Q1

From $m_2 - m_1 = w^T(m_2 - m_1)$ we can construct the Lagrangian function

$$L = w^T(m_2 - m_1) + \lambda(w^T w - 1)$$

Taking the gradient of L we obtain

$$\nabla L = m_2 - m_1 + 2\lambda w.$$

and setting this gradient to zero gives

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

from which follows that $w \propto (m_2 - m_1)$.

Q2. Starting with the numerator on the r.h.s of $J(w)$, we can rewrite it as follows

$$\begin{aligned}(m_2 - m_1)^2 &= (w^T(m_2 - m_1))^2 \\ &= w^T(m_2 - m_1)(m_2 - m_1)^T w \\ &= w^T S_B w.\end{aligned}$$

Similarly, we can rewrite the ~~de~~ denominator of the r.h.s of $J(w)$:

$$\begin{aligned}s_1^2 + s_2^2 &= \sum_{n \in C_1}(y_n - m_1)^2 + \sum_{k \in C_2}(y_k - m_2)^2 \\ &= \sum_{n \in C_1}(w^T(x_n - m_1))^2 + \sum_{k \in C_2}(w^T(x_k - m_2))^2 \\ &= \sum_{n \in C_1} w^T(x_n - m_1)(x_n - m_1)^T w + \sum_{k \in C_2} w^T(x_k - m_2)(x_k - m_2)^T w. \\ &= w^T S_w w.\end{aligned}$$

Q3. The likelihood function is given by

$$p(\{\phi_n, t_n\} | \{\pi_k\}) = \prod_{n=1}^{N}\prod_{k=1}^{K}\{p(\phi_n | C_k)\pi_k\}^{t_{nk}}$$

and taking the logarithm, we obtain

$$\ln p(\{\phi_n, t_n\} | \{\pi_k\}) = \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\{\ln p(\phi_n | C_k) + \ln \pi_k\}$$

In order to maximize the log likelihood with respect to π_k we need to preserve the constraint $\sum_k \pi_k = 1$. This can be done by introducing a Lagrange multiplier λ and maximizing

$$\ln p(\{\phi_n, t_n\} | \{\pi_k\}) + \lambda\left(\sum_{k=1}^{K}\pi_k - 1\right).$$

Setting the derivative with respect to π_k equal to zero. we obtain

$$\sum_{n=1}^{N}\frac{t_{nk}}{\pi_k} + \lambda = 0.$$

re-arrange the gives

$$-z_k \lambda = \sum_{n=1}^{N} t_{nk} = N_k.$$

Summing both sides over k we find that $\lambda = -N$. and easily this to eliminate λ.

we obtain. $$z_k = \frac{N_k}{N}.$$

Q4. Differentiating $\sigma(a)$ we obtain

$$\frac{d\sigma}{da} = \frac{e^{-a}}{(1+e^{-a})^2} = \sigma(a)\left\{\frac{e^{-a}}{1+e^{-a}}\right\} = \sigma(a)\left\{\frac{1+e^{-a}}{1+e^{-a}} - \frac{1}{1+e^{-a}}\right\} = \sigma(a)(1-\sigma(a))$$

Q5. We start by computing the derivative of $E(w)$. w.r.t y_n.

$$\frac{\partial E}{\partial y_n} = \frac{1-t_n}{1-y_n} - \frac{t_n}{y_n} = \frac{y_n(1-t_n) - t_n(1-y_n)}{y_n(1-y_n)} = \frac{t_n - y_n t_n - t_n + y_n t_n}{y_n(1-y_n)} = \frac{y_n - t_n}{y_n(1-y_n)}$$

From the result $$\frac{d\sigma}{da} = \sigma(1-\sigma).$$

we see that.

$$\frac{\partial y_n}{\partial a_n} = \frac{\partial \sigma(a_n)}{\partial a_n} = \sigma(a_n)(1-\sigma(a_n)) = y_n(1-y_n).$$

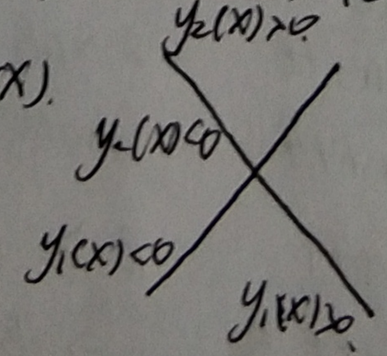
Finally we have.

$$\nabla a_n = \phi_n.$$

where ∇ denotes the gradient with respect to w. Combining these three equations using the chain rule, we obtain

$$\nabla E = \sum_{n=1}^{N} \frac{\partial E}{\partial y_n} \frac{\partial y_n}{\partial a_n} \nabla a_n = \sum_{n=1}^{N} (y_n - t_n)\phi_n.$$

Q6. P1. Since there are 3 classes, $C=3$ and these are 2 discriminant functions $y_1(x)$ and $y_2(x)$. For $x \in C_1$, $y_1(x) \geq 0$. and for $x \in C_2$, $y_2(x) > 0$. This leads to the following problems. How do we ~~clearly~~ classify input patterns x which have the property that $y_1(x) \geq 0$ and $y_2(x) \geq 0$. Clearly they belong to both class A and C_2. Figure 1 illustrates the problem. Making these discriminant lines parallel to each other does not resolve the problem since the intersection $y_1(x) \geq 0$ and $y_2(x) \geq 0$ is non-empty. Note that the intersection is a null set if and only if the two line coincide which means $y_1(x) = y_2(x)$.

