Math 789 Assignment 4

Brendan Drachler

March 3, 2020

Problem 3.1

Starting from Eq. 3.1:

$$y(x) = \mathbf{w}_0 + \mathbf{w}^\top x \tag{1}$$

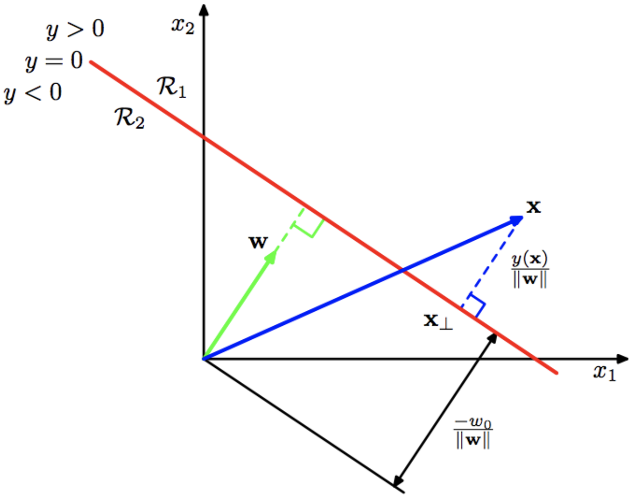
we will consider a point $\tilde{x}$ that lies on the decision boundary such that:

$$y(\tilde{x}) = \mathbf{w}_0 + \mathbf{w}^\top \tilde{x} \tag{2}$$

Now, we will try to understand how we can characterize the distance between our point and the decision boundary.

$$y(x) - y(\tilde{x}) = \mathbf{w}^\top (x - \tilde{x}) = 0 \tag{3}$$

This equation implies that $\mathbf{w}$ is orthogonal to the decision boundary containing point $\tilde{x}$. Now, we can look at this geometrically as Bishop does in Figure 4.1 (in the newest edition of Pattern Recognition and Machine Learning).



We can now calculate the perpendicular distance, L, from the decision boundary to an arbitrary point $\mathbf{x}$. It is clear that the hypotenuse of our triangle is given by:

$$\mathbf{x} = \mathbf{x}_\perp + L\frac{\mathbf{w}}{\|\mathbf{w}\|} \tag{4}$$

In this form, this equation doesn't help us much because we don't know what the form of L is. But we do know that $y(\mathbf{x}_\perp) = 0$, because $\mathbf{x}_\perp$ is on the decision boundary.

Therefore, we can write:

$$\mathbf{x}_\perp = \mathbf{x} - L\frac{\mathbf{w}}{\|\mathbf{w}\|} \tag{5}$$

and

$$y(\mathbf{x}_\perp) = \mathbf{w}_0 + \mathbf{w}^\top \mathbf{x}_\perp = \mathbf{w}_0 + \mathbf{w}^\top(\mathbf{x} - L\frac{\mathbf{w}}{\|\mathbf{w}\|}) = 0 \tag{6}$$

Multiplying out the last term and noting that $\mathbf{w}^\top \mathbf{w} = \|\mathbf{w}\|^2$ leads to:

$$\mathbf{w}_0 + \mathbf{w}^\top \mathbf{x} - L\|\mathbf{w}\| = 0 \tag{7}$$

$$\mathbf{w}_0 + \mathbf{w}^\top \mathbf{x} = L\|\mathbf{w}\| \tag{8}$$

The left hand side of this equation is $y(\mathbf{x})$ which allows us to solve for L:

$$L = \frac{y(\mathbf{x})}{\|\mathbf{w}\|} \tag{9}$$

Problem 3.2

For this example, I will include Fig. 4.2 from the latest edition of Bishop's textbook for reference.

