Data science exam – version 1

Fabrice Rossi

2017

This exam consists in a series of independent exercises. They can be solved in any order. Answers must be justified: a simple "yes" or "no" answer will not be considered as a proper one.

Exercise 1

We study in this exercise a data set $\mathcal{D} = (X_i,Y_i)_{1\leq i\leq 20}$ with $Y_i \in \{1,2\}$. The random pairs (X_i,Y_i) are assumed to by independant and identically distributed copies of a data generating pair (X,Y). Using this data set, an analyst builds two models, g_1 and g_2 whose predictions on $\mathcal{D}$ are given by the following table:

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Y_i	1	1	1	2	1	1	1	1	1	2	2	2	2	1	2	2	2	1	1	1
$g_1(X_i)$	1	1	1	2	1	2	1	2	1	2	1	1	2	1	1	2	2	1	1	1
$g_2(X_i)$	1	1	2	2	1	1	1	1	2	2	1	1	1	1	2	2	1	1	1	2

In this table, each column corresponds to an observation (X_i, Y_i). The top row gives the value of Y_i while the two other ones correspond to model predictions.

Question 1 Using the table, provide an estimation of $\mathbb{P}(Y = 1)$. What general estimation principle is used to compute this estimation?

Question 2 Using the table, compute estimations of $\mathbb{P}(g_1(X) \neq Y)$ and $\mathbb{P}(g_2(X) \neq Y)$.

Question 3 Assume the loss function $l_0(p,t) = \mathbb{I}_{p\neq t}$ is used (p is the prediction, t the true value and $\mathbb{I}_C$ equals 1 when the condition C is fulfilled and 0 when it is not). The corresponding risk is denoted L_0. Using the table, provide an estimation of the risks of g_1 and g_2. What are the limitations (if any) of this estimation? What other strategy could be used to estimate the risks?

Question 4 Let l_1 be the following loss function:

$l_1(p,t)$	$t = 1$	$t = 2$
$p = 1$	0	1
$p = 2$	2	0

Determine the best model among g_1 and g_2 according to the risk associated to l_1. Possible limitations described in the previous question should be disregarded in the present one.

Exercise 2

We assume given a random pair (X,Y) with the following characteristics:

1. Y takes values in $\{-1,1\}$ and $\mathbb{P}(Y = \text{-1}) = \frac{2}{3}$;
2. X takes values in $\{a, b, c\}$ and the conditional distribution of X given Y is specified by the following table:

x	a	b	c
$\mathbb{P}(X = x \mid Y = -1)$	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{2}$
$\mathbb{P}(X = x \mid Y = 1)$	$\frac{1}{3}$	$\frac{1}{6}$	$\frac{1}{2}$

Question 1 Recall the expression of the general theoretical optimal model g_0^* for (X,Y) when using the loss function l_0 defined by $l_0(p,t) = \mathbb{I}_{p \neq t}$ (p is the prediction and t the true value) and assuming the joint distribution of (X,Y) is known.

Question 2 Using the assumptions on (X,Y) compute $g_0^*(x)$ for all $x \in \{a, b, c\}$.

Question 3 Compute $L_0^* = L_0(g_0^*)$.

Question 4 Let l_2 be the following loss function:

$l_2(p,t)$	$t = 1$	$t = -1$
$p = 1$	0	2
$p = -1$	1	0

Compute both g_2^* and L_2^*, respectively the optimal model according to l_2 and its risk.

Exercise 3

We assume given a random pair (X,Y) with the following characteristics:

1. Y takes values in $\{A, B, C\}$ and $\mathbb{P}(Y = A) = \mathbb{P}(Y = B) = \frac{1}{3}$;
2. X takes values in $\{0,1\}^3$ (that is a typical value of X is (1,0,1)). $X_{i,j}$ denotes the j-th coordinate of observation number i;
3. the distribution of (X,Y) satisfies the conditional independence assumptions of the Naive Bayes classifier.

We study a data set $\mathcal{D} = (X_i,Y_i)_{1 \leq i \leq 300}$ for such that $|\{i|Y_i = A\}| = |\{i|Y_i = B\}| = 100$, where $|U|$ is the cardinality of the set U (its number of elements).

