Data Mining - Homework 2

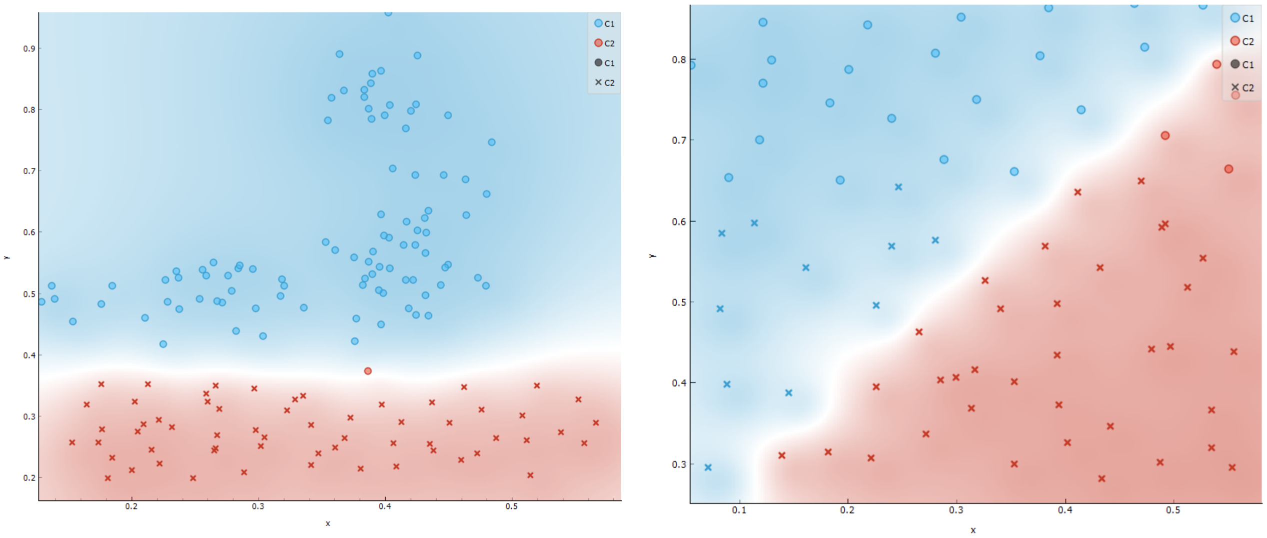
Dmitry Donetskov (ddonetskov@gmail.com)

April 27, 2018

1 First Part

1.1 Classification Tree of the Depth 1 ("Stump")

The classifier has difficulty with the data on Figure 1b because it can split the area by only x or y i.e. by only vertical or horizontal lines whereas the data requires a line based on both x and y.



