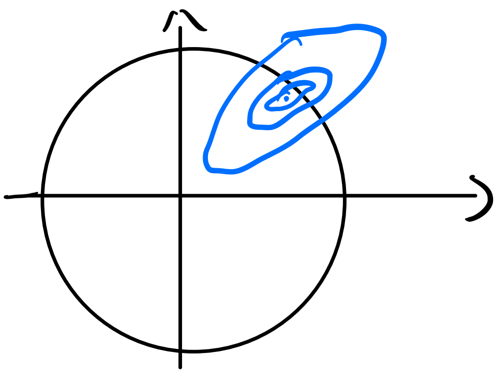
Case 1: if $r = \sum_{j=1} |w_j|$, graphically this corresponds to the following case



$r = \sum_{j=1}^{d} |\hat{w}_j|^p$, so the optimal solution is taken when the constraint and the objective touches each other

In this case, $\mu \neq 0$. Now observe that when $u = -\lambda$, then essentially (1) and (3) are doing the same minimization problem as λr is a constant term. Hence, if $\lambda \neq 0$ in [1], then let $r = \sum_{j=1}^{d} |w_j|^p$ in [2].

Case 2: if $u = 0$. then $\lambda = 0$, $r > \sum_{j=1}^{d} |w_j|^p$



Graphically, this correspond to the case when the constraint includes the w^*, so we don't need to find the "tangency point" and the constraint is loose at the optimal

Hence, to summarize the solution

$$\begin{cases} \text{if } \lambda = 0, & \text{then } r > \sum_{j=1}^{d} |\hat{w}_j|^p \\ \text{if } \lambda > 0, & \text{then } r = \sum_{j=1}^{d} |\hat{w}_j|^p \end{cases}$$

($\hat{w}_j$ depends on the dataset)

Problem 2

$$y = w \cdot x + r, \quad r \sim N(0, \sigma_x^2) \qquad (3)$$

Without knowing anything else besides the assumption in (3), can we compute the maximum likelihood estimate for the linear regression parameters w^* from a given dataset under this noise model?

Answer: No. we cannot.

for a given (x, y)

$$p(y|x, w, \sigma_x) = \prod_{i=1}^{N} p(y_i|x_i; w, \sigma_{x_i})$$

$$\hat{w}_L = \arg\max_{w} \prod_{i=1}^{N} \frac{1}{\sigma_{x_i}\sqrt{2\pi}} \exp\left(-\frac{(y_i - f(x_i; w))^2}{2\sigma_{x_i}^2}\right)$$

The problem is in this setting, we do not know the value of each σ_{x_i} which corresponds to x_i. In other words, we don't have exact knowledge about the distribution of any certain y_i, so we cannot calculate the term containing σ_{x_i} in the above expression.

Problem 3 Now suppose we know the value of the noise variance $\sigma_{x_i}^2$ at every training input x_i for $i = 1, \cdots N$. Can we compute maximum likelihood estimate for w^*?

Answer. Yes, we can.

$$p(y, X, w, \sigma_x) = \prod_{i=1}^{N} p(y_i | x_i ; w, \sigma_{x_i})$$

$$\hat{w}_L = \underset{w}{\operatorname{argmax}} \prod_{i=1}^{N} \frac{1}{\sigma_{x_i}\sqrt{2\pi}} \exp\left(-\frac{(y_i - f(x_i, w))^2}{2\sigma_{x_i}}\right)$$

We know the value of each x_i, y_i, σ_{x_i} for $i \in [N]$

Hence, $\hat{w}_L$ can be calculated using the above formula.

The intuition is we know the distribution of the noise now, so we can measure how close y_i is from $f(x_i, w)$ [the likelihood that y_i can be explained by the parameter]

$$\log p(y | X; w, \sigma) = \frac{1}{N}\sum_{i=1}^{N}\left[-\frac{(y_i - f(x_i, w))^2}{2\sigma_{x_i}^2} - \log \sigma_{x_i}\sqrt{2\pi}\right]$$

$$= -\frac{1}{2N}\sum_{i=1}^{N}\left[\frac{(y_i - f(x_i, w))^2}{2\sigma_{x_i}^2}\right] - \frac{1}{N}\sum_{i=1}^{N}\log \sigma_{x_i} - \log\sqrt{2\pi}$$

Hence $\underset{w}{\operatorname{argmax}} \log p(y | X; w, \sigma) = \underset{w}{\operatorname{argmin}} \sum_{i=1}^{N} \frac{(y_i - f(x_i, w))^2}{2\sigma_{x_i}}$

$g(w) = \sum_{i=1}^{N} \frac{(y_i - f(x_i, w))^2}{2\sigma_{x_i}}$ is convex, optimal point taken when FOC is satisfied.

$$\frac{\partial g(w)}{\partial w_i} = 0 \qquad \text{for } i \in [n]$$

w is the vector that satisfies all of the FOC constraints.

Problem 4: Show that the softmax model is over parameterized. that is, show that for any w_c for $c=1, \cdots, c$, there is a different value that yields exactly the same $p(y|x)$ for every x. Then show for $C=2$ softmax is equivalent to logistic regression.

