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?

Correction

Using *the maximum likehood principle*, the parameter of the Bernoulli distribution of Y is estimated by the frequency of the "success" value. Here $\mathbb{P}(Y = 1) \simeq 0{,}6$.

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

Correction

We estimate again probabilities by frequencies. This gives

$$\mathbb{P}(g_1(X) \neq Y) \simeq 0{,}25,$$
$$\mathbb{P}(g_2(X) \neq Y) \simeq 0{,}35.$$

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?

Correction

The proposed method (risk estimation based on the table) corresponds to empirical risk calculation on the training set. Obviously, the empirical risks are equal to the probabilities of error computed in the previous question, that is

$$\hat{L}_0(g_1) \simeq 0{,}25,$$
$$\hat{L}_0(g_2) \simeq 0{,}35.$$

The main problem with this estimation is that it underestimate the risk and can lead to overfitting. One should use a resampling technique or a least a holdout estimation set.

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.

Correction

We use empirical risk minimization to select the best model.

The empirical risks are

$$\hat{L}_1(g_1) \simeq 0{,}35,$$
$$\hat{L}_1(g_2) \simeq 0{,}5.$$

Therefore the best model is g_1.

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
$\mathbb{P}(X=x\|Y=-1)$	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{2}$
$\mathbb{P}(X=x\|Y=1)$	$\frac{1}{3}$	$\frac{1}{6}$	$\frac{1}{2}$

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.

Correction

The theoretical optimal model minimizing the L_0 risk is given by

$$g_0^*(x) = \begin{cases} 1 & \text{if } \mathbb{P}(Y=-1|X=x) < \mathbb{P}(Y=1|X=x) \\ -1 & \text{if } \mathbb{P}(Y=-1|X=x) \geq \mathbb{P}(Y=1|X=x) \end{cases}$$

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

Correction

As recalled above, we need to compute $\mathbb{P}(Y=y|X=x)$, or, more efficiently, the ratio $\frac{\mathbb{P}(Y=1|X=x)}{\mathbb{P}(Y=-1|X=x)}$. Using the Bayes rule, we know that

$$\mathbb{P}(Y=y|X=x) = \frac{\mathbb{P}(X=x|Y=y)\mathbb{P}(Y=y)}{\mathbb{P}(X=x)},$$

and thus the ratio is given by

$$\begin{aligned} \frac{\mathbb{P}(Y=1|X=x)}{\mathbb{P}(Y=-1|X=x)} &= \frac{\mathbb{P}(X=x|Y=1)\mathbb{P}(Y=1)}{\mathbb{P}(X=x|Y=-1)\mathbb{P}(Y=-1)}, \\ &= \frac{2\mathbb{P}(X=x|Y=1)}{\mathbb{P}(X=x|Y=-1)}, \end{aligned}$$

where the second expression is obtained by using the hypothesis on Y. The following table give the value of the ratio for each value of x, as well as $g_0^*(x)$.

x	a	b	c
$\frac{\mathbb{P}(Y=-1\|X=x)}{\mathbb{P}(Y=1\|X=x)}$	1	4	2
$g_0^*(x)$	-1	-1	-1

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

Correction

We want to compute $L_0(g_0^*) = \mathbb{E}(l_0(g_0^*(X),Y))$. We have

$$\begin{aligned} \mathbb{E}(l_0(g_0^*(X),Y)) &= \sum_{x,y} l_0(g_0^*(x),y))\mathbb{P}(X=x,Y=y), \\ &= \sum_{x} \mathbb{P}(X=x,Y \neq g_0^*(x)), \end{aligned}$$

where the simplification is induced by the properties of l_0.

As $\mathbb{P}(X=x,Y \neq g_0^*(x)) = \mathbb{P}(X=x|Y \neq g_0^*(x))\mathbb{P}(Y \neq g_0^*(x))$, one can compute

$\mathbb{E}(l_0(g_0^*(X),Y))$ using both the (conditional) probabilities from the hypothesis and the optimal model computed above.

This gives $\mathbb{E}(l_0(g_0^*(X),Y) = 0{,}3333333$.

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.

Correction

We follow the exact same reasoning. We know that g_2^* is given by

$$g_2^*(x) = \begin{cases} 1 & \text{if } 2\mathbb{P}(Y=-1|X=x) < \mathbb{P}(Y=1|X=x) \\ -1 & \text{if } 2\mathbb{P}(Y=-1|X=x) \geq \mathbb{P}(Y=1|X=x) \end{cases}$$

Thus we have

x	a	b	c
$\frac{\mathbb{P}(Y=-1\|X=x)}{\mathbb{P}(Y=1\|X=x)}$	1	4	2
$g_2^*(x)$	-1	-1	-1

The computation of L_2^* is essentially similar as L_0^*, with the added costs. One have

$$\begin{aligned} \mathbb{E}(l_2(g_2^*(X),Y)) &= \sum_{x,y} l_2(g_2^*(x),y))\mathbb{P}(X=x,Y=y), \\ &= \sum_{x} l_2(g_2^*(x), -g_2^*(x))\mathbb{P}(X=x,Y=-g_2^*(x)), \end{aligned}$$

using the fact that when $y \neq g_2^*(x)$, then $y = -g_2^*(x)$ because of the only two possible values of Y.

