Data science exam

Fabrice Rossi

2018

This exam consists in a set of three 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

In this exercise, we study a classification problem in which the target variable $\mathbf{Y}$ can take three different values in $\mathcal{Y} = \{A, B, C\}$. From a learning set $\mathcal{D}$, two models have been constructed g_1 and g_2. Their predictions on a new set $\mathcal{D}'$ are summarized by the following confusion matrices (we use the convention that the predicted values are in rows while the true values are in columns):

g_1

	A	B	C
A	57	0	0
B	5	59	1
C	3	6	54

g_2

	A	B	C
A	56	3	1
B	1	61	5
C	8	1	49

Question 1 Using the confusion matrices, compute an estimation of the distribution of $\mathbf{Y}$, i.e. of the probabilities $\mathbb{P}(\mathbf{Y} = \mathbf{y})$ for $\mathbf{y} \in \mathcal{Y}$.

Question 2 What minimal consistency checks between $\mathcal{D}$ and $\mathcal{D}'$ should be done?

Question 3 Compute the accuracy of each model on $\mathcal{D}'$ (the accuracy is the percentage of correct classification).

Question 4 Determine the best model between g_1 and g_2 according to the loss function $l_1(p,v) = \mathbf{1}_{p \neq v}$ using the empirical risk on $\mathcal{D}'$.

Question 5 Is the selected model the best one according to the risk associated to l_1?

Question 6 We define a new loss function l_2 as follows:

$l_2(p,v)$		v A	B	C
	A	0	2	1
p	B	1	0	1
	C	2	1	0

We use the convention that p is the predicted value and v the true value. Compute the empirical risk of each model according to this loss function on $\mathcal{D}'$.

Exercise 2

In this exercise, the data under analysis are binary, with $\mathcal{X} = \{0,1\}^3$ and $\mathcal{Y} = \{0,1\}$. The data probability distribution is partially known and given in the following table:

	X1	X2	X3	P.Y.given.X
1	0	0	0	0.75
2	0	0	1	0.50
3	0	1	0	0.20
4	0	1	1	0.25
5	1	0	0	0.40
6	1	0	1	0.20
7	1	1	0	0.80
8	1	1	1	0.80

In each row of the table corresponds to a value $\mathbf{x} \in \mathcal{X}$. The last column (entitled "P.Y.given.X") gives the conditional probability $\mathbb{P}(\mathbf{Y} = 1|\mathbf{X} = \mathbf{x})$. For instance, the second row of the table specifies that $\mathbb{P}(\mathbf{Y} = 1|\mathbf{X} = (0,0,1)) = 0.5$.

Question 1 Assuming that X and Y are distributed in a way compatible with the table, compute an optimal g_1^* for the loss function $l_1(p,v) = \mathbf{1}_{p \neq v}$. More precisely, give the value of an optimal decision $g_1^*(\mathbf{x})$ for each $\mathbf{x} \in \mathcal{X}$.

Question 2 Is the optimal model for l_1 unique?

We introduce a parametric loss function l_λ given by $l_\lambda(0,1) = \lambda$ and $l_\lambda(1,0) = 1$. $g_{l_\lambda}^*$ is the optimal model for the loss function l_λ.

Question 3 Write the condition on $\mathbb{P}(\mathbf{Y} = 1|\mathbf{X} = \mathbf{x})$ that has to be fulfilled to have $g_{l_\lambda}^*(\mathbf{x}) = 1$.

Question 4 Compute a value of $\lambda > 1$ such that the optimal model for the corresponding l_λ gives always the same decision, that is $g_{l_\lambda}^*(\mathbf{x})$ is constant over $\mathcal{X}$.

We assume in the following questions that $\mathbf{X}$ has a uniform probability distribution on $\mathcal{X}$: $\mathbb{P}(\mathbf{X} = \mathbf{x}) = \frac{1}{8}$ for all $\mathbf{x}$.

Question 5 Compute $\mathbb{P}(\mathbf{Y} = 1)$.

Question 6 Compute the risk of the optimal model g_1^* (the model computed in question 1).

Exercise 3

In this exercise, we study a data set of grades obtained by students. Each observation in the data set is a student described by her/his 16 grades obtained at 16 different exams (given by variables E1 to E16). Exams are identical for all students and therefore grades are comparable between two students. They are expressed by only three possible outcomes: FAIL, PASS and MERIT (which stands for pass with merit). Students were separated into two groups as indicated by the variable Group, with two possible values Standard and Tutored (thus a student is described by 17 variables). There are 241 students in the Standard group and 152 students in the Tutored group.

Question 1 The analyst builds a decision tree on the full data set: the target variable is the Group while the predictive variables are the 16 grades. The fully developed tree has 25 leaves. Discuss briefly this value taking into account the size of the data set.

	Standard	Tutored
Standard	241	0
Tutored	0	152

Table 1: Confusion matrix of the fully developed tree. Each row corresponds to the predicted group of the student while each column corresponds to its true group.

Question 2 Table 1 gives the confusion matrix of the fully developed tree. Comment briefly the results.

Question 3 The analyst decides to use a 5 fold cross-validation to chose the number of leaves in the decision tree. Give a brief justification of this choice.

Question 4 List precautions that should be taken, if any, when building the blocks of the cross-validation. This must be specific to the case under study, not some general rules of thumb.

