Principles of Machine Learning

Classification I

Academic year 2021/2022

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 (o and o, 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 $\boldsymbol{w}^T\boldsymbol{x}_i > 0$ and the value -1 to samples in the decision region defined by $\boldsymbol{w}^T\boldsymbol{x}_i < 0$. A correctly classified sample will receive a label that has the same sign as the quantity $\boldsymbol{w}^T\boldsymbol{x}_i$, whereas misclassified samples will receive a label with the opposite sign. Therefore, the margin $m_i = y_i\left[\boldsymbol{w}^T\boldsymbol{x}_i\right]$ will be positive if $\boldsymbol{x}_i$ is correctly classified, as the true label y_i and $\boldsymbol{w}^T\boldsymbol{x}_i$ have the same sign. If it is misclassified, the margin will be negative, as the true label y_i and the quantity $\boldsymbol{w}^T\boldsymbol{x}_i$ have opposite signs.

EXERCISE ♯2 (SOL): Figure **??** 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.

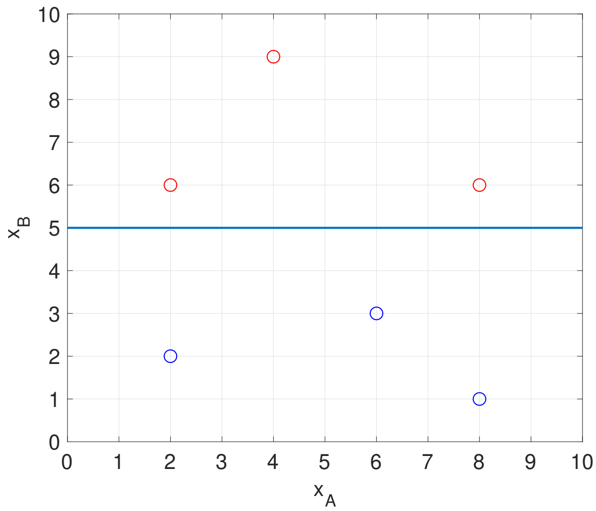

Figure 1: Simple dataset and linear boundary

- A linear boundary is defined by the equation $\boldsymbol{x}^T\boldsymbol{w} = 0$ 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. $\boldsymbol{w} = [-5, 0, 1]^T$. These coefficients are not unique: if they are multiplied by a constant k, they define the same boundary. For instance, $\boldsymbol{w} = [-10, 0, 2]^T$, $\boldsymbol{w} = [-50, 0, 10]^T$ and $\boldsymbol{w} = [25, 0, -5]^T$ define the same boundary.

- The sample $\boldsymbol{x}_1$ with predictors $x_A = 2$ and $x_B = 5$ lies on the boundary and $\boldsymbol{x}_1^T\boldsymbol{w} = [1, 2, 5][-5, 0, 1]^T = 1 \times (-5) + 2 \times 0 + 5 \times 1 = -5 + 0 + 5 = 0$. The sample $\boldsymbol{x}_1$ with predictors $x_A = 8$ and $x_B = 5$ also lies on the boundary and $\boldsymbol{x}_2^T\boldsymbol{w} = [1, 8, 5][-5, 0, 1]^T = 1 \times (-5) + 8 \times 0 + 5 \times 1 = -5 + 0 + 5 = 0$.

- Let's consider the samples belonging to class ◯ and compute the quantity $\boldsymbol{x}^T\boldsymbol{w}$ (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 ○ we obtain (from left to right): $[1,2,2][-5,0,1]^T = 1\times(-5)+2\times 0+2\times 1 = -5+0+2 = -3$, $[1,3,6][-5,0,1]^T = 1\times(-5)+6\times 0+3\times 1 = -5+0+3 = -2$ and $[1,8,1][-5,0,1]^T = 1\times(-5)+8\times 0+1\times 1 = -5+0+1 = -4$, and the distances are 3, 2 and 4 respectively. They have the opposite sign.

- Samples such that $\boldsymbol{x}^T\boldsymbol{w} > 0$ should be labeled as ○, samples where $\boldsymbol{x}^T\boldsymbol{w} < 0$ should be labeled as ○.

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

EXERCISE ♯3 (SOL): Figure **??** 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 ○ and three samples belonging to class ○. The straight line shown in Figure **??** is the boundary of our linear classifier.

