

# MACHINE LEARNING

## CLASSIFICATION I

ACADEMIC YEAR 2023/2024

QUEEN MARY UNIVERSITY OF LONDON

---

# SOLUTIONS

---

**EXERCISE #1 (SOL):** The representation of a binary label is unimportant, in principle we could use any two symbols (○ and ○, A and B, 0 and 1, pear and apple, etc). We can however gain some computational advantages using the numerical values +1 and -1. Specifically, we can assign the value +1 to samples in the decision region defined by  $\mathbf{w}^T \mathbf{x}_i > 0$  and the value -1 to samples in the decision region defined by  $\mathbf{w}^T \mathbf{x}_i < 0$ . A correctly classified sample will receive a label that has the same sign as the quantity  $\mathbf{w}^T \mathbf{x}_i$ , whereas misclassified samples will receive a label with the opposite sign. Therefore, the margin  $m_i = y_i [\mathbf{w}^T \mathbf{x}_i]$  will be positive if  $\mathbf{x}_i$  is correctly classified, as the true label  $y_i$  and  $\mathbf{w}^T \mathbf{x}_i$  have the same sign. If it is misclassified, the margin will be negative, as the true label  $y_i$  and the quantity  $\mathbf{w}^T \mathbf{x}_i$  have opposite signs.

**EXERCISE #2 (SOL):** Figure 1 shows a dataset consisting of three samples belonging to class ○ and three samples belonging to class ○ in a 2D predictor space with attributes  $x_A$  and  $x_B$  and the linear boundary defining a classifier.

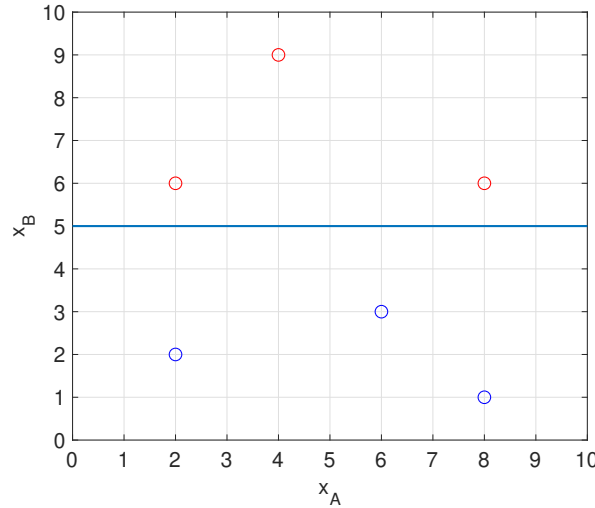


Figure 1: Simple dataset and linear boundary

- A linear boundary is defined by the equation  $\mathbf{x}^T \mathbf{w} = 0$  or equivalently  $w_0 + w_A x_A + w_B x_B = 0$ . The linear boundary is defined by all the points in the attribute space such that  $x_B = 5$ , i.e.  $x_B - 5 = 0$ . Therefore,  $w_0 = -5$ ,  $w_A = 0$  and  $w_B = 1$ , i.e.  $\mathbf{w} = [-5, 0, 1]^T$ . These coefficients are not unique: if they are multiplied by a constant  $k$ , they define the same boundary. For instance,  $\mathbf{w} = [-10, 0, 2]^T$ ,  $\mathbf{w} = [-50, 0, 10]^T$  and  $\mathbf{w} = [25, 0, -5]^T$  define the same boundary.
- The sample  $\mathbf{x}_1$  with predictors  $x_A = 2$  and  $x_B = 5$  lies on the boundary and  $\mathbf{x}_1^T \mathbf{w} = [1, 2, 5][ -5, 0, 1]^T = 1 \times (-5) + 2 \times 0 + 5 \times 1 = -5 + 0 + 5 = 0$ . The sample  $\mathbf{x}_2$  with predictors  $x_A = 8$  and  $x_B = 5$  also lies on the boundary and  $\mathbf{x}_2^T \mathbf{w} = [1, 8, 5][ -5, 0, 1]^T = 1 \times (-5) + 8 \times 0 + 5 \times 1 = -5 + 0 + 5 = 0$ .
- Let's consider the samples belonging to class ○ and compute the quantity  $\mathbf{x}^T \mathbf{w}$  (from left to right):  $[1, 2, 6][ -5, 0, 1]^T = 1 \times (-5) + 2 \times 0 + 6 \times 1 = -5 + 0 + 6 = 1$ ,  $[1, 4, 9][ -5, 0, 1]^T = 1 \times (-5) + 4 \times 0 + 9 \times 1 = -5 + 0 + 9 = 4$  and  $[1, 8, 6][ -5, 0, 1]^T = 1 \times (-5) + 8 \times 0 + 6 \times 1 = -5 + 0 + 6 = 1$ , which numerically are the same as the distances 1, 4 and 1 respectively.