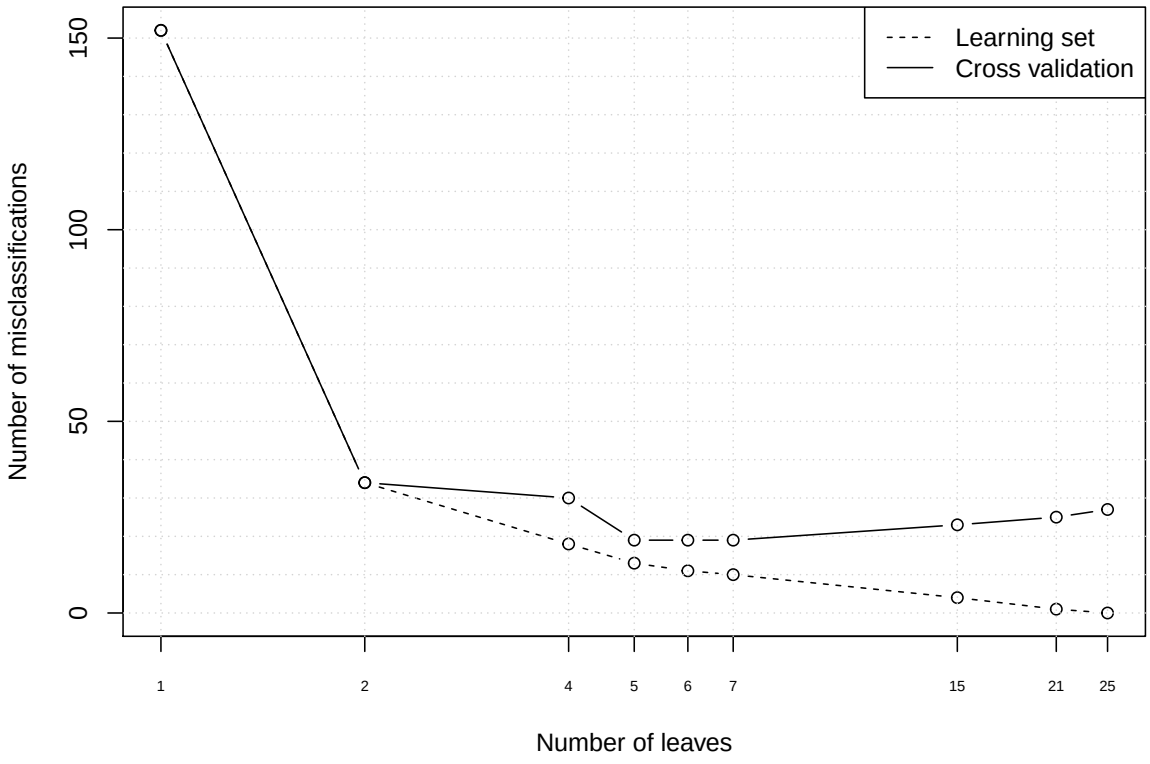

Figure 1: Number of misclassifications as a function of the number of leaves, on the learning set and as estimated by cross-validation. The x axis uses a logarithmic scale.

Question 5 Results from the cross-validation procedure are given on Figure 1. List tree sizes that could be retained by the analyst in order to obtain the best model based on those results. In what sense would this (or these) model(s) be optimal?

Question 6 The analyst decides to build a tree with 4 leaves, as illustrated on Figure 2. Compute the confusion matrix of the tree from the figure.

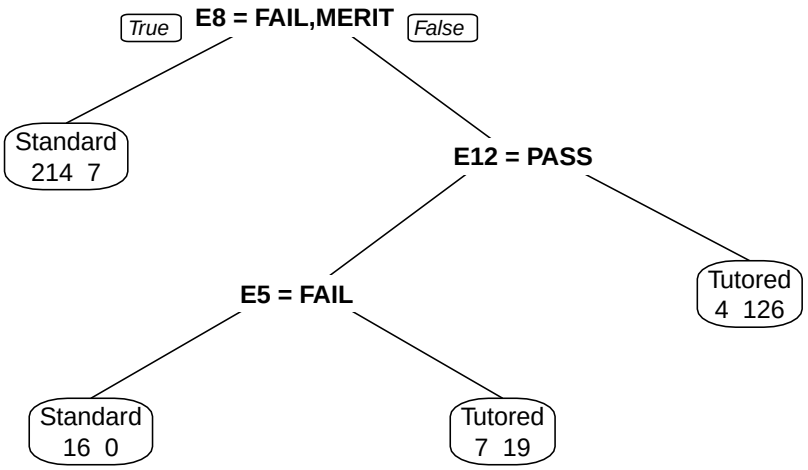

Figure 2: Reduced decision tree. The left hand side branch always corresponds to answering True to the question of the node, while the right hand side branch corresponds to answering False. The first row in each leaf gives the decision associated to the leaf. The second row in each leaf gives the number of students of each group associated to the leaf: the first number corresponds to the Standard group, the second to the Tutored group.

Question 7 Let us consider a student who obtained the following results:

E1	E2	E3	E4	E5	E6	E7	E8
PASS	FAIL	PASS	MERIT	FAIL	PASS	PASS	PASS

E9	E10	E11	E12	E13	E14	E15	E16
FAIL	FAIL	MERIT	PASS	FAIL	FAIL	FAIL	PASS

Computes the group of this student as predicted by the reduced tree.

Question 8 Determine the order in which the nodes of the reduced tree would be pruned by the standard greedy pruning algorithm. Nodes can be referred to using the associated question.

Question 9 Describe briefly a way to estimate the performances of the reduced tree on a future set of students.

Question 10 Assume the students used in the learning set are now tested on 16 new exams with possibly completely different subjects. Could one use the reduced tree to determine the group of those students? What kind of performances should we expect in this case? As in all questions, answers must be justified briefly and precisely.