1. (Decision theory). Consider a case where we have learned a conditional probability distribution $P(y|x)$. Suppose there are only two classes, and let $p_0 = P(y = 0|x)$ and $p_1 = P(y = 1|x)$. Consider the following loss matrix:

$\begin{array}{ccc}
\text{predicted} & \text{true label } y \\
\text{label } \hat{y} & 0 & 1 \\
0 & 0 & 10 \\
1 & 5 & 0 \\
\end{array}$

It can be shown that the decision $\hat{y}$ that minimizes the expected loss is equivalent to setting a specific probability threshold $\theta$ and predicting $\hat{y} = 0$ if $p_1 < \theta$ and $\hat{y} = 1$ if $p_1 \geq \theta$. Please compute the $\theta$ for the above given loss matrix. Show a loss matrix where the threshold is 0.1.

We want to predict $\hat{y} = 1$ if the expected loss of guessing 1 is less than the expected loss of guessing zero. Following the derivation in class, we should predict 1 iff

$$P(y = 0|x)5 < P(y = 1|x)10$$  

(1)

Define $p_1 = P(y = 1|x)$, then this becomes

$$(1 - p_1)5 < p_110$$  

(2)

$$5 - 5p_1 < 10p_1$$  

(3)

$$-15p_1 < -5$$  

(4)

$$p_1 > \frac{5}{15}$$  

(5)

$$p_1 > \frac{1}{3}$$  

(6)

So we should set the threshold to 1/3.

To get a threshold of 0.1, we could use the matrix:

$\begin{array}{ccc}
\text{predicted} & \text{true label } y \\
\text{label } \hat{y} & 0 & 1 \\
0 & 0 & 9 \\
1 & 1 & 0 \\
\end{array}$

2. (Reject Option). In many applications, the classifier is allowed to “reject” a test example rather than classifying it into one of the classes. Consider, for example, a case in which the cost of a misclassification is $10 but the cost of having a human manually make the decision is only $3. We can formulate this as the following loss matrix:

$\begin{array}{ccc}
\text{decision} & \text{true label } y \\
\text{} & 0 & 1 \\
predict 0 & 0 & 10 \\
predict 1 & 10 & 0 \\
reject & 3 & 3 \\
\end{array}$

Suppose $P(y = 1|x)$ is predicted to be 0.2. Which decision minimizes the expected loss?

Show that generally introducing the reject option will lead to two specific thresholds $\theta_0$ and $\theta_1$ such that the optimal decision is to predict 0 if $p_1 < \theta_0$, reject if $\theta_0 \leq p_1 \leq \theta_1$, and predict 1 if $p_1 > \theta_1$.

The expected loss for predicting $y = 1$ is:

$$0.2 \times 0 + 0.8 \times 10 = 8$$
The expected loss for predicting $y = 0$ is:

$$0.2 \times 10 + 0.8 \times 0 = 2$$

The expected loss for reject is:

$$0.2 \times 3 + 0.8 \times 3 = 3$$

The optimal prediction is thus $y = 0$

Now let’s consider the general case. Let’s denote the loss $L(0,1) = c_1$, and $L(1,0) = c_0$. We will assume that $L(0,0) = L(1,1) = 0$. Further let’s denote the loss of reject is $c_r$. Let $p_1 = p(y = 1|x)$. The expect loss of predicting $y = 1$ is:

$$0 \times p_1 + c_0 \times (1 - p_1) = c_0 \times (1 - p_1)$$

The expect loss of predicting $y = 0$ is:

$$c_1 \times p_1 + 0 \times (1 - p_1) = c_1 \times p_1$$

The expect loss of reject is $c_r$.

Thus, we will predict $y = 1$ if $c_0 \times (1 - p_1) < c_1 \times p_1$, and $c_0 \times (1 - p_1) < c_r$, or equivalently

$$\text{predict } y = 1 \text{ if } p_1 > \max\left(\frac{c_0}{c_0 + c_1}, 1 - \frac{c_r}{c_0}\right)$$

We will predict $y = 0$ if $c_1 \times p_1 < c_0 \times (1 - p_1)$, and $c_1 p_1 < c_r$, i.e.,

$$\text{Predict } y = 0 \text{ if } p_1 < \min\left(\frac{c_0}{c_0 + c_1}, \frac{c_r}{c_1}\right)$$

We will reject otherwise.

3. Show that under the modeling assumption used by LDA, we have

$$p(y|x) = \frac{1}{1 + \exp(-w^T x)}$$

Please clearly express $w$ using parameters of the LDA model including $\pi, \mu_0, \mu_1$ and $\Sigma$.

