CS229 Problem Set 1

Tianyu Du

Sunday 7th July, 2019

1 Question 1: Linear Classifiers

1.1 Question 1(a)

Lemma 1.1.

$$g'(z) = g(z)\left(1 - g(z)\right) \tag{1.1}$$

Lemma 1.2. For every $x, z \in \mathbb{R}^n$,

$$\sum_i \sum_j z_i x_i z_j x_j = (x^T z)^2 \geq 0 \tag{1.2}$$

$$\begin{aligned}
\nabla_\theta J(\theta) &= -\frac{1}{n}\sum_{i=1}^{n}\left(y^{(i)}\frac{g'(\theta^T x^{(i)})}{g(\theta^T x^{(i)})} - (1-y^{(i)})\frac{g'(\theta^T x^{(i)})}{1-g(\theta^T x^{(i)})}\right)x^{(i)} && (1.3)\\
&= -\frac{1}{n}\sum_{i=1}^{n}\left(y^{(i)}(1-g(\theta^T x^{(i)})) - (1-y^{(i)})g(\theta^T x^{(i)})\right)x^{(i)} && (1.4)\\
&= -\frac{1}{n}\sum_{i=1}^{n}\left(y^{(i)} - g(\theta^T x^{(i)})\right)x^{(i)} && (1.5)\\
\implies \frac{\partial J(\theta)}{\partial \theta_j} &= -\frac{1}{n}\sum_{i=1}^{n}\left(y^{(i)} - g(\theta^T x^{(i)})\right)x_j^{(i)} \ \forall j \in [d] && (1.6)\\
\implies \forall k \in [d],\ \frac{\partial^2 J(\theta)}{\partial \theta_j \partial \theta_k} &= -\frac{1}{n}\sum_{i=1}^{n}\left(-g(\theta^T x^{(i)})(1-g(\theta^T x^{(i)}))x_k^{(i)}\right)x_j^{(i)} && (1.7)\\
&= \frac{1}{n}\sum_{i=1}^{n}\left(g(\theta^T x^{(i)})(1-g(\theta^T x^{(i)}))x_j^{(i)}x_k^{(i)}\right) && (1.8)
\end{aligned}$$

Therefore, $H_J(\theta)$ can be constructed from the array of second order derivatives of $J(\theta)$ as

$$H_J(\theta)_{j,k} := \frac{1}{n}\sum_{i=1}^{n}\left(g(\theta^T x^{(i)})(1-g(\theta^T x^{(i)}))x_j^{(i)}x_k^{(i)}\right) \tag{1.9}$$

Notice that since $g(\theta^T x^{(i)}) \in (0,1)$, therefore $g(\theta^T x^{(i)})(1-g(\theta^T x^{(i)})) > 0$ for every θ and $x^{(i)}$.

Proof. Show that $H_J(\theta) \succeq 0$: let $z = (z_1, \ldots, z_d) \in \mathbb{R}^d$, then

$$z^T H_J(\theta) \in \mathbb{R}^{1\times d} \tag{1.10}$$

Then the β^{th} column of $z^T H_J(\theta)$ is

$$z^T H_J(\theta)_\beta = \frac{1}{n}\sum_{\alpha=1}^{d}\sum_{i=1}^{n} g(\theta^T x^{(i)})(1-g(\theta^T x^{(i)}))z_\alpha x_\alpha^{(i)} x_\beta^{(i)} \tag{1.11}$$

Therefore

$$z^T H_J(\theta) z = \frac{1}{n} \sum_{\beta=1}^{d} \sum_{\alpha=1}^{d} \sum_{i=1}^{n} g(\theta^T x^{(i)})(1 - g(\theta^T x^{(i)})) z_\alpha x_\alpha^{(i)} x_\beta^{(i)} z_\beta \tag{1.12}$$

$$= \frac{1}{n} \sum_{i=1}^{n} g(\theta^T x^{(i)})(1 - g(\theta^T x^{(i)})) \sum_{\beta=1}^{d} \sum_{\alpha=1}^{d} z_\alpha x_\alpha^{(i)} x_\beta^{(i)} z_\beta \tag{1.13}$$

$$= \frac{1}{n} \sum_{i=1}^{n} \underbrace{g(\theta^T x^{(i)})(1 - g(\theta^T x^{(i)}))}_{>0 \ \because g(\cdot)\in(0,1)} \underbrace{(z^T x)^2}_{\geq 0} \geq 0 \tag{1.14}$$