In the case of $c = 3$, we will have 3 classes but only $(c - 1) = (3 - 1) = 2$ linear discriminant functions. This is the example given on the left side of Bishop's figure. We know that:

- $\mathbf{x} \in C_1, y_1(\mathbf{x}) > 0$
- $\mathbf{x} \in C_2, y_2(\mathbf{x}) > 0$

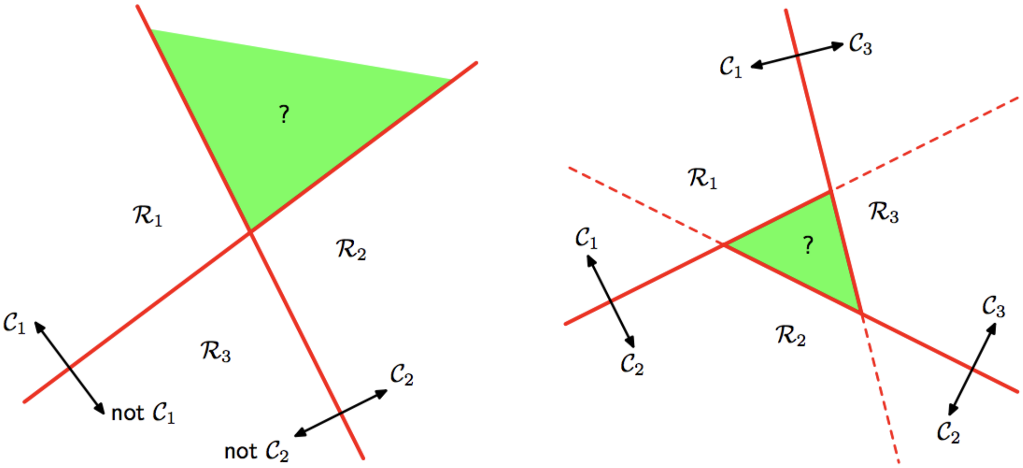
With this in mind, it is clear that the lower triangle has the property that $y_1(\mathbf{x}) < 0$ and $y_3(\mathbf{x}) < 0$. Therefore, that region is not ambiguous. The ambiguity stems from the upper region where $y_1(\mathbf{x}) > 0$ and $y_3(\mathbf{x}) > 0$. It is not clear that values in that region should belong to any of the classes because there is no way to differentiate them from each other.

However, we can define our number of linear discriminant functions in a different way by declaring that when $c = 3$, we have $c(c-1)/2 = 3(3-1)/2 = 3$ linear discriminant functions. We know that when:

- $(y_{12}(\mathbf{x}) > 0)$ and $(y_{13}(\mathbf{x}) < 0)$ then $(\mathbf{x} \in C_1)$
- $(y_{12}(\mathbf{x}) < 0)$ and $(y_{23}(\mathbf{x}) < 0)$ then $(\mathbf{x} \in C_2)$
- $(y_{13}(\mathbf{x}) > 0)$ and $(y_{23}(\mathbf{x}) > 0)$ then $(\mathbf{x} \in C_3)$

This leaves an ambiguous region when:

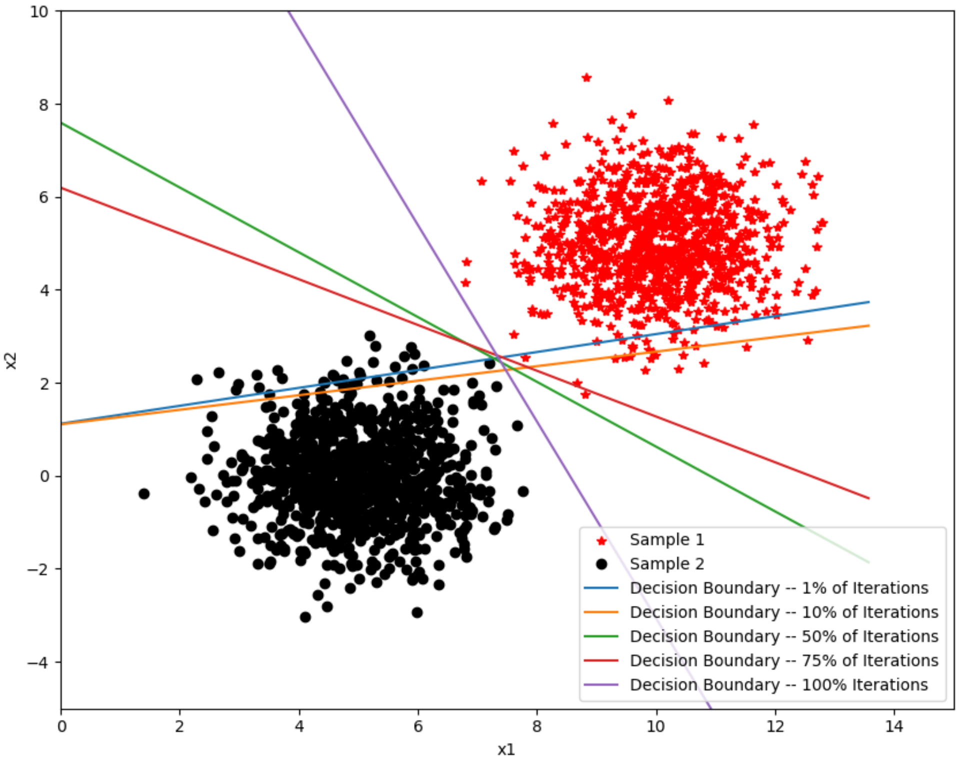
- $(y_{12}(\mathbf{x}) < 0)$ and $(y_{13}(\mathbf{x}) < 0)$ and $(y_{23}(\mathbf{x}) > 0)$



Problem 3.6

For this problem, I generated two normally distributed datasets with mean, $\mu_1 = (10, 5)$ and $\mu_2 = (5, 0)$ and a variance equal to the identity matrix.

The gradient descent algorithm produces the following results. Instead of giving the lines after a set number of iterations, I chose to give them as a percentage of the total iterations. I chose this method of displaying the results because the total number of iterations is arbitrary based on what learning rate one chooses. In some instances, a given learning rate can cause there to be many thousands of iterations before convergence.



The decision boundary is clearly performing very poorly for a low number of iterations but quickly improves.

Problem 3.7

Plug $y(x) = wx + w_0$ into Bishop Eq. 3.113:

$$E = 3\int_0^3 \left[w^2x^2 + (w_0 - 1)^2 + 2\,(w_0 - 1)\,wx\right] dx + \int_4^5 \left[w^2x^2 + (w_0 + 1)^2 + 2\,(w_0 + 1)\,wx\right] dx \tag{10}$$

This trivially leads to:

$$E = \frac{142}{3}w^2 + 36ww_0 + 10w_0^2 - 18w - 16w_0 + 10 \tag{11}$$

Now, we want to minimize the function with respect to w and w_0.

$$\frac{\partial E}{\partial w} = \frac{284}{3}w + 36w_0 - 18 = 0 \tag{12}$$

$$\frac{\partial E}{\partial w_0} = 36w + 20w_0 - 16 = 0 \tag{13}$$

Solving these two equations for w and w_0 leads to $w = -0.36161$ and $w_0 = 1.4509$. Thus, the linear discriminant function separating these two classes is $y(x) = -0.36161x + 1.4509$. Does this linear discriminant function work at $y(x) = 0$?

$$y(x) = -0.36161x + 1.4509 = 0 \tag{14}$$

$$x \approx 4.0123 \tag{15}$$

$x \approx 4.0123$ fails to separate the boundaries because it falls within class C_2.

A single layer perceptron could separate the two classes because they are clearly linearly separable and the perceptron convergence theorem states that convergence will be reached in a finite number of steps. Though, the number of steps can still be substantial compared to least-squares. However, there is no guarantee that least-squares will converge on a boundary that linearly separates the two classes.