Pf: Let θ be a fixed vector, consider $w_c - \theta$

$$p(y=c|x) = \frac{\exp[(w_c-\theta)x]}{\sum_{k=1}^{C}\exp[(w_c-\theta)x]}$$

$$= \frac{\exp[w_c \cdot x] / \exp(\theta x)}{\sum_{k=1}^{C}\exp(w_c \cdot x) / \exp(\theta x)}$$

$$= \frac{\exp(w_c \cdot x)}{\sum_{k=1}^{C}\exp(w_c \cdot x)} = \text{softmax}(w_c \cdot x)$$

Hence, if $(w_1, w_2, \cdots w_c)$ minimizes the log loss, then $(w_1-\theta, w_2-\theta, \cdots, w_c-\theta)$ also minimizes the log loss.

Since choice of θ is arbitrary, we can set $\theta = w_k$ for $k \in [C]$ then the k-th row of W will become $w_k - w_k = \vec{0}$. We just need to optimize the other $C-1$ parameters of the softmax.

For softmax when $C=2$, we have

$$p(y=1|x) = \frac{\exp(w_1 \cdot x)}{\exp(w_1 \cdot x) + \exp(w_2 \cdot x)}$$

$$= \frac{\exp(W_1 x)/\exp(W_0 x)}{[\exp(W_1 \cdot x)\ /\exp(W_0 \cdot x)] + 1}$$

$$= \frac{\exp[(W_1 - W_0)x]}{1 + \exp[(W_1 - W_0)x]}$$

$$= \frac{1}{1 + e^{-(W_1 - W_0)x}}$$

This is exactly the form of logistic regression.

Problem 5:

(a) the log loss of the linear C-way softmax classification model using feature mapping $\phi(x)$

Answer:

Let $$p(y^{(i)}=c\mid x)=\frac{\exp(W_c\cdot x^{(i)})}{\sum_{j=1}^{C}\exp(W_j\cdot x^{(i)})}$$

$$L(W)=-\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{C}1\{y^{(i)}=k\}\cdot\log(p(y^{(i)}=c\mid x^{(i)}))$$

(b) its gradient with respect to W

WLOG, consider $\frac{\partial L(W)}{\partial W_1}$

$$\frac{\partial L(W)}{\partial W_1}=-\frac{1}{N}\sum_{i=1}^{N}\Big[\frac{\partial}{\partial W_1}\Big[1\{y^{(i)}=1\}\cdot\log\Big(\frac{e^{W_1\cdot x^{(i)}}}{\sum_{j=1}^{C}e^{W_j\cdot x^{(i)}}}\Big)+\cdots+1\{y^{(i)}=c\}\cdot\log\Big(\frac{e^{W_c x^{(i)}}}{\sum_{j=1}^{C}e^{W_j\cdot x^{(i)}}}\Big)\Big]$$

First deal with:

$$\frac{\partial\Big[1\{y^{(i)}=1\}\cdot\log\Big(\frac{e^{W_1\cdot x^{(i)}}}{\sum_{j=1}^{C}e^{W_j\cdot x^{(i)}}}\Big)\Big]}{\partial W_1}$$

$$=\frac{\partial\Big[1\{y^{(i)}=1\}\cdot\Big[W_1\cdot x^{(i)}-\log\sum_{j=1}^{C}e^{W_j\cdot x^{(i)}}\Big]\Big]}{}$$

$$= 1\{y^{(i)}=1\} \cdot x^{(i)} - 1\{y^{(i)}=1\} \cdot \frac{\exp(w_1 \cdot x^{(i)}) \cdot x^{(i)}}{\sum_{j=1}^{C} \exp(w_j \cdot x^{(i)})}$$

Then deal with $k \neq 1$, WLOG, $k=2$.

$$\frac{\partial \left[1\{y^{(i)}=2\} \cdot \log \left(\frac{e^{w_2 \cdot x^{(i)}}}{\sum_{j=1}^{C} e^{w_j \cdot x^{(i)}}} \right) \right]}{\partial w_1}$$

$$= \frac{\partial \left[1\{y^{(i)}=2\} \cdot \left(w_2 \cdot x^{(i)} - \log \sum_{j=1}^{C} e^{w_j \cdot x^{(i)}} \right) \right]}{\partial w_1}$$

$$= -1\{y^{(i)}=2\} \cdot \frac{\exp(w_1 \cdot x^{(i)}) \cdot x^{(i)}}{\sum_{j=1}^{C} \exp(w_j \cdot x^{(i)})}$$

Hence $\frac{\partial L(w)}{\partial w_1} = -\frac{1}{N} \sum_{i=1}^{N} \left[1\{y^{(i)}=1\} \cdot x^{(i)} - \frac{\exp(w_1 \cdot x^{(i)}) \cdot x^{(i)}}{\sum_{j=1}^{C} \exp(w_j \cdot x^{(i)})} \right]$

Hence, the new parameter w_{t+1} is

$$w_{t+1} = w_t + \frac{1}{N} \cdot \alpha \cdot \sum_{i=1}^{N} x^{i} \cdot \left(1\{y^{(i)}=k\} - \frac{\exp(w_k \cdot x^{(i)})}{\sum_{j=1}^{C} \exp(w_j \cdot x^{(i)})} \right)$$

α is the learning rate

Problem 6 is written in haichuan_wang_sol2_P6.ipynb.
Problem 8 is written in haichuan_wang_sol2_P8.ipynb.