Prepare for Test #2

If you have any questions on the solutions, please contact Barbora Hladká at hladka@ufal.mff.cuni.cz.

1) Binary classifier evaluation

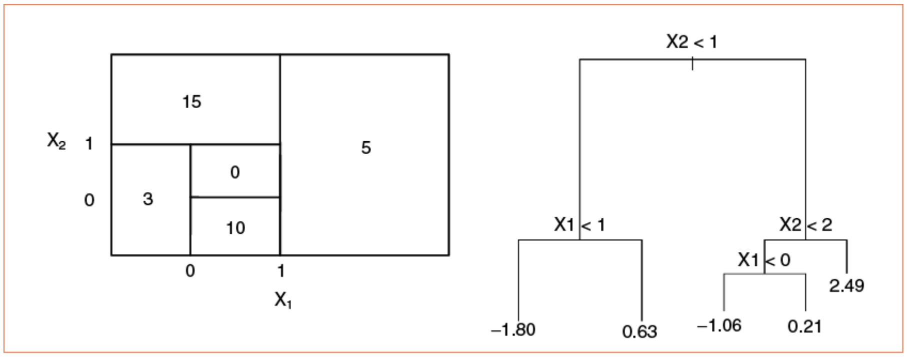
A binary classifier was evaluated using a set of 1000 test examples in which 50 % of all examples are negative. It was found that the classifier has 60 % sensitivity and 70 % accuracy.

a) Write the whole confusion matrix.

b) Compute the classifier's precision.

c) Compute the classifier's specificity.

2) Decision Trees



a) Sketch the tree corresponding to the partition of the feature space illustrated in the left-hand panel of the figure. The numbers inside the boxes indicate the average value of the target attribute within each region. (The right-hand panel of the figure is only for illustration.)

3) Clustering

Suppose that we have four examples with the following *distance matrix*:

$$\begin{array}{c|ccccc} & \mathbf{x_1} & \mathbf{x_2} & \mathbf{x_3} & \mathbf{x_4} & \mathbf{x_5} \\ \hline \mathbf{x_1} & & 2 & 6 & 10 & 9 \\ \mathbf{x_2} & 2 & & 5 & 9 & 8 \\ \mathbf{x_3} & 6 & 5 & & 4 & 5 \\ \mathbf{x_4} & 10 & 9 & 4 & & 3 \\ \mathbf{x_5} & 9 & 8 & 5 & 3 & \end{array}$$

For instance, the distance between the first and second examples $d(\mathbf{x}_1, \mathbf{x}_2)$ is 2, and the distance between the second and fourth observations $d(\mathbf{x}_2, \mathbf{x}_4)$ is 9.

a) On the basis of this distance matrix, sketch the dendrogram that results from hierarchical aglomerative clustering these four examples using *single linkage*. Be sure to indicate on the plot the height at which each fusion occurs, as well as the examples corresponding to each leaf in the dendrogram.

b) Suppose that we cut the dendogram obtained in (a) so that two clusters result. Draw the cut in your dendrogram. Which observations are in each cluster?

4) Evaluation of a simple Linear Regression model

Work with the following five training examples:

x	0	1	3	5	6
y	5	4	3	2	1

What is the proportion of the explained variance in the target y by the linear regression on x (i.e. what is the coefficient of determination) if the regression parameters are $\theta_0 = 5$ and $\theta_1 = -1$?

a) Write the formula for the coefficient of determination.

b) Compute the coefficient of determination for the given model of simple linear regression. Write also all results that you need for computing the coefficient of determination using the formula in a).

5) Logistic regression

Suppose that we collect data for a group of students in a statistics class with two features A_1 (= hours studied) and A_2 (= undergrad Grade Point Average) and a target binary attribute Y (= receive a grade A). We fit a logistic regression model and produce estimated parameters:

$$\hat{\theta_0} = -6,\ \hat{\theta}_1 = 0.05,\ \hat{\theta}_2 = 1.$$

Estimate the probability that a student who studies for 50 hours and has an undergrad GPA of 3.5 gets an A in the class. Show how you compute it.