TD 1: Decision Trees

Ensemble learning from theory to practice

Exercise 1 (Reminders):

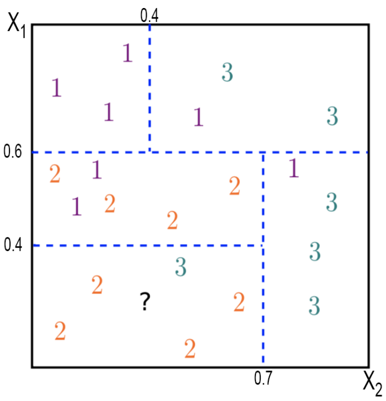

Figure 1: The spatial segmentation of a classification decision tree.

1. Thanks to the spatial segmentation, draw the visual tree.
2. What is the input space? What is the output space?
3. Write the f formula associated.
4. How many split there are?
5. How many nodes?
6. How many leaves?
7. What will be the predicted value associated to the new data symbolized by the question mark?
8. Is this tree accurate?
9. Compute the MSE (Mean Squared Error measure) to evaluate the quality of the decision tree predictor
$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$
10. What do you think about the result? What is disturbing in the previous question (clue: train data vs test data)?

Exercise 2 (Questions about understanding basic concepts of the course):

1. A student has trained a decision tree and notices that its performance is better on the train set than the test set. Should he increase or decrease the depth of the tree?
2. A student has dataset with n instances and p features.
   (a) what is the maximum number of leaves of a decision tree on this dataset?
   (b) what is the maximum depth of a decision tree on this dataset?

Exercise 3 (Understand the splitting process idea by practice with Gini impurity):
Consider the following dataset containing for 10 plants the length and width of their sepals. We want to discriminate plants that belong to the species Iris virginica (+) from others (-).

Label	+	+	+	+	+	+	-	-	-	-
Length (cm)	6.7	6.7	6.3	6.5	6.2	5.9	6.1	6.4	6.6	6.8
Width(cm)	3.3	3	2.5	3	3.4	3	2.8	2.9	3	2.8

1. Calculate the Gini impurity for all possible separation points using the **length** of the sepals as the separating variable.
   Note that, the Gini impurity of an R region is defined as:
   $$Imp(R) := \sum_{c=1}^{C} p_c(R)(1 - p_c(R)) \tag{1}$$
   where, $p_c(R) := \frac{1}{|R|} \sum_{i:\boldsymbol{x_i} \in R} \mathbb{1}_{\{y_i=c\}}$. Thus, if all the instances of a region belong to the same class, the impurity of this region is equal to 0; conversely, if a region contains as many instances of each of the C classes, the right part of the product is $1 - p_c(R) = 1 - \frac{1}{C}$, or $\frac{1}{2}$ in the case of a binary classification.
2. Calculate the Gini impurity for all possible separation points using the **width** of the sepals as the separating variable.
3. What is the first node of a decision tree trained on this dataset with the Gini impurity?

Exercise 4 (Understand the splitting process idea):
We will show that minimizing the quadratic risk amounts to minimizing the variance in each hyper-rectangle of the input space partition.

1. For all region R_k , write the minimization empirical quadratic risk problem where the predictor is a decision tree.
2. Write R_k in function of 2 subspaces $R_L(j, s)$ and $R_R(j, s)$ obtained after splitting a region R_k via a split (j, s).
3. How are these two subspaces relative to each other?
4. Write the new risk formula based on these two subspaces.
5. Write the variance formula in a node k (corresponding to a region R_k), then in each child nodes of k.
6. Show that minimizing the empirical quadratic risk formula amounts to minimizing the variance in each hyper-rectangle of the input space partition.