CS 229, Fall 2018

Problem Set #1 Solutions: Supervised Learning

1.

(a)

$$
\begin{aligned}
\frac{\partial J(\theta)}{\partial \theta_j} &= -\frac{1}{m}\sum_{i=1}^{m} y^{(i)}\frac{g(\theta^T x^{(i)})[1-g(\theta^T x^{(i)})]}{g(\theta^T x^{(i)})}x_j^{(i)} - (1-y^{(i)})\frac{g(\theta^T x^{(i)})[1-g(\theta^T x^{(i)})]}{1-g(\theta^T x^{(i)})}x_j^{(i)} \\
&= -\frac{1}{m}\sum_{i=1}^{m} y^{(i)}[1-g(\theta^T x^{(i)})]x_j^{(i)} - (1-y^{(i)})g(\theta^T x^{(i)})x_j^{(i)} \\
&= \frac{1}{m}\sum_{i=1}^{m}[g(\theta^T x^{(i)}) - y^{(i)}]x_j^{(i)}
\end{aligned}
$$

$$\nabla_\theta J(\theta) = \frac{1}{m}X^T(g(X\theta) - Y)$$

$$H_{jk} = \frac{\partial^2 J(\theta)}{\partial\theta_j\partial\theta_k} = \frac{1}{m}\sum_{i=1}^{m} g(\theta^T x^{(i)})[1-g(\theta^T x^{(i)})]x_j^{(i)}x_k^{(i)}$$

$$H = \frac{1}{m}[X^T\cdot g(X\theta)\cdot(1-g(X\theta))]X$$

$$
\begin{aligned}
z^T Hz &= \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{n} g(\theta^T x^{(i)})[1-g(\theta^T x^{(i)})]x_j^{(i)}x_k^{(i)}z_j z_k \\
&= \frac{1}{m}\sum_{i=1}^{m} g(\theta^T x^{(i)})[1-g(\theta^T x^{(i)})][(x^{(i)})^T z]^2 \geq 0
\end{aligned}
$$

(c)

$$
\begin{aligned}
p(y=1|x) &= \frac{p(x|y=1)p(y=1)}{p(x|y=1)p(y=1)+p(x|y=0)p(y=0)} \\
&= \frac{\exp\{-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\}\phi}{\exp\{-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\}\phi + \exp\{-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)\}(1-\phi)} \\
&= \frac{1}{1+\exp\{\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1) - \frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)\}\frac{1-\phi}{\phi}} \\
&= \frac{1}{1+\exp\{-[(\Sigma^{-1}(\mu_1-\mu_0))^T x + \frac{1}{2}(\mu_0+\mu_1)^T\Sigma^{-1}(\mu_0-\mu_1) - \ln(\frac{1-\phi}{\phi})]\}}
\end{aligned}
$$

$$\theta = \Sigma^{-1}(\mu_1-\mu_0)$$

$$\theta_0 = \frac{1}{2}(\mu_0+\mu_1)^T\Sigma^{-1}(\mu_0-\mu_1) - \ln(\frac{1-\phi}{\phi})$$

(d)

$$\mu_{y^{(i)}} = 1\{y^{(i)}=0\}\mu_0 + 1\{y^{(i)}=1\}\mu_1$$

$$p(x^{(i)}|y^{(i)};\mu_0,\mu_1,\Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}\exp\Big\{-\frac{1}{2}(x^{(i)}-\mu_{y^{(i)}})^T\Sigma^{-1}(x^{(i)}-\mu_{y^{(i)}})\Big\}$$

$$p(y^{(i)};\phi) = \phi^{1\{y^{(i)}=1\}}(1-\phi)^{1-1\{y^{(i)}=1\}}$$

$$\begin{aligned}
\ell &= \sum_{i=1}^m \log p(x^{(i)}|y^{(i)};\mu_0,\mu_1,\Sigma) + \sum_{i=1}^m \log p(y^{(i)};\phi) \\
&= \sum_{i=1}^m \log \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\Big\{-\frac{1}{2}(x^{(i)}-\mu_{y^{(i)}})^T\Sigma^{-1}(x^{(i)}-\mu_{y^{(i)}})\Big\} + \sum_{i=1}^m \log \phi^{1\{y^{(i)}=1\}}(1-\phi)^{1-1\{y^{(i)}=1\}} \\
&= -\frac{mn}{2}\log(2\pi) - \frac{m}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^m (x^{(i)}-\mu_{y^{(i)}})^T\Sigma^{-1}(x^{(i)}-\mu_{y^{(i)}}) \\
&\quad + \sum_{i=1}^m 1\{y^{(i)}=1\}\log\phi + \Big(m - \sum_{i=1}^m 1\{y^{(i)}=1\}\Big)\log(1-\phi)
\end{aligned}$$