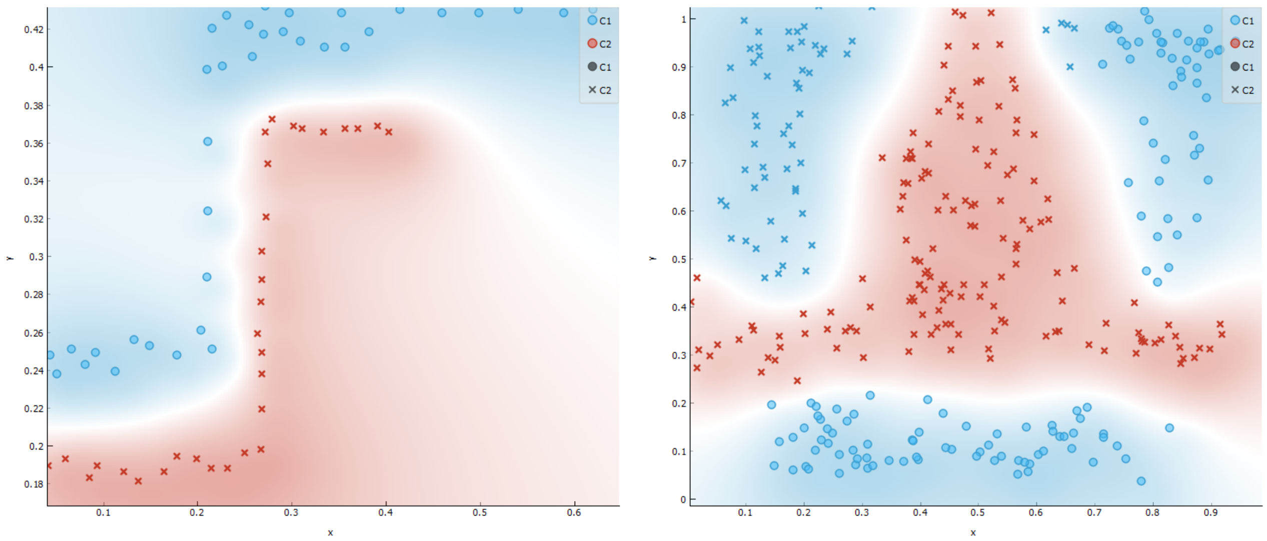
(a) success: AUC = 0.993

(b) failure: AUC = 0.633

Figure 1: Classification Tree of the Depth 1 ("Stump")

1.2 Classification Tree of the Depth 3

The classifier has difficulty with correctly classifying the upper left corner's data on Figure 2b because it can split the area by only three horizontal/vertical cuts whereas the data requires more cuts.



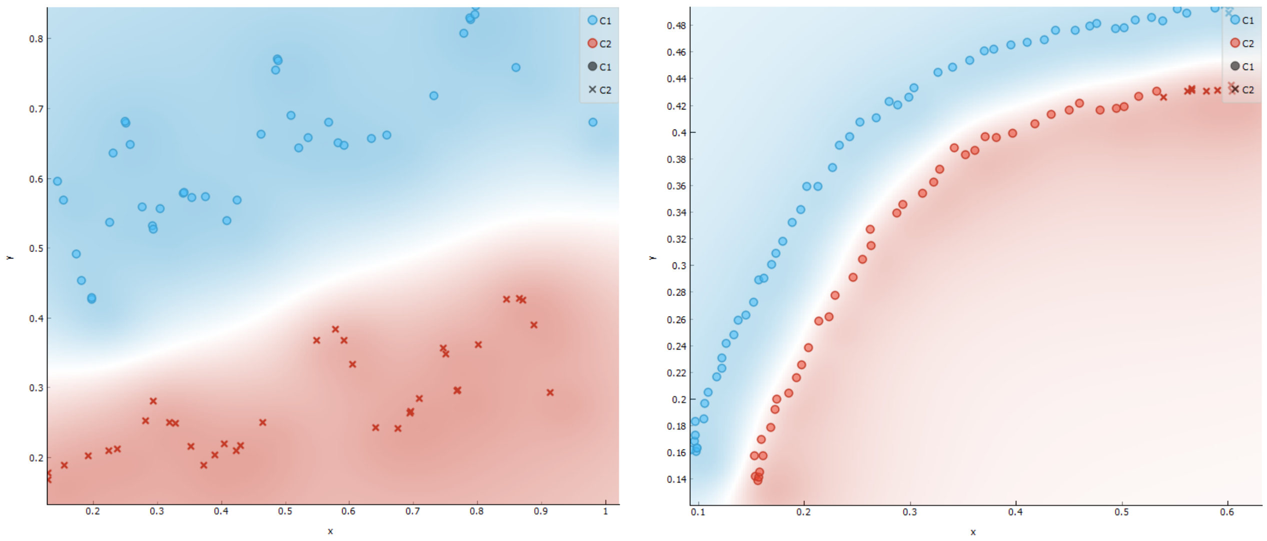
(a) success: AUC = 0.954

(b) failure: AUC = 0.782

Figure 2: Classification Tree of the Depth 3

1.3 Logistic Regression

The classifier is based on the linear combinations of x and y which means it has difficulties with separating data if it's not possible to separate it with a straight line: Figure 3b.



