Q1.1. $P(Y=1) = P(Y=0) = \frac{1}{2}$. $\pi_1 = \frac{1}{2}$

(a). $X|Y=0 \sim N(0,1)$. $f_0 = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$

$X|Y=1 \sim \frac{1}{2}N(-5,1) + \frac{1}{2}N(5,1)$. $f_1 = \frac{1}{2} \times \frac{1}{\sqrt{2\pi}} e^{\frac{-(x+5)^2}{2}} + \frac{1}{2} \times \frac{1}{\sqrt{2\pi}} e^{\frac{-(x-5)^2}{2}}$

$\therefore \frac{1-\pi_1}{\pi_1} = 1$.

$$\frac{f_1(x)}{f_0(x)} = \frac{1}{2} \cdot \frac{e^{\frac{-(x+5)^2}{2}} + e^{\frac{-(x-5)^2}{2}}}{e^{-\frac{x^2}{2}}}$$

$$= \frac{1}{2}\left(e^{-\frac{(x+5)^2+x^2}{2}} + e^{-\frac{(x-5)^2+x^2}{2}} \right)$$

$\therefore$ Bayes ~~rate~~ classifier:

$$h^*(x) = \begin{cases} 1 & \text{if } \frac{1}{2}\left(e^{-\frac{(x+5)^2+x^2}{2}} + e^{-\frac{(x-5)^2+x^2}{2}}\right) > 1 \\ 0 & \end{cases}$$

Bayes risk:

$$R^*(h) = P\left(\frac{1}{2}\left(e^{-\frac{(x+5)^2+x^2}{2}} + e^{-\frac{(x-5)^2+x^2}{2}}\right) > 1,\ Y=0\right) + P\left(\frac{1}{2}\left(e^{-\frac{(x+5)^2+x^2}{2}} + e^{-\frac{(x-5)^2+x^2}{2}}\right) \le 1,\ Y=1\right).$$

$$= P\left(\frac{1}{2}\left(e^{-\frac{(x+5)^2+x^2}{2}} + e^{-\frac{(x-5)^2+x^2}{2}}\right) > 1 \mid X \sim N(0,1)\right)$$
$$+ P\left(\frac{1}{2}\left(e^{-\frac{(x+5)^2+x^2}{2}} + e^{-\frac{(x-5)^2+x^2}{2}}\right) \le 1 \mid X \sim \frac{1}{2}N(-5,1) + \frac{1}{2}N(5,1)\right)$$

(b). Linear classifier that minimize the risk: LDA.

$h^*(x) = \mathrm{argmax}_k\, \delta_k(x) = \mathrm{Sign}(\delta_1(x) - \delta_0(x))$

where $\delta_1(x) = x^T \Sigma^{-1} \mu_1 - \frac{1}{2}\mu_1^T \Sigma^{-1} \mu_1 + \log \pi_1$

$\mu_1 = 0$

$\delta_1(x) = \log \pi_1 = -\log(2)$.

$\delta_0(x) = \log \pi_0 = -\log(2)$.

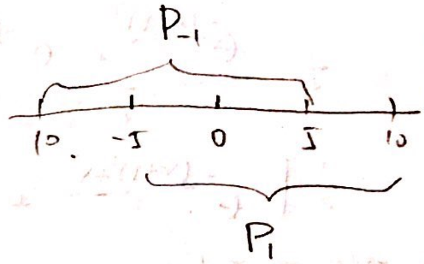
$\therefore \delta_1(x) - \delta_0(x) = 0$.

Q1.2. (a) $\pi_1 = \frac{1}{2}$

$X|Y=-1 \sim \text{Uniform}(-10, 5)$. $\quad P_{-1} = \begin{cases} \frac{1}{15} & \text{if } -10 \le X \le 5 \\ 0 & \text{otherwise} \end{cases}$

$X|Y=1 \sim \text{Uniform}(-5, 10)$ $\quad P_1 = \begin{cases} \frac{1}{15} & \text{if } -5 \le X \le 10 \\ 0 & \text{o.w.} \end{cases}$

$$\frac{\pi_1 P_1(X)}{\pi_1 P_1(X) + (1-\pi_1) \cdot P_{-1}(X)} < \frac{1}{2}.$$



① if $X < -10$. $\quad \frac{P_1(X)}{P_{-1}(X)} = $ undefined.