Hence, $H_J(\theta) \succeq 0$ is shown by showing $z^T H_J(\theta) z$ for every $z \in \mathbb{R}^d$. ■

1.2 Question 1(c)

Proof. By Bayes' theorem,

$$p(y=1|x;\phi,\mu_0,\mu_1,\Sigma) = \frac{p(x|y=1;\phi,\mu_0,\mu_1,\Sigma)p(y=1;\phi,\mu_0,\mu_1,\Sigma)}{p(x;\phi,\mu_0,\mu_1,\Sigma)} \tag{1.15}$$

Define

$$\begin{align} z &:= \frac{p(x|y=1;\phi,\mu_0,\mu_1,\Sigma)p(y=1;\phi,\mu_0,\mu_1,\Sigma)}{p(x;\phi,\mu_0,\mu_1,\Sigma)} \tag{1.16}\\ \Theta &:= \{\phi,\mu_0,\mu_1,\Sigma\} \tag{1.17}\end{align}$$

Conditioned on particular x, y is either 0 or 1, therefore,

$$\begin{align} p(y=0|x;\Theta) &= 1-z \tag{1.18}\\ \implies \frac{z}{1-z} &= \frac{p(y=1|x;\Theta)}{p(y=0|x;\Theta)} \tag{1.19}\\ &= \frac{p(x|y=1;\Theta)p(y=1;\Theta)}{p(x|y=0;\Theta)p(y=0;\Theta)} \tag{1.20}\\ &= \frac{\phi}{1-\phi}\frac{\exp\left(-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right)}{\exp\left(-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)\right)} \tag{1.21}\\ \implies \log\frac{z}{1-z} &= \log\frac{\phi}{1-\phi} \tag{1.22}\\ &+\left(-\frac{1}{2}x^T\Sigma^{-1}x+\mu_1^T\Sigma^{-1}x-\frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1\right) \tag{1.23}\\ &-\left(-\frac{1}{2}x^T\Sigma^{-1}x+\mu_0^T\Sigma^{-1}x-\frac{1}{2}\mu_0^T\Sigma^{-1}\mu_0\right) \tag{1.24}\\ &= \log\frac{\phi}{1-\phi}+\left((\mu_1-\mu_0)^T\Sigma^{-1}x+\frac{1}{2}\mu_0^T\Sigma^{-1}\mu_0-\frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1\right) \tag{1.25}\\ \implies \frac{z}{1-z} &= \exp\underbrace{\left(\log\frac{\phi}{1-\phi}+\left((\mu_1-\mu_0)^T\Sigma^{-1}x+\frac{1}{2}\mu_0^T\Sigma^{-1}\mu_0-\frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1\right)\right)}_{=:\Delta} \tag{1.26}\\ \implies z &= \frac{\exp(\Delta)}{1+\exp(\Delta)} = \frac{1}{1+\exp(-\Delta)} \tag{1.27}\end{align}$$

Therefore

$$\frac{p(x|y=1;\phi,\mu_0,\mu_1,\Sigma)p(y=1;\phi,\mu_0,\mu_1,\Sigma)}{p(x;\phi,\mu_0,\mu_1,\Sigma)} = \frac{1}{1+\exp(-(\theta^Tx+\theta_0))} \tag{1.28}$$

where

$$\begin{align} \theta &= (\Sigma^{-1})^T(\mu_1-\mu_0) \tag{1.29}\\ \theta_0 &= \log\frac{\phi}{1-\phi}+\frac{1}{2}\mu_0^T\Sigma^{-1}\mu_0-\frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 \tag{1.30}\end{align}$$

$\blacksquare$

1.3 Question 1(d)

1.3.1 ϕ

Proof.

