Homework 4

Questions: https://github.com/SUSTech-ML-Course/Machine-Learning-Course/blob/05b4e0bd4063b9d486f34ce557720b519947cbe3/lesson-homework/hw4/HW4.md

Question 1

Using Lagrange multiplier, we need to maximize

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

So we calculate the derivatives with respect to λ and $\mathbf{w}$

$$\frac{\partial L(\lambda, \mathbf{w})}{\partial \lambda} = \mathbf{w}^T\mathbf{w} - 1$$
$$\frac{\partial L(\lambda, \mathbf{w})}{\partial \mathbf{w}} = \mathbf{m_2} - \mathbf{m_1} + 2\lambda\mathbf{w}$$

Set the derivatives to 0, we have $\mathbf{w}^T\mathbf{w} = 1$ and

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

Question 2

$$\begin{aligned}
J(\mathbf{w}) = & \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2} \\
= & \frac{||\mathbf{w}^T(\mathbf{m_2} - \mathbf{m_1})||^2}{\sum_{n \in C_1}(\mathbf{w}^T\mathbf{x_n} - m_1)^2 + \sum_{n \in C_2}(\mathbf{w}^T\mathbf{x_n} - m_2)^2} \\
= & \frac{\left[\mathbf{w}^T(\mathbf{m_2} - \mathbf{m_1})\right]\left[\mathbf{w}^T(\mathbf{m_2} - \mathbf{m_1})\right]^T}{\sum_{n \in C_1}(\mathbf{w}^T\mathbf{x_n} - m_1)^2 + \sum_{n \in C_2}(\mathbf{w}^T\mathbf{x_n} - m_2)^2} \\
= & \frac{\mathbf{w}^T\mathbf{S_B}\mathbf{w}}{\sum_{n \in C_1}(\mathbf{w}^T\mathbf{x_n} - m_1)^2 + \sum_{n \in C_2}(\mathbf{w}^T\mathbf{x_n} - m_2)^2} \\
= & \frac{\mathbf{w}^T\mathbf{S_B}\mathbf{w}}{\mathbf{w}^T\mathbf{S_{w1}}\mathbf{w} + \mathbf{w}^T\mathbf{S_{w2}}\mathbf{w}} \\
= & \frac{\mathbf{w}^T\mathbf{S_B}\mathbf{w}}{\mathbf{w}^T\mathbf{S_w}\mathbf{w}}
\end{aligned}$$

Question 3

The likelyhood function

$$
\begin{aligned}
p(\{\phi_{\mathbf{n}}, t_n\}|\pi_1, \pi_2, \ldots, \pi_K) = & \quad \prod_{n=1}^{N}\prod_{k=1}^{K}[p(\boldsymbol{\phi_n}|C_k)p(C_k)]^{t_{nk}} \\
= & \quad \prod_{n=1}^{N}\prod_{k=1}^{K}[\pi_k p(\phi_n|C_k)]^{t_{nk}} \\
ln\ p(\{\phi_{\mathbf{n}}, t_n\}|\pi_1, \pi_2, \ldots, \pi_K) = & \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\left[\ln \pi_k + \ln p(\phi_n|C_k)\right] \\
\propto & \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk} \ln \pi_k
\end{aligned}
$$

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

$$
\begin{aligned}
L &= \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk} \ln \pi_k + \lambda(\sum_{k=1}^{K}\pi_k - 1) \\
\frac{\partial L}{\partial \pi_k} &= \sum_{n=1}^{N}\frac{t_{nk}}{\pi_k} + \lambda = 0 \\
\Rightarrow \quad \pi_k &= -\frac{\sum_{n=1}^{N} t_{nk}}{\lambda} = -\frac{N_k}{\lambda} \qquad\qquad\qquad\qquad (*) \\
\sum_{k=1}^{K}\pi_k &= -\frac{\sum_{k=1}^{K}\sum_{n=1}^{N} t_{nk}}{\lambda} = -\sum_{k=1}^{K}\frac{N_k}{\lambda} \\
1 &= -\frac{\sum_{k=1}^{K} N_k}{\lambda} = -\frac{N}{\lambda} \\
\Rightarrow \quad \lambda &= -N
\end{aligned}
$$

