DATA MINING

CLASSIFICATION I

ACADEMIC YEAR 2019/2020

QUEEN MARY UNIVERSITY OF LONDON

EXERCISES

EXERCISE ♯1. Consider the simple dataset shown in Figure 1, consisting of three samples belonging to the class ◯ and three samples belonging to the class ◯ in a 2D predictor space with features x_A and x_B.

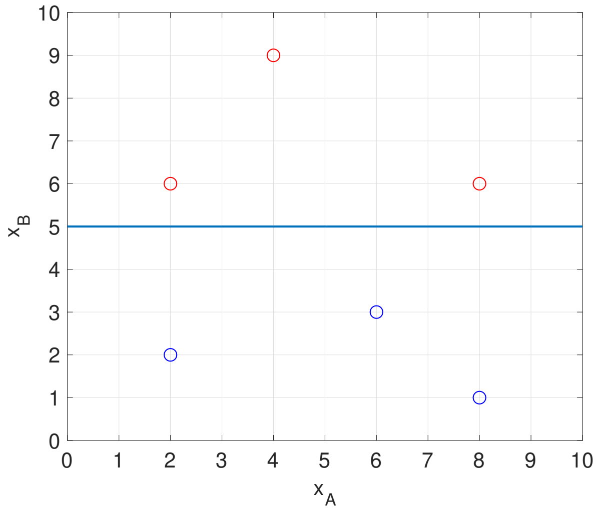

Figure 1: Simple dataset and linear boundary

Assume that we use a classifier whose boundary is the straight line shown in Figure 1.

- Find the coefficients $\boldsymbol{w}$ for the equation $\boldsymbol{w}^T\boldsymbol{x} = 0$ representing the linear boundary of the classifier, where $\boldsymbol{x} = [1, x_A, x_B]$ is the extended predictor vector. Are these coefficients unique?
- Select two samples $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ on the classifier boundary and show that $\boldsymbol{w}^T\boldsymbol{x}_1 = 0$ and $\boldsymbol{w}^T\boldsymbol{x}_2 = 0$.
- For every sample $\boldsymbol{x}_i$ belonging to the class ◯, compute the quantity $\boldsymbol{w}^T\boldsymbol{x}_i$ and compare its value with the distance from the sample $\boldsymbol{x}_i$ to the boundary.
- Carry the previous comparison for every sample $\boldsymbol{x}_i$ belonging to the class ◯.
- Given an arbitrary sample $\boldsymbol{x}$, how would our classifier use the result of the computation $\boldsymbol{w}^T\boldsymbol{x} = 0$ to classify it?
- Define a new classifier by a linear boundary with coefficients $\boldsymbol{w}' = k\boldsymbol{w}$, where k is an arbitrary constant. How would this classifier compare with one defined by $\boldsymbol{w}$?

EXERCISE ♯2. 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 the class ◯ and three samples belonging to the class ◯. Using the straight line shown in Figure 2 as the boundary of our linear classifier, repeat the steps described in Exercise 1 for this new scenario.

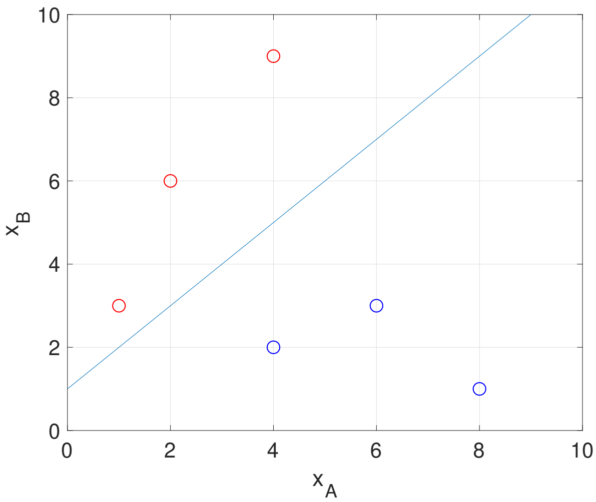

Figure 2: Simple dataset and linear boundary

EXERCISE ♯3. Figure 3 shows four samples belonging to a dataset with predictors x_A, x_B and x_C. As you can see, two samples belong to the class • and the other two samples to the class •.

