$x \in \mathbb{R}^n \quad X \in \mathbb{R}^{m \times n} \quad \theta \in \mathbb{R}^n \quad Y \in \mathbb{R}^m$

(a) [10 points] In lecture we saw the average empirical loss for logistic regression:

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m} y^{(i)} \log(h_\theta(x^{(i)})) + (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)})),$$

where $y^{(i)} \in \{0, 1\}$, $h_\theta(x) = g(\theta^T x)$ and $g(z) = 1/(1 + e^{-z})$.
Find the Hessian H of this function, and show that for any vector z, it holds true that

$$z^T H z \geq 0.$$

$$\frac{\partial}{\partial \theta_j} J(\theta) = \left(y \frac{1}{g(\theta^T x)} - (1-y)\frac{1}{1-g(\theta^T x)}\right) \frac{\partial}{\partial \theta_j} g(\theta^T x)$$

$$= \left(y \frac{1}{g(\theta^T x)} - (1-y)\frac{1}{1-g(\theta^T x)}\right) g(\theta^T x)(1-g(\theta^T x)) \frac{\partial}{\partial \theta_j}(\theta^T x)$$

$$= [y(1-g(\theta^T x)) - (1-y)g(\theta^T x)] x_j$$

$$h(x) = \frac{1}{1+e^{-\theta^T x}}$$

$$= [y - \cancel{y g(\theta^T x)} - g(\theta^T x) + \cancel{y g(\theta^T x)}] x_j$$

$$= (y - h_\theta(x)) x_j$$

$$\frac{\partial}{\partial \theta_j} J(\theta) = -\frac{1}{m}\sum_{i=1}^{m} (y^{(i)} - h_\theta(x^{(i)})) x_j^{(i)}$$

$$\nabla J(\theta) = -\frac{1}{m} X^T (Y - h_\theta(X)) \longrightarrow -\frac{1}{m}\begin{bmatrix} \sum_{i=1}^{m} (y^{(i)} - h_\theta(x^{(i)})) x_1^{(i)} \\ \sum_{i=1}^{m} (y^{(i)} - h_\theta(x^{(i)})) x_2^{(i)} \\ \vdots \\ \sum_{i=1}^{m} (y^{(i)} - h_\theta(x^{(i)})) x_n^{(i)} \end{bmatrix}$$

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

$$H = \frac{1}{m} X^T h_\theta(X)(1 - h_\theta(X)) X$$

$h_\theta(X) \in \mathbb{R}^m$

$H \in \mathbb{R}^{n \times n}$

$$\frac{1}{m}\begin{bmatrix} \sum_{i=1}^{m} h_\theta(x^{(i)})(1-h_\theta(x^{(i)})) x_1^{(i)} x_1^{(i)} & \cdots \\ \vdots & \ddots \\ \cdots & \sum_{i=1}^{m} h_\theta(x^{(i)})(1-h_\theta(x^{(i)})) x_n^{(i)} x_n^{(i)} \end{bmatrix}$$

$$z^T H z = \frac{1}{m}\sum_{j=1}^{n}\sum_{k=1}^{n} H_{jk} z_j z_k$$

$$= \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{n} h_\theta(x^{(i)})\left(1-h_\theta(x^{(i)})\right) x_j^{(i)} x_k^{(i)} z_j z_k$$

$$= \frac{1}{m}\sum_{i=1}^{m} h_\theta(x^{(i)})\left(1-h_\theta(x^{(i)})\right)\sum_{j=1}^{n}\sum_{k=1}^{n} x_j^{(i)} x_k^{(i)} z_j z_k$$

$$= \frac{1}{m}\sum_{i=1}^{m} h_\theta(x^{(i)})\left(1-h_\theta(x^{(i)})\right)\left[(x^{(i)})^T z\right]^2$$

$\because h_\theta(x^{(i)}) \& 1-h_\theta(x^{(i)}) > 0$

$\therefore z^T H z \geq 0$

(c) [5 points] Recall that in GDA we model the joint distribution of (x, y) by the following equations:

$$\begin{aligned} p(y) &= \begin{cases} \phi & \text{if } y = 1 \\ 1-\phi & \text{if } y = 0 \end{cases} \\ p(x|y=0) &= \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)\right) \\ p(x|y=1) &= \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right), \end{aligned}$$

where ϕ, μ_0, μ_1, and Σ are the parameters of our model.

