Universidad de la República **Introducción a la Ciencias de Datos**
Facultad de Ingeniería **Curso 2020**

Practice: Introduction to Statistical Learning and Modeling

1. Think about three multivariate methods. For each one, say:

   a) if it is a non supervised or a supervised approach,if it is for doing regression or classification or both.

   b) is there any restriction about the distribution or the kind of variables to use?

   c) which are the different parameters used for the method?

2. What are the advantages and disadvantages of a very flexible (versus a less flexible) approach for regression or classification? Under what circumstances might a more flexible approach be preferred to a less flexible approach? When might a less flexible approach be preferred?

3. *a*) Read pages 39 to 42 about K-nearest neighbors of the book *An Introduction to Statistical Learning with Applications in R.*

   b) The table below provides a training data set containing six observations, three predictors, and one qualitative response variable.

x_1	x_2	x_3	y
0	3	0	Red
2	0	0	Red
0	1	3	Red
0	1	2	Green
-1	0	1	Green
1	1	1	Red

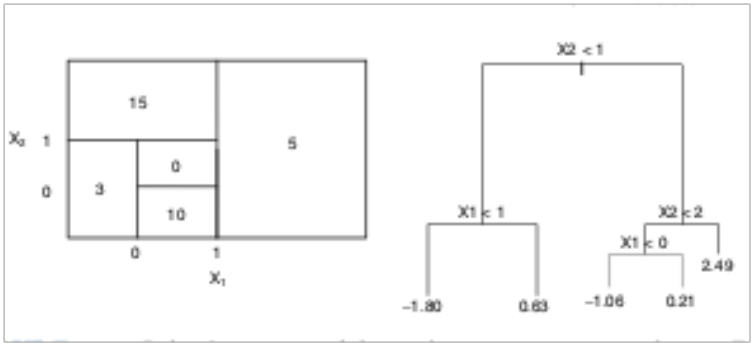
   Suppose we wish to use this data set to make a prediction for Y when $x_1 = x_2 = x_3 = 0$ using K-nearest neighbors.

   1) With the euclidean distance, what is the prediction with $K = 1$ and with $K = 3$ for the test point $(0, 0, 0)$?
   2) If the Bayes decision boundary in this problem is highly non-linear, then would we expect the best value for K to be large or small? Why?

4. This question relates to the plots in Figure 1.

   a) Sketch the tree corresponding to the partition of the predictor space illustrated in the left-hand panel of Figure 1. The numbers inside the boxes indicate the mean of Y within each region.

   b) Create a diagram similar to the left-hand panel of Figure 1, using the tree illustrated in the right-hand panel of the same figure. You should divide up the predictor space into the correct regions, and indicate the mean for each region.



5. In this problem, you will perform K-means clustering manually, with $K = 2$, on a small example with $n = 6$ observations and $p = 2$ features. The observations are as follows.

Obs.	x_1	x_2
1	1	4
2	1	3
3	0	4
4	5	1
5	6	2
6	4	0

   a) Plot the observations.

   b) Randomly assign a cluster label to each observation. You can use the sample() command in R to do this. Report the cluster labels for each observation.

   c) Compute the centroid for each cluster.

   d) Assign each observation to the centroid to which it is closest, in terms of Euclidean distance. Report the cluster labels for each observation.

   e) Repeat (c) and (d) until the answers obtained stop changing.

   f) In your plot from (a), color the observations according to the cluster labels obtained.