CS229 Fall 2018 Homework

Tien-Dat Do

June 3, 2025

Problem set #1: Supervised learning

Problem 1: Linear Classifiers (logistic regression and GDA)

(a) We have

$$
\begin{aligned}
J(\theta) &= -\frac{1}{m}\sum_{i=1}^{m}\left(y^{(i)}\log(h_\theta(x^{(i)})) + (1-y^{(i)})\log(1-h_\theta(x^{(i)}))\right) \\
&= -\frac{1}{m}\sum_{i=1}^{m}\left(y^{(i)}\log(g(\theta^\top x^{(i)})) + (1-y^{(i)})\log(1-g(\theta^\top x^{(i)}))\right) \\
\Rightarrow \frac{\partial J(\theta)}{\partial \theta_j} &= -\frac{1}{m}\sum_{i=1}^{m}\left(y^{(i)}\frac{g(\theta^\top x^{(i)})(1-g(\theta^\top x^{(i)}))}{g(\theta^\top x^{(i)})}x_j^{(i)} - (1-y^{(i)})\frac{g(\theta^\top x^{(i)})(1-g(\theta^\top x^{(i)}))}{1-g(\theta^\top x^{(i)})}x_j^{(i)}\right) \\
&= -\frac{1}{m}\sum_{i=1}^{m}\left(y^{(i)}(1-g(\theta^\top x^{(i)})) - (1-y^{(i)})g(\theta^\top x)\right)x_j^{(i)} \\
&= \frac{1}{m}\sum_{i=1}^{m}\left(g(\theta^\top x^{(i)}) - y^{(i)}\right)x_j^{(i)} \\
\Rightarrow \nabla J(\theta) &= \frac{1}{m}X^\top\left(g(X\theta) - Y\right)
\end{aligned}
$$

Again, we have

$$
\begin{aligned}
\Rightarrow \frac{\partial J(\theta)}{\partial \theta_k \partial \theta_j} &= \frac{1}{m}\sum_{i=1}^{m} g(\theta^\top x^{(i)})(1-g(\theta^\top x^{(i)}))x_j^{(i)}x_k^{(i)} = H_{jk} \\
\Rightarrow H &= \frac{1}{m}X^\top D X
\end{aligned}
$$

with $D = diag\left(g(X\theta)(1-g(X\theta))\right)$
$\forall z \in \mathbb{R}^m$, we have

$$z^\top H z = z^\top X^\top D X z = (Xz)^\top D(Xz)$$

Easily to see that D is PSD, so $z^\top H z \geq 0 \Rightarrow H \succeq 0$

(b) Coding

(c) We have

$$\begin{aligned}
p(y=1|x) &= \frac{p(x|y=1)p(y=1)}{p(x)} = \frac{p(x|y=1)p(y=1)}{p(x|y=0)p(y=0)+p(x|y=1)p(y=1)} \\
&= \frac{\exp\left(-\frac{1}{2}(x-\mu_1)^\top\Sigma^{-1}(x-\mu_1)\right)\phi}{\exp\left(-\frac{1}{2}(x-\mu_0)^\top\Sigma^{-1}(x-\mu_0)\right)(1-\phi)+\exp\left(-\frac{1}{2}(x-\mu_1)^\top\Sigma^{-1}(x-\mu_1)\right)\phi} \\
&= \frac{1}{\exp\left(-\frac{1}{2}(x-\mu_0)^\top\Sigma^{-1}(x-\mu_0)+\frac{1}{2}(x-\mu_1)^\top\Sigma^{-1}(x-\mu_1)\right)\frac{1-\phi}{\phi}+1} \\
&= \frac{1}{1+\exp\left((\mu_0-\mu_1)^\top\Sigma^{-1}x-\frac{1}{2}\mu_0^\top\Sigma^{-1}\mu_0+\frac{1}{2}\mu_1^\top\Sigma^{-1}\mu_1+\ln\frac{1-\phi}{\phi}\right)} \\
&= \frac{1}{1+\exp(-(\theta^\top x+\theta_0))}
\end{aligned}$$

with

$$\begin{aligned}
\theta &= \Sigma^{-1}(\mu_1-\mu_0)^\top \\
\theta_0 &= \frac{1}{2}\mu_0^\top\Sigma^{-1}\mu_0-\frac{1}{2}\mu_1^\top\Sigma^{-1}\mu_1-\ln\frac{1-\phi}{\phi} \\
&= \frac{1}{2}(\mu_0-\mu_1)^\top\Sigma^{-1}(\mu_0+\mu_1)-\ln\frac{1-\phi}{\phi}
\end{aligned}$$

