Luis Buenaventura.

CMSC 25025
Problem Set 3

1. Classification

1.1 a) Suppose $\mathbb{P}(Y=1) = \mathbb{P}(Y=0) = \frac{1}{2}$

$X|Y=0 \sim N(0,1)$ & $X|Y=1 \sim \frac{1}{2}N(-5,1) + \frac{1}{2}N(5,1)$ (Assuming iid N)

$\Rightarrow X|Y=0 \sim N(0,1)$ & $X|Y=1 \sim N(-\frac{5}{2}, \frac{1}{4}) + N(\frac{5}{2}, \frac{1}{4})$

$\Rightarrow X|Y=0 \sim N(0,1)$ & $X|Y=1 \sim N(0, \frac{1}{4})$

We define the Bayes Classifier as:

$$h^*(x) = \begin{cases} 1 \text{ if } \frac{(x-\mu_1)^2}{\sigma_1^2} < \frac{(x-\mu_0)^2}{\sigma_0^2} + 2\log(\frac{\pi_1}{1-\pi_1}) + \log\left(\frac{|\sigma_0^2|}{|\sigma_1^2|}\right) \\ 0 \end{cases}$$

$\Rightarrow \pi_1 = \mathbb{P}(Y=1) = \frac{1}{2}$, $\mu_1 = 0 = \mu_0$, $\sigma_0^2 = 1$, $\sigma_1^2 = \frac{1}{2}$

So, we have:

$$h^*(x) = \begin{cases} 1 \text{ if } 2x^2 < x^2 + \log(2) \\ 0 \quad \text{o.w.} \end{cases} \Leftrightarrow \begin{cases} 1 \text{ if } |x| < \sqrt{\log(2)} = 0.545 \\ 0 \quad \text{o.w.} \end{cases}$$

The Bayes risk can be described as

$$\begin{aligned} R(h^*) = \mathbb{P}(h^*(x) \neq Y) &= \mathbb{P}(h^*(x)=1, Y=0) + \mathbb{P}(h^*(x)=0, Y=1) \\ &= \mathbb{P}(h^*(x)=1|Y=0)\mathbb{P}(Y=0) + \mathbb{P}(h^*(x)=0|Y=1)\mathbb{P}(Y=1) \\ &= \mathbb{P}(x \in X|Y=0, |x| < \sqrt{\log 2})\mathbb{P}(Y=0) + \mathbb{P}(x \in X|Y=1, x \geq \sqrt{\log 2})\mathbb{P}(Y=1) \\ &= \frac{1}{2}\left[\left(\int_{-\sqrt{\log 2}}^{\sqrt{\log 2}} \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} dx\right) + \left(1 - \int_{-\sqrt{\log 2}}^{\sqrt{\log 2}} \frac{2e^{-x^2}}{\sqrt{\pi}} dx\right)\right] \\ &= 0.949523 \end{aligned}$$

b) There is no traditional linear classifier that minimizes the risk, since the decision boundary $\{x \in X : \delta_0(x) = \delta_1(x)\}$ is true for all $x \in X$. This results from the fact that $X|Y=1 \sim N(0, \frac{1}{4})$ and $X|Y=0 \sim N(0,1)$ both have mean 0. So, our linear classifier is $\hat{h}(x) = 1$ for $\forall x$.

So, there is a Bayes Risk of 0.5 since we will be wrong 50% of the time.

1.2 Suppose that $\mathbb{P}(Y=1) = \mathbb{P}(Y=-1) = \frac{1}{2}$ and $X|Y=-1 \sim U(-10,5)$ and $X|Y=1 \sim U(-5,10)$.

a) We define the Bayes Classifier as:

$$h^*(x) = \begin{cases} 1 & \text{if } \frac{\mathbb{P}(x|Y=1)}{\mathbb{P}(x|Y=-1)} > \frac{1-\pi_1}{\pi_1} \\ -1 & \text{ow.} \end{cases}$$

$$\Rightarrow h^*(x) = \begin{cases} 1 & \text{if } \mathbb{P}(x|Y=1) > \mathbb{P}(x|Y=-1) \\ -1 & \text{ow.} \end{cases}$$

$$\rightarrow h^*(x) = \begin{cases} 1 & \text{if } 5 < x < 10 \\ -1 & \text{ow} \end{cases}$$