- As for the samples belonging to class  $\circ$  we obtain (from left to right):  $[1, 2, 2] \begin{bmatrix} -5 \\ 0 \\ 1 \end{bmatrix}^T = 1 \times (-5) + 2 \times 0 + 2 \times 1 = -5 + 0 + 2 = -3$ ,  $[1, 3, 6] \begin{bmatrix} -5 \\ 0 \\ 1 \end{bmatrix}^T = 1 \times (-5) + 6 \times 0 + 3 \times 1 = -5 + 0 + 3 = -2$  and  $[1, 8, 1] \begin{bmatrix} -5 \\ 0 \\ 1 \end{bmatrix}^T = 1 \times (-5) + 8 \times 0 + 1 \times 1 = -5 + 0 + 1 = -4$ , and the distances are 3, 2 and 4 respectively. They have the opposite sign.
- Samples such that  $\mathbf{x}^T \mathbf{w} > 0$  should be labeled as  $\circ$ , samples where  $\mathbf{x}^T \mathbf{w} < 0$  should be labeled as  $\circ$ .
- If  $k > 0$ , we would use the same rule, if  $k < 0$  we would change change it as follows: samples such that  $\mathbf{x}^T \mathbf{w} > 0$  should be labeled as  $\circ$ , samples where  $\mathbf{x}^T \mathbf{w} < 0$  should be labeled as  $\circ$ .

**EXERCISE #3 (SOL):** Figure 2 shows a simple dataset in a 2D predictor space with features  $x_A$  and  $x_B$ . The dataset consists of three samples belonging to class  $\circ$  and three samples belonging to class  $\circ$ . The straight line shown in Figure 2 is the boundary of our linear classifier.

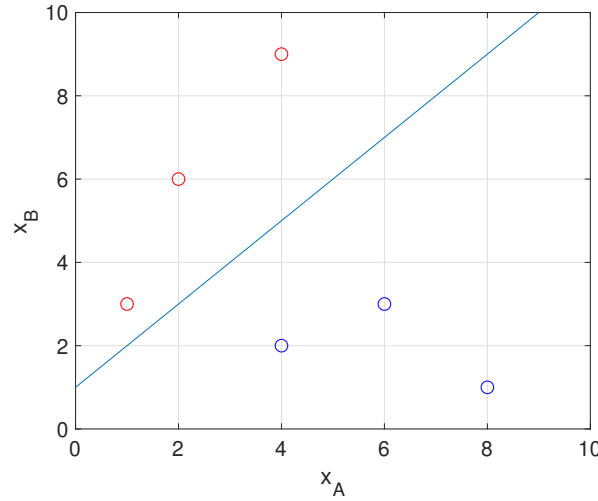


Figure 2: Simple dataset and linear boundary

- The linear boundary is defined by the equation  $x_B = x_A + 1$ , or equivalently  $x_B - x_A - 1 = 0$ . Therefore,  $w_0 = -1$ ,  $w_A = -1$  and  $w_B = 1$ , i.e.  $\mathbf{w} = [-1, -1, 1]^T$ . These coefficients are not unique: if they are multiplied by a constant  $k$ , they define the same boundary.
- The sample  $\mathbf{x}_1$  with predictors  $x_A = 1$  and  $x_B = 2$  lies on the boundary and  $\mathbf{x}_1^T \mathbf{w} = [1, 1, 2] \begin{bmatrix} -1 \\ -1 \\ 1 \end{bmatrix}^T = 1 \times (-1) + 1 \times (-1) + 1 \times 2 = -1 - 1 + 2 = 0$ . The sample  $\mathbf{x}_2$  with predictors  $x_A = 6$  and  $x_B = 7$  also lies on the boundary and  $\mathbf{x}_2^T \mathbf{w} = [1, 6, 7] \begin{bmatrix} -1 \\ -1 \\ 1 \end{bmatrix}^T = 1 \times (-1) + 6 \times (-1) + 7 \times 1 = -1 - 6 + 7 = 0$ .
- Let's consider the samples belonging to class  $\circ$  and compute the quantity  $\mathbf{x}^T \mathbf{w}$  (from left to right):  $[1, 1, 3] \begin{bmatrix} -1 \\ -1 \\ 1 \end{bmatrix}^T = 1 \times (-1) + 1 \times (-1) + 3 \times 1 = -1 - 1 + 3 = 1$ ,  $[1, 2, 6] \begin{bmatrix} -1 \\ -1 \\ 1 \end{bmatrix}^T = 1 \times (-1) + 2 \times (-1) + 6 \times 1 = -1 - 2 + 6 = 3$  and  $[1, 4, 9] \begin{bmatrix} -1 \\ -1 \\ 1 \end{bmatrix}^T = 1 \times (-1) + 4 \times (-1) + 9 \times 1 = -1 - 4 + 9 = 4$ .
- As for the samples belonging to class  $\circ$  we obtain (from left to right):  $[1, 4, 2] \begin{bmatrix} -1 \\ -1 \\ 1 \end{bmatrix}^T = 1 \times (-1) + 4 \times (-1) + 2 \times 1 = -1 - 4 + 2 = -3$ ,  $[1, 6, 3] \begin{bmatrix} -1 \\ -1 \\ 1 \end{bmatrix}^T = 1 \times (-1) + 6 \times (-1) + 3 \times 1 = -1 - 6 + 3 = -4$  and  $[1, 8, 1] \begin{bmatrix} -1 \\ -1 \\ 1 \end{bmatrix}^T = 1 \times (-1) + 8 \times (-1) + 1 \times 1 = -1 - 8 + 1 = -8$ .
- Samples such that  $\mathbf{x}^T \mathbf{w} > 0$  should be labeled as  $\circ$ , samples where  $\mathbf{x}^T \mathbf{w} < 0$  should be labeled as  $\circ$ .

