1) Cross-entropy error measure

(a) More generally, if we are learning from ±1 data to predict a noisy target $P(y|x)$ with candidate hypothesis h, show that that the maximum likelihood method reduces to the task of finding h that minimizes

$$E_{in}(w) = \sum_{n=1}^{N} [y_n = +1]\, ln \frac{1}{h(x_n)} + [y_n = -1]\, ln \frac{1}{1 - h(x_n)}$$

We first note that the PMF of $p(y|x)$ can be written as

$$p(y|x) = \begin{cases} (h(x))^{[y=+1]} \cdot (1 - h(x))^{[y=-1]} & y = -1,\, 1 \\ 0 & otherwise \end{cases}$$

Computing the MLE of $L(w)$

$$\begin{aligned} L(w) &= \prod_{n=1}^{N} p(y_n|x_n) \\ ln(L(w)) &= ln(\prod_{n=1}^{N} p(y_n|x_n)) \\ &= \sum_{n=1}^{N} ln(p(y_n|x_n)) \\ &= \sum_{n=1}^{N} ln((h(x_n))^{[y_n=+1]} \cdot (1 - h(x_n))^{[y_n=-1]}) \\ &= \sum_{n=1}^{N} [y_n = +1] \cdot ln(h(x_n)) + [y_n = -1] \cdot ln(1 - h(x_n)) \end{aligned}$$

We negate $-ln(L(w))$ to obtain the following

$$\begin{aligned} -ln(w) &= -\sum_{n=1}^{N} [y_n = +1] \cdot ln(h(x_n)) + [y_n = -1] \cdot ln(1 - h(x_n)) \\ &= \sum_{n=1}^{N} [y_n = +1] \cdot ln(\frac{1}{h(x_n)}) + [y_n = -1] \cdot ln(\frac{1}{1 - h(x_n)}) \end{aligned}$$

(b) For the case $h(x) = \theta(w^T x)$, argue that minimizing the in-sample error in part (a) is equivalent to minimizing the following equation:

$$\frac{1}{N}\sum_{n=1}^{N} ln(1 + e^{-y_n w^T x_n}) \qquad (1)$$

WLOG we suppose $y_n = +1$ for the nth training sample. The error obtained for the single training example becomes

$$ln(\frac{1}{h(x_n)}) \quad = \quad ln(1 + e^{-w^T x})$$

Equation (1) reduces down to $\frac{1}{N}ln(1+e^{-w^T x})$ for the error for the nth training sample for when $y_n = +1$ which is equivalent save for the scaling factor $\frac{1}{N}$. We conclude that minimizing the equation in part (a) is equivalent to minimizing equation (1)

2) For Logistic regression, show that

$$\begin{aligned} \nabla_w E_{in}(w) &= -\frac{1}{N}\sum_{n=1}^{N} \frac{y_n \cdot x_n}{1 + e^{y_n w^T x_n}} \\ &= \frac{1}{N}\sum_{n=1}^{N} -y_n \cdot x_n \theta(-y_n w^T x_n) \end{aligned}$$

Argue that a 'misclassified' example contributes more to the gradient than a correctly classified one

We first show

$$\begin{aligned} \nabla E_{in}(w) &= -\frac{1}{N}\sum_{n=1}^{N} \frac{y_n \cdot x_n}{1 + e^{y_n w^T x_n}} \\ &= \frac{1}{N}\sum_{n=1}^{N} -\frac{y_n \cdot x_n}{1 + e^{y_n w^T x_n}} \\ &= \frac{1}{N}\sum_{n=1}^{N} -y_n \cdot x_n \cdot \theta(-y_n w^T x_n) \end{aligned}$$

We analyze the ith training sample. WLOG we suppose $y_i = +1$. Now suppose $w^T x_i \to \infty$ which corresponds when the prediction $\hat{y} = 1$ we then have

$$\begin{aligned}\theta(-w^T x_i) &= -x_i \cdot \frac{e^{-\infty}}{1+e^{-\infty}} \\ &= 0\end{aligned}$$

Now suppose $w^T x_i \rightarrow -\infty$ which corresponds to when the prediction $\hat{y} = -1$ we then have

$$\begin{aligned}\theta(w^T x_i) &= -x_i \cdot \frac{e^{\infty}}{1+e^{\infty}} \\ &= \infty\end{aligned}$$

We thus conclude that a misclassified example contributes more to the gradient than a correctly classified one.

