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 $\mathbf{y} \in \mathcal{Y}$.

Correction

One simply needs to compute the frequencies of three possible values of $\mathbf{Y}$. This is done by summing the values by columns in either of the confusion matrices and then by dividing the quantities by the total number of examples. This gives:

	A	B	C
1	0.35	0.35	0.30

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

Correction

One should verify that the empirical distributions of $\mathbf{Y}$ are comparable on both sets.

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

Correction

The accuracy is computed by summing the diagonal terms and dividing them by the total number of examples. This gives for g_1 0.9189189 and for g_2 0.8972973.

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 $\mathcal{D}'$.

Correction

The best model is here the one with the largest accuracy as we are using the binary loss function. Thus the best model is g_1.

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

Correction

As the model is selected on a validation set, we can expect it to be better than the other model according to the true risk and not only the empirical risk.

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

$l_2(p,v)$		v		
		A	B	C
p	A	0	2	1
	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 $\mathcal{D}'$.

Correction

This is again a simple calculation. One can compute the term by term product of the loss matrix and of the confusion matrix, sum the obtained values and divide them by the total number of examples.

We get g_1 0.0972973 and for g_2 0.1621622. Therefore g_1 is the best model for the loss function l_2.

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 $\mathbf{x} \in \mathcal{X}$.

Correction

As the loss function is the standard binary loss, the optimal decision is obtained by maximizing $\mathbb{P}(\mathbf{Y} = \mathbf{y}|\mathbf{X} = \mathbf{x})$ over $\mathbf{y}$ for a given $\mathbf{x}$. Here, we can compare directly $\mathbb{P}(\mathbf{Y} = 1|\mathbf{X} = \mathbf{x})$ and $\mathbb{P}(\mathbf{Y} = 0|\mathbf{X} = \mathbf{x})$ in the table, which leads to

	X1	X2	X3	g(X)
1	0	0	0	1
2	0	0	1	1
3	0	1	0	0
4	0	1	1	0
5	1	0	0	0
6	1	0	1	0
7	1	1	0	1
8	1	1	1	1

Question 2 Is the optimal model for l_1 unique?

Correction

Based on the table, we can see that there are values of $\mathbf{x}$ for which $\mathbb{P}(\mathbf{Y} = 1|\mathbf{X} = \mathbf{x}) = \frac{1}{2}$. In this situation one can decide that $g_1^*(\mathbf{x}) = 1$ or $g_1^*(\mathbf{x}) = 0$ with no effect on the performances. Thus we can compute several distinct optimal models.

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 $g_{l_\lambda}^*(\mathbf{x}) = 1$.

Correction

$g_{l_\lambda}^*(\mathbf{x}) = 1$ if and only if this decision leads to the smallest possible cost. The expected cost for deciding wrongly that $g_{l_\lambda}^*(\mathbf{x}) = 1$ is $l_\lambda(1,0)\mathbb{P}(\mathbf{Y} = 0|\mathbf{X} = \mathbf{x})$ while it's $l_\lambda(0,1)\mathbb{P}(\mathbf{Y} = 1|\mathbf{X} = \mathbf{x})$ when deciding $g_{l_\lambda}^*(\mathbf{x}) = 0$. Thus $g_{l_\lambda}^*(\mathbf{x}) = 1$ if and only

$$l_\lambda(1,0)\mathbb{P}(\mathbf{Y} = 0|\mathbf{X} = \mathbf{x}) \leq l_\lambda(0,1)\mathbb{P}(\mathbf{Y} = 1|\mathbf{X} = \mathbf{x}).$$

This can be rewritten into

$$\mathbb{P}(\mathbf{Y} = 1|\mathbf{X} = \mathbf{x}) \geq \frac{l_\lambda(1,0)}{l_\lambda(1,0) + l_\lambda(0,1)} = \frac{1}{1+\lambda}.$$

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 $\mathcal{X}$.

Correction

$g_{l_\lambda}^*(\mathbf{x})$ is commuted by comparing $\mathbb{P}(\mathbf{Y} = 1|\mathbf{X} = \mathbf{x})l_\lambda(0,1)$ with $\mathbb{P}(\mathbf{Y} = 0|\mathbf{X} = \mathbf{x})l_\lambda(1,0)$.

When λ grows, the following ratio grows

$$\frac{\mathbb{P}(\mathbf{Y}=1|\mathbf{X}=\mathbf{x})\lambda}{\mathbb{P}(\mathbf{Y}=0|\mathbf{X}=\mathbf{x})}.$$

If the ratio is always larger than 1 regardless of the value of $\mathbf{x}$, then the optimal model is constant with $g^*_{l_\lambda}(\mathbf{x})=1$. If we look at values taken by the ratio for $\lambda=1$ for different values of $\mathbf{x}$, we see that the smallest value is 0.25. Thus the smallest value of λ that ensure the ratio is always higher than 1 is the inverse of 0.25, that is 4.

