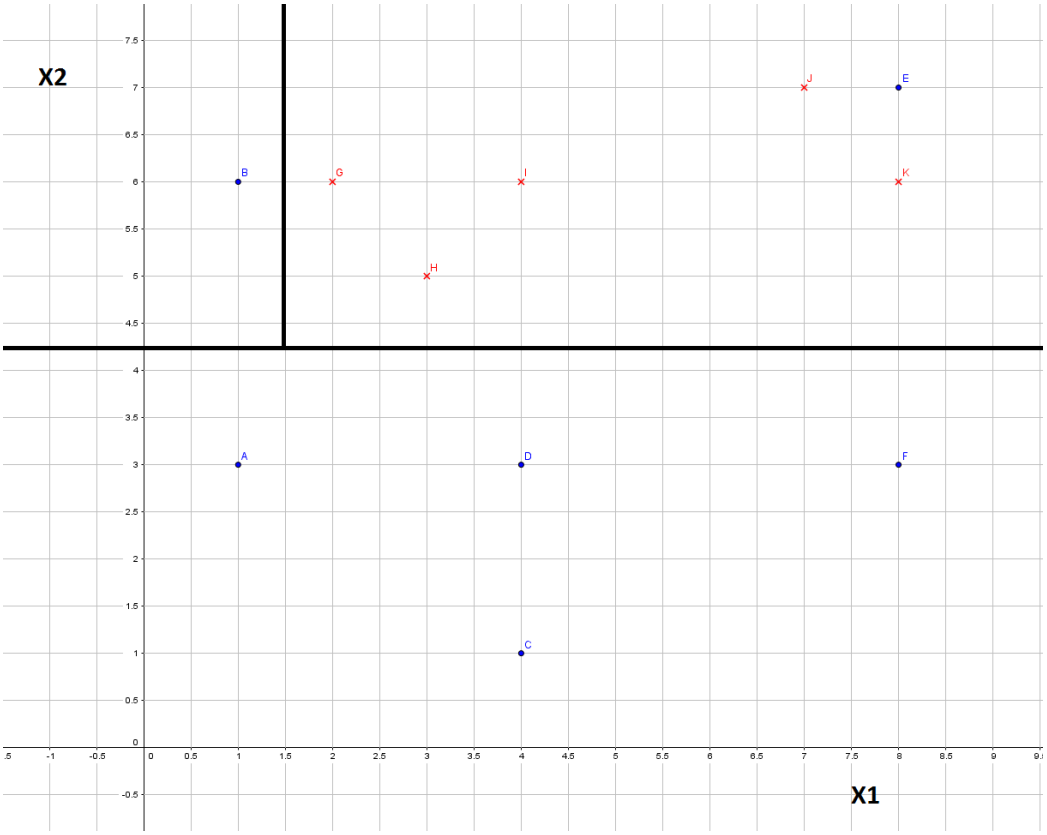
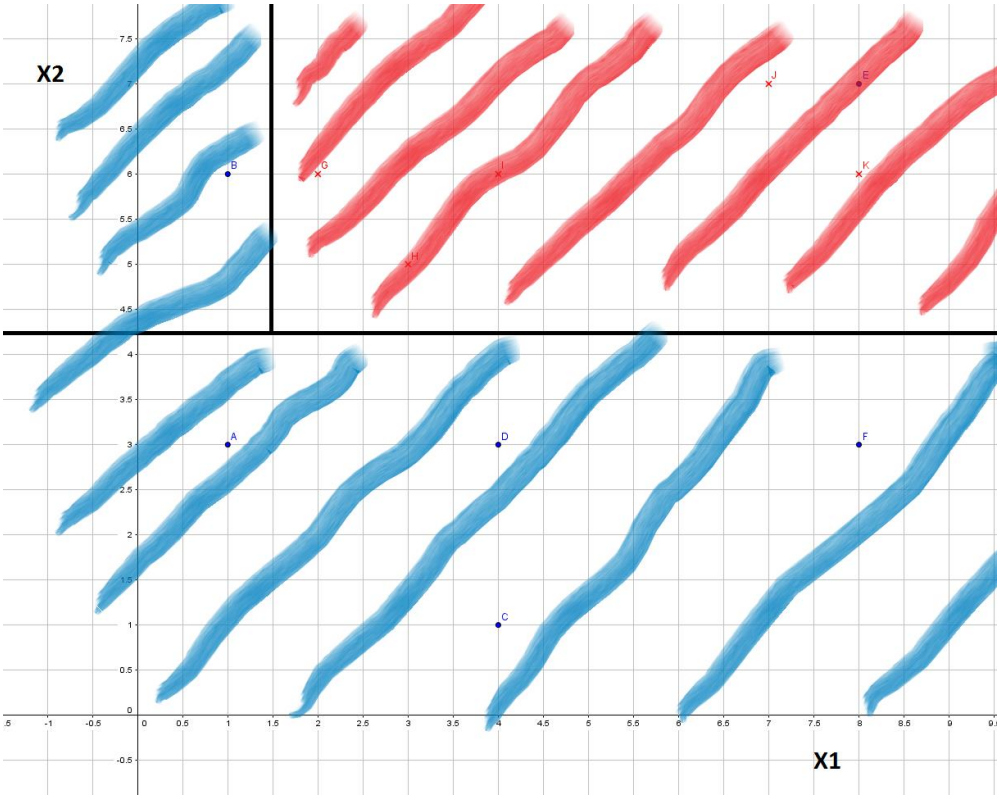
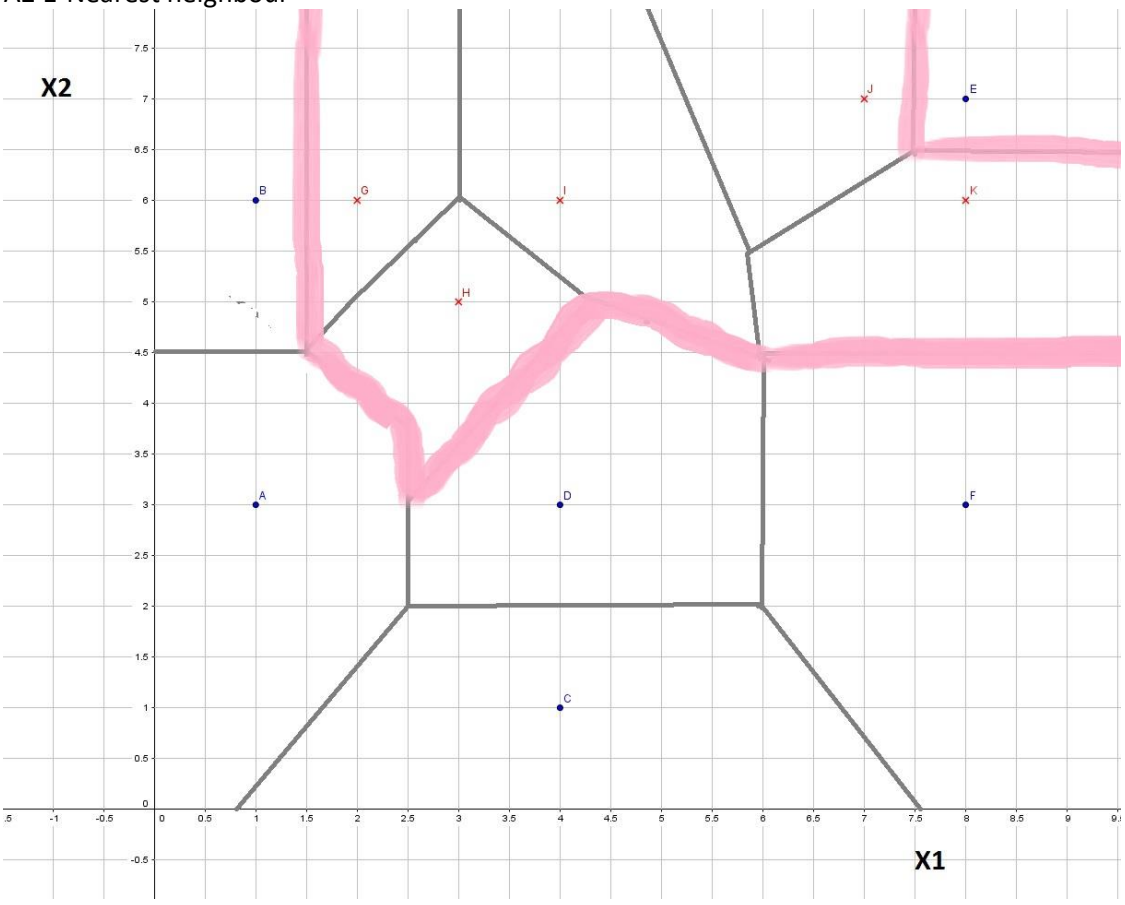
Written Assignment 4
Machine Learning
Isabel Walter
Worked together with Carine Candel