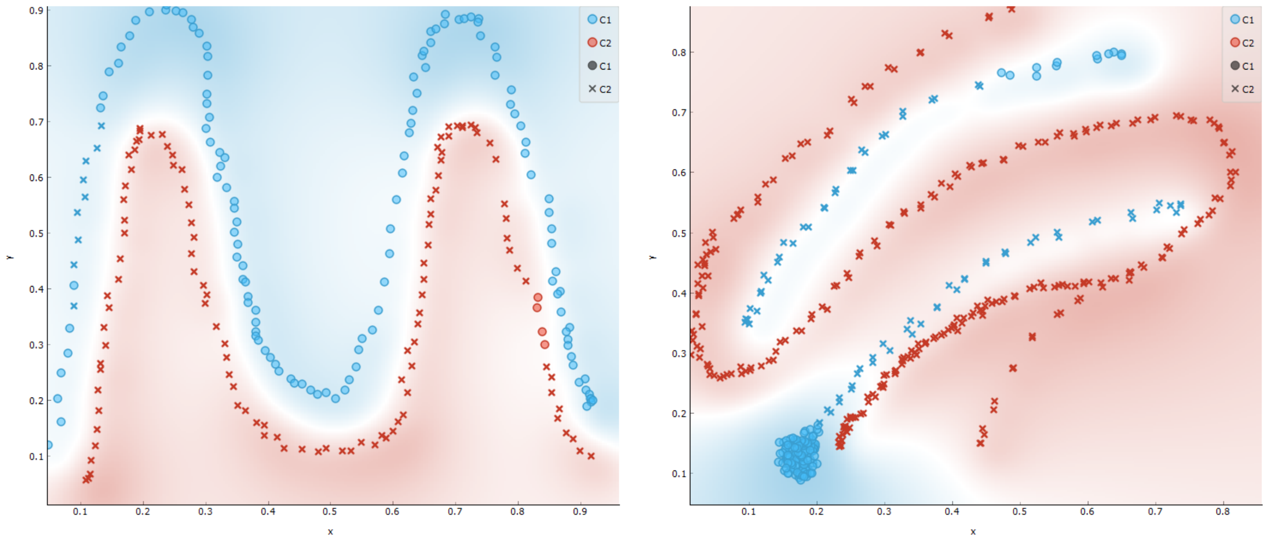
(a) success: AUC = 1.0

(b) failure: AUC = 0.844

Figure 3: Logistic Regression

1.4 SVM with RBF

The classifier is a more sophisticated one than the previous ones as it can deal with data which can effectively be separated only with a non-linear function. However, it has difficulties with situations like on Figure 4b where the classes are not balanced, they are close to each other somewhere in space, and one class (C1) is heavily concentrated in one area of space. The AUC is quite high for this case because of that heavy concentration of C1's cases at the bottom but the C1's cases above are not properly classified, their number might be much less than of those at the bottom but they might be of the same importance or even of more in which case it's a considerable failure of the classifier.