Where the decision boundary corresponds to $-5 \le x \le 5$

The Bayes Risk corresponds to:

$$\begin{aligned} \mathbb{P}(Y \neq h^*(x)) &= \mathbb{P}(Y=1, h^*(x)=-1) + \mathbb{P}(Y=-1, h^*(x)=1) \\ &= \mathbb{P}(Y=1, h^*(x)=-1) + 0 \\ &= \mathbb{P}(h^*(x)=-1 \mid Y=1)\, \mathbb{P}(Y=1) \\ &= \mathbb{P}(-5 \le x < 5 \mid Y=1)\, \mathbb{P}(Y=1) + \cancel{\mathbb{P}(x<-5 \mid Y=1)}^{\,0} \\ &= \left(\int_{-5}^{5} \frac{dx}{15}\right) \frac{1}{2} = \left(\frac{2}{3}\right)\frac{1}{2} = \frac{1}{3} \end{aligned}$$

b) $$h_5(x) = \begin{cases} 1 & \text{if } \operatorname{sgn}(x-5) > 0 \\ -1 & \text{if } \operatorname{sgn}(x-5) \le 0 \end{cases}$$

This essentially the same classifier as part (a), so the Bayes Risk is the same.

c) We wish to compute the Hinge Risk

$R_\phi(\beta) = \mathbb{E}(1 - Y\beta X)_+ = 2/3$

2. Logistic Regression

a) We define the log-likelihood function to be

$\ell(\beta_0, \beta) = \sum_{i=1}^{n} [y_i(\beta_0 + x_i^T\beta) - \log(1 + e^{\beta_0 + x_i^T\beta})]$

We use the simplification:

$x_i \leftarrow (1, x_i^T)^T \quad \beta \leftarrow (\beta_0, \beta^T)^T$ to simplify notation:

$\ell(\beta) = \sum_{i=1}^{n} [y_i (x_i^T\beta) - \log(1 + e^{x_i^T\beta})]$

Also, recall $\pi_i(x_i, \beta^{(k)}) = \frac{e^{x_i^T\beta^{(k)}}}{1 + e^{x_i^T\beta^{(k)}}}$

We move to the $(k+1)$th-step:

We know from our class notes that

$$\hat{\beta}^{(k+1)} = \hat{\beta}^{(k)} - \left(\frac{\partial^2 \ell(\hat{\beta}^{(k)})}{\partial\beta\partial\beta^T}\right)^{-1} \frac{\partial \ell(\hat{\beta}^{(k)})}{\partial\beta}$$

So, we need to compute:

$\frac{\partial \ell(\hat{\beta}^{(k)})}{\partial\beta} = \sum_{i=1}^{n} \left(y_i x_i - \frac{e^{x_i^T\beta}}{1 + e^{x_i^T\beta}} x_i\right)\Big|_{\beta = \hat{\beta}^{(k)}} = X^T(y - \pi_i(x_i, \hat{\beta}^{(k)}))$ as expected.

and $\frac{\partial \ell(\hat{\beta}^{(k)})}{\partial\beta\partial\beta^T} = \sum_{i=1}^{n} -\left[\frac{x_i^T e^{x_i^T\beta}(1 + e^{x_i^T\beta}) - x_i^T e^{x_i^T\beta}(e^{x_i^T\beta})}{(1 + e^{x^T\beta})^2}\right] x_i$

$$= \sum_{i=1}^{n} - x_i^T \left[\frac{e^{x_i^T\beta} \quad (1)}{(1 + e^{x_i^T\beta})(1 + e^{x_i^T\beta})}\right] x_i$$

$$= -X^T \begin{bmatrix} \pi_1(1-\pi_1) & & 0 \\ & \ddots & \\ 0 & & \pi_n(1-\pi_n) \end{bmatrix} X = -X^T W X$$

So, we get that:

$$\hat{\beta}^{(k+1)} = \hat{\beta}^{(k)} - \left(\frac{\partial^2 \ell(\hat{\beta}^{(k)})}{\partial\beta\partial\beta^T}\right)^{-1} \frac{\partial \ell(\hat{\beta}^{(k)})}{\partial\beta}$$

$$= \hat{\beta}^{(k)} + (X^T W X)^{-1} X^T(y - \pi_1^{(k)})$$

$$= (X^T W X)^{-1}(X^T W X)\hat{\beta}^{(k)} + (X^T W X)^{-1} X^T(y - \pi_1^{(k)})$$

$$= (X^T W X)^{-1} X^T W (X\hat{\beta}^{(k)} + W^{-1}(y - \pi_1^{(k)}))$$

This gives us iteratively reweighted least squares.

b) We wish to show that if the data is perfectly separable then the maximum conditional log-likelihood does not exist for the log-regression model.