Substitute it back into $(*)$, we have

$$
\pi_k = \frac{N_k}{N}
$$

Question 4

$$
\begin{aligned}
\sigma(a) &= \frac{1}{1+e^{-a}} \\
\frac{d\sigma}{da} &= \frac{e^{a}}{(1+e^{-a})^2} \\
&= \frac{1}{1+e^{-a}}\frac{e^{a}}{1+e^{-a}} \\
&= \sigma(1-\sigma)
\end{aligned}
$$

Question 5

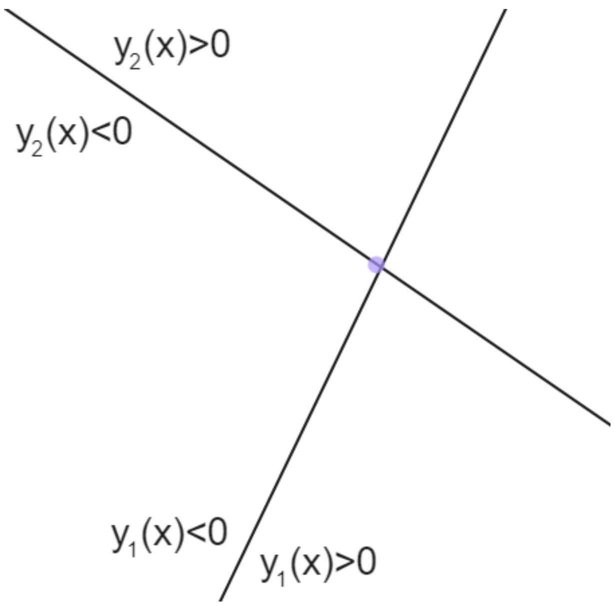
$$
\begin{aligned}
\nabla E(\boldsymbol{w}) &= -\nabla \sum_{n=1}^{N} [t_n \ln y_n + (1-t_n)\ln(1-y_n)] \\
&= -\sum_{n=1}^{N} \nabla [t_n \ln y_n + (1-t_n)\ln(1-y_n)] \\
&= -\sum_{n=1}^{N} \frac{d[t_n \ln y_n + (1-t_n)\ln(1-y_n)]}{dy_n} \frac{dy_n}{da_n} \frac{da_n}{d\mathbf{w}} \\
&= -\sum_{n=1}^{N} (\frac{t_n}{y_n} - \frac{1-t_n}{1-y_n}) y_n(1-y_n)\boldsymbol{\phi_n} \\
&= -\sum_{n=1}^{N} \frac{t_n - y_n}{y_n(1-y_n)} y_n(1-y_n)\boldsymbol{\phi}_n \\
&= -\sum_{n=1}^{N} (t_n - y_n)\boldsymbol{\phi}_n \\
&= \sum_{n=1}^{N} (y_n - t_n)\boldsymbol{\phi_n}
\end{aligned}
$$

Question 6

1. Since $c = 3$, we have two discriminant functions $y_1(\mathbf{x})$ and $y_2(\mathbf{x})$

   For $\mathbf{x} \in C_1, y_1(\mathbf{x}) > 0$ and $\mathbf{x} \in C_2, y_2(\mathbf{x}) > 0$, if we have $y_1(\mathbf{x}) > 0$ and $y_2(\mathbf{x}) > 0$, then we have $\mathbf{x} \in C_1 \cap C_2$

