Assignment 1

This document contains answers to the "theoretical" questions in the assignment. Please look at this file in combination with the code for more clarity.

1.2:

Looking at the data set we have a clear separation between the two classes (A and B).

For this linear classification problem there are an infinite number of possible boundaries (as long as they separate the data set it is a solution to the problem). All three lines (red, green and blue) are possible solutions to the problem. They are however, probably not the optimal one we are looking for.
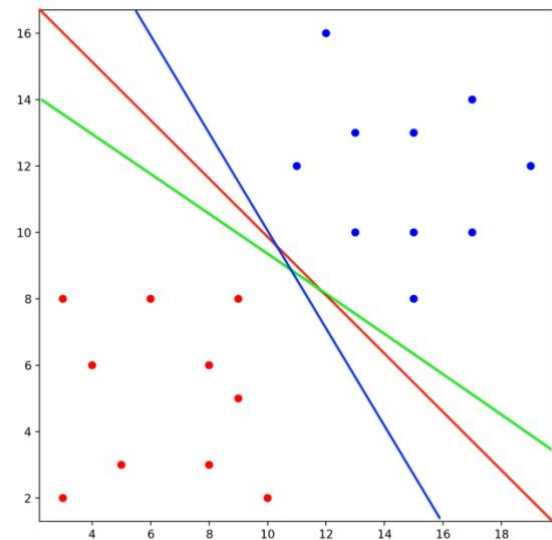
1.2a

As for which of my boundaries I think would generalize the best I would suggest the red line, as this is the line which has the biggest distance between both parts of the dataset (maximizes the margin between A and B).

Although note that I don't think that the red line *actually* maximizes the margin.

Figure 1 Three suggestions for linear decision boundaries for our classification problem

1.3

Looking at the two attributes we see that both classes have a lot of data points with the value eight as its AttrY (see Figure 2). This means that separating the two classes based on this attribute will cause problems (atleast if we base the model on this set only) This makes the attribute hard to justify.

If we look at AttrX instead we see a more clear distinction between the classes when looking at our scatter plot, unlike AttrY the classes have no overlapping values. This makes this attribute more fit for distinguishing the two classes.
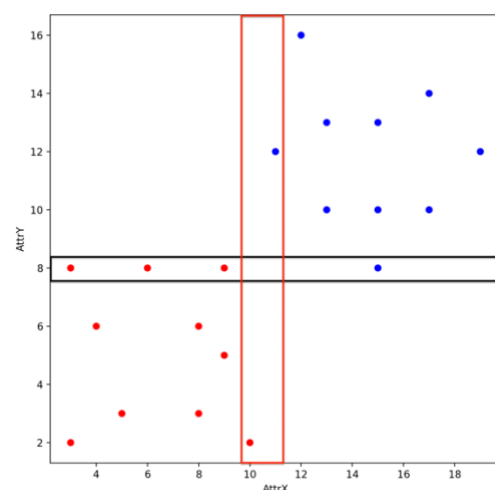
To test this hypothesis I decided to use the following attribute evaluation methods in Weka: GainRatioAttributeEval,

Figure 2 Overlapping AttrY values (Black Box), no overlapping values (Red Box)

InfoGainAttributeEval and OneAttributeEval. I tested them with and without cross validation to see if it impacted the results. By using the a1_training.csv dataset I got the following (I only discuss the final result, the complete evaluation output can be found in the hand-in folder:

Evaluation results
First I used the GainRatioAttributeEval method for evaluation. For this classification method we got a ranking favoring AttrX, this supports my claim which also favored this attribute. The scores for the test were:

- AttrX: 1
- AttrY: 0.764

We see a clear favoring of AttrX however the difference between the two attributes were small. Adding cross validation to the evaluation didn´t change the ranking but it did raise the value of AttrY slightly (0.764 -> 0.771). I suspect however that the lack of impact is because of the size of the dataset. As with the above method InfoGainAttributeEval method also favors AttrX although for this method AttrY scored lower (0.758) further supporting my claim. Lastly, in the OneRAttributeEval the difference between AttrX and AttrY was miniscule (AttrX: 100 and AttrY 95).