Name:.....

Data Mining and Visualization

Assignment 5

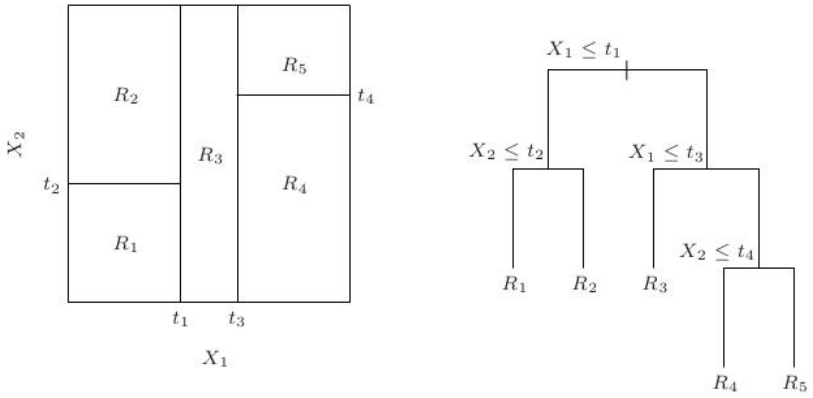
Regression Trees, Boosting and Random Forrests

Assignments constitute a part of the examination and must be returned in time. You are asked to hand in the solutions during a week following the week on which the assignment has been discussed in classes. You should submit your work to CANVAS.

Problem 1 – Basic concept for binary trees. In the lecture, it was presented that the partition sets for a binary tree have the form

$$R_m = \bigcap_{k=1}^{K_1} \{x \in \mathbb{R}^p : x_{i_{1k}} > c_{1k}\} \cap \bigcap_{k=1}^{K_2} \{x \in \mathbb{R}^p : x_{i_{2k}} \leq c_{2k}\} \tag{1}$$

Notationally, this is a complicated expression but it can be read easily from the graph of a binary tree. Consider the following tree and the corresponding partition



1. For each R_m of the partition sets on the graphs identify explicitly (1) for this particular simple example of a binary tree.
2. How in words would you explain the meaning of numbers K_1 and K_2? If p is the dimension of x-space argue that it is enough to consider $K_1 \leq p$ and $K_2 \leq p$.

Problem 2 – Regression fit by a binary tree Suppose that we want to make a regression fit using a binary tree. In the lecture we mentioned that the best least square fit over a particular region of the partition is the average value of the response over the points in the region. Here we want to provide the argument for this claim. We also argue that each additional split in a tree improve the performance of the tree as measured by the sum of squares.

1. Let X be an arbitrary random variable. Then argue that the quantity a that minimize
$$E(X-a)^2$$
is equal to $E(X)$.

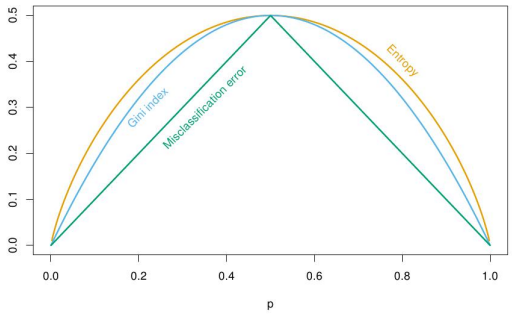
2. Using the above fact, argue that to minimize the least squares:
$$N_m \, Q_m(T) = \sum_{x_i \in R_m} |y_i - \hat{c}_m|^2$$
where N_m is the number of the inputs in R_m one can take
$$\hat{c}_m = \frac{\sum_{x_i \in R_m} y_i}{N_m}$$

3. Let X and Y be two random variables.
   (a) Explain why
   $$Var(Y) = E(Var(Y|X)) + Var(E(Y|X)).$$
   (b) This relation can be interpreted that on average the conditioning on some variable is reducing variability as expressed by variance. Please, explain this statement based on the above relation.
   (c) Can you related the above equation with the within clusters and between cluster variability as discussed in the lecture on clustering methods?

4. Use the above fact to explain that in the binary tree each split is reducing the mean square error.

Problem 3 – Binary tree for classification. Three measures of improvement in a binary tree for the classification problem have been mentioned in the lecture.

1. Explain what it means that a function is convex (concave down).
2. Argue that all three measures are convex. Here are their graphs



3. Can you give an argument that this property guarantees that each split in a tree leads to an improvement (reduction of the measures)?