(d) We have

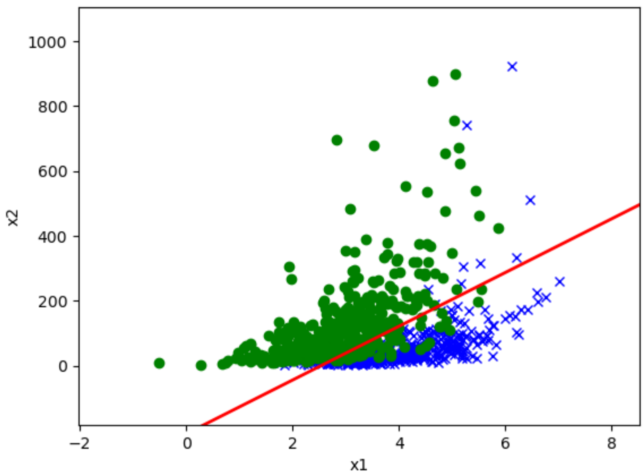
$$\begin{aligned}
\ell(\phi,\mu_0,\mu_1,\Sigma) &= \log\prod_{i=1}^m p(x^{(i)},y^{(i)}) \\
&= \log\prod_{i=1}^m p(x^{(i)}|y^{(i)})p(y^{(i)}) \\
&= \sum_{i=1}^m \log(p(x^{(i)}|y^{(i)}))+\sum_{i=1}^m \log(p(y^{(i)})) \\
&= \sum_{i=1}^m \log\left(\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x^{(i)}-\mu_{y^{(i)}})^2}{2\sigma^2}\right)\right) \\
&\quad + \sum_{i=1}^m \left(y^{(i)}\log\phi+(1-y^{(i)})\log(1-\phi)\right) \\
&= -m\log(\sqrt{2\pi}\sigma)-\sum_{i=1}^m\frac{(x^{(i)}-\mu_{y^{(i)}})^2}{2\sigma^2} \\
&\quad + \sum_{i=1}^m\left(y^{(i)}\log\phi\right)+\sum_{i=1}^m\left((1-y^{(i)})\log(1-\phi)\right) \\
&= -m\log(\sqrt{2\pi}\sigma)-\sum_{i=1}^m\frac{(x^{(i)}-\mu_{y^{(i)}})^2}{2\sigma^2} \\
&\quad + \log\phi\sum_{i=1}^m\left(1\{y^{(i)}=1\}\right)+\log(1-\phi)\sum_{i=1}^m\left(1-1\{y^{(i)}=1\}\right)
\end{aligned}$$

By basic calculus, we have

$$
\begin{aligned}
\frac{\partial \ell}{\partial \phi} &= \frac{1}{\phi}\sum_{i=1}^{m}\left(1\{y^{(i)}=1\}\right) - \frac{1}{1-\phi}\left(m - \sum_{i=1}^{m} 1\{y^{(i)}=1\}\right) \\
\frac{\partial \ell}{\partial \mu_0} &= \frac{1}{\sigma^2}\sum_{i=1}^{m} 1\{y^{(i)}=0\}(x^{(i)}-\mu_0) \\
\frac{\partial \ell}{\partial \mu_1} &= \frac{1}{\sigma^2}\sum_{i=1}^{m} 1\{y^{(i)}=1\}(x^{(i)}-\mu_1) \\
\frac{\partial \ell}{\partial \sigma} &= \frac{-m}{\sigma} + \sum_{i=1}^{m}\frac{(x^{(i)}-\mu_{y^{(i)}})^2}{\sigma^3} \Rightarrow \frac{\partial \ell}{\partial \Sigma} = \frac{\partial \ell}{\partial \sigma}\frac{\partial \sigma}{\partial \Sigma} = -m + \sum_{i=1}^{m}\frac{(x^{(i)}-\mu_{y^{(i)}})^2}{\sigma^2}
\end{aligned}
$$