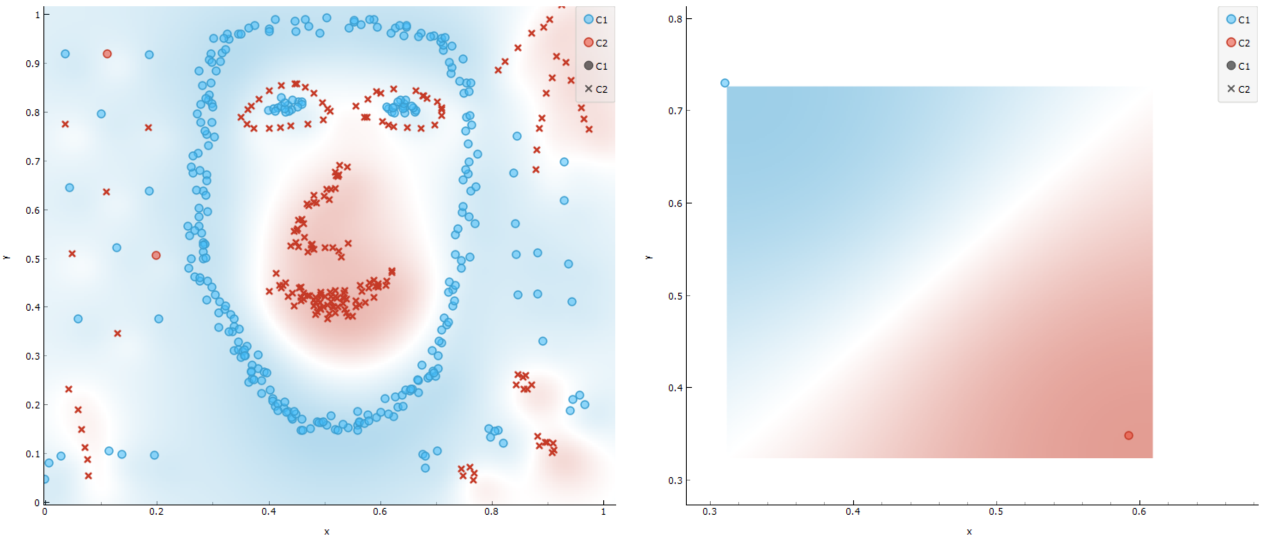
(a) success: AUC = 0.980

(b) failure: AUC = 0.890

Figure 4: SVM with RBF

1.5 Random Forest (50 Trees)

The random forrest of 50 tree does a very good job at classifying. So, I have had a difficulty at finding that data which would show a weak side if this classifier... Ok. I have found that an extremely small set of data may cause even so smart the classifier fail: see Figure 5b.



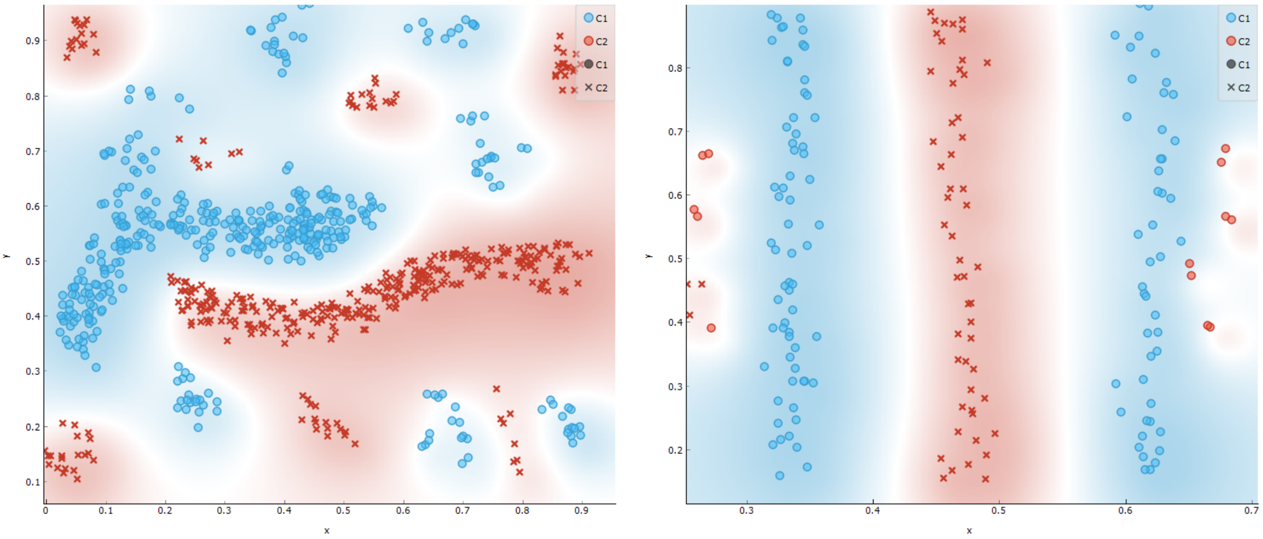
(a) success: AUC = 0.984

(b) failure: AUC = 0.5

Figure 5: Random Forest (50 Trees)

1.6 Nearest Neighbors Classifier (k = 5)

The classifier has difficulties with correctly classifying outliers: those cases which are far from their centroids: check the outer C2's cases which are wrongly classified as C1.



(a) success: AUC = 1.0

(b) failure: AUC = 0.967

Figure 6: Nearest Neighbors Classifier (k = 5)