Suppose we have already fit ϕ, μ_0, μ_1, and Σ, and now want to predict y given a new point x. To show that GDA results in a classifier that has a linear decision boundary, show the posterior distribution can be written as

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

where $\theta \in \mathbb{R}^n$ and $\theta_0 \in \mathbb{R}$ are appropriate functions of ϕ, Σ, μ_0, and μ_1.

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

$$= \frac{p(x|y=1)\,p(y=1)}{p(x|y=1)\,p(y=1) + p(x|y=0)\,p(y=0)}$$

$$= \frac{\phi \exp\left(-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right)}{\phi \exp\left(-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right) + (1-\phi)\exp\left(-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)\right)}$$

divided by numerator

$$\frac{1-\phi}{\phi}\exp\left\{\frac{1}{2}\left[(x-\mu_1)^T\Sigma^{-1}(x-\mu_1) - (x-\mu_0)^T\Sigma^{-1}(x-\mu_0)\right]\right\}$$

$$\left[x^T\Sigma^{-1}x - x^T\Sigma^{-1}\mu_1 - \mu_1^T\Sigma^{-1}x + \mu_1^T\Sigma^{-1}\mu_1 - \right. \quad \because (\Sigma^{-1})^T = \Sigma^{-1}$$

$$\left.(x^T\Sigma^{-1}x - x^T\Sigma^{-1}\mu_0 - \mu_0^T\Sigma^{-1}x + \mu_0^T\Sigma^{-1}\mu_0)\right]$$

$$= x^T\Sigma^{-1}(\mu_0-\mu_1) + (\mu_0^T-\mu_1^T)\Sigma^{-1}x + (\mu_1^T\Sigma^{-1}\mu_1 - \mu_0^T\Sigma^{-1}\mu_0)$$

$$= 2(\mu_0-\mu_1)^T\Sigma^{-1}x - (\mu_0-\mu_1)^T\Sigma^{-1}(\mu_0+\mu_1)$$

$$= 2(\Sigma^{-1}(\mu_0-\mu_1))^T\Sigma^{-1}x - (\mu_0-\mu_1)^T\Sigma^{-1}(\mu_0+\mu_1)$$

$$\therefore p(y|x) = \frac{1}{1+\exp\left(-\left[(\Sigma^{-1}(\mu_1-\mu_0))^T x - \frac{1}{2}(\mu_1-\mu_0)^T\Sigma^{-1}(\mu_1+\mu_0) - \ln\frac{1-\phi}{\phi}\right]\right)}$$

Let

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

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

(d) [7 points] For this part of the problem only, you may assume n (the dimension of x) is 1, so that $\Sigma = [\sigma^2]$ is just a real number, and likewise the determinant of Σ is given by $|\Sigma| = \sigma^2$. Given the dataset, we claim that the maximum likelihood estimates of the parameters are given by

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

The log-likelihood of the data is

$$\begin{aligned}
\ell(\phi, \mu_0, \mu_1, \Sigma) &= \log \prod_{i=1}^{m} p(x^{(i)}, y^{(i)}; \phi, \mu_0, \mu_1, \Sigma) \\
&= \log \prod_{i=1}^{m} p(x^{(i)}|y^{(i)}; \mu_0, \mu_1, \Sigma) p(y^{(i)}; \phi).
\end{aligned}$$