Problem 3.11

Starting from Bishop Eq. 3.90:

$$\sum_{n=1}^{N} \left(\mathbf{w}^{\mathrm{T}}\mathbf{x}^n + w_0 - t^n\right)\mathbf{x}^n = 0 \tag{16}$$

Multiplying through by $\mathbf{x}^n$, utilizing Bishop Eq. 3.91, and considering that t^n takes on the value $N/N1$ for class C_1 and $-N/N2$ for class C_2 we can simplify our expression to:

$$\sum_{n=1}^{N}\mathbf{x_n}\mathbf{x_n}^T\mathbf{w} - \mathbf{w}^T\mathbf{m}\cdot(N\mathbf{m}) - \left(\sum_{n\in C_1}\frac{N}{N_1}\mathbf{x_n} + \sum_{n\in C_2}\frac{-N}{N_2}\mathbf{x_n}\right) = 0 \tag{17}$$

We can now use Bishop Eq. 3.98 to simplify further.

$$\sum_{n=1}^{N}\mathbf{x_n}\mathbf{x_n}^T\mathbf{w} - N\mathbf{m}\mathbf{m}^T\mathbf{w} = N\left(\mathbf{m}_1 - \mathbf{m}_2\right) \tag{18}$$

We now have the right hand side of Bishop Eq. 3.93. We need to work the left hand side further.

$$LHS = w[\sum_{n=1}^{N}\mathbf{x_n}\mathbf{x_n}^T - N\left(\frac{N_1}{N}\mathbf{m}_1 + \frac{N_2}{N}\mathbf{m}_2\right)^2] \tag{19}$$

$$= w[\sum_{n=1}^{N}\mathbf{x_n}\mathbf{x_n}^T - \frac{N_1^2}{N}\left\|\mathbf{m}_1\right\|^2 - \frac{N_2^2}{N}\left\|\mathbf{m}_2\right\|^2 - 2\frac{N_1N_2}{N}\mathbf{m}_1\mathbf{m}_2^T] \tag{20}$$

$$\begin{aligned}
&= w[\sum_{n=1}^{N}\mathbf{x_n}\mathbf{x_n}^T + (N_1 - 2N_1)\left\|\mathbf{m_1}\right\|^2 + (N_2 - 2N_2)\left\|\mathbf{m_2}\right\|^2 + \frac{N_1N_2}{N}\left\|\mathbf{m}_1 - \mathbf{m}_2\right\|^2] \\
&= w[\sum_{n=1}^{N}\mathbf{x_n}\mathbf{x_n}^T + N_1\left\|\mathbf{m_1}\right\|^2 - 2\mathbf{m_1}\sum_{n\in C_1}x_n^T + N_2\left\|\mathbf{m_2}\right\|^2 - 2\mathbf{m_2}\sum_{n\in C_2}x_n^T + \frac{N_1N_2}{N}\mathbf{S_B}] \\
&= w[\sum_{n\in C_1}\mathbf{x_n}\mathbf{x_n}^T + N_1\left\|\mathbf{m}_1\right\|^2 - 2\mathbf{m}_1\sum_{n\in C_1}x_n^T] \\
&= w[\sum_{n\in C_1}\left(\mathbf{x_n}\mathbf{x_n}^T + \left\|\mathbf{m_1}\right\|^2 - 2\mathbf{m_1}x_n^T\right) + \sum_{n\in C_2}\left(\mathbf{x_n}\mathbf{x_n}^T + \left\|\mathbf{m_2}\right\|^2 - 2\mathbf{m_2}\mathbf{x_n}^T\right) + \frac{N_1N_2}{N}\mathbf{S_B}] \\
&= w[\sum_{n\in C_1}\left\|\mathbf{x_n} - \mathbf{m_1}\right\|^2 + \sum_{n\in C_2}\left\|\mathbf{x_n} - \mathbf{m_2}\right\|^2 + \frac{N_1N_2}{N}\mathbf{S_B}] \\
&= w[\mathbf{S_w} + \frac{N_1N_2}{N}\mathbf{S_B}]
\end{aligned} \tag{21}$$

This validates Bishop Eq. 3.93.

$$\begin{gathered}
LHS = RHS \\
w[\mathbf{S_w} + \frac{N_1N_2}{N}\mathbf{S_B}] = N\left(\mathbf{m}_1 - \mathbf{m}_2\right)
\end{gathered} \tag{22}$$