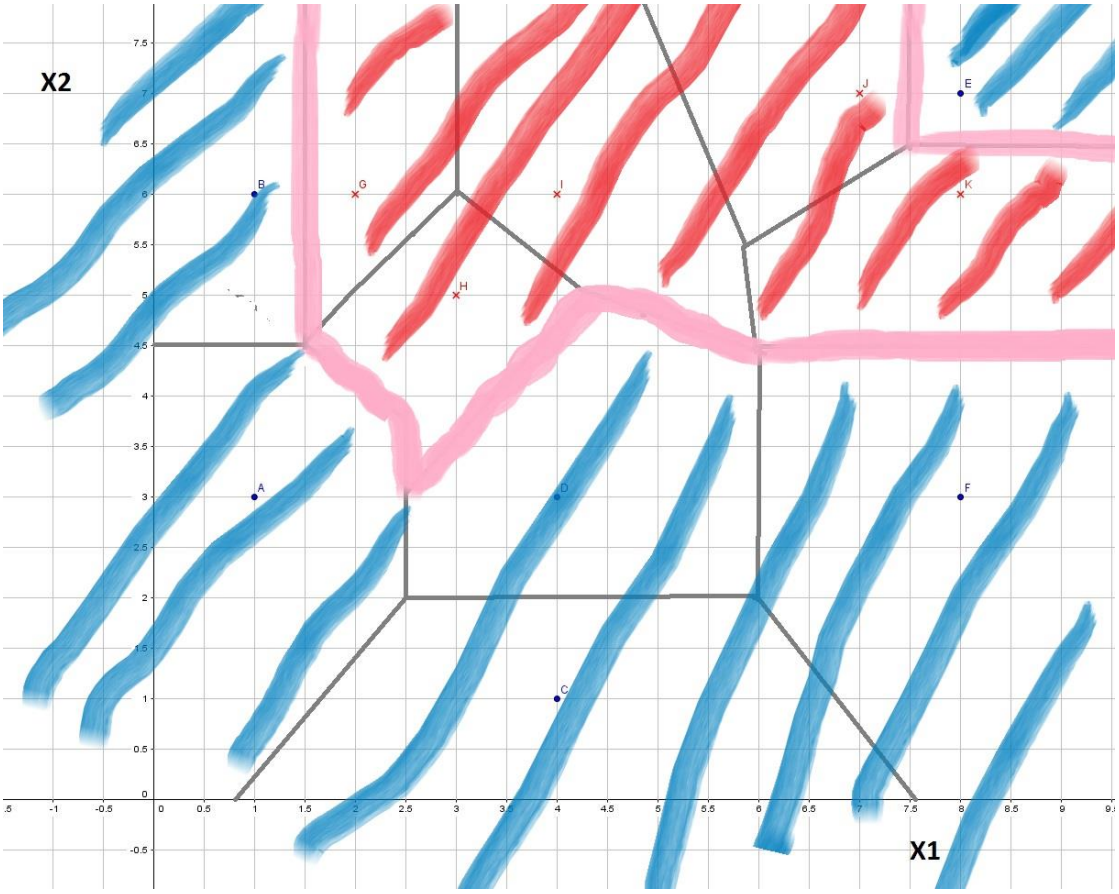
1.
A1 Decision Tree
Red: y = 1
Blue: y = 0



