Faculté des Sciences appliquées
ELEN062-1 : Introduction to Machine Learning

Classification Algorithms

Teachers :
Louis Wehenkel and Pierre Geurts

Group :
LAMBERMONT Romain
LOUIS Arthur

23 octobre 2022

1 Decision trees

1.1 Impact of the tree's depth on the decision boundary

a) The lower the depth of the tree, the lower the decision boundary will be. Indeed, when the depth is equal to 1, the boundary is a simple line and as the depth increases, the boundary becomes more complex, fitting the data better. This phenomenon can be seen in the following figure 1.

b) On one hand, when the depth is 1, the separation between areas is a simple line. The model is too complex to be fit in this way so the depth 1 is clearly underfitting the problem. For the depth 2, the model is more complex and fits the data better but it is still underfitting the problem. On the other hand, when the depth is 8, the model is too complex and the decision boundary is overfitting the problem. When the hyperparameter `depth` is set to `None`, the nodes are expanded until all leaves are pure or until all leaves contain less than `min_samples_split` samples, which is the minimum number of samples required to split an internal node. The best depth is 4 because it is the one that fits the problem the best. As demonstrated in the next point, the depth 4 is the one that gives the best accuracy.

c) As previously said, when the depth becomes large, the model overfits the data, leading to a better confidence on the training set but a lower confidence on the test set.

1.2 Tests accuracy report

We report the results obtained for the different depths in the following table. The models were tested on 5 iterations of the dataset. We can clearly see that the best depth is 4, as it gives the best accuracy while keeping the standard deviation low. When the depth gets past 4, the model overfits the data leading to decreased accuracy and increased standard deviation (the model is fitting the noise in the data).

`max_depth`	Mean	Standard Deviation
1	0.68	0.031
2	0.80	0.021
4	0.83	0.014
8	0.81	0.016
None	0.77	0.02

TABLE 1 – Decision trees results

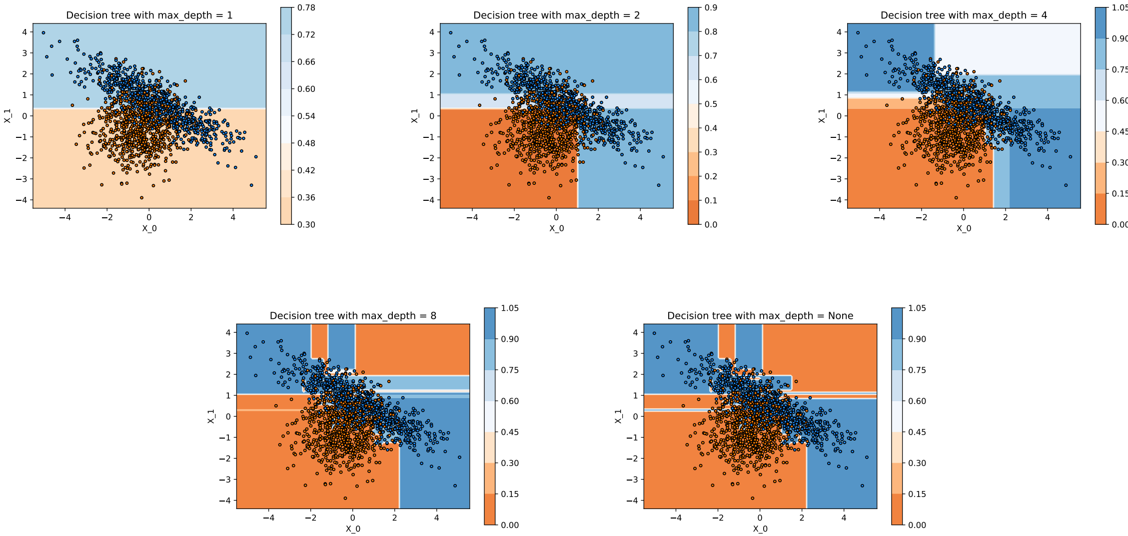

FIGURE 1 – Decision trees with different depths

2 K-nearest neighbors

2.1 Impact of the number of neighbors on the decision boundary

The different boundaries are represented in figure 2. The lower the number of neighbors, the more the decision boundary is affected by the noise in the data (it starts to overfit). When the number of neighbors increases, e.g for 5, 25 and 125, there's less overfitting and the decision boundary is more stable. When the number of neighbors gets even greater, e.g 625, the model starts to underfit the data, as it is not complex enough to fit the data. And in the worst case, when the number of neighbors is equal to the number of samples, the model simply makes the same prediction for every sample, which is the mean of the training set, giving the worst accuracy possible.

2.2 Tests accuracy report

`n_neighbors`	Mean	Standard Deviation
1	0.78	0.024
5	0.81	0.013
25	0.84	0.013
125	0.85	0.021
625	0.81	0.010
1200	0.48	0.014

Table 2 – Decision trees results

Reporting to the table, we can see that the best number of neighbors is 25. Indeed, it gives good accuracy while keeping the standard deviation low. When the number of neighbors is too low, the model overfits the data and when it is too high, the model underfits the data.