The values taken by the X_i are summarized in the following table:

	$X_{i,1}$	$X_{i,2}$	$X_{i,3}$
$Y_i = A$	85	74	85
$Y_i = B$	38	57	50
$Y_i = C$	43	93	8

The table reads as follows: each row corresponds to the subset of observations for which Y_i takes the given value. For instance the first row corresponds to observations for which $Y_i = A$. Then each column gives the number of such observations for which the coordinate associated to the column equals 1. For instance, the upper left corner value 85 says that among the 100 observations for which $Y_i = A$, 85 observations have a 1 as their first coordinate.

Question 1 Estimate from the data all the (conditional) probabilities needed to design a Naive Bayes classifier on those data.

Question 2 Using the loss function l_0 defined by $l_0(p,t) = \mathbb{I}_{p \neq t}$ (p is the prediction and t the true value), compute the optimal decision of the Naive Bayes classifier for $u = (1, 0, 1)$ and for $v = (0,0,1)$.

Exercise 4

The data set studied in this exercise comprises 1473 objects described by 9 explanatory (X_1 to X_9) and 1 target variable Y. Y takes values in $\{1, 2, 3\}$. Explanatory variables X_1 and X_4 are numerical variables while all other variables are nominal ones (with numbered categorical values, from 0 to 4, but no special numerical meaning should be assigned to said values).

Values taken by Y on the data set are summarized on the following table:

	1	2	3
Y	624	345	504

Question 1 The data scientist builds a full (unprunned) decision tree on the data. The tree ends up with 644 leaves and its confusion matrix on the data set is given by following table:

	1	2	3
1	603	9	4
2	5	321	22
3	16	15	478

In the confusion matrix, predictions are in row, with true values are in column.

Comment briefly the results.

Question 2 The analysts implements a ten fold cross-validation on the tree in order to estimate its performances during the pruning process. Results are shown on Figure 1. Comment briefly the figure. Compare in particular the results from the figure and the confusion matrix.

Question 3 The analysts decides to prune the tree up to leaving only 7 leaves. Justify briefly her choice.

Question 4 The pruned tree is represented on Figure 2. Use this figure to compute the confusion matrix of the pruned tree. Are the results compatible with the ones from Figure 1?