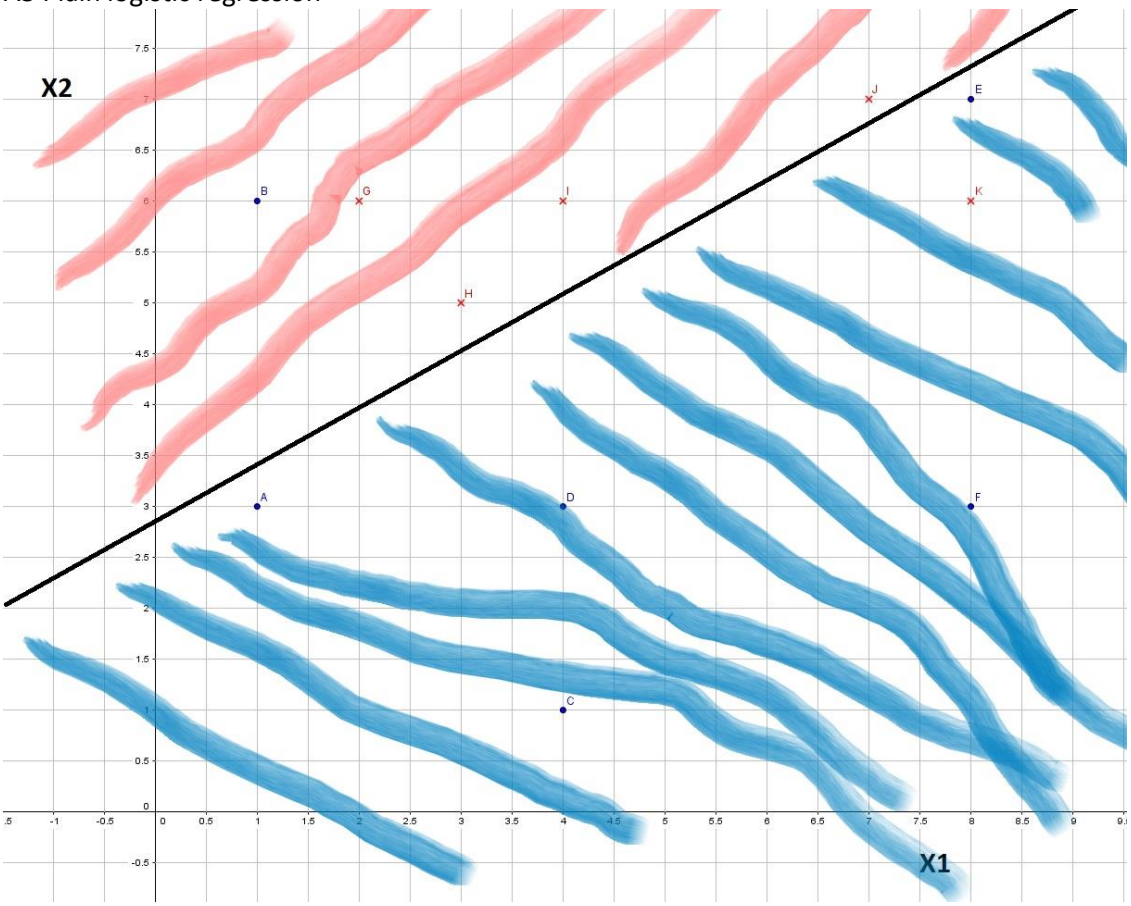


A2 1-Nearest neighbour

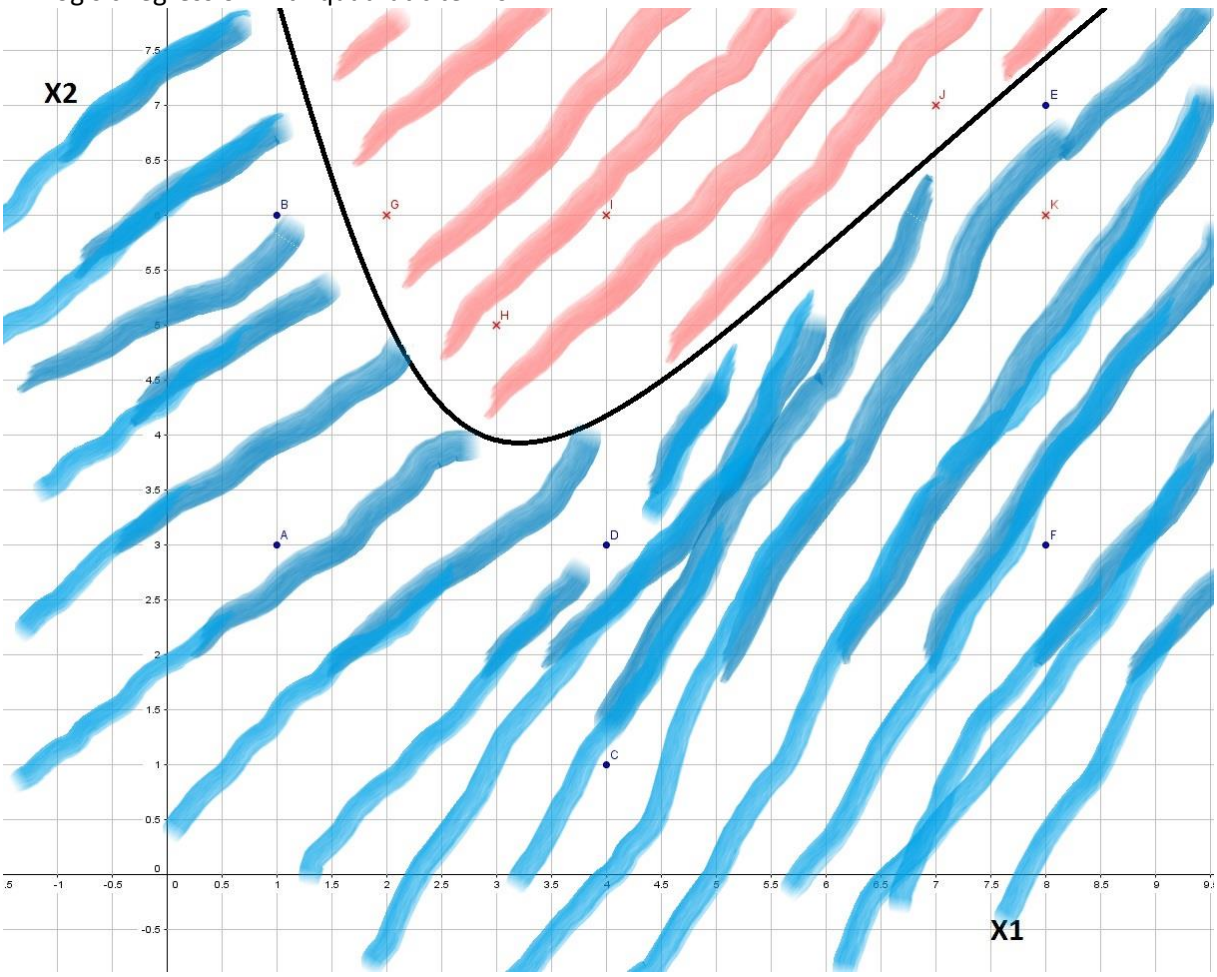




A3 Plain logistic regression



A4 Logisic regression with quadratic terms



B

I think that one boundary is definitely better than another. For example, the 1 nearest neighbour boundary seems to be overfitting because it fits (too) well to the training data. However, it will probably not fit new data that well. Another example is with the plain logistic regression and the logistic regression with quadratic terms. The former seems to have higher bias than the latter (e.g. when you look at point B → very wrongly predicted with plain logistic regression). I would say that the logistic regression with quadratic terms forms a better boundary. Furthermore, I think that the decision tree boundary is pretty good, since it only predicts one point wrongly, which could be an outlier.

