Machine Learning | **Spring 2017**

Solution : Homework 1

Lecturer: Yang Yang | *Homework taker: Li Xu*

Due Time: March 12

Problem 1. For parameter w, try to prove that logistic regression function $y = \frac{1}{1+e^{-(\boldsymbol{w^T x}+b)}}$ is non-convex, but its logarithmic likelihood function $l(\boldsymbol{w}) = \sum_{i=1}^{m}(-y_i(\boldsymbol{w^T x_i}+b)+\ln(1+e^{\boldsymbol{w^T x_i}+b}))$ is convex.

Solution: To prove $y = \frac{1}{1+e^{-(\boldsymbol{w^T x}+b)}}$ is non-convex, we just need to prove $g(z) = \frac{1}{1+e^{-z}}$ is non-convex as $z = \boldsymbol{w^T x} + b$ is linear.

Proof. $g(z) = \frac{1}{1+e^{-z}}$ is non-convex □

$$
\begin{aligned}
g'(z) &= \frac{d}{dz}\frac{1}{1+e^{-z}} \\
&= \frac{1}{1+e^{-z}}(1-\frac{1}{1+e^{-z}}) \\
&= g(z)(1-g(z))
\end{aligned}
$$

$$
\begin{aligned}
g''(z) &= \frac{d}{dz}g(z) \\
&= g'(z)(1-g(z)) - g(z)g'(z) \\
&= g'(z) - 2g(z)g'(z) \\
&= g(z)(1-g(z))(1-2g(z))
\end{aligned}
$$

As the range of g(z) is from $(0,+\infty)$, g”(z) is not constant greater than 0. So $g(z) = \frac{1}{1+e^{-z}}$ is non-convex.

Similarly, to prove $l(\boldsymbol{w}) = \sum_{i=1}^{m}(-y_i(\boldsymbol{w^T x_i}+b)+\ln(1+e^{\boldsymbol{w^T x_i}+b}))$ is convex, we just need to prove $\ln(1+e^z)$ is convex as both $-y_i(\boldsymbol{w^T x_i}+b$ and $\boldsymbol{w^T x_i}+b$ is linear to $\boldsymbol{w}$.

Proof. $g(z) = \ln(1+e^z)$ is convex □

$$
g'(z) = \frac{d}{dz}\ln(1+e^z) = \frac{e^z}{1+e^z}
$$

$$
g''(z) = \frac{d}{dz}\frac{e^z}{1+e^z} = \frac{e^z(1+e^z)-e^{2z}}{(1+e^z)^2} = \frac{e^z}{(1+e^z)^2} > 0
$$

■

Problem 2. Using the technique of Lagrange multipliers, show that minimization of the regularized error function

$$
\frac{1}{2}\sum_{n=1}^{N}\{t_n - \boldsymbol{w^T\phi}(x_n)\}^2 + \frac{\lambda}{2}\sum_{j=1}^{M}|w_j|^q
$$

is equivalent to minimizing the unregularized sum-of-squares error

$$E_D(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N}\{t_n - \boldsymbol{w^T}\boldsymbol{\phi}(x_n)\}^2$$

subject to the constraint $\sum_{j=1}^{M}|w_j|^q \leq \eta$.Discuss the relationship between the parameters η and λ.

Solution: We first define $E(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N}\{t_n - \boldsymbol{w^T}\boldsymbol{\phi}(x_n)\}^2 + \frac{\lambda}{2}\sum_{j=1}^{M}|w_j|^q$

To minimize $E(\boldsymbol{w})$, we have

$$\frac{\partial E}{\partial \boldsymbol{w}} = 0$$

To minimize $E_D(\boldsymbol{w})$ subject to $\sum_{j=1}^{M}|w_j|^q \leq \eta$. We define:

$$g(\boldsymbol{w}) = \frac{1}{2}(\sum_{j=1}^{M}|w_j|^q - \eta)$$

Using Lagrange multipliers, let

$$L(\boldsymbol{w}, \lambda) = E_D(\boldsymbol{w}) + \lambda \cdot g(\boldsymbol{w})$$

and solve

$$\nabla_{\boldsymbol{w}} L(\boldsymbol{w}, \lambda) = 0$$

with $\lambda \cdot g(\boldsymbol{w}) = 0$.

As $L(\boldsymbol{w}, \lambda) = E(\boldsymbol{w} - \lambda\eta)$ and $\lambda\eta$ is not relevant to $\boldsymbol{w}$.

$$\nabla_{\boldsymbol{w}} L(\boldsymbol{w}, \lambda) = 0$$

is equivalent to

$$\frac{\partial E}{\partial \boldsymbol{w}} = 0$$

So minimize $E(\boldsymbol{w})$ is equivalent to minimize $E_D(\boldsymbol{w})$ subject to $\sum_{j=1}^{M}|w_j|^q \leq \eta$.