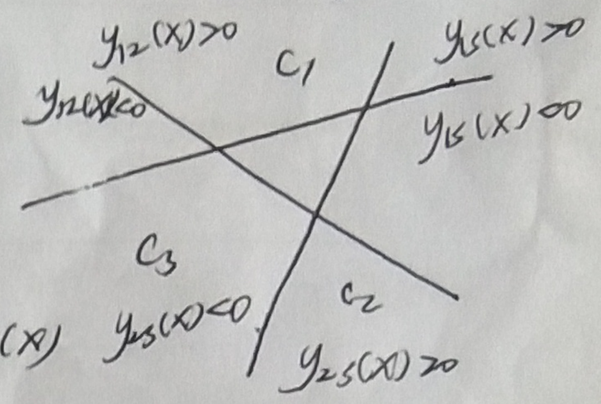


P2. Since there are 3 classes, $C(C-1)/2 = 3$ and there are three discriminant functions, $y_{12}(x)$, $y_{23}(x)$ and $y_{13}(x)$. The classification structure is as follows.

1. If $y_{12}(x) > 0$ and $y_{13}(x) > 0$, then $x \in C_1$.
2. If $y_{12}(x) < 0$. and $y_{23}(x) > 0$. then $x \in C_2$
3. If $y_{13}(x) < 0$, and $y_{23}(x) < 0$. then $x \in C_3$

This leads to the following problems as illustrated in figure. The following regions are unclassified.

1. $y_{12}(x) < 0$ and $y_{23}(x) > 0$.
2. $y_{12}(x) > 0$ and $y_{23}(x) > 0$ and $y_{13}(x) < 0$.



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

Q7. First, lets calculate the linear discriminants for the points belonging to the two convex hulls. For points in the convex hull of $\{x^n\}$, the linear discriminant is:

$$y(x) = \hat{w}^T x^n + w_0 \quad (2)$$

Substituting (1) in (2). we get

$$y(x) = \hat{w}^T \left(\sum_n \alpha_n x^n\right) + w_0 \quad (3)$$

Since α_n is a scalar quantity, we can bring the summation in (3) outside →

$$y(x) = \sum_n \alpha_n (\hat{w}^T x^n) + w_0 = \sum_n \alpha_n (\hat{w}^T x^n + w_0) \quad (4)$$

where wave made us of the fact that $\sum_n \alpha_n = 1$. Similarly, we can develope the linear discriminant for the ~~problem~~ points belonging to the convex hull of $\{z^m\}$.

$$y(z) = \sum_m \beta_m (\hat{w}^T z^m + w_0) \quad (5) \text{ where } \beta_m \geq 0 \text{ and } \sum_m \beta_m = 1$$

Convex hulls intersect: If the convex hulls intersect, there must be at least one point in common, between $\{x\}$ and $\{z\}$. Lets call that point x_z. Since x_z belongs to both convex hulls there must be a set of $\{\alpha_n\}$ and $\{\beta_m\}$ that give rise to x_z. The linear discriminant for x_z can now be written in two separate but equivalent ways. From (4) and (5), we get

$$y(x_z) = \sum_n \alpha_n (\hat{w}^T x^n + w_0) = \sum_m \beta_m (\hat{w}^T z^m + w_0) \quad (6)$$

For linear separability, we must have.

$$y(x^n) = \hat{w}^T x^n + w_0 > 0. \qquad y(z^m) = \hat{w}^T z^m + w_0 < 0. \quad (7)$$

From the non-negativity and simplex constraints on α and β, (6) and (7), we have a contradiction. The linear discriminant. $y(x_z)$ has to be simultaneously greater than and less than zero. which is important.

~~Pattern~~ Patterns are linearly seperable:

If the patterns are linearly separable, we know that

$$y(x^n) = \hat{w}^T x^n + w_0 > 0$$
$$y(z^m) = \hat{w}^T z^m + w_0 < 0 \quad (8)$$

Assume that there is a point ~~xz~~ lying in the intersection of the convex hulls. From (6) above. $y(xz) = \sum_n \alpha_n (\hat{w}^T x^n + w_0) = \sum_m \beta_m (\hat{w}^T z^m + w_0)$ (9)

The equality in (9) is not possible given the fact from (8) that the patterns are linearly separable.