$$\frac{\partial \ell}{\partial \phi} = \frac{1}{\phi}\sum_{i=1}^m 1\{y^{(i)}=1\} + \frac{1}{\phi-1}(m - \sum_{i=1}^m 1\{y^{(i)}=1\})$$

$$\frac{\partial \ell}{\partial \mu_{y^{(i)}}} = \Sigma^{-1}\sum_{i=1}^m (x^{(i)}-\mu_{y^{(i)}})$$

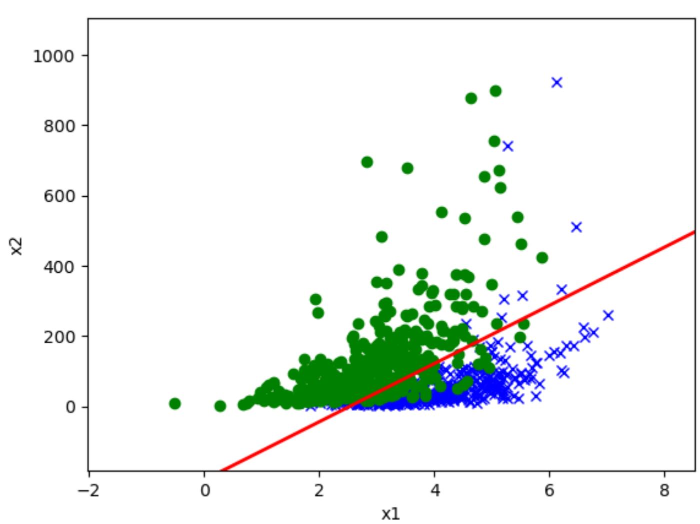
$$\frac{\partial \mu_{y^{(i)}}}{\partial \mu_0} = 1\{y^{(i)}=0\},\ \frac{\partial \mu_{y^{(i)}}}{\partial \mu_1} = 1\{y^{(i)}=1\}$$

$$\begin{aligned}
\frac{\partial \ell}{\partial \mu_0} &= \frac{\partial \ell}{\partial \mu_{y^{(i)}}}\frac{\partial \mu_{y^{(i)}}}{\partial \mu_0} = \Sigma^{-1}\sum_{i=1}^m \Big(x^{(i)}1\{y^{(i)}=0\} - \mu_0 1\{y^{(i)}=0\}\Big) \\
\frac{\partial \ell}{\partial \mu_1} &= \frac{\partial \ell}{\partial \mu_{y^{(i)}}}\frac{\partial \mu_{y^{(i)}}}{\partial \mu_1} = \Sigma^{-1}\sum_{i=1}^m \Big(x^{(i)}1\{y^{(i)}=1\} - \mu_1 1\{y^{(i)}=1\}\Big)
\end{aligned}$$

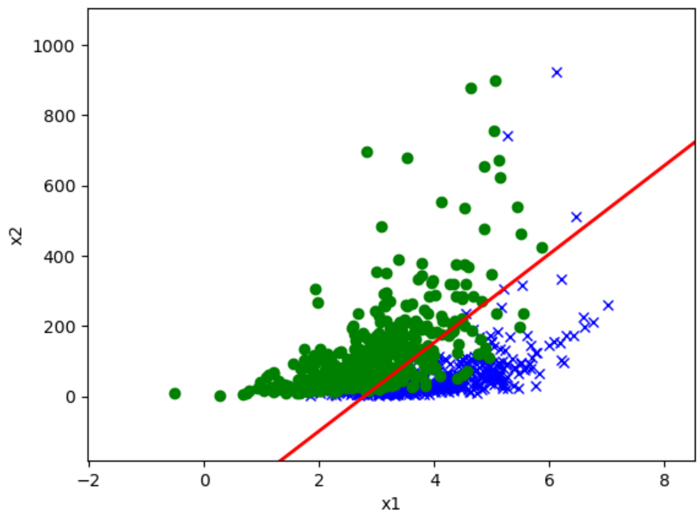
$$\frac{\partial \ell}{\partial \Sigma} = -\frac{m}{2}\Sigma^{-1} + \frac{1}{2}\Sigma^{-1}\Big(\sum_{i=1}^m (x^{(i)}-\mu_{y^{(i)}})(x^{(i)}-\mu_{y^{(i)}})^T\Big)\Sigma^{-1}$$