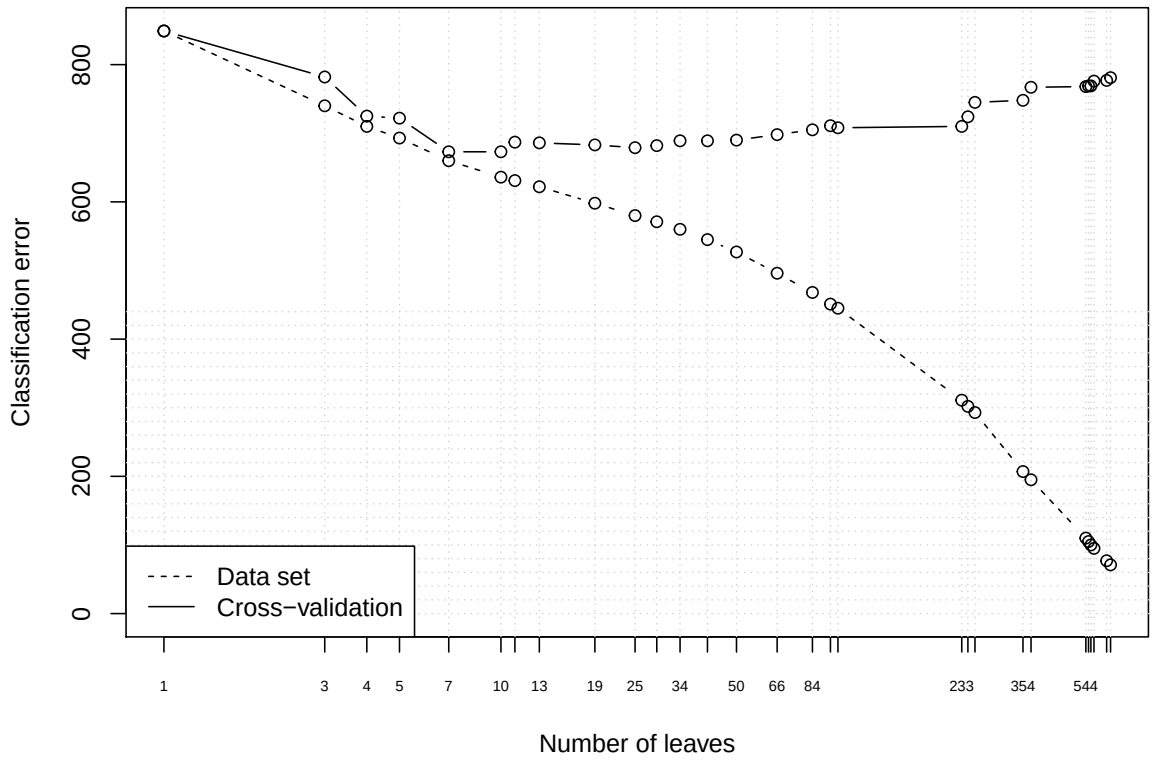

Figure 1: Behavior of the number of misclassified examples as a function of the number of leaves in the tree (during pruning), both on the data set and as estimated by a 10 fold cross-validation method. The x-axis uses a logarithmic scale.

Question 5 For each of the leaves of the pruned tree, construct an artificial data point (by choosing the values of its coordinates) in such a way that it will fall into the associated leaf when submitted to the tree.

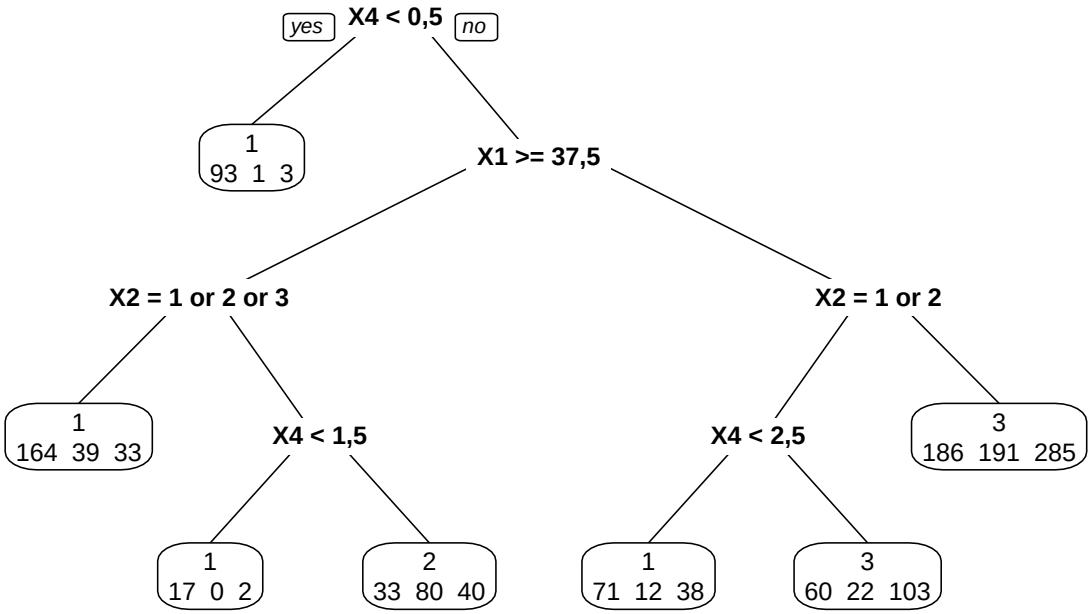

Figure 2: Pruned decision tree. At each node, the left branch corresponds to a "yes" answer to the question of the node, while the right branch correspond to "no". In a leaf, the upper value is the predicted class for the leaf, while the lower values correspond to the number of examples from the data set that fall into this leaf, attributed to each class (in the natural class order, 1, 2 and 3).