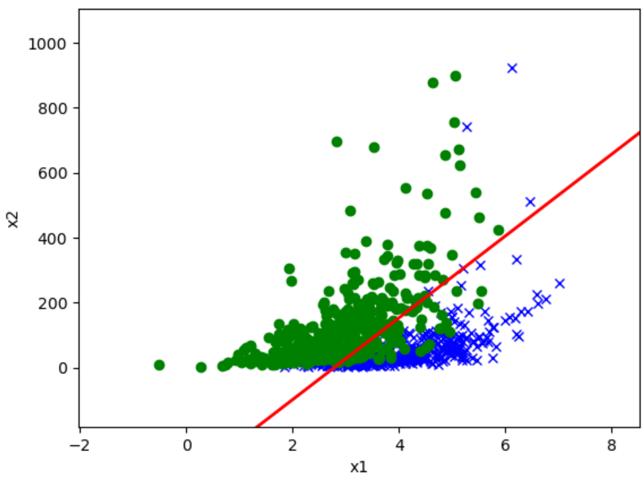
$$
\begin{cases}
\frac{\partial \ell}{\partial \phi} = 0 \\
\frac{\partial \ell}{\partial \mu_0} = 0 \\
\frac{\partial \ell}{\partial \mu_1} = 0 \\
\frac{\partial \ell}{\partial \Sigma} = 0
\end{cases}
\Leftrightarrow
\begin{cases}
\phi = \frac{\sum_{i=1}^{m} 1\{y^{(i)}=1\}}{m} \\
\mu_0 = \frac{\sum_{i=1}^{m} 1\{y^{(i)}=0\}x^{(i)}}{\sum_{i=1}^{m} 1\{y^{(i)}=0\}} \\
\mu_1 = \frac{\sum_{i=1}^{m} 1\{y^{(i)}=1\}x^{(i)}}{\sum_{i=1}^{m} 1\{y^{(i)}=1\}} \\
\Sigma = \sigma^2 = \sum_{i=1}^{m} \frac{(x^{(i)}-\mu_{y^{(i)}})(x^{(i)}-\mu_{y^{(i)}})^\top}{m}
\end{cases}
$$

(e) Coding

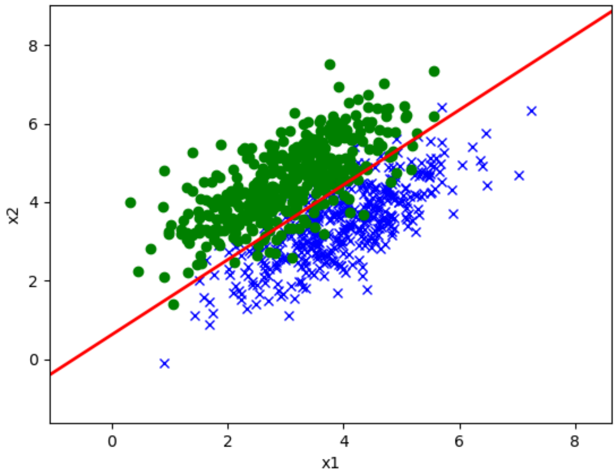


Logistic Regression

(f)

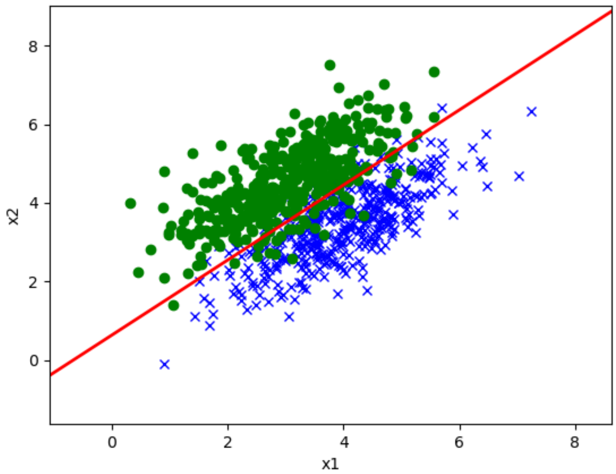


GDA



Logistic Regression

(g)



GDA

On dataset 1, logistic perform better than GDA, because x|y may be not Gaussian distribution