$$\begin{cases}
\frac{\partial \ell}{\partial \phi} &= 0 \\
\frac{\partial \ell}{\partial \mu_0} &= 0 \\
\frac{\partial \ell}{\partial \mu_1} &= 0 \\
\frac{\partial \ell}{\partial \Sigma} &= 0
\end{cases}
\Rightarrow
\begin{cases}
\phi &= \frac{1}{m}\sum_{i=1}^m 1\{y^{(i)}=1\} \\
\mu_0 &= \frac{\sum_{i=1}^m 1\{y^{(i)}=0\}x^{(i)}}{\sum_{i=1}^m 1\{y^{(i)}=0\}} \\
\mu_1 &= \frac{\sum_{i=1}^m 1\{y^{(i)}=1\}x^{(i)}}{\sum_{i=1}^m 1\{y^{(i)}=1\}} \\
\Sigma &= \frac{1}{m}\sum_{i=1}^m (x^{(i)}-\mu_{y^{(i)}})(x^{(i)}-\mu_{y^{(i)}})^T
\end{cases}$$

(f)

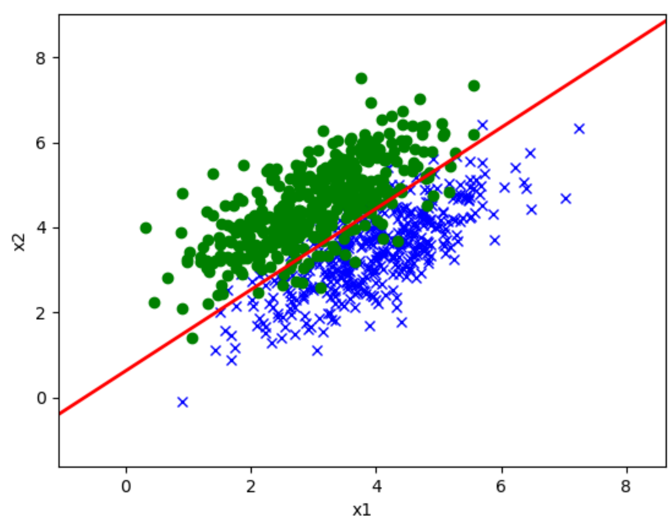


logistic regression

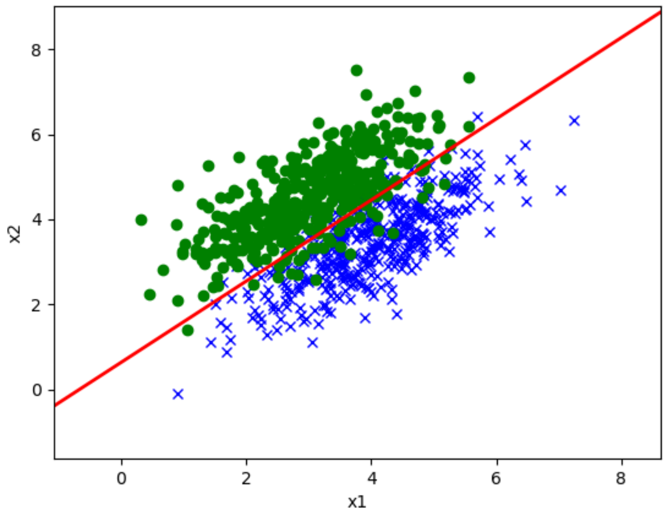


GDA

(g)



logistic regression



GDA

On Dataset 1 GDA perform worse than logistic regression.

Because $p(x|y)$ may be not Gaussian distribution.

(h)

Box-Cox transformation.

2.