Problem 4 – ROC curve. The ROC curve when applied to the sensitivity and specificity can serve as visual comparison of different classification methods. The curve can be also used to compare two distributions.

1. Plot the ROC curve for the exponential and uniform distribution on $[0, 1]$.
2. Argue that the ROC curve for comparison of two distributions always starts at $(0, 0)$ and end up in $(1, 1)$.
3. Explain that if the ROC curve is the straight diagonal line, then the distributions are identical.

The remaining problems are based on a simple simulated data set. The model is inspired by the example from the lecture on Random Forest. Namely, the features space is two dimensional, i.e. $p = 2$, each of the features having a standard Gaussian distribution with pairwise correlation 0.95, the response Y was generated according to

$$P(Y = 1|x_1 \leq 0.5) = 0.2,$$
$$P(Y = 1|x_1 > 0.5) = 0.8.$$

Here is a sample of size twenty from the model

```
      y    x1    x2
 [1,] 0 -0.31  0.17
 [2,] 0 -0.07  0.41
 [3,] 1  1.05  1.01
 [4,] 0  1.38  1.09
 [5,] 1  0.20 -0.29
 [6,] 0 -0.33 -0.41
 [7,] 0  0.12  0.07
 [8,] 0 -1.47 -1.73
 [9,] 0 -1.88 -1.55
[10,] 0  1.67  1.28
[11,] 0 -0.06 -0.38
[12,] 0 -0.83 -0.42
[13,] 1  1.33  0.84
[14,] 0 -0.47 -0.75
[15,] 0 -0.72 -0.93
[16,] 0  0.02  0.34
[17,] 0  0.05  0.39
[18,] 1  0.51  0.24
[19,] 1  0.41  0.23
[20,] 1 -0.60 -0.33
```

and the following is R-code that has produced these data

```
y1=rbinom(20,1,0.2)
rho=0.95

x1=rnorm(20)
x=rnorm(20)
x2=rho*x1+sqrt(1-rho^2)*x

y=(x1<0.5)*y1+(x1>0.5)*(1-y1)

x1=round(x1,2)
x2=round(x2,2)
z=cbind(y,x1,x2)
```

Problem 5 – Optimal predictor. Intuitively, it appears rather clear that the optimal predictor for the so simulated data is

$$G(x) = G(x_1, x_2) = \begin{cases} 1 & : x_1 \geq 0.5 \\ 0 & : x_1 < 0.5 \end{cases}$$

1. Can you provide a formal argument for the optimality of such a predictor?
2. Can you derive the optimal error rate?
3. Apply the predictor to the given data set? What is the observed error rate?

Problem 6 – R-code. Give a short explanation of the provided R-code, by describing what each line of the code is doing.

Suppose that the following commands would be executed in R

```
var(x1)
var(x2)
corr(x1,x2)
corr(x,x2)
corr(x1,x)
```

Give approximate values to the numbers you would expect to see.

Problem 7 – Classification tree. For the given data, sketch their scatter plot, marking all 20 points with the values of Y variable.

1. Use misclassification error to build the optimal 'stump' for your data.
2. Repeat this with Gini index and deviance (entropy).
3. How much of the improvement you obtain by performing this step?
4. What is the observed misclassification error for this classfier?

Problem 8 – Boosting. One can use the simple stump from the previous step to obtain boost to the prediction.

1. Evaluate the weights for the boosting algorithm after applying your stump for classification of the data.
2. Explain the first step of the boosting algorithm based on your computation, i.e. provide the new re-weighted data for which you would evaluate the next stump.

Problem 9 – Random Forrest. Explain how would you apply random forrest methodology to the data, i.e. answer the following questions

1. Explain how would you bootstrap the tree.
2. Give an example of a tree that could belong to a random forest.

Problem 10 – Extra Problem (solution is not required except for PhD students)

In the lecture we have discussed the data simulated from chi-square distribution to illustrate boosting. More precisely, the response is deterministic

$$Y = \begin{cases} 1 & : \sum_{j=1}^{10} X_j^2 > 9.34 \\ -1 & : \text{ otherwise} \end{cases}$$

X_j are iid standard normal, 9.34 is the median of the chi-square distribution with 10 degrees of freedom. The weak classifier is a simple tree with two sets in the partition – a **"stump"**. The splitting variable and the split point are based on minimizing the mean square error. Explain how the boosting could be implemented for this problem and try to write a simple code in R to run the boosting in this illustrative example.