Problem 0.

(i) Describe the CART algorithm for regression trees. In particular, answer the following questions:

(a) Using a square loss function, what is the predicted value in each region?

(b) Explain how a greedy algorithm helps us to choose the split variable and the split point.

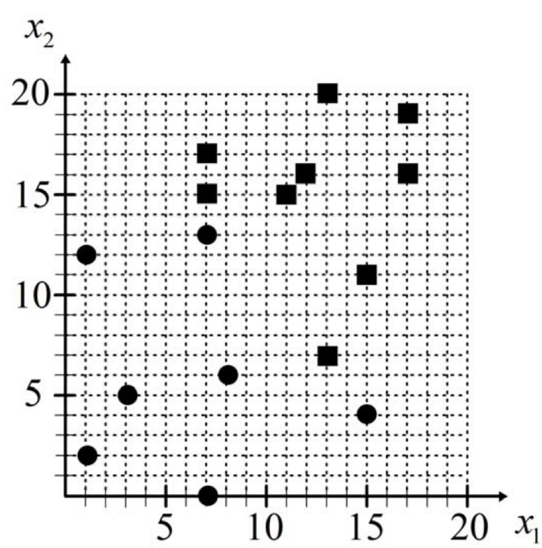
(c) Why isn't it a good idea to grow a very large tree?

(d) What is cost complexity pruning? What is it used for?

(d) What is weakest link pruning?

(ii) Which criteria would you use to grow a classification tree? Give its (their) expression(s) and discuss how it (they) relate to misclassification error.

Problem 1.
Below is a small classification training set (for 2 classes in $\mathbb{R}^2$) displayed in graphical and tabular forms (circles are class 0 and squares are class 1).



y	x_1	x_2
−1	1	2
−1	1	12
−1	3	5
−1	7	0
−1	7	13
−1	8	6
−1	15	4
1	7	15
1	7	17
1	11	15
1	12	16
1	13	7
1	13	20
1	15	11
1	17	16
1	17	19

(i) Using empirical misclassification rate as your splitting criterion and standard forward selection, find a reasonably simple binary tree classifier that has training error rate 0. Carefully describe it below, using as many nodes as you need.

At the root node: split on x_1/x_2 (**circle the correct one of these**) at the value ________

Classify to Class 0 if ______________ (creating Node $\#1$)

Classify to Class 1 otherwise (creating Node $\#2$)

At node ________: split on x_1/x_2 (**circle the correct one of these**) at the value ________

Classify to Class 0 if ______________ (creating Node $\#3$)

Classify to Class 1 otherwise (creating Node $\#4$)

At node ________: split on x_1/x_2 (**circle the correct one of these**) at the value ________

Classify to Class 0 if ______________ (creating Node $\#5$)

Classify to Class 1 otherwise (creating Node $\#6$)

At node ________: split on x_1/x_2 (**circle the correct one of these**) at the value ________

Classify to Class 0 if ______________ (creating Node $\#7$)

Classify to Class 1 otherwise (creating Node $\#8$)

(ii) Draw in the final set of rectangles corresponding to your binary tree on the graph on the previous page.

(iii) For every sub-tree T of your full binary tree above, list in the table below the size (number of leaves $|T|$) of the sub-tree T, and the training error rate of its associated classifier. We recall that the training error rate is defined as

$$E := n^{-1} \sum_{m=1}^{|T|} |R_m| Q_m(T) \,,$$

where n denotes the total number of observations, $|R_m|$ the number of observations in the terminal region R_m and $Q_m(T)$ a measure of impurity, taken as the misclassification error here.

Full tree pruned at nodes $\#$	Pruned tree size $\lvert T \rvert$	E
None (full tree)		

(iv) Using the values in your table from (iii), find for every $\alpha > 0$ a sub-tree of your full tree minimizing the cost-complexity criterion

$$C_\alpha(T) = \alpha\, E + |T| \,.$$

Plot $C_\alpha(T)$ as a function of α for each sub-tree.

Problem 2.
Consider the following dataset

x_1	x_2	x_3	y
0	0	0	0
0	0	1	0
0	1	0	1
0	1	1	1
1	0	0	1
1	0	1	1
1	1	0	0
1	1	1	0

(i) Can we represent this boolean function with a decision tree? In other words, is there a decision tree with 0 training error on this dataset?

(ii) Give a simple expression of y as a function of x_1, x_2 and x_3.

(ii) Can the CART algorithm find this tree? Explain.

Problem 3. *Bagging with linear statistics*
Consider the following quote: "The more linear is an estimator, [...] the less effective bagging will be. And vice-versa, the more effective bagging proves to be, the less linear is the problem. For example, estimators derived from linear least squares regression and ridge regression [...] should not receive much variance reduction through bagging. On the other hand, highly non-linear methods such as decision trees and neural networks should benefit substantially" in *On Bagging and non-linear estimation*, Friedman & Hall (2000). We illustrate this quote on a simple example.

Let $\mathcal{L}_n := \{X_1, \dots, X_n\}$ where the X_j are i.i.d. with mean μ and variance σ^2. Let $\bar{X}_1^*$ and $\bar{X}_2^*$ be two bootstrap realizations of the sample mean,

$$\bar{X}_i^* = \frac{1}{n}\sum_{k=1}^{n} X_{ik}^*, \qquad i = 1, 2\,,$$

where $(X_{ik}^* | \mathcal{L}_n) = X_j$ with probability $1/n$, for $j, k = 1, \dots, n$, and $i = 1, 2$.

(i) Show that the correlation $\mathrm{Corr}(\bar{X}_1^*, \bar{X}_2^*) = n/(2n-1) \approx 1/2$.

(ii) Derive the variance of the bagged mean $\bar{X}_{bag} = B^{-1}\sum_{b=1}^{B} \bar{X}_b^*$, and show that as $B \to \infty$, this term tends to the variance of $\bar{X}$.