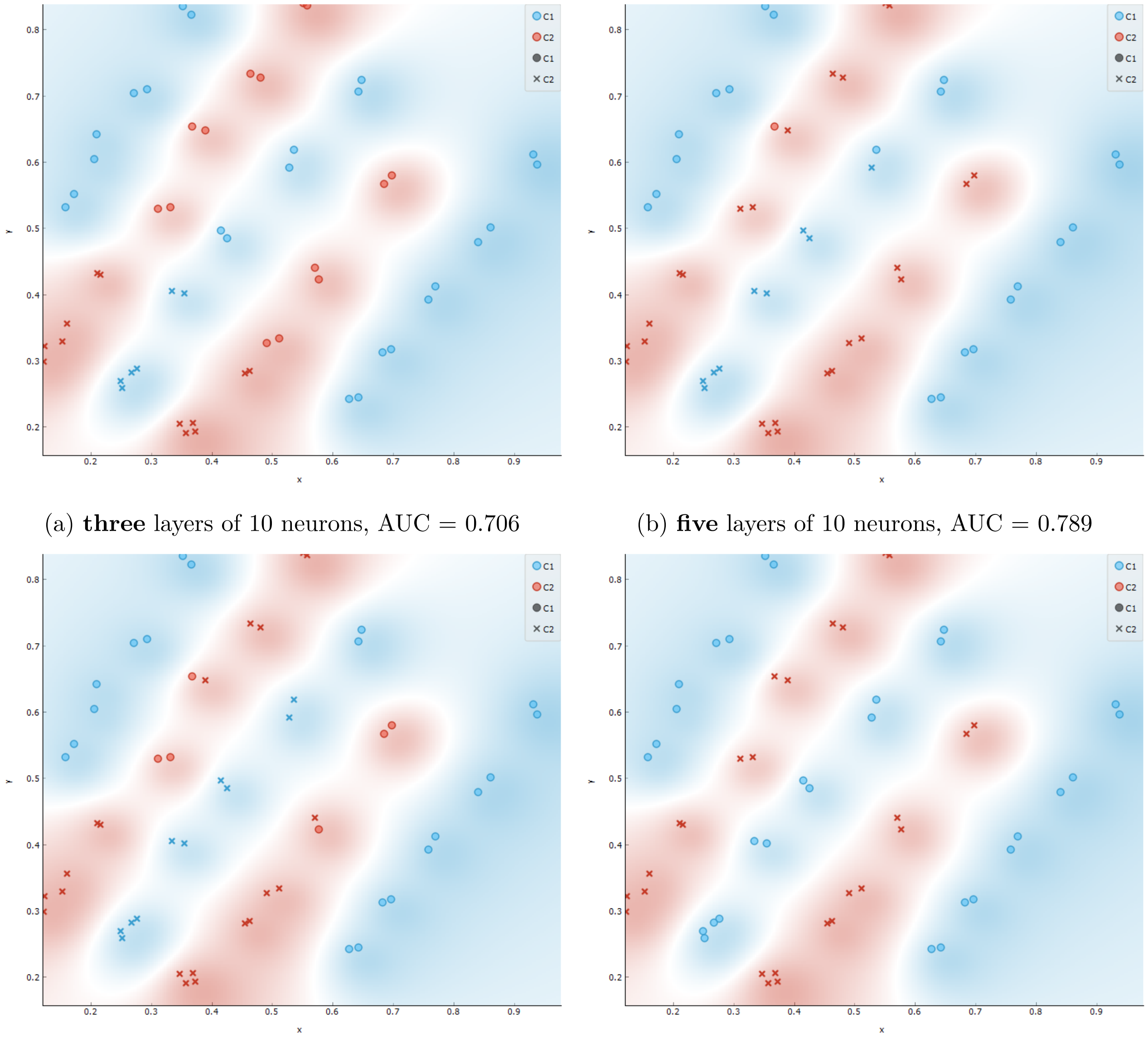
2 Second Part

The effect of the neural network's layer numbers to the quality of classification is shown below.

Figure 7a shows that the neural network with three layers of ten neurons has difficulties with correct classification of C2 whereas adding two more layers like the existing ones to the network allows their correct classification (Figure 7b).

Keeping the same number of layers but increasing the number of neurons per layer also helps as shown on Figure 7c (increased the number of neurons per layer from 10 to 20) and Figure 7d (increased the number of neurons per layer from 20 to 50).

The conclusion is that the more difficult data is the more neurons and layers are requires to 'learn' from data. One just needs to make sure that a neural network won't grow impractically large to become either computationally difficult or just to have memorized all data (overfitting).



(a) **three** layers of 10 neurons, AUC = 0.706

(b) **five** layers of 10 neurons, AUC = 0.789

(c) **three** layers of 20 neurons, AUC = 0.794

(d) **three** layers of 50 neurons, AUC = 1.0

Figure 7: Neural Network's Classifier