Problem 3.12

Our goal is to utilize the following two equations:

$$\mathbf{m}_k = \frac{1}{N_k} \sum_{\mathbf{n} \in \mathcal{C}_k} \mathbf{x}^n \tag{23}$$

$$\mathbf{S}_T = \sum_{n=1}^{N} (\mathbf{x}^n - \mathbf{m}) (\mathbf{x}^n - \mathbf{m})^{\mathbf{T}} \tag{24}$$

to prove that $\mathbf{S}_T$ can be decomposed into two independent components of the covariance matrix given by:

$$\mathbf{S}_T = \mathbf{S}_W + \mathbf{S}_B \tag{25}$$

where $\mathbf{S}_W$ and $\mathbf{S}_B$ are given by:

$$\mathbf{S}_B = \sum_{k=1}^{c} N_k (\mathbf{m}_k - \mathbf{m}) (\mathbf{m}_k - \mathbf{m})^{\mathbf{T}} \tag{26}$$

$$\mathbf{S}_W = \sum_{k=1}^{c} \mathbf{S}_k = \sum_{k=1}^{c} (\mathbf{x}^n - \mathbf{m}_k)(\mathbf{x}^n - \mathbf{m}_k)^{\mathbf{T}} \tag{27}$$

We will begin by expanding $\mathbf{S}_T$ as follows:

$$\mathbf{S}_T = \sum_{k=1}^{c} \sum_{n=1}^{N} (\mathbf{x}^n - \mathbf{m}_k + \mathbf{m}_k - \mathbf{m}) (\mathbf{x}^n - \mathbf{m}_k + \mathbf{m}_k - \mathbf{m})^{\mathbf{T}} \tag{28}$$

We can now expand the inside of the summation by grouping the first two terms and the last two terms of each member of the product.

$$\begin{aligned}\mathbf{S}_T = \sum_{k=1}^{c} \sum_{n=1}^{N} (\mathbf{x}^n - \mathbf{m}_k)(\mathbf{x}^n - \mathbf{m}_k)^{\mathbf{T}} + \sum_{k=1}^{c} \sum_{n=1}^{N} (\mathbf{m}_k - \mathbf{m})(\mathbf{m}_k - \mathbf{m})^{\mathbf{T}} \\ + \sum_{k=1}^{c} \sum_{n=1}^{N} [(\mathbf{x}^n - \mathbf{m}_k)(\mathbf{m}_k - \mathbf{m})^{\mathbf{T}} + (\mathbf{m}_k - \mathbf{m})(\mathbf{x}^n - \mathbf{m}_k)^{\mathbf{T}}]\end{aligned} \tag{29}$$

Notice that the first term is $\mathbf{S}_W$ and the second term is $\mathbf{S}_B$. We still need to deal with the last two terms by utilizing the first equation in this section.

$$\mathbf{S}_T = \mathbf{S}_W + \mathbf{S}_B + \sum_{k=1}^{c}\sum_{n=1}^{N}[(\mathbf{x}^n - \mathbf{m}_k)(\mathbf{m}_k - \mathbf{m})^{\mathbf{T}} + (\mathbf{m}_k - \mathbf{m})(\mathbf{x}^n - \mathbf{m}_k)^{\mathbf{T}}] \tag{30}$$

Let's focus on the following term that shows up in the left over pieces of the expression:

$$\sum_{n=1}^{N}(\mathbf{x}^n - \mathbf{m}_k) = N_k\mathbf{m}_k - \sum_{n=1}^{N}\mathbf{m}_k = N_k\mathbf{m}_k - N_k\mathbf{m}_k = 0$$

$$\Downarrow$$

$$\sum_{k=1}^{c}\sum_{n=1}^{N}[(\mathbf{x}^n - \mathbf{m}_k)(\mathbf{m}_k - \mathbf{m})^{\mathbf{T}} + (\mathbf{m}_k - \mathbf{m})(\mathbf{x}^n - \mathbf{m}_k)^{\mathbf{T}}] = 0 \tag{31}$$

Therefore, $\mathbf{S}_T = \mathbf{S}_W + \mathbf{S}_B$.