For the relationship between λ and η. As we have $\lambda \cdot g(\boldsymbol{w}) = 0$, if $\lambda \neq 0$, $\eta = \sum_{j=1}^{M}|w_j|^q$. ∎

Problem 3. Consider a data set in which each data point t_n is associated with a weighting factor $r_n > 0$, so that the sum-of-squares error function becomes

$$E_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N} r_n\{t_n - \mathbf{w}^T\phi(\mathbf{x}_n)\}^2.$$

where $\phi(\mathbf{x}_n)$ is basis function. Find an expression for the solution $\mathbf{w}^*$ that minimizes this error function.

Solution: To minimize

$$E_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N} r_n\{t_n - \mathbf{w}^T\phi(\mathbf{x}_n)\}^2.$$

$\mathbf{w}$ should satisfy:

$$\frac{\partial}{\partial \mathbf{w}} E_D(\mathbf{w}) = -\sum_{n=1}^{N} r_n\{t_n - \mathbf{w}^T\phi(\mathbf{x}_n)\}\phi(\mathbf{x}_n) = 0$$

Solve for $\mathbf{w}$:

$$\sum_{n=1}^{N} r_n t_n \phi(\mathbf{x}_n) = (\sum_{n=1}^{N} r_n \phi(\mathbf{x}_n)\phi(\mathbf{x}_n)^T)\mathbf{w}$$
$$\mathbf{w} = (\sum_{n=1}^{N} r_n \phi(\mathbf{x}_n)\phi(\mathbf{x}_n)^T)^{-1}(\sum_{n=1}^{N} r_n t_n \phi(\mathbf{x}_n))$$

■

Problem 4. ......

Solution: We define $\mu_{ic} = P(y_i = c|x_i, W), y_{ic} = 1\{y_i = c\}$

(a)

$$l(W) = log \prod_{i=1}^{n}\prod_{c=1}^{C} y_{ic} log\mu_{ic} = \sum_{i=1}^{n}(\sum_{c=1}^{C} y_{ic} w_c^T x_i - log \sum_{c'=1}^{C} exp(w_{c'}^T x_i))$$

(b)

$$\begin{aligned}
g_c(W) &= \frac{\partial}{\partial w_c}\sum_{i=1}^{n}(\sum_{c=1}^{C} y_{ic} w_c^T x_i - log \sum_{c'=1}^{C} exp(w_{c'}^T x_i)) \\
&= \sum_{i=1}^{n}(\frac{\partial}{\partial w_c}\sum_{c=1}^{C} y_{ic} w_c^T x_i - \frac{\partial}{\partial w_c} log \sum_{c'=1}^{C} exp(w_{c'}^T x_i)) \\
&= \sum_{i=1}^{n}(y_{ic}x_i - \frac{\frac{\partial}{\partial w_c}\sum_{c'=1}^{C} exp(w_{c'}^T x_i)}{\sum_{c'=1}^{C} exp(w_{c'}^T x_i)}) \\
&= \sum_{i=1}^{n}(y_{ic}x_i - \frac{exp(w_c^T x_i)x_i}{\sum_{c'=1}^{C} exp(w_{c'}^T x_i)}) \\
&= \sum_{i=1}^{n}(y_{ic} - \mu_{ic})x_i
\end{aligned}$$

(c) $\delta_{cc'}$ denotes the Dirac delta function and is equal to one if $c' = c$ and zero otherwise.

$$\begin{aligned}
H_{c,c'}(W) &= \frac{\partial}{\partial w_c} g_{c'}(W) \\
&= \frac{\partial}{\partial w_c}\sum_{i=1}^{n}(y_{ic'}x_i - \frac{exp(w_{c'}^T x_i)x_i}{\sum_{c''=1}^{C} exp(w_{c''}^T x_i)}) \\
&= -\sum_{i=1}^{n}\frac{\partial}{\partial w_c}\frac{exp(w_{c'}^T x_i)x_i}{\sum_{c''=1}^{C} exp(w_{c''}^T x_i)} \\
&= -\sum_{i=1}^{n}(\delta_{cc'}\mu_{ic'}\mu_{ic})x_i x_i^T \\
&= \sum_{i=1}^{n}\mu_{ic}(\mu_{ic'} - \delta_{cc'})x_i x_i^T
\end{aligned}$$

■

Problem 5. ......

Solution: program is in the folder "B16037910007-LiXu-hw1-program" using 1. Stochastic Gradient Descent 2. Batch Gradient Ascent Method 3. Newtons method 4. Normal Equation

comparing RMSE: Normal Equation has best result. ■