$$
\begin{aligned}
\frac{\partial}{\partial \phi}\ell(\phi,\cdot) &= \frac{\partial}{\partial \phi}\sum_{i=1}^{n}\underbrace{\log p(x^{(i)}|y^{(i)};\mu_0,\mu_1,\Sigma)}_{\perp\phi} + \log p(y^{(i)};\phi) && (1.31)\\
&= \frac{\partial}{\partial \phi}\sum_{i=1}^{n}\log p(y^{(i)};\phi) && (1.32)\\
&= \frac{\partial}{\partial \phi}\sum_{i=1}^{n}\log \phi^{y^{(i)}}(1-\phi)^{1-y^{(i)}} && (1.33)\\
&= \frac{\partial}{\partial \phi}\sum_{i=1}^{n} y^{(i)}\log\phi + (1-y^{(i)})\log(1-\phi) && (1.34)\\
&= \sum_{i=1}^{n} y^{(i)}\frac{1}{\phi} - (1-y^{(i)})\frac{1}{1-\phi} && (1.35)
\end{aligned}
$$

The first order condition of maximizing likelihood becomes

$$
\begin{aligned}
&\sum_{i=1}^{n} y^{(i)}\frac{1}{\phi} - (1-y^{(i)})\frac{1}{1-\phi} = 0 && (1.36)\\
\implies &\sum_{i=1} \frac{y^{(i)}}{\phi} + \frac{y^{(i)}}{1-\phi} - \frac{1}{1-\phi} = 0 && (1.37)\\
\implies &\sum_{i=1}^{n} y^{(i)}\frac{1-\phi+\phi}{\phi(1-\phi)} - \frac{1}{1-\phi} = 0 && (1.38)\\
\implies &\sum_{i=1}^{n} y^{(i)}\frac{1}{\phi(1-\phi)} = n\frac{1}{1-\phi} && (1.39)\\
\implies &\phi = \frac{1}{n}\sum_{i=1}^{n} y^{(i)} && (1.40)
\end{aligned}
$$

■

1.3.2 μ_0

Proof.

$$\frac{\partial}{\partial \mu_0}\ell(\mu_0,\cdot) = \frac{\partial}{\partial \mu_0}\sum_{i=1}^{n}\log p(x^{(i)}|y^{(i)};\mu_0,\mu_1,\Sigma) + \underbrace{\log p(y^{(i)};\phi)}_{\perp\mu_0} \tag{1.41}$$

$$= \frac{\partial}{\partial \mu_0}\sum_{i=1}^{n}\log p(x^{(i)}|y^{(i)};\mu_0,\mu_1,\Sigma) \tag{1.42}$$

$$= \frac{\partial}{\partial \mu_0}\sum_{i=1}^{n}\Bigg\{\overbrace{y^{(i)}\log\left[\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\exp\left(-\frac{1}{2}(x^{(i)}-\mu_1)^T\Sigma^{-1}(x^{(i)}-\mu_1)\right)\right]}^{\perp\mu_0} \tag{1.43}$$

$$+ (1-y^{(i)})\log\left[\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\exp\left(-\frac{1}{2}(x^{(i)}-\mu_0)^T\Sigma^{-1}(x^{(i)}-\mu_0)\right)\right]\Bigg\} \tag{1.44}$$

$$= \frac{\partial}{\partial \mu_0}\sum_{i=1}^{n}(1-y^{(i)})\left(\underbrace{\log\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}}_{\perp\mu_0} - \frac{1}{2}(x^{(i)}-\mu_0)^T\Sigma^{-1}(x^{(i)}-\mu_0)\right) \tag{1.45}$$

$$= \frac{\partial}{\partial \mu_0}(-1)\sum_{i=1}^{n}(1-y^{(i)})\frac{1}{2}(x^{(i)}-\mu_0)^T\Sigma^{-1}(x^{(i)}-\mu_0) = 0 \tag{1.46}$$

$$\implies \sum_{i=1}^{n}(1-y^{(i)})\Sigma^{-1}(x^{(i)}-\mu_0) = 0 \tag{1.47}$$

$$\tag{1.48}$$

$$\implies \sum_{i=1}^{n}\Sigma^{-1}(1-y^{(i)})x^{(i)} = \sum_{i=1}^{n}\Sigma^{-1}(1-y^{(i)})\mu_0 \tag{1.49}$$

$$\implies \sum_{i=1}^{n}(1-y^{(i)})x^{(i)} = \sum_{i=1}^{n}(1-y^{(i)})\mu_0 \tag{1.50}$$

$$\implies \mu_0 = \frac{\sum_{i=1}^{n}(1-y^{(i)})x^{(i)}}{\sum_{i=1}^{n}(1-y^{(i)})} = \frac{\sum_{i=1}^{n}\mathbb{1}\{y^{(i)}=0\}x^{(i)}}{\sum_{i=1}^{n}\mathbb{1}\{y^{(i)}=0\}} \tag{1.51}$$

