Solutions to Homework # 1

Bishop 1.1: Plug in $x = d/2$ into 1.41, we get

$$\prod_{i=1}^{d} \int_{-\infty}^{+\infty} e^{-x_i^2} dx_i = \pi^{\frac{d}{2}} = S_d \int_0^{\infty} e^{-r^2} r^{d-1} dr \Rightarrow S_d = \frac{\pi^{\frac{d}{2}}}{\int_0^\infty e^{-r^2} r^{d-1} dr}$$

Using $r = \sqrt{u}$ to replace r, we get

$$\begin{aligned} S_d &= \frac{\pi^{\frac{d}{2}}}{\int_0^\infty e^{-u} \times u^{\frac{d-1}{2}} \times \frac{1}{2} \times u^{-\frac{1}{2}} du} \\ &= \frac{\pi^{\frac{d}{2}}}{\frac{1}{2} \int_0^\infty u^{\frac{d}{2}-1} e^{-u} du} \\ &= \frac{2\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2})} \end{aligned}$$

If we plug in $d = 2$, we get $S_d = \frac{2\pi}{\Gamma(1)} = 2\pi$, similarly, when $d = 3$,$S_d = \frac{2\pi^{\frac{3}{2}}}{\Gamma(\frac{3}{2})} = 4\pi$, both verify that S_d is the surface area of the unit sphere.

Bishop 1.2: We can do integration based on the following formula:
$V_d = \int_0^a S_d \times r^{d-1} dr = S_d \times \frac{r^d}{d} \mid_0^a = \frac{S_d a^d}{d}$, the geometric meaning of $S_d \times r^{d-1} \times dr$ is the volume of the thin shell whose bottom surface area is given by S_d, the whole hypersphere's volume is the sum of all such thin shells. Thus, $\frac{\text{volume of sphere}}{\text{volume of cube}} = \frac{\pi^{\frac{d}{2}}}{d 2^{d-1} \Gamma(\frac{d}{2})}$. Using Stirling's approximation, we can get $\frac{\text{volume of sphere}}{\text{volume of cube}} = \frac{\pi^{\frac{d}{2}}}{d 2^{d-1} e^{-(\frac{d}{2}-1)} (\frac{d}{2}-1)^{\frac{d-1}{2}}}$, when $d \to \infty$, the ratio becomes:

$$\begin{aligned} \lim_{d\to\infty} \frac{\pi^{\frac{d}{2}} e^{\frac{d}{2}-1}}{d 2^{d-1} (\frac{d}{2}-1)^{\frac{d}{2}-\frac{1}{2}}} &= \lim_{d\to\infty} \frac{2}{d} \lim_{d\to\infty} (\frac{\pi}{4})^{\frac{d}{2}} \lim_{d\to\infty} \frac{e^{\frac{d}{2}-1}}{(\frac{d}{2}-1)^{\frac{d-1}{2}}} \\ &= 0 \times 0 \times (\lim_{d\to\infty} \frac{e}{\frac{d}{2}-1})^{\frac{d}{2}-1} \times \lim_{d\to\infty} \frac{1}{(\frac{d}{2}-1)^{\frac{1}{2}}} \\ &= 0 \times 0 \times 0 \times 0 \\ &= 0 \end{aligned}$$

Similarly,

$$\begin{aligned} \frac{\text{distance from center of the cube to a corner}}{\text{distance from center to a face}} &= \frac{\frac{\sqrt{d \times 4a^2}}{2}}{a} \\ &= \sqrt{d} \end{aligned}$$

As $d \to \infty$, the above ratio $\to \infty$.

Bishop 3.2: There are several possible ways in which to generalize the concept of linear discriminant functions from two classes to c classes. One possibility would be to use $(c-1)$ linear discriminant functions, such that $y_k(\mathbf{x}) > 0$ for inputs **x** in class C_k and $y_k(\mathbf{x}) < 0$ for inputs not in class C_k. By drawing a simple example in two dimensions for $c = 3$, show that this approach can lead to regions of **x**-space for which the classification is ambiguous. Another approach would be to use one discriminant function $y_{jk}(\mathbf{x})$ for each possible pair of classes C_j and C_k , such that $y_{jk}(\mathbf{x}) > 0$ for patterns in class C_j and $y_{jk}(\mathbf{x}) < 0$ for patterns in class C_k. For c classes, we would need $c(c-1)/2$ discriminant functions. Again, by drawing a specific example in two dimensions for $c = 3$, show that this approach can also lead to ambiguous regions.