- If  $k > 0$ , we would use the same rule, if  $k < 0$  we would change change it as follows: samples such that  $\mathbf{x}^T \mathbf{w} > 0$  should be labeled as  $\bigcirc$ , samples where  $\mathbf{x}^T \mathbf{w} < 0$  should be labeled as  $\bigcirc$ .

**EXERCISE #4 (SOL):** Figure 3 shows four samples belonging to a dataset with predictors  $x_A$ ,  $x_B$  and  $x_C$ . The plane represents a linear boundary in a 3D predictor space.

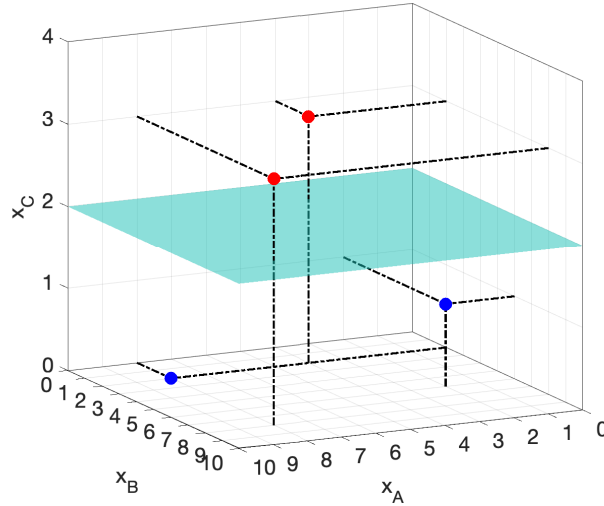


Figure 3: Simple dataset and linear boundary

- A linear boundary is defined by the equation  $\mathbf{x}^T \mathbf{w} = [1, x_A, x_B, x_C][w_0, w_A, w_B, w_C]^T = 0$  or equivalently  $w_0 + w_A x_A + w_B x_B + w_C x_C = 0$ . The linear boundary is defined by the equation  $x_C = 2$ , or equivalently  $x_C - 2 = 0$ . Therefore,  $w_0 = -2$ ,  $w_A = 0$ ,  $w_B = 0$  and  $w_C = 1$ , or  $\mathbf{w} = [-2, 0, 0, 1]^T$ . These coefficients are not unique: if they are multiplied by a constant  $k$ , they define the same boundary.
- The sample  $\mathbf{x}_1$  with predictors  $x_A = 1$ ,  $x_B = 2$  and  $x_C = 2$  lies on the boundary and  $\mathbf{x}_1^T \mathbf{w} = [1, 1, 2, 2][-2, 0, 0, 1]^T = 1 \times (-2) + 1 \times 0 + 2 \times 0 + 2 \times 1 = -2 + 0 + 0 + 2 = 0$ . The sample  $\mathbf{x}_2$  with predictors  $x_A = 6$ ,  $x_B = 7$  and  $x_C = 2$  also lies on the boundary and  $\mathbf{x}_2^T \mathbf{w} = [1, 6, 7, 2][-2, 0, 0, 1]^T = 1 \times (-2) + 6 \times 0 + 7 \times 0 + 2 \times 1 = -2 + 0 + 0 + 2 = 0$ .
- Let's consider the samples belonging to class  $\bigcirc$  and compute the quantity  $\mathbf{x}^T \mathbf{w}$  (from left to right):  $[1, 8, 8, 3][-2, 0, 0, 1]^T = 1 \times (-2) + 8 \times 0 + 8 \times 0 + 3 \times 1 = -2 + 0 + 0 + 3 = 1$  and  $[1, 4, 2, 3][-2, 0, 0, 1]^T = 1 \times (-2) + 4 \times 0 + 2 \times 0 + 3 \times 1 = -2 + 0 + 0 + 3 = 1$ .
- As for the samples belonging to class  $\bigcirc$  we obtain (from left to right):  $[1, 8, 2, 0][-2, 0, 0, 1]^T = 1 \times (-2) + 8 \times 0 + 2 \times 0 + 0 \times 1 = -2 + 0 + 0 + 0 = -2$  and  $[1, 2, 6, 1][-2, 0, 0, 1]^T = 1 \times (-2) + 2 \times 0 + 6 \times 0 + 1 \times 1 = -2 + 0 + 0 + 1 = -1$ .
- Samples such that  $\mathbf{x}^T \mathbf{w} > 0$  should be labeled as  $\bigcirc$ , samples where  $\mathbf{x}^T \mathbf{w} < 0$  should be labeled as  $\bigcirc$ .
- If  $k > 0$ , we would use the same rule, if  $k < 0$  we would change change it as follows: samples such that  $\mathbf{x}^T \mathbf{w} > 0$  should be labeled as  $\bigcirc$ , samples where  $\mathbf{x}^T \mathbf{w} < 0$  should be labeled as  $\bigcirc$ .

