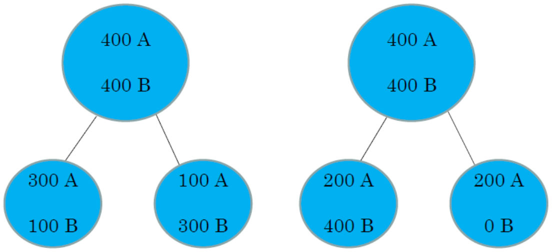
Universidad de la República
Facultad de Ingeniería

PRÁCTICO 3 : CART - MÉTODOS DE AGREGACIÓN

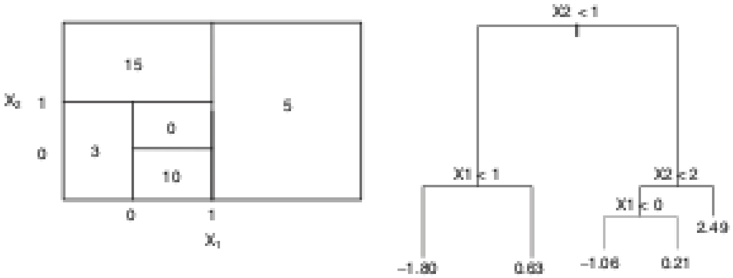
1. Prove that if x is categorical with m levels that there is $2^{m-1} - 1$ possible splits.

2. Prove that the three expressions of Gini index are the same.

3. Compute $\Delta\, i(t, s)$ for these two partitions using classification error, Gini index and entropy



4. Let consider the following figure:



   a) Sketch the tree corresponding to the partition of the predictor space illustrated in the left-hand panel of the figure.

   b) Create a diagram similar to the left-hand panel of the figure, using the tree illustrated in the right-hand panel of the same figure.

5. Suppose we produce 3 bootstrapped samples form a set containing black and white classes. We then apply a classification treee to each bootstrapped sample and, for a new value $\mathbf{x_0}$ produce 3 estimates of $\mathbb{P}(\text{black}|\mathbf{x_0})$: 0.1, 0.1 and 0.9. What is the final prediction under majority vote approach? What is the final prediction if we average the probabilities?

6. From the dataset spam of the kernlab library:

   a) Compute and draw the default tree T provided by rpart and the decision stump. Look at T$frame and examine it.

   b) 1) Compute and draw the optimal tree T_1 with associate cp parameter given by cross-validation error.

   2) Compute and draw the optimal tree T_2 with associate cp parameter given by the 1-SE rule.

   3) Compare T, T_{max}, T_1 and T_2 in learning and in test samples.

   c) Apply Bagging and Random Forest (default) and compare the prediction errors with a single tree.

   d) Study the evolution of the OOB error with respect to ntree using do.trace.

   e) Calculate the variable importance of the spam variables for Random Forest (default).

   f) Calculate the importance of spam variables for stumps Random Forest.

   g) Illustrate the influence of the mtry parameter on the OOB error and on the variable importance.

7. Use the Carseats dataset of the tree library.

   a) Split the data set into a training set and a test set.

   b) Fit a regression tree to the training set. Plot the tree, and interpret the results. What test error rate do you obtain?

   c) Use cross-validation in order to determine the optimal level of tree complexity. Does pruning the tree improve the test error rate?

   d) Use the Bagging approach in order to analyze this data. What test error rate do you obtain?

   e) Use Random Forests to analyze this data. What test error rate do you obtain? Use the importance() function to determine which variables are most important. Describe the effect of m, the number of variables considered at each split, on the error rate obtained.

   f) Answer the same questions if the variable Sales is discretized as follows: 1 if the Sales variable is higher than 8, 0 otherwise.