This gives $\mathbb{E}(l_2(g_2^*(X),Y) = 0{,}3333333$.

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.

Correction

We have been given the distribution of Y, therefore we only need the conditional distribution of the X_i given Y. There are given directly by the table is one simply divides each value by 100. This corresponds to a frequency based estimate of e.g. $\mathbb{P}(X_{i,1} = 1|Y_i = A)$ by the fraction

$$\frac{|\{i|Y_i = A \text{ and } X_{i,1} = 1\}|}{|\{i|Y_i = A\}|} = 0{,}85.$$

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 $v = (0,0,1)$.

Correction

As already recalled in the previous exercise, the optimal decision is obtained here by maximizing the a posteriori probability of Y given X, that is:

$$g_0^*(x) = \arg\max_y \mathbb{P}(Y = x|X = x).$$

Using the NB hypothesis, we have

$$\mathbb{P}(Y = x|X = x) = \frac{\mathbb{P}(Y = y)\prod_{j=1}^3 \mathbb{P}(X_j = x_j|Y = y)}{\mathbb{P}(X = x)}$$

For a given x, $\mathbb{P}(X = x)$ is fixed. In addition, the distribution of Y is uniform, thus g_0^* is given by

$$g_0^*(x) = \arg\max_y \prod_{j=1}^3 \mathbb{P}(X_j = x_j|Y = y).$$

As explained above, each of the conditional probability $\mathbb{P}(X_j = x_j|Y = y)$ can be estimated easily from the summary table. Let us first consider $u = (1,0,1)$.

We have

$$\prod_{j=1}^{3} \mathbb{P}(X_j = u_j|Y = A) = \frac{85 \times 26 \times 85}{100^3},$$
$$\prod_{j=1}^{3} \mathbb{P}(X_j = u_j|Y = B) = \frac{38 \times 43 \times 50}{100^3},$$
$$\prod_{j=1}^{3} \mathbb{P}(X_j = u_j|Y = C) = \frac{43 \times 7 \times 8}{100^3},$$

and thus $g_0^*(u) = A$.

Similar calculations give $g_0^*(v) = B$.

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.

Correction

As expected for a full unprunned tree, the results are very good in the sense where the number of classification errors is very limited. Appart from that it seems that the larger class might be easier to separate from the two others than those one from another.

However, because of the possible overfitting, this might not be very meaningful.