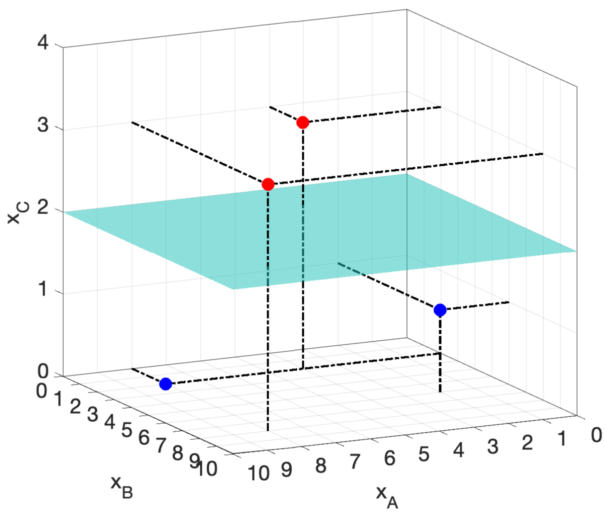

Figure 3: Simple dataset and linear boundary

Consider the linear classifier represented by the surface shown in Figure 3 and repeat all the steps described in Exercise 1 for this new scenario. Note that the extended vector $\boldsymbol{x}$ should now be defined as $\boldsymbol{x} = [1, x_A, x_B, x_C]$, and the coefficient vector describing the classifier's linear boundary will need to be redefined accordingly.

EXERCISE ♯4. Consider a linear classifier defined by the coefficients vector $\boldsymbol{w}$, where samples $\boldsymbol{x}_i$ such that $\boldsymbol{w}^T\boldsymbol{x}_i \geq 0$ are labelled as ◯ (otherwise, they are labelled as ◯). A convenient way to quantify the linear classifier's certainty that a sample $\boldsymbol{x}_i$ belongs to the class ◯ is to use the logistic function and compute the value:

$$p(\boldsymbol{x}_i; \boldsymbol{w}) = \frac{e^{\boldsymbol{w}^T\boldsymbol{x}_i}}{1 + e^{\boldsymbol{w}^T\boldsymbol{x}_i}}$$

- Show that $p(\boldsymbol{x}_i; \boldsymbol{w})$ is 0.5 for samples on the boundary and that as samples move away from the boundary, either $p(\boldsymbol{x}_i; \boldsymbol{w}) \to 0$ or $p(\boldsymbol{x}_i; \boldsymbol{w}) \to 1$. How would you interpret these numerical values?
- Obtain the likelihood $L(\boldsymbol{w})$ of the classifier with coefficients $\boldsymbol{w}$ defined in Figure 1 for the dataset shown.
- Create a new classifier by moving the boundary of the previous classifier one unit of x_B down, i.e. the new classifier is defined by the boundary $x_B = 4$. Obtain the likelihood $L(\boldsymbol{w}')$ of the new classifier, where $\boldsymbol{w}'$ are the new coefficients.
- Obtain the likelihood of the classifier defined by $\boldsymbol{w}'$ for the dataset shown in Figure 2.

EXERCISE ♯5. Figure 4 shows a dataset consisting of samples belonging to classes • and • in a predictor space with features x_A and x_B.

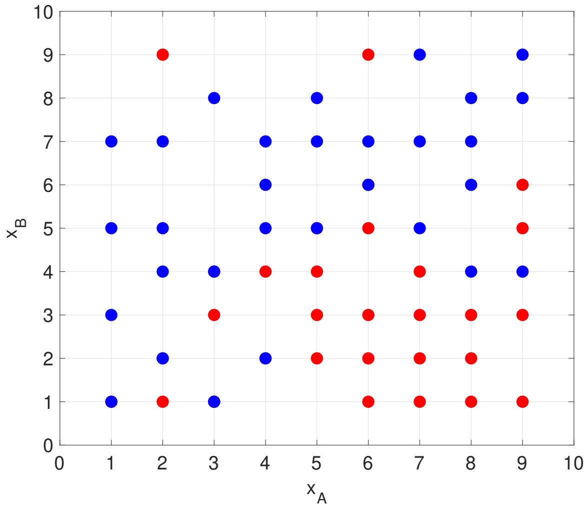

Figure 4

- Sketch the boundaries of three kNN classifiers, where $k = 1, 3$ and 7 respectively. How does the boundary change as k increases? What would the boundary be for $k = 53$?
- Sketch the boundary of a kNN classifier, where $k = 2$ and 4. Why is this choice problematic?

EXERCISE ♯6. Given the dataset shown in Figure 4, obtain the confusion matrix of the classifiers defined by the boundaries $x_B = 0.5$, $x_B = 1.5$, $x_B = 3.5$, $x_B = 5.5$, $x_B = 7.5$ and $x_B = 9.5$. Use the resulting rates to sketch the ROC curve of the family of classifier $x_B = c$, where c is the calibration parameter.

EXERCISE ♯7. Repeat the previous exercise for the family of classifiers defined by $x_B = x_A + c$, where c is the calibration parameter. Obtain the confusion matrix for the boundaries defined by the values $c = -8.5, -4.5, -1.5, 1.5, 4.5, 8.5$ and compare the estimated ROC curve with the ROC curve obtained in the previous exercise. Which family of classifiers represent better the distribution of data?

EXERCISE ♯8. Consider the following dataset, where x will be used as a predictor and y as a label:

x	y
-2	A
-1	A
0	A
1	A
2	A
1	B
2	B
3	B
4	B
5	B

We will assume that the value of x is distributed following a Gaussian distribution for both classes A and B.