$P_1(X) = 0$
$P_{-1}(X) = 0$

② if $-10 \le X \le -5$ $\quad \frac{P_1(X)}{P_{-1}(X)} = 0$

$P_1(X) = 0$
$P_0(X) = \frac{1}{15}$

③ if $-5 \le X \le 5$ $\quad \frac{P_1(X)}{P_{-1}(X)} = \frac{\frac{1}{15}}{\frac{1}{15}} = 1.$

$P_1(X) = \frac{1}{15}$
$P_0(X) = \frac{1}{15}$

④ if $5 \le X < 10$ $\quad \frac{P_1(X)}{P_{-1}(X)} = $ undefined.

$P_1(X) = \frac{1}{15}$.
$P_0(X) = 0$.

$\therefore$ for $\frac{P_1(X)}{P_{-1}(X)} > \frac{1-\pi_1}{\pi_1} = 1$, none of ~~the~~ above intervals satisfy.

$h^*(X) = \begin{cases} 1 & \text{if } X \ge 5, \\ 0 & \text{o.w} \end{cases}$

$R(h^*) = P(\text{misclassify})$. $\quad$ will make an error if $X \in [-5, 5]$

$= [P[X \in [-5,5] \mid y=-1] \cdot P(y=-1) + P[X \in [-5,5] \mid y=1] \cdot P(y=1)] \cdot P\text{error} \mid X \in [-5,5]$

$= \left(\frac{2}{3} \times \frac{1}{2} + \frac{2}{3} \times \frac{1}{2} \right) \times \frac{1}{2}.$

$= \frac{1}{3}.$

(b). $h_\alpha(X) = \text{Sign}(X - \alpha)$

$\alpha = 0.$

$\therefore h_1^*(X) = \begin{cases} 1 & \text{if } X > 0 \\ 0 & \text{o.w.} \end{cases}$

$R(h_1^*) = P(y=-1, X>0) + P(y=1, X<0)$

$= \frac{1}{3} + \frac{1}{3}$

$= \frac{2}{3}.$

Q2 (a). Show that if Newton's method is applied to logistic regression, log-likelihood, it leads to the reweighted least square algorithm.

Let $y_i \in (0,1)$. $P_1(X;\theta) = P(X;\theta)$.
$P_0(X;\theta) = 1 - P(X;\theta)$.

$$\ell(\beta) = \sum_{i=1}^{n} (y_i \cdot \ln p(x_i;\beta) + (1-y_i)\ln(1-p(x_i;\beta))$$
$$= \sum_{i=1}^{n} \{ y_i \beta^t x_i - \ln(1+\exp(\beta^t x_i)) \}$$

maximize $\ell(\beta)$
$\Rightarrow$ minimize ~~log~~ $-\ell(\beta)$.

$$\frac{d\ell(\beta)}{d\beta} = \sum_{i=1}^{n} \{ x_i (y_i - p(x_i, \beta)) \} = 0$$

Use Newton's method to solve this.

$$\beta' = \beta - \left(\frac{\partial^2 \ell(\beta)}{\partial\beta\,\partial\beta^t} \right)^{-1} \frac{\partial \ell(\beta)}{\partial \beta}$$

where Hession of log-likelihood is:

$$\frac{\partial^2 \ell(\beta)}{\partial\beta\,\partial\beta^t} = -\sum_{i=1}^{N} x_i x_i^T p(x_i,\beta)(1-p(x_i,\beta)).$$

Let matrix X be the data, $\vec{y}$ be the labels, $\vec{p}$ be the probability and W be the weighting matrix.

$$\beta' = \beta - (X^T W X)^{-1} X^T (\vec{y} - \vec{p})$$
$$= (X^T W X)^{-1} X^T W \vec{z}$$

where W is a diagonal matrix with ith diagonal $p(x_i;\beta)(1-p(x_i;\beta))$.

$$\vec{z} = X\beta + W^{-1}(\vec{y} - \vec{p})$$

$\Rightarrow$ Weighted least square step.