3) Consider the feature transform $\Phi(x) = (1, x_1^2, x_2^2)$. What kind of boundary in χ does a hyperplane $\tilde{w}$ in Z correspond to in the following cases? Draw a picture that illustrates an example of each case.

(a) $\tilde{w}_1 > 0$, $\tilde{w}_2 < 0$

We let $\tilde{w}_0 = 1$, $\tilde{w}_1 = 1$, $\tilde{w}_2 = -1$

$$\begin{aligned}sign(\begin{bmatrix}\tilde{w}_0 & \tilde{w}_1 & \tilde{w}_2\end{bmatrix}\begin{bmatrix}1 \\ x_1^2 \\ x_2^2\end{bmatrix}) &= \tilde{w}_0 + \tilde{w}_1 \cdot x_1^2 + \tilde{w}_2 x_2^2 > 0 \\ &= x_1^2 - x_2^2 > 1\end{aligned}$$

This decision boundary is a hyperbola

(b) $\tilde{w}_1 > 0$, $\tilde{w}_2 = 0$

We let $\tilde{w}_0 = -1$, $\tilde{w}_1 = 1$, $\tilde{w}_2 = 0$

$$\begin{aligned}sign(\begin{bmatrix}\tilde{w}_0 & \tilde{w}_1 & \tilde{w}_2\end{bmatrix}\begin{bmatrix}1 \\ x_1^2 \\ x_2^2\end{bmatrix}) &= \tilde{w}_0 + \tilde{w}_1 \cdot x_1^2 > 0 \\ &= \tilde{w}_1 \cdot x_1^2 > -\tilde{w}_0 \\ &= x_1^2 > 1\end{aligned}$$

This decision boundary is the vertical lines located at $x_1 = \pm 1$

(c) $\tilde{w}_1 > 0$, $\tilde{w}_2 > 0$, $\tilde{w}_0 < 0$

We let $\tilde{w}_0 = -1$, $\tilde{w}_1 = 1$, $\tilde{w}_2 = 1$

$$\begin{aligned} sign(\begin{bmatrix}\tilde{w}_0 & \tilde{w}_1 & \tilde{w}_2\end{bmatrix}\begin{bmatrix}1\\x_1^2\\x_2^2\end{bmatrix}) &= \tilde{w}_0 + \tilde{w}_1 \cdot x_1^2 + \tilde{w}_2 \cdot x_2^2 > 0 \\ &= x_1^2 + x_2^2 > 1 \end{aligned}$$

This is a circle of radius 1

(d) $\tilde{w}_1 > 0$, $\tilde{w}_2 > 0$, $\tilde{w}_0 > 0$

We let $\tilde{w}_0 = 1$, $\tilde{w}_1 = 1$, $\tilde{w}_2 = 1$

$$\begin{aligned} sign(\begin{bmatrix}\tilde{w}_0 & \tilde{w}_1 & \tilde{w}_2\end{bmatrix}\begin{bmatrix}1\\x_1^2\\x_2^2\end{bmatrix}) &= \tilde{w}_0 + \tilde{w}_1 \cdot x_1^2 + \tilde{w}_2 \cdot x_2^2 > 0 \\ &= x_1^2 + x_2^2 > -1 \end{aligned}$$

This boundary cannot be created

4) Recall the objective function for linear regression can be expressed as

$$E(w) = \frac{1}{N}\left\|Xw - y\right\|^2$$

This solution holds only when $X^T X$ is nonsingular. To overcome this problem, the following objective function is commonly minimized instead:

$$E_2(w) = \left\|Xw - y\right\|^2 + \lambda\left\|w\right\|^2$$

(a) Derive the optimal w that minimize $E_2(w)$.

$$\begin{aligned}
\nabla_w E_2(w) &= 0 \\
\nabla_w(\left\|Xw - y\right\|^2 + \lambda \left\|w\right\|^2) &= 0 \\
\nabla_w \left\|Xw - y\right\|^2 + \nabla_w \lambda \left\|w\right\|^2 &= 0 \\
2X^T(Xw - y) + 2\lambda w &= 0 \\
X^T(Xw - y) + \lambda w &= 0 \\
X^T Xw - X^T y + \lambda w &= 0 \\
X^T Xw + \lambda w &= X^T y \\
(X^T X + \lambda I)w &= X^T y \\
w &= (X^T X + \lambda I)^{-1} X^T y
\end{aligned}$$

(b) Explain how this new objective function can overcome the singularity problem of $X^T X$

$$\begin{aligned}
v^T(X^T X + \lambda I)v &= \\
&= v^T X^T Xv + v^T \lambda v \\
&= v^T X^T Xv + \lambda v^T v \\
&= \left\|Xv\right\|^2 + \lambda \left\|v\right\|^2
\end{aligned}$$