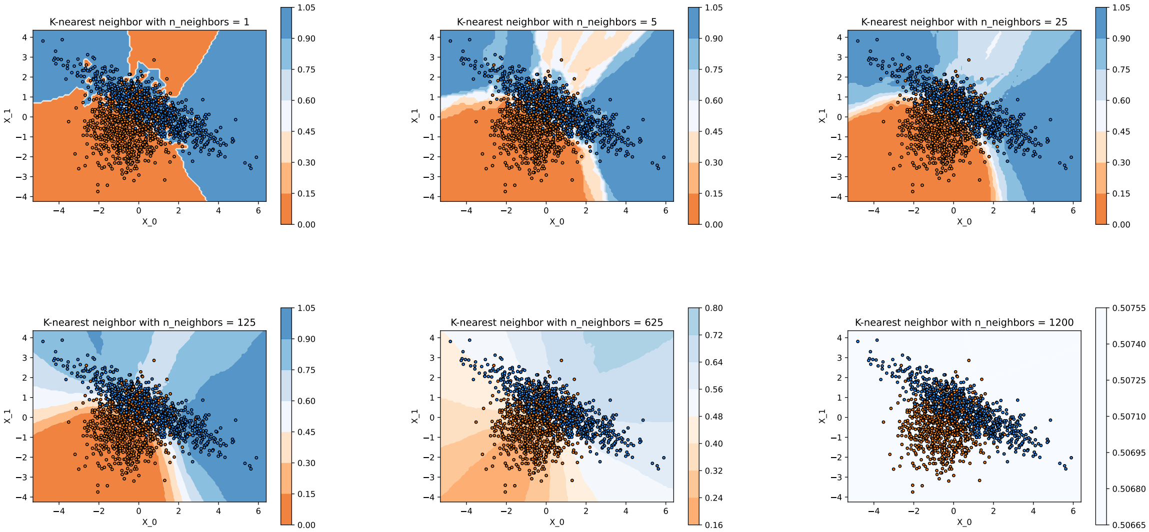

FIGURE 2 – K-nearest neighbors with different n_neighbors values

3 Quadratic/Linear discriminant analysis

3.1 Decision boundaries

In the 2 models, the decision boundaries is the curve where :

$$P(y=1|x) = P(y=0|x) \tag{1}$$

$$\frac{f_0(x)\pi_0}{\sum_{l=0}^{1} f_l(x)\pi_l} = \frac{f_1(x)\pi_1}{\sum_{l=0}^{1} f_l(x)\pi_l} \tag{2}$$

$$\tag{3}$$

With :

$$f_k(x) = \frac{1}{2\pi\sqrt{|\Sigma_k|}} \exp\left(-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)\right) \tag{4}$$

Dividing both sides by $P(y=1|x)$ and taking the logarithm, we get the following equation :

$$\log\frac{P(y=0|x)}{P(y=1|x)} = \log\frac{f_0(x)\pi_0}{f_1(x)\pi_1} = 0 \tag{5}$$

$$\log\frac{f_0(x)}{f_1(x)} + \log\frac{\pi_0}{\pi_1} = 0 \tag{6}$$

$$-\frac{1}{2}(\mu_0+\mu_1)^T\Sigma^{-1}(\mu_0-\mu_1) + x^T\Sigma^{-1}(\mu_0+\mu_1) + \log\frac{\pi_0}{\pi_1} = 0 \tag{7}$$

3.2

3.2.1 Quadratic/Linear discriminant analysis

As we can see, LDA is a special case of QDA where $\Sigma_0 = \Sigma_1 = \Sigma$. This is why the decision boundary is linear in LDA and quadratic in QDA. It is also the reason why the curve of iso probbility is parallel line in LDA. We can also see that QDA is much more confident in general than LDA.

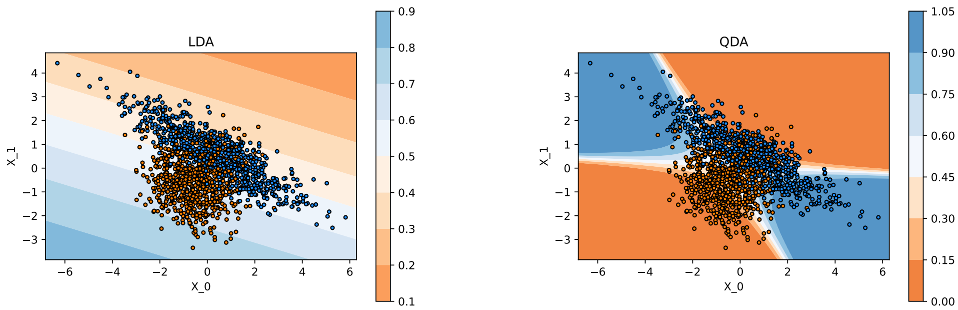

FIGURE 3 – Quadratic/Linear discriminant analysis

3.2.2 Tests accuracy report

`LDA QDA`	Mean	Standard Deviation
LDA1	0.85	0.068
QDA1	0.86	0.057
LDA2	0.72	0.053
QDA2	0.76	0.062

TABLE 3 – Decision trees results

We can see that QDA is better than LDA in general. But the results are really close for dataset 1 because this dataset is create thanks to a circular gaussian distribution. So the matrix of covariance is diagonal and the decision boundary is linear.

4 Comparison

4.1 Tune hyperparameters

To find the best hyperparameters, we tried different values for the hyperparameters and we kept the best ones. The best hyperparameters are the ones that give the best accuracy while keeping the standard deviation low and have a good time of execution so the lower depth in decision trees and the less neighbors in KNN. To sum up, to find the best hyperparameters, we look at 3 criterias : accuracy, standard deviation and time of execution.