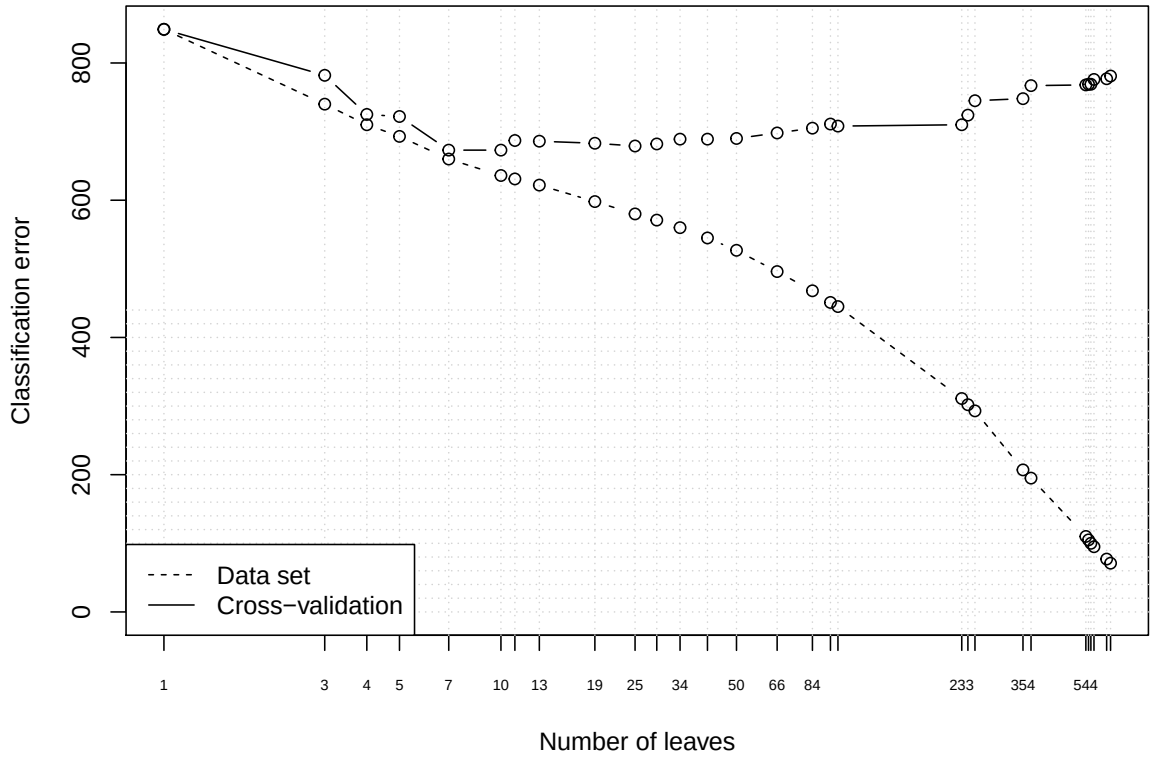

Figure 1: Behavior of the number of misclassified examples as a function of the number of leaves in the tree (during pruning), both on the data set and as estimated by a 10 fold cross-validation method. The x-axis uses a logarithmic scale.

Question 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.

Correction

We observe the standard behavior of learning algorithms. Indeed, the cross-validation estimates are valid estimates of the risk and thus they show the classical increase of the risk when the tree starts to overfit. This happens after a relatively small number of leafs (7). On the contrary, the risk estimated on the training set keeps on decreasing as in a typical overfitting scenario. A very important point is that the classification error is here around 650 over 1473 objects which is quite bad and thus the overfitting shown in the confusion matrix is massive.

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

Correction

This corresponds to the smallest risk as estimated by cross-validation.

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?

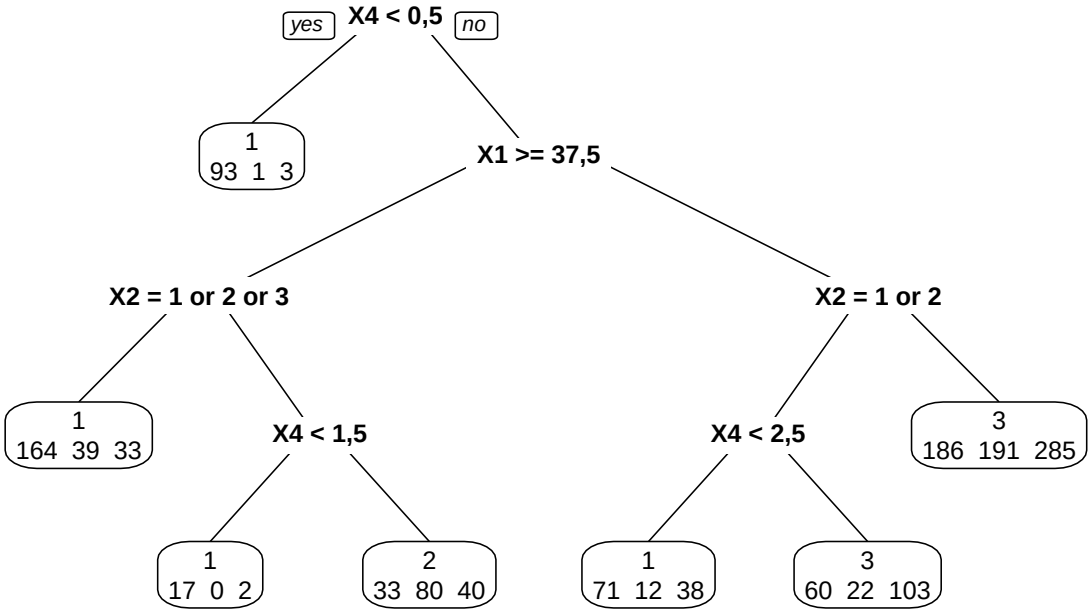

Figure 2: Pruned decision tree. At each node, the left branch corresponds to a "yes" answer to the question of the node, while the right branch correspond to "no". In a leaf, the upper value is the predicted class for the leaf, while the lower values correspond to the number of examples from the data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

Correction

We obtain the following confusion matrix

	1	2	3
1	345	52	76
2	33	80	40
3	246	213	388

The performances are quite low, as expected from the cross-validation estimates. Indeed, before overfitting kicks in the cross-validation risk estimates are very close to the risk on the learning set (on Figure 1) and thus we expect roughly 650 classification errors. Figure 2 gives 660 classification errors and thus results are compatible.

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.

Correction

Nothing complicated here...