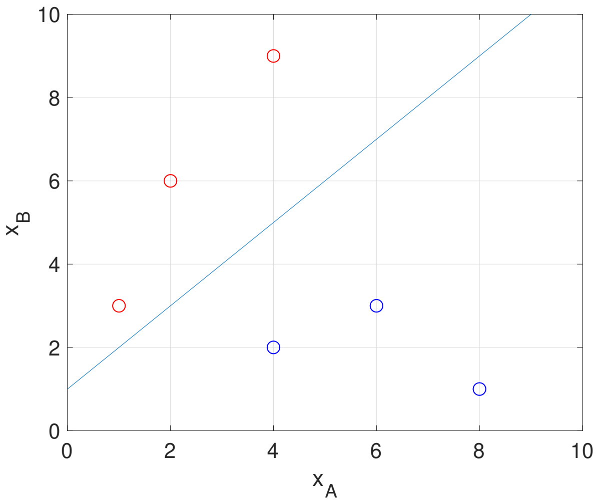

Figure 2: Simple dataset and linear boundary

- 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. $\boldsymbol{w} = [-1,-1,1]^T$. These coefficients are not unique: if they are multiplied by a constant k, they define the same boundary.

- The sample $\boldsymbol{x}_1$ with predictors $x_A = 1$ and $x_B = 2$ lies on the boundary and $\boldsymbol{x}_1^T\boldsymbol{w} = [1,1,2][-1,-1,1]^T = 1\times(-1)+1\times(-1)+1\times 2 = -1-1+2 = 0$. The sample $\boldsymbol{x}_1$ with predictors $x_A = 6$ and $x_B = 7$ also lies on the boundary and $\boldsymbol{x}_2^T\boldsymbol{w} = [1,6,7][-1,-1,1]^T = 1\times(-1)+6\times(-1)+7\times 1 = -1-6+7 = 0$.

- Let's consider the samples belonging to class ○ and compute the quantity $\boldsymbol{x}^T\boldsymbol{w}$ (from left to right): $[1,1,3][-1,-1,1]^T = 1\times(-1)+1\times(-1)+3\times 1 = -1-1+3 = 1$, $[1,2,6][-1,-1,1]^T = 1\times(-1)+2\times(-1)+6\times 1 = -1-2+6 = 3$ and $[1,4,9][-1,-1,1]^T = 1\times(-1)+4\times(-1)+9\times 1 = -1-4+9 = 4$.

- As for the samples belonging to class ○ we obtain (from left to right): $[1,4,2][-1,-1,1]^T = 1\times(-1)+4\times(-1)+2\times 1 = -1-4+2 = -3$, $[1,6,3][-1,-1,1]^T = 1\times(-1)+6\times(-1)+3\times 1 = -1-6+3 = -4$ and $[1,8,1][-1,-1,1]^T = 1\times(-1)+8\times(-1)+1\times 1 = -1-8+1 = -8$.

- Samples such that $\boldsymbol{x}^T\boldsymbol{w} > 0$ should be labeled as ○, samples where $\boldsymbol{x}^T\boldsymbol{w} < 0$ should be labeled as ○.

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

EXERCISE ♯4 (SOL): Figure **??** 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.

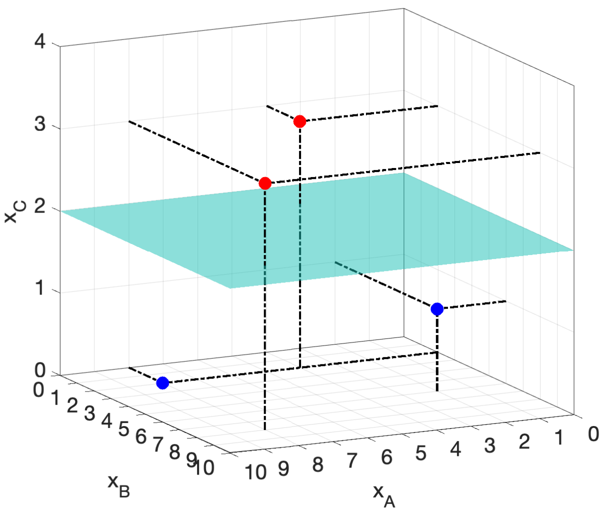

Figure 3: Simple dataset and linear boundary

- A linear boundary is defined by the equation $\boldsymbol{x}^T\boldsymbol{w} = [1, x_A, x_B, x_C][w_0, w_A, w_B, w_C]^T = 0$ 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 $\boldsymbol{w} = [-2, 0, 0, 1]^T$. These coefficients are not unique: if they are multiplied by a constant k, they define the same boundary.