Part 1: Since there are 3 classes, $c = 3$ and there are 2 discriminant functions $y_1(\mathbf{x})$ and $y_2(\mathbf{x})$. For $\mathbf{x} \in C_1, y_1(\mathbf{x}) > 0$ and for $\mathbf{x} \in C_2, y_2(\mathbf{x}) > 0$. This leads to the following problem. How do we classify input patterns **x** which have the property that $y_1(\mathbf{x}) > 0$ AND $y_2(\mathbf{x}) > 0$. Clearly they belong to both class C_1 AND C_2. Figure 1 illustrates the problem. Making the discriminant lines parallel to each other does not resolve the problem since the intersection $y_1(\mathbf{x}) > 0$ 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 $y_1(\mathbf{x}) = y_2(\mathbf{x})$.

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

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

This leads to the following problems as illustrated in Figure 2. The following regions are unclassified.

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

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

Bishop 3.4: Given a set of data points $\{\mathbf{x}^n\}$ we can define the convex *hull* to be the set of points **x** given by

$$\mathbf{x} = \sum_n \alpha_n \mathbf{x}^n \tag{1}$$

where $\alpha_n \geq 0$ and $\sum_n \alpha_n = 1$. Consider a second set of points $\{\mathbf{z}\}^m$ and its corresponding convex hull. The two sets of points will be linearly separable if there exists a vector $\hat{\mathbf{w}}$ and a scalar w_0 such that $\hat{\mathbf{w}}^T \mathbf{x}^n + w_0 > 0$ for all $\mathbf{x}^n$, and $\hat{\mathbf{w}}^T \mathbf{z}^m + w_0 < 0$ for all $\mathbf{z}^m$. Show that, if their convex hulls intersect, the two

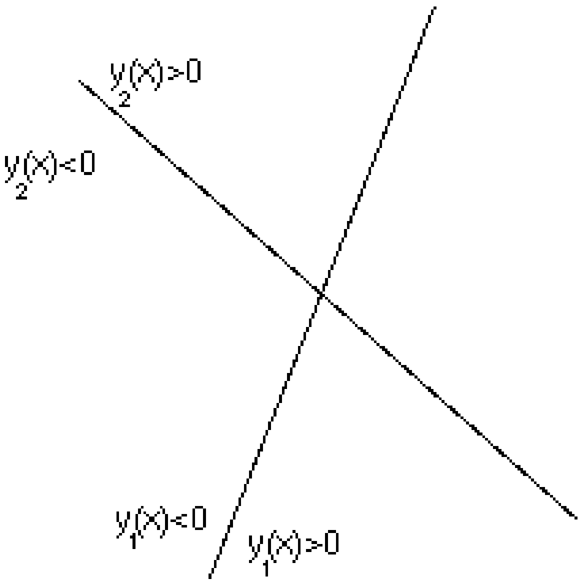

Figure 1: The two dividing linear discriminant boundaries clearly leave a region of space classified into two classes.

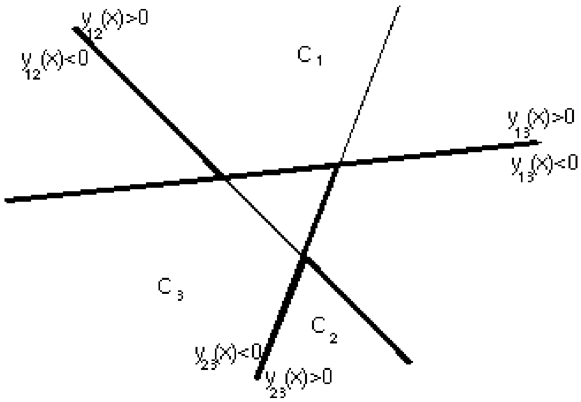

Figure 2: The three dividing linear discriminant boundaries clearly leave a region of space unclassified.

sets of points cannot be linearly separable, and conversely that, if they are linearly separable, their convex hulls do not intersect.

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

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

Substituting (1) in (2), we get

$$y(\mathbf{x}) = \hat{\mathbf{w}}^T (\sum_n \alpha_n \mathbf{x}^n) + w_0. \tag{3}$$

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

$$\begin{aligned} y(\mathbf{x}) &= \sum_n \alpha_n \left(\hat{\mathbf{w}}^T \mathbf{x}^n \right) + w_0 \\ &= \sum_n \alpha_n \left(\hat{\mathbf{w}}^T \mathbf{x}^n + w_0 \right) \end{aligned} \tag{4}$$

where we've made us of the fact that $\sum_n \alpha_n = 1$. Similarly, we can develop the linear discriminant for the points belonging to the convex hull of $\{\mathbf{z}^m\}$:

$$y(\mathbf{z}) = \sum_m \beta_m \left(\hat{\mathbf{w}}^T \mathbf{z}^m + w_0 \right) \tag{5}$$

where $\beta_m \geq 0$ and $\sum_m \beta_m = 1$.