(h) Box-Cox transformation

Problem 2: Incomplete, Positive-Only label

(a) Let $x^{(i)} = x, y^{(i)} = y, t^{(i)} = t$. Then we have

$$\begin{aligned} p(y=1|t=1,x) &= \frac{p(y=1,t=1,x)}{p(t=1,x)} \\ &= \frac{p(t=1|y=1,x)p(y=1|x)p(x)}{p(t=1|x)p(x)} \\ &= \frac{p(t=1|y=1,x)p(y=1|x)}{p(t=1|x)} \\ &= p(y=1|t=1) \quad \text{(base on the assumption)} \end{aligned}$$

$$\Rightarrow p(t=1|x) = p(y=1|x)\frac{p(t=1|y=1,x)}{p(y=1|t=1)}$$

$$\begin{aligned}
p(t=1|y=1,x) &= \frac{p(x|t=1,y=1)p(t=1|y=1)}{p(x|y=1)} \\
&= \frac{p(x|t=1,y=1)}{p(x|y=1,t=1)p(t=1|y=1)+p(x|y=1,t=0)p(t=0|y=1)} \quad \text{(LOTP)} \\
&= \frac{p(x|t=1,y=1)}{p(x|t=1,y=1)} = 1
\end{aligned}$$

(because the probability that a labeled example is negative is 0)
So we have

$$p(t=1|x) = \frac{p(y=1|x)}{p(y=1|t=1)}$$
$$\Rightarrow \alpha = p(y=1|t=1)$$

(b) We have

$$h(x) \approx p(y=1|x) = \alpha p(t=1|x) \approx \alpha \quad \forall x \in V_+$$

Problem 3: Poisson Regression

(a) We have

$$\begin{aligned}
p(y;\lambda) &= \frac{e^{\lambda}\lambda^y}{y!} \\
&= \frac{1}{y!}\exp(y\log\lambda - \lambda)
\end{aligned}$$

Let

$$\begin{cases}
b(y) = \dfrac{1}{y!} \\
\eta = \log(\lambda) \\
T(y) = y \\
a(\eta) = \exp(\eta)
\end{cases}$$

(b) The canonical response function

$$g(\eta) = E(T(y);\eta) = \frac{\partial}{\partial\eta}a(\eta) = \exp(\eta)$$

(c) We have

$$\begin{aligned}
\mathcal{L} = \log(p(y^{(i)}|x^{(i)},\theta)) &= \log\left(\frac{1}{y^{(i)}!}\exp(y^{(i)}\log\lambda^{(i)} - \lambda^{(i)})\right) \\
&= -\log(y^{(i)}!) + y^{(i)}\log\lambda^{(i)} - \lambda^{(i)} \\
&= -\log(y^{(i)}!) + y^{(i)}\theta^\top x^{(i)} - \exp(\theta^\top x^{(i)})
\end{aligned}$$

$$\begin{aligned}
\frac{\partial}{\partial\theta}\mathcal{L} &= y^{(i)}x^{(i)} - x^{(i)}\exp(\theta^\top x^{(i)}) \\
&= x^{(i)}(y^{(i)} - \exp(\theta^\top x^{(i)})) \\
\Rightarrow \frac{\partial}{\partial\theta_j}\mathcal{L} &= x_j^{(i)}(y^{(i)} - \exp(\theta_j x_j^{(i)}))
\end{aligned}$$

So the update rule is

$$\theta_j := \theta_j - \alpha x_j^{(i)}(y^{(i)} - \exp(\theta_j x_j^{(i)}))$$

Problem 4: Convexity of Generalized Linear Models

(a) Since $p(y;\eta)$ is PDF so $\int p(y;\eta)dy = 1$. We have

$$\begin{aligned}
0 = \frac{\partial}{\partial \eta}\int p(y;\eta)dy &= \int \frac{\partial}{\partial \eta} p(y;\eta)dy \\
&= \int \left(y - \frac{\partial}{\partial \eta}a(\eta)\right) p(y;\eta)dy \\
&= \int yp(y;\eta)dy - \frac{\partial}{\partial \eta}a(\eta) \\
&= \mathbb{E}[Y|X;\theta] - \frac{\partial}{\partial \eta}a(\eta) \\
\Rightarrow \mathbb{E}[Y|X;\theta] &= \frac{\partial}{\partial \eta}a(\eta)
\end{aligned}$$

(b) We have

$$\begin{aligned}
\frac{\partial^2}{\partial \eta^2}a(\eta) &= \frac{\partial}{\partial \eta}\left(\int yp(y;\eta)dy\right) \\
&= \frac{\partial}{\partial \eta}\left(\exp(-a(\eta))\int yb(y)\exp(\eta y)dy\right) \\
&= -\frac{\partial}{\partial \eta}a(\eta)p(y;\eta) + \int y^2 p(y;\eta)dy \\
&= -\mathbb{E}^2[X|Y;\theta] + \mathbb{E}[X^2|Y;\theta] \\
&= \mathrm{Var}[X|Y;\theta]
\end{aligned}$$