Let $\pi, \mu_1, \mu_0$ and $\Sigma$ be the parameters of the LDA model, and we have:

$$p(y = 1) = \pi$$

$$p(x|y = 1) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp -\frac{(x - \mu_1)^T \Sigma^{-1} (x - \mu_1)}{2}$$

$$p(x|y = 0) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp -\frac{(x - \mu_0)^T \Sigma^{-1} (x - \mu_0)}{2}$$
Derive a gradient ascent algorithm for maximum likelihood estimation of multi-class logistic regression.

4. We want to take partial gradient with respect to $w$.

Note that we have:

\[
\frac{\partial}{\partial w} \log p(y = 1|\mathbf{x}) = \frac{\partial}{\partial w} \log \left( \frac{p(y = 1|\mathbf{x})}{p(y = 0|\mathbf{x})} \right)
\]

The likelihood function can be written as:

\[
L(\mathbf{w}) = \prod_{k=1}^{K} \prod_{y_i = k} p(y_i = k|\mathbf{x}_i)
\]

Note that the second product is over all instances such that $y_i = k$. Take the log of the likelihood function, we have:

\[
l(\mathbf{w}) = \sum_{k=1}^{K} \sum_{y_i = k} \log p(y_i = k|\mathbf{x}_i)
\]

We want to take partial gradient with respect to $w_c$. Before doing that, let’s break $l$ into two parts:

\[
l(\mathbf{w}) = \sum_{k \neq c} \sum_{y_i = k} \log p(y_i = k|\mathbf{x}_i) + \sum_{y_i = c} \log p(y_i = c|\mathbf{x}_i)
\]

Note that we have

\[
p(y_i = k|\mathbf{x}_i) = \frac{\exp(\mathbf{w}_k \cdot \mathbf{x}_i)}{\sum_{j=1}^{K} \exp(\mathbf{w}_j \cdot \mathbf{x}_i)}
\]

Take the log:

\[
\log p(y_i = k|\mathbf{x}_i) = \mathbf{w}_k \cdot \mathbf{x}_i - \log(\sum_{j=1}^{K} \exp(\mathbf{w}_j \cdot \mathbf{x}_i))
\]
Let \( z_i = \sum_{j=1}^{K} \exp(w_j \cdot x_i) \), we have:

\[
\log p(y_i = k | x_i) = w_k \cdot x_i - \log z_i
\]

Plug this into \( l \), we have:

\[
l(w) = \sum_{k \neq c} \sum_{y_i = k} (w_k \cdot x_i - \log z_i) + \sum_{y_i = c} (w_c \cdot x_i - \log z_i)
\]

Now take the partial gradient:

\[
\frac{\partial}{\partial w_c} l = - \sum_{k \neq c} \sum_{y_i = k} \frac{1}{z_i} \frac{\partial z_i}{\partial w_c} + \sum_{y_i = c} \left( x_i - \frac{1}{z_i} \frac{\partial z_i}{\partial w_c} \right)
\]

\[
= \sum_{y_i = c} x_i - \sum_{k=1}^{K} \sum_{y_i = k} \frac{1}{z_i} \frac{\partial z_i}{\partial w_c}
\]

Note that the second double summation can be simplified to \( \sum_{i=1}^{N} \), where \( N \) is the total number of points. We now plug in

\[
\frac{\partial z_i}{\partial w_c} = x_i \exp(w_c \cdot x_i)
\]

and \( z_i = \sum_{j=1}^{K} \exp(w_j \cdot x_i) \), we have:

\[
\frac{\partial}{\partial w_c} l = \sum_{k=c} x_i - \sum_{i=1}^{N} \frac{\exp(w_c \cdot x_i)}{\sum_{j=1}^{K} \exp(w_j \cdot x_i)} x_i
\]

We will use \( \tilde{y}_{ic} \) to denote

\[
\frac{\exp(w_c \cdot x_i)}{\sum_{j=1}^{K} \exp(w_j \cdot x_i)}
\]

and use \( y_{ic} \) to denote a binary indicator variable such that \( y_{ic} = 1 \) if \( y_i = c \) and 0 otherwise.

Putting these new notations to use, we arrive at:

\[
\frac{\partial}{\partial w_c} l = \sum_{i=1}^{N} y_{ic} x_i - \tilde{y}_{ic} x_i = \sum_{i=1}^{N} (y_{ic} - \tilde{y}_{ic}) x_i
\]

Therefore, a gradient ascent update rule will be:

\[
w \leftarrow w + \lambda \sum_{i=1}^{N} (y_{ic} - \tilde{y}_{ic}) x_i
\]