■

1.3.3 μ_1

Proof.

$$
\begin{aligned}
\frac{\partial}{\partial \mu_1}\ell(\mu_1, \cdot) &= \frac{\partial}{\partial \mu_1}\sum_{i=1}^{n}\log p(x^{(i)}|y^{(i)};\mu_0,\mu_1,\Sigma) + \underbrace{\log p(y^{(i)};\phi)}_{\perp \mu_1} && (1.52)\\
&= \frac{\partial}{\partial \mu_1}\sum_{i=1}^{n}\log p(x^{(i)}|y^{(i)};\mu_0,\mu_1,\Sigma) && (1.53)\\
&= \frac{\partial}{\partial \mu_1}\sum_{i=1}^{n}\left\{y^{(i)}\log\left[\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\exp\left(-\frac{1}{2}(x^{(i)}-\mu_1)^T\Sigma^{-1}(x^{(i)}-\mu_1)\right)\right]\right. && (1.54)\\
&\left. + \underbrace{(1-y^{(i)})\log\left[\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\exp\left(-\frac{1}{2}(x^{(i)}-\mu_0)^T\Sigma^{-1}(x^{(i)}-\mu_0)\right)\right]}_{\perp \mu_1}\right\} && (1.55)\\
&= \frac{\partial}{\partial \mu_1}\sum_{i=1}^{n}y^{(i)}\left(\underbrace{\log\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}}_{\perp \mu_1} - \frac{1}{2}(x^{(i)}-\mu_1)^T\Sigma^{-1}(x^{(i)}-\mu_1)\right) && (1.56)\\
&= \frac{\partial}{\partial \mu_1}(-1)\sum_{i=1}^{n}y^{(i)}\frac{1}{2}(x^{(i)}-\mu_1)^T\Sigma^{-1}(x^{(i)}-\mu_1) = 0 && (1.57)\\
&\implies \sum_{i=1}^{n}y^{(i)}\Sigma^{-1}(x^{(i)}-\mu_1) = 0 && (1.58)\\
&\implies \sum_{i=1}^{n}\Sigma^{-1}y^{(i)}x^{(i)} = \sum_{i=1}^{n}\Sigma^{-1}y^{(i)}\mu_1 && (1.59)\\
&\implies \sum_{i=1}^{n}y^{(i)}x^{(i)} = \sum_{i=1}^{n}y^{(i)}\mu_1 && (1.60)\\
&\implies \mu_1 = \frac{\sum_{i=1}^{n}y^{(i)}x^{(i)}}{\sum_{i=1}^{n}y^{(i)}} = \frac{\sum_{i=1}^{n}\mathbb{1}\{y^{(i)}=1\}x^{(i)}}{\sum_{i=1}^{n}\mathbb{1}\{y^{(i)}=1\}} && (1.61)
\end{aligned}
$$

∎

1.3.4 Σ^{-1}

Proof. TODO ∎

2 Question 2: Incomplete, Positive-Only Labels

2.1 Question 2(c)

Proof.

$$\begin{aligned}
p(t^{(i)} = 1|y^{(i)} = 1, x^{(i)}) &= \frac{p(y^{(i)} = 1|t^{(i)} = 1, x^{(i)})p(t^{(i)} = 1, x^{(i)})}{p(y^{(i)} = 1, x^{(i)})} && (2.1) \\
&= \frac{p(y^{(i)} = 1|t^{(i)} = 1, x^{(i)})p(t^{(i)} = 1, x^{(i)})}{p(y^{(i)} = 1|t^{(i)} = 1, x^{(i)})p(t^{(i)} = 1, x^{(i)}) + p(y^{(i)} = 1|t^{(i)} = 0, x^{(i)})p(t^{(i)} = 0, x^{(i)})} && (2.2) \\
&= \frac{\alpha p(t^{(i)} = 1, x^{(i)})}{\alpha p(t^{(i)} = 1, x^{(i)}) + 0p(t^{(i)} = 0, x^{(i)})} && (2.3) \\
&= \frac{\alpha p(t^{(i)} = 1, x^{(i)})}{\alpha p(t^{(i)} = 1, x^{(i)})} = 1 && (2.4)
\end{aligned}$$

■