(c) Give one data point (x, y), the NLL is

$$\begin{aligned}
\ell(x,y,\theta) &= -\log(p(y;\eta)) = -\log b(y) - \eta y + a(\eta) \\
&= -\log b(y) - y\theta^\top x + a(\theta^\top x) \\
\Rightarrow \frac{\partial}{\partial \theta}\ell(x,y,\theta) &= -yx + xa'(\theta^\top x) \\
\Rightarrow \frac{\partial^2}{\partial \theta^2}\ell(x,y,\theta) &= x^\top x a''(\theta^\top x)
\end{aligned}$$

The loss function is

$$\mathcal{L}(\theta) = \sum_{i=1}^{m} \ell(x^{(i)}, y^{(i)}, \theta)$$

The Hessian matrix of loss function is

$$\begin{aligned}
H = \frac{\partial^2}{\partial \theta^2}\mathcal{L}(\theta) &= \sum_{i=1}^{m} \left(x^{(i)}\right)^\top x^{(i)} a''(\theta^\top x^{(i)}) \\
&= X^\top D X
\end{aligned}$$

where

$$\begin{cases} X \text{ is the original data matrix} \\ D \text{ is diagonal matrix with } D_{ii} = a''(\theta^\top x^{(i)}) \end{cases}$$

$\forall z \in \mathbb{R}^n$, we have

$$z^\top H z = z^\top X^\top D X z = (Xz)^\top D X z \geq 0$$

So H is PSD.

Problem 5: Locally weighted linear regression

(a) We have

$$X = \begin{bmatrix} — \left(x^{(1)}\right)^\top — \\ — \left(x^{(2)}\right)^\top — \\ \vdots \\ — \left(x^{(n)}\right)^\top — \end{bmatrix}, y = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(n)} \end{bmatrix}$$

From that,

$$X\theta - y = \begin{bmatrix} \left(x^{(1)}\right)^\top \theta - y^{(1)} \\ \left(x^{(2)}\right)^\top \theta - y^{(2)} \\ \vdots \\ \left(x^{(n)}\right)^\top \theta - y^{(n)} \end{bmatrix}$$

So we choose W such that

$$W_{ij} = \begin{cases} 0 \text{ if } i \neq j \\ \dfrac{1}{2} w^{(i)} \text{ if } i = j \end{cases}$$

(b) We have

$$\begin{aligned} J(\theta) &= (X\theta - y)^\top W (X\theta - y) \\ &= \theta^\top X^\top W X \theta - \theta^\top X^\top W y - y^\top W X \theta + y^\top W y \\ &= \theta^\top X^\top W X \theta - 2\theta^\top X^\top W y + y^\top W y \\ \Rightarrow \nabla_\theta J(\theta) &= 2X^\top W X \theta - 2X^\top W y \end{aligned}$$

With

$$\begin{aligned} \nabla_\theta J(\theta) = 0 &\Leftrightarrow 2X^\top W X \theta - 2X^\top W y = 0 \\ &\Leftrightarrow \left(X^\top W X\right) \theta = X^\top W y \\ &\Leftrightarrow \theta = \left(X^\top W X\right)^{-1} X^\top W y \end{aligned}$$

(c) For each data point (x, y), we have

$$\begin{aligned} \ell(\theta, x, y) &= \log p(y|x; \theta) \\ &= -\frac{1}{2}\log(2\pi) - \log(\sigma) - \frac{\left(y - \theta^\top x\right)^2}{2\sigma^2} \end{aligned}$$

Easy to see that

$$w^{(i)} = \frac{1}{\left(\sigma^{(i)}\right)^2}$$

Likelihood estimate of θ is

$$\begin{aligned}
\mathcal{L}(\theta) &= \sum_{i=1}^{m} \log p(y^{(i)}|x^{(i)};\theta) \\
&= -m\frac{1}{2}\log(2\pi) - m\log(\sigma^{(i)}) - \frac{1}{2}\sum_{i=1}^{m} w^{(i)}\left(y^{(i)} - \theta^\top x^{(i)}\right)^2
\end{aligned}$$

Maximize $\mathcal{L}(\theta)$ is equivalent to minimize

$$\frac{1}{2}\sum_{i=1}^{m} w^{(i)}\left(y^{(i)} - \theta^\top x^{(i)}\right)^2$$

So finding the maximum likelihood estimate is actually solving a weighted linear regression.

(d) It seems like underfitting

(e) $\tau = 0.05$