**EXERCISE #5 (SOL):**

- If  $x_i$  lies on the boundary, by definition  $x_i^T w = 0$  and therefore  $e^{w^T x_i} = 1$ . Hence,  $p(x_i) = 1/(1+1) = 0.5$ .
- As we move away from the boundary on the positive side,  $x_i^T w \rightarrow \infty$  and  $e^{w^T x_i} \rightarrow \infty$  and  $p(x_i) \rightarrow 1$ . On the negative side,  $x_i^T w \rightarrow -\infty$  and  $e^{w^T x_i} \rightarrow 0$  and  $p(x_i) \rightarrow 0/(1+0) = 0$ . The quantity  $p(x_i)$  can be seen as the certainty of the classifier that the sample belongs to the label associated with the positive region.
- The likelihood  $L(w)$  of the classifier defined in Exercise 2 on the dataset shown in Figure 1 is:

$$\begin{aligned} L(w) &= \frac{e^1}{1+e^1} \frac{e^4}{1+e^4} \frac{e^1}{1+e^1} \left(1 - \frac{e^{-3}}{1+e^{-3}}\right) \left(1 - \frac{e^{-2}}{1+e^{-2}}\right) \left(1 - \frac{e^{-4}}{1+e^{-4}}\right) \\ &= 0.73 \times 0.98 \times 0.73 \times 0.95 \times 0.88 \times 0.98 = 0.43 \end{aligned}$$

- The distances to the boundary are now 2, 5 and 2 (○ class) and -2, -1 and -3 (○ class). The new likelihood is:

$$\begin{aligned} L(w') &= \frac{e^2}{1+e^2} \frac{e^5}{1+e^5} \frac{e^2}{1+e^2} \left(1 - \frac{e^{-2}}{1+e^{-2}}\right) \left(1 - \frac{e^{-1}}{1+e^{-1}}\right) \left(1 - \frac{e^{-3}}{1+e^{-3}}\right) \\ &= 0.88 \times 0.99 \times 0.88 \times 0.88 \times 0.73 \times 0.95 = 0.47 \end{aligned}$$

- If we use the likelihood as our metric to rank classifiers, the second classifier  $w'$  has a higher certainty and would be preferred..

**EXERCISE #6 (SOL):** Figure 4 shows a dataset consisting of samples belonging to classes ● and ● in a predictor space with attributes  $x_A$  and  $x_B$ .

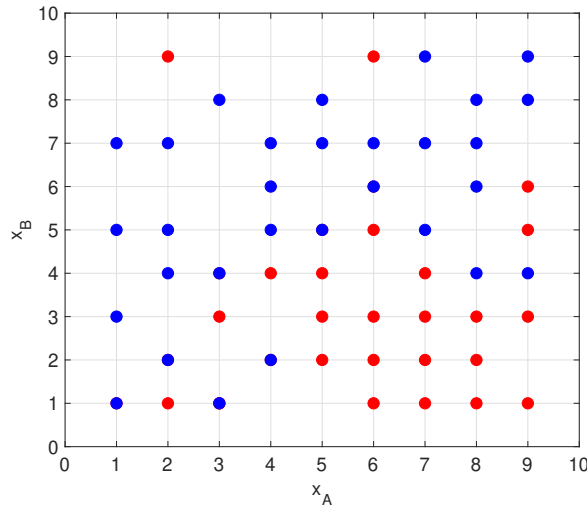


Figure 4

- As  $k$  increases, the complexity of the boundary decreases. Since there are 53 samples, 30 ● samples and 23 ● samples, for  $k = 53$  kNN would label every samples as ●.
- There would be many more locations in the predictor space where we cannot decide how to classify a sample, as 50% of the neighbours would belong to either class. In other words, the decision boundary would increase in size.