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

Correction

We have $\mathbb{P}(\mathbf{Y}=1)=\sum_{\mathbf{x}\in\mathcal{X}}\mathbb{P}(\mathbf{Y}=1|\mathbf{X}=\mathbf{x})\mathbb{P}(\mathbf{X}=\mathbf{x})$. As the distribution of $\mathbf{X}$ is uniform on $\mathcal{X}$, we have $\mathbb{P}(\mathbf{Y}=1)=\frac{1}{8}\sum_{\mathbf{x}\in\mathcal{X}}\mathbb{P}(\mathbf{Y}=1|\mathbf{X}=\mathbf{x})$. Thus, to compute $\mathbb{P}(\mathbf{Y}=1)$, one can simply sum the values probabilities in the last column of the table and divide the results by 8. This gives $\mathbb{P}(\mathbf{Y}=1)=0.4875$.

Question 6 Compute the risk of the optimal model g^*_1 (the model computed in question 1).

Correction

The risk is given by

$$\begin{aligned}
R_{l_1}(g^*_1) &= \sum_{\mathbf{x},\mathbf{y}} l_1(g^*_1(\mathbf{x}),\mathbf{y})\mathbb{P}(\mathbf{X}=\mathbf{x},\mathbf{Y}=\mathbf{y}),\\
&= \frac{1}{8}\sum_{\mathbf{x},\mathbf{y}} l_1(g^*_1(\mathbf{x}),\mathbf{y})\mathbb{P}(\mathbf{Y}=\mathbf{y}|\mathbf{X}=\mathbf{x}),\\
&= \frac{1}{8}\sum_{\mathbf{x}} \mathbb{P}(\mathbf{Y}\neq g^*_1(\mathbf{x})|\mathbf{X}=\mathbf{x}).
\end{aligned}$$

Using the the optimal model obtained previously and the conditional probability table, one get the risk 0.275.

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.

Correction

The tree has 25 while the data set contains Sexprnrow(HouseVotes84), thus the average number of observations in each leaf is 15.72. This seems relatively small but not to the point where the tree would be guaranteed to overfit.

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

	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.

Correction

While the tree was small enough to hope for no overfitting, the perfect results obtained on the learning set (0 error) seem to indicate that there might be some overfitting. One cannot rule out the possiblity of facing a very easy classification problem, but an educated guess would be to always suspect overfitting with such a low error. In any case, performances on the learning set are always overestimated and it is very unlikely that the tree would do as well on another data set.

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.

Correction

Cross-validation is a valid method for estimating the true risk of a model. As the tree obtains perfect performances on the learning set, it might be a good candidate for a low risk and it is therefore interesting to perform a cross-validation to get an accurate estimate. The number of folds is on the lower end of the standard range (5 to 10) which seems a good choice according to the small size of the data set. A larger number of folds would increase significantly the variance of the CV estimator coinsdering the small size of said blocs.

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.

Correction

We know that the data set is slighlty imbalanced as there are 241 students in the Standard group and 152 students in the Tutored group. One should therefore use a stratified cross-validation to ensure that those proportions are preserved in the blocs.

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?