Suppose X is singular and v is a non-zero vector s.t $Xv = 0$ then

$$\begin{aligned}
\left\|Xv\right\|^2 + \lambda \left\|v\right\|^2 &= \\
&= \lambda \left\|v\right\|^2
\end{aligned}$$

$\lambda \left\|v\right\|^2 > 0$ for all vectors non-zero vectors v and thus the matrix $(X^T X + \lambda I)$ is positive definite. one of the properties of positive definite is invertibility and thus we conclude $(X^T X + \lambda I)$ is an invertible matrix for $\lambda > 0$

5) In logistic regression, the objective function can be written as

$$E(w) = \frac{1}{N} \sum_{n=1}^{N} ln(1 + e^{-y_n w^T x_n})$$

(a) Compute the first-order deriative $\nabla_w E(w)$. You will need to provide intermediate steps of derivation.

$$
\begin{aligned}
\nabla_w E(w) &= \nabla_w \frac{1}{N}\sum_{n=1}^{N} ln(1+e^{-y_n w^T x_n}) \\
&= \frac{1}{N}\sum_{n=1}^{N} \nabla_w ln(1+e^{-y_n w^T x_n}) \\
&= \frac{1}{N}\sum_{n=1}^{N} \frac{1}{1+e^{-y_n w^T x_n}} \cdot \nabla_w (1+e^{-y_n w^T x_n}) \\
&= \frac{1}{N}\sum_{n=1}^{N} \frac{1}{1+e^{-y_n w^T x_n}} \cdot e^{-y_n w^T x_n} \nabla_w (-y_n w^T x_n) \\
&= \frac{1}{N}\sum_{n=1}^{N} \frac{1}{1+e^{-y_n w^T x_n}} \cdot e^{-y_n w^T x_n} \cdot -y_n \cdot x_n \\
&= -\frac{1}{N}\sum_{n=1}^{N} \frac{y_n \cdot x_n}{1+e^{-y_n w^T x_n}} \cdot \frac{1}{e^{y_n w^T x_n}} \\
&= -\frac{1}{N}\sum_{n=1}^{N} \frac{y_n \cdot x_n}{1+e^{y_n w^T x_n}} \\
&= \frac{1}{N}\sum_{n=1}^{N} -y_n \cdot x_n \cdot \theta(-y_n w^T x_n)
\end{aligned}
$$

(b) Once the optimal w is obtain, it will be used to make predictions as follows:

$$
Predicted\, class\, of\, x \quad = \quad \begin{cases} 1 & \theta(w^T x) \geq 0.5 \\ -1 & \theta(w^T x) < 0.5 \end{cases}
$$

Explain why the decision boundary of logistic regression is still linear, though the linear signal $w^T x$ is passed through a nonlinear function θ to compute the outcome of prediction.

We note the logistic sigmoid function has an inverse for every b where $0 < b < 1$ where the form of the inverse is $\theta^{-1}(b) = ln(\frac{b}{1-b})$. We apply the inverse to $\theta(w^T x) = 0.5$ which is where the decision boundary lies

$$\begin{aligned}
\theta(w^T x) &= 0.5 \\
\theta^{-1}(\theta(w^T x)) &= \theta^{-1}(0.5) \\
w^T x &= ln(\frac{0.5}{1-0.5}) \\
w^T x &= 0
\end{aligned}$$

which shows that the decision boundary is linear

(c) Is the decision boundary still linear if the prediction rule is changed to the following? Justify briefly.

$$Predicted\, class\, of\, x \quad = \quad \begin{cases} 1 & if\, \theta(w^T x) \geq 0.9 \\ -1 & if\, \theta(w^T x) < 0.9 \end{cases}$$

We again apply the logistic sigmoid inverse function to the decision boundary which is where $\theta(w^T x) = 0.9$

$$\begin{aligned}
\theta(w^T x) &= 0.9 \\
\theta^{-1}(\theta(w^T x)) &= \theta^{-1}(0.9) \\
w^T x &= log(\frac{0.9}{1-.9}) \\
w^T x &= log(9)
\end{aligned}$$

Which shows that the decision boundary is linear

(d) In light of your answers to the above two questions, what is the essential property of logistic regression that results in the linear decision boundary?

The essential property of logistic regression that results in the linear decision boundary is the fact that the system is linear in terms of x and w