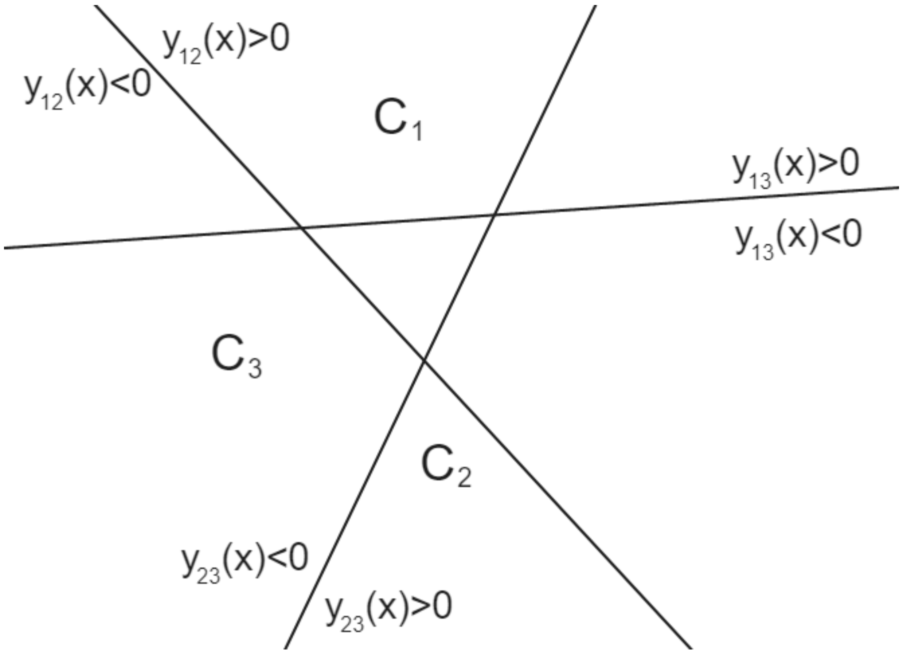
   Here we have the example figure

   According to the example, making the discriminant lines parallel to each other does not resolve the problem, since the intersection of $y_1(\mathbf{x}) > 0$ and $y_2(\mathbf{x}) > 0$ is non-empty. Note that the intersection is a null set if and only if the two lines coincide which means $y_1(\mathbf{x}) = y_2(\mathbf{x})$

2. Since $c = 3$, $c(c-1)/2 = 3$, we have three discriminant functions $y_{12}(\mathbf{x}), y_{13}(\mathbf{x}), y_{23}(\mathbf{x})$, the classification structure is below
   - If $y_{12}(\mathbf{x}) > 0$ and $y_{13}(\mathbf{x}) > 0$, then $\mathbf{x} \in C_1$
   - If $y_{12}(\mathbf{x}) < 0$ and $y_{23}(\mathbf{x}) > 0$, then $\mathbf{x} \in C_2$
   - If $y_{13}(\mathbf{x}) < 0$ and $y_{23}(\mathbf{x}) < 0$, then $\mathbf{x} \in C_3$

   This leads to the problems illustrated in the figure below

   The following regions are unclassified
   - $y_{12}(\mathbf{x}) < 0$ and $y_{13}(\mathbf{x}) > 0$
   - $y_{12}(\mathbf{x}) > 0$ and $y_{23}(\mathbf{x}) > 0$ and $y_{13}(\mathbf{x}) < 0$

   The intersections are null sets if and only if $y_{12}(\mathbf{x}) = y_{13}(\mathbf{x}) = y_{23}(\mathbf{x})$

Question 7

1. If the convex hull of $\{\mathbf{x_n}\}$ and $\{\mathbf{y_n}\}$ intersects, there will be a point $\mathbf{z}$ which can be written as $\mathbf{z} = \sum_n \alpha_n \mathbf{x_n}$ and also $\mathbf{z} = \sum_n \beta_n \mathbf{y_n}$

   Then we have

$$\begin{aligned} \widehat{\mathbf{w}}^T\mathbf{z} + w_0 &= \widehat{\mathbf{w}}^T(\sum_n \alpha_n \mathbf{x_n}) + w_0 \\ &= (\sum_n \alpha_n \widehat{\mathbf{w}}^T \mathbf{x_n}) + (\sum_n \alpha_n) w_0 \\ &= \sum_n \alpha_n (\widehat{\mathbf{w}}^T \mathbf{x_n} + w_0) \qquad\qquad (*) \end{aligned}$$

   We can use contradiction to prove $\{\mathbf{x_n}\}$ and $\{\mathbf{y_n}\}$ are not linearly separable:

   Suppose $\{\mathbf{x_n}\}$ and $\{\mathbf{y_n}\}$ are linearly separable, we have $\widehat{\mathbf{w}}^T\mathbf{x_n} + w_0 > 0$ and $\widehat{\mathbf{w}}^T\mathbf{y_n} + w_0 < 0$, for $\forall \mathbf{x_n}, \mathbf{y_n}$

   Together with $\alpha_n \geqslant 0$ and $(*)$, we gett $\widehat{\mathbf{w}}^T\mathbf{z} + w_0 > 0$

   If we calculate $\widehat{\mathbf{w}}^T\mathbf{z} + w_0$ from the perspective of $\{\mathbf{y_n}\}$ in the same way, we can get $\widehat{\mathbf{w}}^T\mathbf{z} + w_0 < 0$

Hence the contradiction occurs, $\{\mathbf{x_n}\}$ and $\{\mathbf{y_n}\}$ are not linearly separable

2. The second statement is the contrapositive of the proven first statement, they are equivalent