Machine Learning Homework 4

Q1

We use the Lagrange multiplier to enforce the constraint $W^T W = 1$. We now need to maximize:

$$L(\lambda, w) = w^T(m_2 - m_1) + \lambda(w^T w - 1) \tag{1}$$

We calculate the derivatives:

$$\frac{\partial L(\lambda, w)}{\partial \lambda} = w^T w - 1 \tag{2}$$

And,

$$\frac{\partial L(\lambda, w)}{\partial w} = m_2 - m_1 + 2\lambda w \tag{3}$$

We set the derivatives above equal to 0, which give

$$w = -\frac{1}{2\lambda}(m_2 - m_1) \propto (m_2 - m_1) \tag{4}$$

Q2

$$\begin{aligned} J(w) &= \frac{(m_2 - m_1)^2}{s_1^2 + s_2 + 2} \\ &= \frac{\|w^T(m_2 - m_1)\|^2}{\sum_{n \in C_1}(w^T x_n - m_1)^2 + \sum_{n \in C_2}(w^T x_n - m_2)^2} \end{aligned} \tag{5}$$

The numerator can be further written as:

numerator=

$$[w^t(m_2 - m_1)][w^T(m_2 - m_1)]^T = w^T S_B w \tag{6}$$

Where we have defined:

$$S_B = (m_2 - m_1)(m_2 - m_1)^T \tag{7}$$

Denominator =

$$\begin{aligned} &\sum_{n \in C_1} [w^T(x_n - m_1)]^2 + \sum_{n \in C_2} [w^T(x_n - m_2)]^2 \\ &= w^T S_{w1} w + w^T S_{w_2} w \\ &= w^T S_w w \end{aligned} \tag{8}$$

Where we have defined

$$S_W = \sum_{n \in C_1} (x_n - m_1)(x_n - m_1)^T + \sum_{n \in C_2} (x_n - m_2)(x_n - m_2)^T \tag{9}$$

Q3

The likelihood function is

$$\begin{aligned} p(\{\emptyset_n, t_n\}|\pi_1, \pi_2, \ldots, \pi_K) &= \prod_{n=1}^{N} \prod_{k=1}^{K} [p(\emptyset_n|C_k)p(C_k)]^{t_{nk}} \\ &= \prod_{n=1}^{N} \prod_{k=1}^{K} [\pi_k p(\emptyset_n|C_k)]^{t_n k} \end{aligned} \tag{10}$$

By using logarithm likelihood:

$$\ln p = \sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} [\ln \pi_k + \ln p(\emptyset_n|C_k)] \propto \sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \ln \pi_k \tag{11}$$

Since there is a constraint on π_k, so we need to add a Lagrange Multiplier to the expression, which becomes:

$$L = \sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \ln \pi_k + \lambda(\sum_{k=1}^{K} \pi_k - 1) \tag{12}$$

Then we calculate the derivative of the expression above with regard to π_k

$$\frac{\partial L}{\partial \pi_k} = \sum_{n=1}^{N} \frac{t_{nk}}{\pi_k} + \lambda \tag{13}$$

Set the derivative to 0, then we can get:

$$\pi_k = -(\sum_{n=1}^{N} t_{nk})/\lambda = -\frac{N_k}{\lambda} \tag{14}$$

And if we perform summation on both sides with regard to k, we can see that:

$$1 = -(\sum_{k=1}^{K} N_k)/\lambda = -\frac{N}{\lambda} \tag{15}$$

Q4

The logistic form:

$$\sigma(a) = \frac{1}{1 + \exp(-a)} \tag{16}$$

And we can calculate its derivative with regard to a.

$$\begin{aligned} \frac{d\sigma(a)}{da} = \frac{\exp(a)}{[1 + \exp(-a)]^2} &= \frac{\exp(a)}{1 + \exp(-a)} \cdot \frac{1}{1 + \exp(-a)} \\ &= [1 - \sigma(a)] \cdot \sigma(a) \end{aligned} \tag{17}$$

Q5

The process is as below

$$
\begin{aligned}
\nabla E(w) &= -\nabla \sum_{n=1}^{N} \{t_n \ln y_n + (1-t_n)\ln(1-y_n)\} \\
&\quad - \sum_{n=1}^{N} \nabla \{t_n \ln y_n + (1-t_n)\ln(1-y_n)\} \\
&\quad - \sum_{n=1}^{N} \frac{d\{t_n \ln y_n + (1-t_n)\ln(1-y_n)\}}{dy_n} \frac{dy_n}{da_n} \frac{da_n}{dw} \\
&= \sum_{n=1}^{N} (y_n - t_n)\emptyset_n
\end{aligned}
\tag{18}
$$

Q6

For the case where $c = 3$ and the x-space dimension is 2,

The first scheme requires two linear discriminant functions, and constructs a set of non-parallel linear discriminant functions y_1 and y_2, then there exists x satisfying $y_1(x) > 0, y_2(x) > 0$, Then the scheme is ambiguous.

The second scheme requires three linear discriminant functions, and constructs a set of linear discriminant functions y_{12}, y_{13} and y_{23} that intersect each other and have different intersection points, so that the enclosed triangle The region satisfies $y_{12}(x) > 0, y_{13}(x) > 0, y_{23}(x) > 0$, then the scheme is ambiguous.

Q7

If the convex hull of $\{x_n\}$ and $\{y_n\}$ intersects, we know that there will be a point z which can be written as $z = \sum_n \alpha_n x_n$ and also $z = \sum_n \beta_n y_n$. Hence we can obtain:

$$
\begin{aligned}
\hat{w}^T z + w_0 &= \hat{w}^T (\sum_n \alpha_n x_n) = w_0 \\
&= (\sum_n \alpha_n \hat{w}^T x_n) + (\sum_n \alpha_n) w_0 \\
&= \sum_n \alpha_n (\hat{w}^T x_n + w_0)(*)
\end{aligned}
\tag{19}
$$

Where we have used $\sum_n \alpha_n = 1$. And if $\{x_n\}$ and $\{y_n\}$ are linearly separable

, we have $\hat{w}^T x_n + w_0 > 0$ and $\hat{w}^T y_n + w_0 < 0$, for $\forall x_N, y_n$. Together with $\alpha_n \geq 0$ and $(*)$, we know that $\hat{w}^T z + w_0 > 0$ And if we calculate $\hat{w}^T z + w_0$ from the perspective of $\{y_n\}$ following the same procedure, we can obtain $\hat{w}^T z + w_0 < 0$. Hence contradictory occurs. In other words, they are not linearly separable if their convex hulls intersect.

We have already proved the first statement, i.e., "convex hulls intersect" gives "not linearly separable", and what the second part wants us to prove is that "linearly separable" gives "convex hulls do not intersect". This can be done simply by contrapositive.

The true converse of the first statement should be if their convex hulls do not intersect, the data sets should be linearly separable. This is exactly what Hyperplane Separation Theorem shows us.