Therefore, I would combine the logistic regression with quadratic terms and the decision tree boundary to invent a new boundary method.

2.

Only the 3-means clustering.

1. Initialize means: μ1 = 1 , μ2 = 3 and μ3 = 8

2. Assign each example to c1 (with μ1) , c2 (with μ2) or c3 (with μ3)

Data	Cluster
1	C1
2	Either C1 or C2 → choose C2
3	C2
3	C2

4	C2
5	C2
5	C2
7	C3
10	C3
11	C3
13	C3
14	C3
15	C3
17	C3
20	C3
21	C3

3. recalculate μ1, μ2 and μ3.
μ1 = 1/1 = 1
μ2 = (2 + 3 + 3 + 4 + 5 + 5)/6 = 3.6667
μ3 = (7 + 10 + 11 + 13 + 14 + 15 + 17 + 20 + 21)/9 = 14.2222

4. Calculate cost
J = 1/16 * ((1-1)^2+(2-3.6667)^2 + (3-3.6667)^2 + (3-3.6667)^2 + (4-3.6667)^2 + (5-3.6667)^2 + (5-3.6667)^2 + (7 – 14.2222)^2 + (10 – 14.2222)^2 + (11 – 14.2222)^2 + (13 – 14.2222)^2 + (13 – 14.2222)^2 + (14 – 14.2222)^2 + (15 – 14.2222)^2 + (17 – 14.2222)^2 + (20 – 14.2222)^2 + (21 – 14.2222)^2) = 11.05556