By maximizing ℓ with respect to the four parameters, prove that the maximum likelihood estimates of ϕ, μ_0, μ_1, and Σ are indeed as given in the formulas above. (You may assume that there is at least one positive and one negative example, so that the denominators in the definitions of μ_0 and μ_1 above are non-zero.)

$$p(x^{(i)}|y^{(i)}; \mu_0, \mu_1, \Sigma)$$

$$= \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x^{(i)} - \mu_{y^{(i)}})^T \Sigma^{-1} (x^{(i)} - \mu_{y^{(i)}})\right)$$

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

$$\ell = \sum_{i=1}^{m} \log p(y^{(i)}; \phi) + \sum_{i=1}^{m} \log p(x^{(i)}|y^{(i)}; \mu_0, \mu_1, \Sigma)$$

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

$$\frac{\partial}{\partial \mu_0}\ell = \left[\sum_{i=1}^{m} \left(-\frac{1}{2}(x^{(i)} - \mu_{y^{(i)}})^T \Sigma^{-1} (x^{(i)} - \mu_{y^{(i)}})\right)\right]'$$

$$= \left[\sum_{i=1}^{m} \left(-\frac{1}{2}(x^{(i)2} + \mu_{y^{(i)}}^2 - 2x^{(i)}\mu_{y^{(i)}})\right)\right]'$$

$$= \sum_{i=1}^{m} 1\{y=0\}\mu_0 - x^{(i)} 1\{y=0\}$$

$$\frac{\partial}{\partial \mu_1}\ell = \sum_{i=1}^{m} 1\{y=1\}\mu_1 - x^{(i)} 1\{y=1\}$$

$$\frac{\partial}{\partial \Sigma} l = \left(-\frac{1}{2}\sum_{i=1}^{m} \log|\Sigma| - \frac{1}{2}\sum_{i=1}^{m}(x-\mu_{y^{(i)}})^T\Sigma^{-1}(x-\mu_{y^{(i)}})\right)'$$

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

$$\left.\begin{aligned}
&\frac{\partial}{\partial \phi} l = 0\\
&\frac{\partial}{\partial \mu_0} l = 0\\
&\frac{\partial}{\partial \mu_1} l = 0\\
&\frac{\partial}{\partial \Sigma} l = 0
\end{aligned}\right\} \Rightarrow$$

$$\frac{1}{\phi}\sum_{i=1}^{m} 1\{y^{(i)}=1\} - \frac{1}{1-\phi}\sum_{i=1}^{m}(1-1\{y^{(i)}=1\}) = 0$$

$$(1-\phi)\sum_{i=1}^{m} 1\{y^{(i)}=1\} = \phi\left(m - \sum_{i=1}^{m} 1\{y^{(i)}=1\}\right)$$

$$\boxed{\phi = \frac{1}{m}\sum_{i=1}^{m} 1\{y^{(i)}=1\}}$$

$$\sum_{i=1}^{m} 1\{y=0\}\mu_0 - x^{(i)}1\{y=0\} = 0$$

$$\mu_0\sum_{i=1}^{m} 1\{y=0\} - \sum_{i=1}^{m} x^{(i)}1\{y=0\} = 0$$

$$\boxed{\mu_0 = \frac{\sum_{i=1}^{m} x^{(i)}1\{y=0\}}{\sum_{i=1}^{m} 1\{y=0\}} \quad \mu_1 = \frac{\sum_{i=1}^{m} x^{(i)}1\{y=1\}}{\sum_{i=1}^{m} 1\{y=1\}}}$$

$$-\frac{m}{2}\Sigma^{-1} + \frac{1}{2}\Sigma^{-1}\sum_{i=1}^{m}(x-\mu_{y^{(i)}})(x-\mu_{y^{(i)}})^T\Sigma^{-1} = 0$$

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

(h)

The dataset where GDA performed wrose is far away from Gaussian distribution. GDA has stronger assumptions than Logistic Regression's. It will perform better if dataset if Gaussian distributed.