- Estimate the parameters of the Gaussian distributions (likelihood of x).
- Build a Bayes classifier for the previous dataset.
- Build a new Bayes classifier with the same likelihoods but different priors, namely $P_A = 0.1$ and $P_B = 0.9$.

EXERCISE ♯9. Figure 5 shows a dataset consisting of samples belonging to three classes •, • and • in a predictor space with features x_A and x_B.

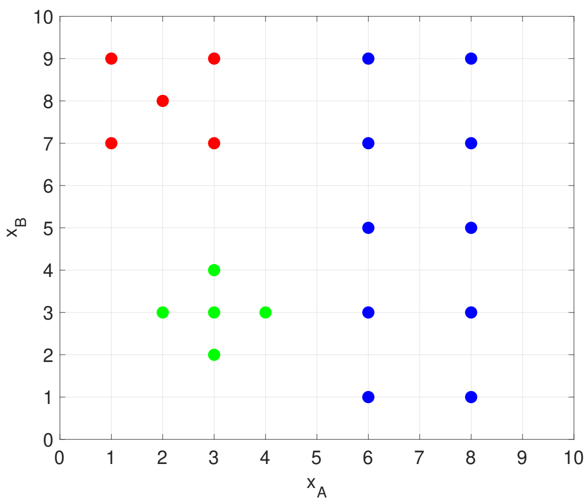

Figure 5

Assuming Gaussian and independent (*naive*) likelihoods for x_A and x_B, define a Bayes classifier and sketch its boundaries.

EXERCISE ♯10. Create a classification tree for each of the datasets shown in Figure 6. Clearly identify and describe the metric that you are using every time you partition a classification region and the stop criterion that you have considered.

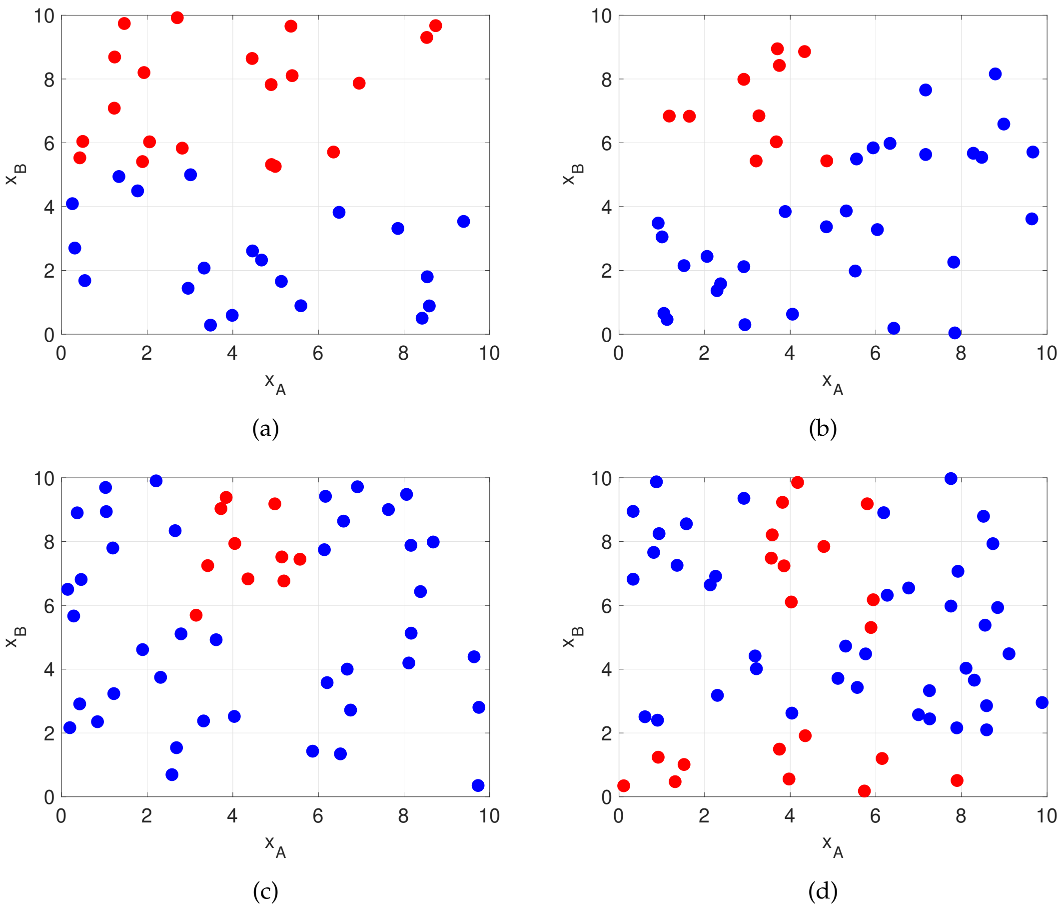

Figure 6