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

$$y(\mathbf{xz}) = \sum_n \alpha_n \left(\hat{\mathbf{w}}^T \mathbf{x}^n + w_0 \right) = \sum_m \beta_m \left(\hat{\mathbf{w}}^T \mathbf{z}^m + w_0 \right). \tag{6}$$

For linear separability, we must have

$$\begin{aligned} y(\mathbf{x}^n) &= \hat{\mathbf{w}}^T \mathbf{x}^n + w_0 > 0 \\ \text{and} y(\mathbf{z}^m) &= \hat{\mathbf{w}}^T \mathbf{z}^m + w_0 < 0. \end{aligned} \tag{7}$$

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

Patterns are linearly separable: If the patterns are linearly separable, we know that

$$\begin{aligned} y(\mathbf{x}^n) &= \hat{\mathbf{w}}^T \mathbf{x}^n + w_0 > 0 \\ \text{and} y(\mathbf{z}^m) &= \hat{\mathbf{w}}^T \mathbf{z}^m + w_0 < 0. \end{aligned} \tag{8}$$

Assume that there is a point **xz** lying in the intersection of the convex hulls. From (6) above

$$y(\mathbf{xz}) = \sum_n \alpha_n \left(\hat{\mathbf{w}}^T \mathbf{x}^n + w_0\right) = \sum_m \beta_m \left(\hat{\mathbf{w}}^T \mathbf{z}^m + w_0\right). \tag{9}$$

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

Bishop 3.7 Ignore the prior probabilities' ratio question, it is not clear.

The sum-of-squares error function is

$$\begin{aligned}
E &= 3\int_0^3 \{y(x)-1\}^2 dx + \int_4^5 \{y(x)+1\}^2 dx \\
&= 3\int_0^3 [w^2x^2 + (w_0-1)^2 + 2(w_0-1)wx]dx + \int_4^5 [w^2x^2 + (w_0+1)^2 + 2(w_0+1)wx]dx \\
&= 3[w^2\frac{x^3}{3} \mid_0^3 + (w_0-1)^2 x \mid_0^3 + (w_0-1)wx^2 \mid_0^3] + [w^2\frac{x^3}{3} \mid_4^5 + (w_0+1)^2 x \mid_4^5 + (w_0+1)wx^2 \mid_4^5] \\
&= \frac{142}{3}w^2 + 36ww_0 + 10w_0^2 - 18w - 16w_0 + 10
\end{aligned}$$

$$\begin{aligned}
\min E &= \frac{142}{3}w^2 + 36ww_0 + 10w_0^2 - 18w - 16w_0 + 10 \\
\frac{\partial E}{\partial w} &= \frac{142 \times 2}{3}w + 36w_0 - 18 = 0 \\
\frac{\partial E}{\partial w_0} &= 36w + 20w_0 - 16 = 0
\end{aligned}$$

Solve the above linear equations, we get $w = -0.36161$,$w_0 = 1.450893$.Thus, the linear desciminant function is $y(x) = -0.36161x + 1.450893$. Since $y(0) = 1.450893, y(3) = 0.366063$,class 1 will be projected to the interval $[y(3), y(0)]$; $y(4) = 0.004453, y(5) = -0.35716$,class 2 will be projected to the interval $[y(5), y(4)]$,clearly, y(0)=0 cannot seperate the above line segements.

Since the two classes are linearly seperable from Figure 3.16, a single-layer perceptron will find a solution to separates the two classes exactly by the perceptron convergence theorem. The least-square algorithm can solve many problem in closed form, thus saving learning time. However,it sometimes will fail to seperate classes that are linearly seperable. The perceptron algorithm can find a solution if the classes are linearly seperatable, however, it cannot find solution for non-linear seperable problem and the learning time may become an issue.

Bishop 3.11: Using the definitions of the between-class and within-class covariance matrices given by (3.84) and (3.85) respectively, together with (3.91) and (3.92) and the choice of target values described in Section 3.6.2, show that the expression (3.90) which minimizes the sum-of-squares error function can be written in the form (3.93).

$$\mathbf{S}_B \quad = \quad (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^T, \text{ and } \mathbf{S}_W = \sum_{n \in C_1} (\mathbf{x}^{(n)} - \mathbf{m}_1)(\mathbf{x}^{(n)} - \mathbf{m}_1)^T + \sum_{n \in C_2} (\mathbf{x}^{(n)} - \mathbf{m}_2)(\mathbf{x}^{(n)} - \mathbf{m}_2)^T.$$