If the data are perfectly separable then we can find a linear decision boundary which perfectly classifies the data.
Since it classifies perfectly
$\log \pi_1$ or $\log 1-\pi_1$ would not converge.
So, there is no unique solution to this problem.

Also, the IRLS algorithm would diverge.
Since, if $y=1$ or $y=0$, $\pi_1 = 1$ or $\pi = 1$, respectively, we would see
$w_{ii}^{(k)} = \pi_1(x_i, \hat{\beta}^{(k)})(1-\pi(x_i, \hat{\beta}^{(k)})) \to 0$.
So, $(x_i^T w_{ii}^{(k)} x_i)^{-1} \to \infty$ which would cause $\hat{\beta}^{(k)} \to \infty$ as $k \to \infty$.
So, the algorithm would not converge.

c) We wish to give a derivation of Newton's Algorithm for Ridge Logistic Regression using $\lambda \|\beta\|^2$.
Recall the Ridge Lostic Reglession equation:

$$\hat{\beta}_0, \hat{\beta} = \underset{\beta_0, \beta}{\operatorname{argmin}} \left\{ -\sum_{i=1} y_i(\beta_0 + x_i^T\beta) - \log(1+\exp(\beta_0 + x_i^T\beta)) + \lambda \|\beta\|^2 \right\}$$

Using the same notation as before $(\beta_0, \beta)^T \to \beta$ and $(x_i^T, 1)^T \to x_i$
So,

$$(\hat{\beta}) = \underset{\beta}{\operatorname{argmin}} \left\{ -\sum_{i=1}^{n} [y_i(x_i^T\beta) - \log(1+\exp(x_i^T\beta))] + \lambda \|\beta\|_2^2 \right\}$$

Again we want to derive: $\beta^{k+1} \leftarrow \beta^k - \left[\frac{\partial^2 \ell(\beta)}{\partial\beta\,\partial\beta^T}\right]^{-1} \frac{\partial \ell(\beta)}{\partial\beta}$

So, $$\frac{\partial \ell(\beta)}{\partial \beta} = -\sum_{i=1}^{n} [y_i x_i - \frac{e^{x_i^T\beta}}{1+e^{x_i^T\beta}} x_i] + 2\lambda\beta = -X^T[y - \pi_i] + 2\lambda\beta$$

And $$\frac{\partial^2 \ell(\beta)}{\partial\beta\,\partial\beta^T} = -\sum_{i=1}^{n} x_i^T \left(\frac{e^{x_i^T\beta} \ (1)}{(1+e^{x_i^T\beta})(1+e^{x_i^T\beta})} \right) x_i + 2\lambda I = X^T W X + 2\lambda I$$

similar to (a)

So, $\hat{\beta}^{(k+1)} = \hat{\beta}^{(k)} - \left(\frac{\partial^2 \ell(\hat{\beta}^{(k)})}{\partial\beta\,\partial\beta^T} \right)^{-1} \frac{\partial \ell(\hat{\beta}^{(k)})}{\partial \beta}$

$$= \beta^{(k)} + \left[X^T W X + 2\lambda I\right]^{-1} \left[X^T(y - \pi_1) - 2\lambda\beta^{(k)}\right]$$

$$= \left[X^T W X + 2\lambda I\right]^{-1} \left[X^T W X + \cancel{2\lambda I}\right]\beta^{(k)}$$
$$+ \left[X^T W X + 2\lambda I\right]^{-1} X^T (y - \pi_1) - \cancel{\left(X^T W X + 2\lambda I\right)^{-1} 2\lambda\beta^{(k)}}$$

$$\hat{\beta}^{(k+1)} = \left[X^T W X + 2\lambda I\right]^{-1} X^T W \left[X\beta^{(k)} + W^{-1}(y - \pi_1^{(k)})\right]$$

From this step we have derived the Newton algorithm for Ridge logistic Regression.
The step is almost exactly the same as in the derivation of the IRLS algorithm, except the inverse term includes the penalty.