- The sample $\boldsymbol{x}_1$ with predictors $x_A = 1$, $x_B = 2$ and $x_C = 2$ lies on the boundary and $\boldsymbol{x}_1^T\boldsymbol{w} = [1, 1, 2, 2][-2, 0, 0, 1]^T = 1 \times (-2) + 1 \times 0 + 2 \times 0 + 2 \times 1 = -2 + 0 + 0 + 2 = 0$. The sample $\boldsymbol{x}_2$ with predictors $x_A = 6$, $x_B = 7$ and $x_C = 2$ also lies on the boundary and $\boldsymbol{x}_2^T\boldsymbol{w} = [1, 6, 7, 2][-2, 0, 0, 1]^T = 1 \times (-2) + 6 \times 0 + 7 \times 0 + 2 \times 1 = -2 + 0 + 0 + 2 = 0$.

- Let's consider the samples belonging to class ◯ and compute the quantity $\boldsymbol{x}^T\boldsymbol{w}$ (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 $[1, 4, 2, 3][-2, 0, 0, 1]^T = 1 \times (-2) + 4 \times 0 + 2 \times 0 + 2 \times 1 = -2 + 0 + 0 + 3 = 1$.

- As for the samples belonging to class ◯ 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 $\boldsymbol{x}^T\boldsymbol{w} > 0$ should be labeled as ◯, samples where $\boldsymbol{x}^T\boldsymbol{w} < 0$ should be labeled as ◯.

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

EXERCISE ♯5 (SOL):

- If $\boldsymbol{x}_i$ lies on the boundary, by definition $\boldsymbol{x}_i^T\boldsymbol{w} = 0$ and therefore $e^{\boldsymbol{w}^T\boldsymbol{x}_i} = 1$. Hence, $p(\boldsymbol{x}_i) = 1/(1+1) = 0.5$.

- As we move away from the boundary on the positive side, $\boldsymbol{x}_i^T\boldsymbol{w} \to \infty$ and $e^{\boldsymbol{w}^T\boldsymbol{x}_i} \to \infty$ and $p(\boldsymbol{x}_i) \to 1$. On the negative side, $\boldsymbol{x}_i^T\boldsymbol{w} \to -\infty$ and $e^{\boldsymbol{w}^T\boldsymbol{x}_i} \to 0$ and $p(\boldsymbol{x}_i) \to 0/(1+0) = 0$. The quantity $p(\boldsymbol{x}_i)$ can be seen as the certainty of the classifier that the sample belongs to the label associated with the positive region.

- The likelihood $L(\boldsymbol{w})$ of the classifier defined in Exercise 2 on the dataset shown in Figure 1 is:

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

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

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

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

EXERCISE ♯6 (SOL): Figure **??** shows a dataset consisting of samples belonging to classes • and • in a predictor space with attributes x_A and x_B.

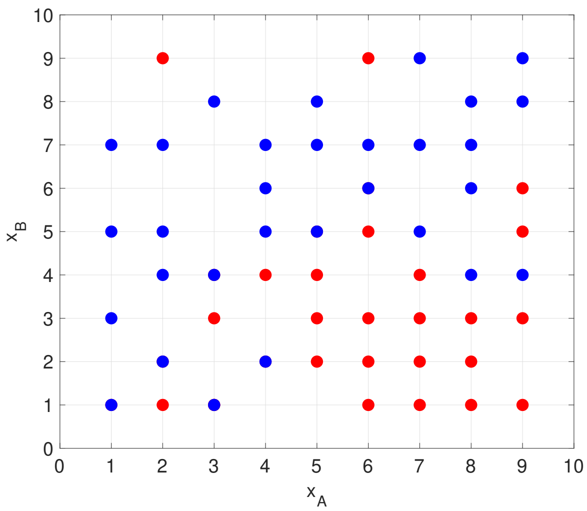

Figure 4

- 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.