When the partial derivative of 3.88 is taken w.r.t. $\mathbf{w}$ and the result set to zero, we obtain

$$\sum_{n=1}^{N} (\mathbf{w}^T \mathbf{x}^{(n)} + w_0 - t^{(n)}) \mathbf{x}^{(n)} = 0, \tag{10}$$

where $w_0 = -\mathbf{w}^T \mathbf{m}$, and $t^{(n)} = \frac{N}{N_1}$ if $\mathbf{x}^{(n)} \in C_1$ and $t^{(n)} = -\frac{N}{N_2}$ if $\mathbf{x}^{(n)} \in C_2$ and $\mathbf{m} = \frac{1}{N}(N_1 \mathbf{m}_1 + N_2 \mathbf{m}_2)$. Substituting the expressions for w_0, $t^{(n)}$ into (10), we get

$$\sum_{n=1}^{N} [\mathbf{w}^T \mathbf{x}^{(n)} - \mathbf{w}^T \mathbf{m} - t^{(n)}] \mathbf{x}^{(n)} = 0.$$

This can be further expanded as

$$\sum_{n \in C_1} [\mathbf{w}^T \mathbf{x}^{(n)} - \mathbf{w}^T \mathbf{m} - \frac{N}{N_1}] \mathbf{x}^{(n)} + \sum_{n \in C_2} [\mathbf{w}^T \mathbf{x}^{(n)} - \mathbf{w}^T \mathbf{m} + \frac{N}{N_2}] \mathbf{x}^{(n)} = 0.$$

More algebra yields

$$[\sum_n \mathbf{x}^{(n)} \mathbf{x}^{(n)^T} - N \mathbf{m}\mathbf{m}^T] \mathbf{w} = N(\mathbf{m}_1 - \mathbf{m}_2).$$

Using the definitions of $\mathbf{S}_w$and $\mathbf{S}_B$,we get

$$\mathbf{S}_W + \frac{N_1 N_2}{N} \mathbf{S}_B = \sum_{n \in C_1} \mathbf{x}^{(n)} \mathbf{x}^{(n)^T} - N_1 \mathbf{m}_1 \mathbf{m}_1^T + \sum_{n \in C_2} \mathbf{x}^{(n)} \mathbf{x}^{(n)^T} - N_2 \mathbf{m}_2 \mathbf{m}_2^T + \frac{N_1 N_2}{N} (\mathbf{m}_2 \mathbf{m}_2^T - \mathbf{m}_2 \mathbf{m}_1^T - \mathbf{m}_1 \mathbf{m}_2^T + \mathbf{m}_1 \mathbf{m}_1^T).$$

Using the fact that $N_2 \mathbf{m}_2 = N\mathbf{m} - N_1 \mathbf{m}_1$,we can expand the last term involving $\mathbf{m}_1$and $\mathbf{m}_2$.

$$\mathbf{m}_2 \mathbf{m}_2^T - \mathbf{m}_2 \mathbf{m}_1^T - \mathbf{m}_1 \mathbf{m}_2^T + \mathbf{m}_1 \mathbf{m}_1^T = \mathbf{m}_2 \mathbf{m}_2^T - \frac{(N\mathbf{m} - N_1 \mathbf{m}_1)}{N_2} \mathbf{m}_1^T - \frac{(N\mathbf{m} - N_2 \mathbf{m}_2)}{N_1} \mathbf{m}_2^T + \mathbf{m}_1 \mathbf{m}_1^T.$$

From this, we get

$$N_1 N_2 (\mathbf{m}_2 \mathbf{m}_2^T - \mathbf{m}_2 \mathbf{m}_1^T - \mathbf{m}_1 \mathbf{m}_2^T + \mathbf{m}_1 \mathbf{m}_1^T) = N_1 N_2 \mathbf{m}_2 \mathbf{m}_2^T + N_1 N_2 \mathbf{m}_1 \mathbf{m}_1^T - N_1 (N\mathbf{m} - N_1 \mathbf{m}_1) \mathbf{m}_1^T - N_2 (N\mathbf{m} - N_2 \mathbf{m}_2) \mathbf{m}_2^T$$

Finally, we can simplify

$$-N_1 \mathbf{m}_1 \mathbf{m}_1^T - N_2 \mathbf{m}_2 \mathbf{m}_2^T + \frac{N_1 N_2}{N} (\mathbf{m}_2 \mathbf{m}_2^T - \mathbf{m}_2 \mathbf{m}_1^T - \mathbf{m}_1 \mathbf{m}_2^T + \mathbf{m}_1 \mathbf{m}_1^T) = -N \mathbf{m}\mathbf{m}^T$$

using which we may write

$$(\mathbf{S}_W + \frac{N_1 N_2}{N} \mathbf{S}_B) \mathbf{w} = (\sum_n \mathbf{x}^{(n)} \mathbf{x}^{(n)^T} - N \mathbf{m}\mathbf{m}^T) \mathbf{w} = N(\mathbf{m}_1 - \mathbf{m}_2).$$