The three projects are to be handed in together with following exam. The exam and the projects should be separate documents.

Clearly explain how each answer has been produced and give good motivations for all calculations.

Those students who have been working together must state who they have been working with and what is to be regarded as joint work. Cite all references that you use according to the standards of a scientific paper. No part of your answers may be copied neither from books or papers nor from anything available on the internet.

1. The following dataset has two numerical predictors and a boolean response r in the last column.

x	y	r
3.874093	1.993441	0
38.065621	3.158144	0
151.304958	5.955080	0
252.845972	0.074435	0
1.537322	0.098415	1
388.705739	20.633984	0
1.151664	0.007804	1
1.557184	0.186299	1
12.459530	3.792753	0
1.073156	0.002964	1
4.348563	2.075805	0
14.912576	31.046062	1
153.259792	6.194166	0

   Preprocess this dataset in the simplest possible way so that the response can be perfectly predicted using just one single test of the form $z < c$, where c is a constant and z is a new feature generated by your preprocessing.

2. A bank has decided to use machine learning to determine if a person should be granted a loan. Assume that the database of the bank contains the following data, where the last column shows the credit history of a potential client.

Age	Married	Unemployed	Credit
39	yes	yes	bad
25	no	yes	bad
35	yes	no	good
19	yes	yes	bad
25	yes	no	good
20	no	no	bad
34	no	no	good
26	no	no	good
22	no	no	bad
18	yes	no	good

Use gini impurity heuristics as in Random Forest to construct just one binary decision tree that perfectly classifies the credit in this data set. Clearly explain all calculations.

3. (a) Assume that the maximum height of a tree is two, that is a tree may only consist of the root and its two children. Use Random Forest with mtry = 1 to construct a forest of 3 trees for the credit dataset above. You are not supposed to use any ready made implementation of Random Forest.

   (b) How is the credit worthiness of the following new customers classified by your forest? Clearly explain why you obtain a given classification.

Age	Married	Unemployed	Credit
28	yes	yes	?
23	yes	no	?
27	no	no	?

4. The following data set describes whether a given coordinate is on your own property or not.

(x, y)	Class
(10.5,5)	unowned
(7.5,6)	unowned
(2,5)	owned
(3,4)	unowned
(3.5,8)	owned
(6.5,6.5)	owned
(1.5,6)	owned
(6.5,8)	owned
(8,4)	unowned
(10.5,9)	owned

   (a) Construct a neural net with only one tanh node that perfectly differentiates between owned and unowned land. Describe how you find the weights in the neural net.

(b) When looking at a new map, it was found that also the points $(5, 9)$ and $(5, 10)$ are unowned. Construct a new neural net that gives perfect classification for both the data above and these new points.

5. Assume that the bank in the example above no longer wishes to use age discrimination and that the dataset then becomes as follows.

Married	Unemployed	Credit
no	no	good
yes	yes	bad
yes	yes	bad
yes	no	good
yes	no	good
no	no	good
no	no	bad
no	yes	bad
yes	no	good
no	no	bad

Show how naive Bayes method would classify the credit for each of the following three new potential clients. Remember to clearly explain all of your calculations.

Married	Unemployed	Credit
yes	yes	?
yes	no	?
no	no	?