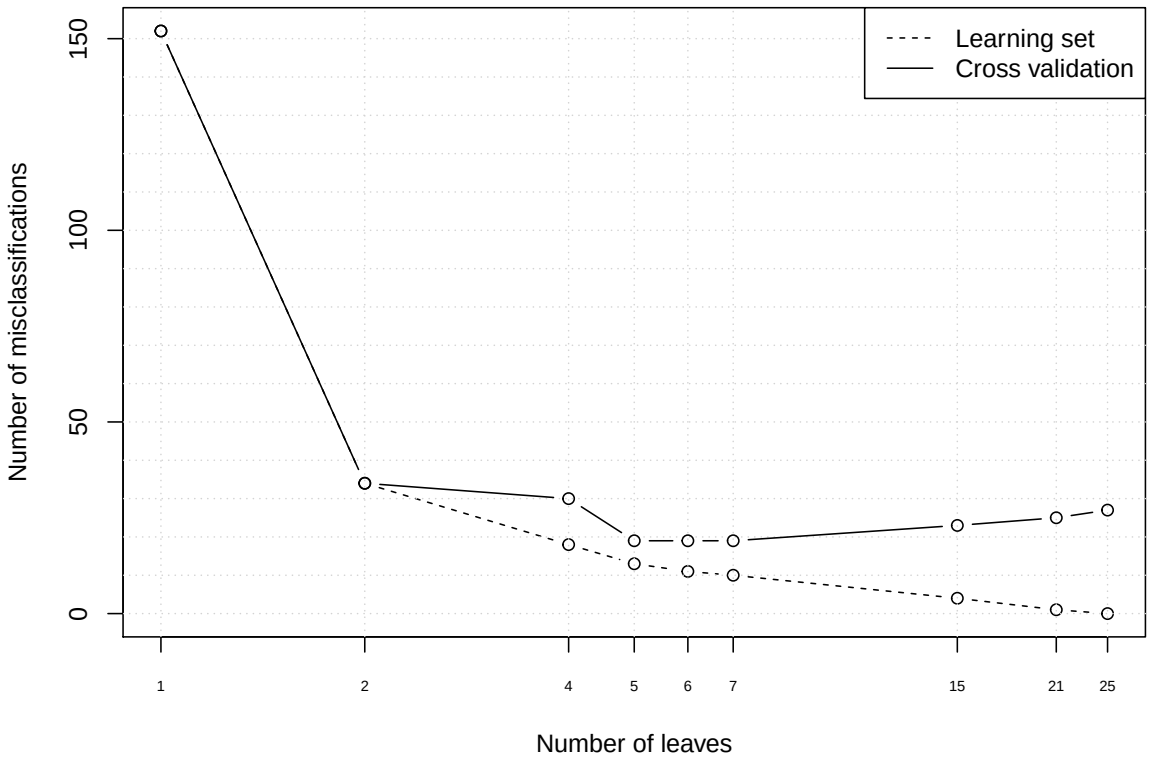

Figure 1: Number of misclassifications as a function of the number of leaves, on the learning set and as estimated by cross-validation. The x axis uses a logarithmic scale.

Correction

As the cross-validation estimator of the risk is valid, one should always retain the model with the lowest CV risk. The model obtained this way is optimal in the sense of the (estimated) risk: it should give the best performances on new data. Here, one can see that trees with 5, 6 and 7 leaves have almost identical CV estimated risks and they could all be retained. In this type of situation, selecting the least complex model is a conservative choice and thus the tree with 5 leaves should be retained.

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.

Correction

		Standard	Tutored
We obtain	Standard	230	7
	Tutored	11	145

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

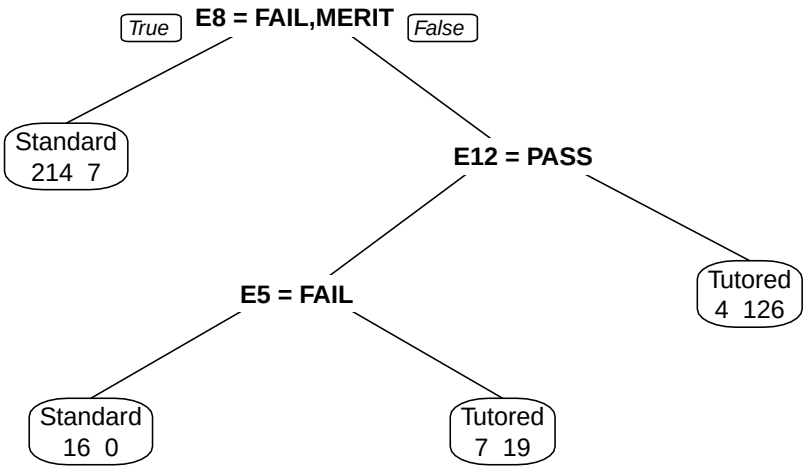

Figure 2: Reduced decision tree. The left hand side branch always corresponds to answering True to the question of the node, while the right hand side branch corresponds to answering False. The first row in each leaf gives the decision associated to the leaf. The second row in each leaf gives the number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

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.

Correction

This student is assigned to group Standard.

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.

Correction

In the standard greedy pruning algorithm, one starts with the node whose replacement by a leaf increases the least the empirical risk. Here, the structure of the tree is such that the only candidate if the node E5=FAIL and thus no particular care has to be exercised. Once E5=FAIL, is pruned the only candidate is E12=PASS and then we end up with a single node.

In fact the question should have been pruned from the exam!

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

Correction

As the tree has been constructed on the full data set, we would need to use another independent data set to evaluate its risk. Another solution would be to split the original data into a learning and a test sets, and then to run cross-validation on the learning set to the select the tree complexity. In a second step, a tree of the selected complexity would be constructed on the learning set and evaluated on the test set.

Notice that because the analyst did not use the cross-validation estimate to select the best number of leaves, the risk estimate it provides for the tree with 4 leaves remains predictive. It gives an idea of the future performances of the tree and it is not biased in the underestimation direction by the optimizing process. So an acceptable estimate of the future performances might be the cv estimated risk. Notice that proceeding this way is dangerous because the actual hand picking of a complexity parameter (here the number of leaves) is an optimization technique and thus deciding whether it induces some risk underestimation is tricky. In the present example, it is unlikely to do so because of the small size of the trees and the relative stability of the risk with the number of leaves. But this might not be the case in other situations.

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.

Correction

As the exams are new and completely different from the previous ones, the students are described by new variables. There is absolutely no reason to believe the tree trained for the original variables can be use with good performances on the new variables. In fact, the tree was able to reach good performances using only 3 exams and one might imagine that those exams are on subjects that benefit most of the tutoring. Thus is the new exams are on unrelated subjects it might even be impossible to separate the tutored students from the other, even by learning a new tree!