(b). $\log L(\beta^{(k)}; \bar{x}) = \sum_{i=1}^{n} \log \frac{1}{1+e^{-k\beta^T \bar{x}_i}}$

where $\bar{x}_i^T = \tilde{y}_i x_i^T$ is the i^{th} row of $\bar{x}$.

Since β gives complete separation among the sample point

$\beta^T \bar{x}_i > 0$ for $i = 1, \dots n$.

$\therefore -\beta^T \bar{x} < 0$

as $k \to \infty$, $e^{-k\beta^T \bar{x}_i} \to 0$

$\therefore \sum_{i=1}^{n} \log \frac{1}{1+e^{-k\beta^T \bar{x}_i}} \to \sum_{i=1}^{n} \log 1 \to n \cdot 0 = 0.$ $\to$ absolute maximum of the equation.

$\therefore$ the log likelihood is maximum and equal to 0 on the boundary of the parameter space

$\therefore$ For any finite β, the log likelihood is strictly negative since $k\beta^T \bar{x}_i > 0$.

$\therefore$ the MLE, $\hat{\beta}$ of the logistic regression model vector β does not exist. when there's complete separation among the sample points

Since IRLS is used to find the mle, and in this case the likelihood is larger when β is larger. So, it will find the point with largest ~~β~~ β.

$\therefore$ β won't converge. but direction of β, $\frac{\hat{\beta}}{|\hat{\beta}|}$ converges.

(c). $l(\beta) = \sum_i y_i \log p(x_i) + (1-y_i)\log(1-p(x_i))$

with ridge penalty $\quad l^\lambda(\beta) = l(\beta) - \lambda\|\beta\|^2.$

$$\frac{d}{d\beta} \cdot l^\lambda(\beta) = \sum_i x_i \{ y_i - p(x_i) \} - 2\lambda\beta.$$

$$= \frac{d}{d\beta} l(\beta) - 2\lambda\beta.$$

Negative of Hessian of penalized

$\Omega^\lambda(\beta) = \Omega(\beta) + 2\lambda I.$ where $\Omega = XW(\beta)X.$

$W(\beta)$ is a $n\times n$ diagonal matrix with $W_{ii} = p(x_i)\cdot(1-p(x_i))$.

$$\therefore \frac{d}{d\beta} l^\lambda(\beta)\cdot(\hat\beta^\lambda) = \frac{d}{d\beta} l^\lambda(\beta)\cdot(\beta_0) - (\hat\beta^\lambda - \beta_0)\cdot\Omega^\lambda(\beta_0) + o(\|\hat\beta^\lambda - \beta_0\|).$$

Let $\frac{d}{d\beta} l^\lambda(\beta)\cdot(\hat\beta^\lambda) = 0.$

$$\hat\beta^\lambda = \beta_0 + \{\Omega(\beta_0) + 2\lambda I\}^{-1}\{\frac{d}{d\beta}\cdot l(\beta_0) - 2\lambda\beta_0\}.$$

$$= (\Omega(\beta_0) + 2\lambda I)^{-1}(\frac{d}{d\beta} l(\beta_0) + \Omega(\beta_0)\beta_0\}.$$

MLE of unrestricted estimate : $\hat\beta = \beta_0 + \Omega^{-1}(\beta_0)\frac{d}{d\beta} l(\beta_0)$

$\therefore$ first order estimate of $\hat\beta^\lambda$ is

$$\hat\beta^\lambda = \{\Omega(\beta_0) + 2\lambda I^{-1}\}\cdot\Omega(\beta_0)\cdot\hat\beta.$$