Data Science
2019/20

Exam version A
January 13th, 2020
Duration: 2 hours

Rules:

- No consultation allowed. Calculator is not necessary.
- True or false: 50% discount when wrong (e.g. +0.2 correct, –0.1 wrong)
- Delivery just this sheet, with your identification and answers.
- Withdrawals: 1 hour after starting time. Room entries: up to 30 minutes of starting time.

Consider a dataset composed by 350 records, described by 6 variables (B is Boolean), and classified according to the decision tree in Figure 1, except for the records signalled as errors in each leaf of the tree. As you can see, there are 224 records classified as 'good' and 126 classified as 'bad'.
Each leaf in the tree shows the label, the number of records classified with the label, and the number of errors covered in the leaf.

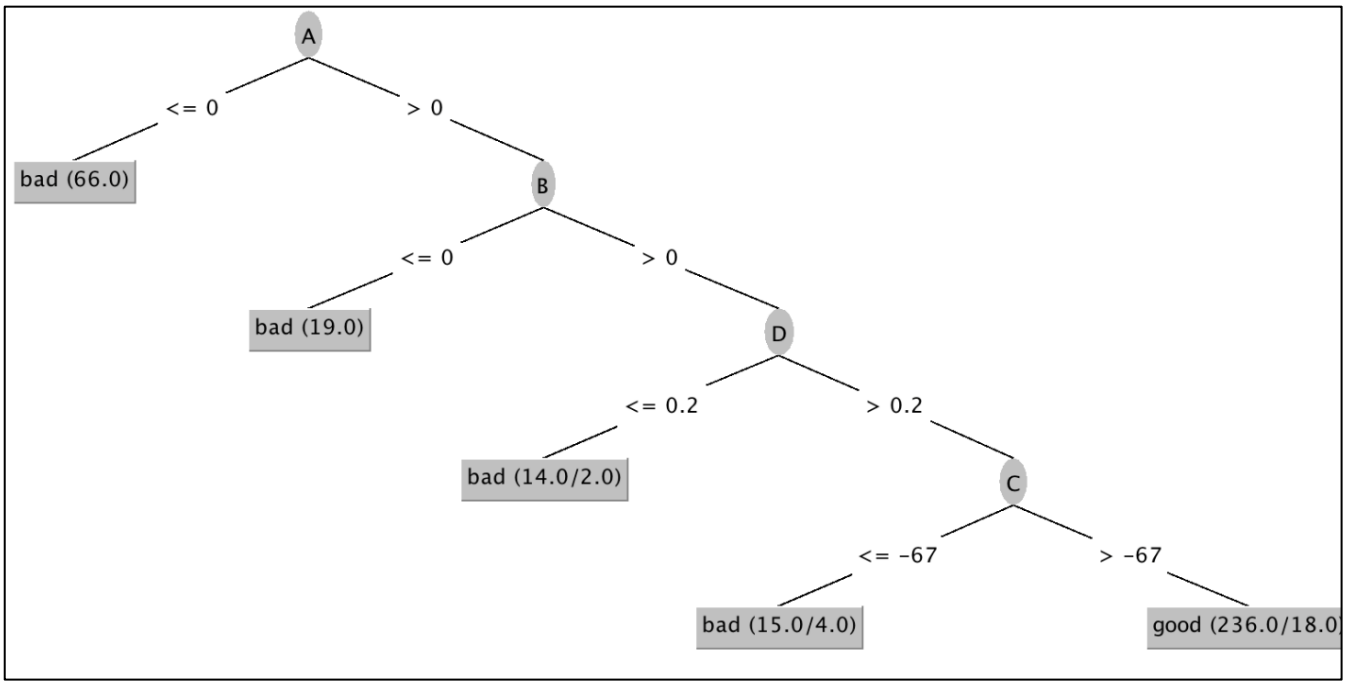

Figure 1 Decision tree trained from the dataset

Figure 2 show the boxplots for each one of the variables.

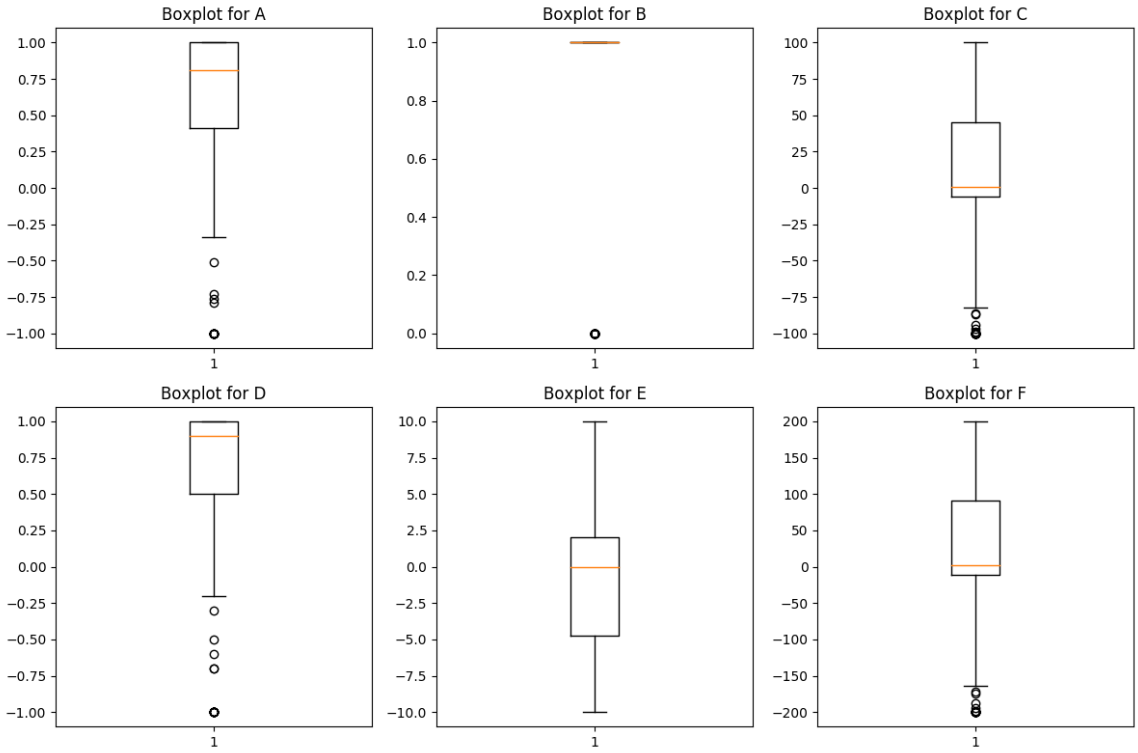

Figure 2 Boxplots for variables A, B, C, D, F and E

Part I

Please mark the following statements as **T**rue or **F**alse on the first sheet. Correct answers count 0.3 each, and wrong ones -0.15.

A. Data exploration and visualization (5 statements – 1.5 v)

1. Boxplots, such as the ones in Figure 2, can be used to understand variables distribution.
2. Those boxplots can prove the correlation between C and F.
3. Variable B is balanced.
4. Variable E shows a high number of outlier values.
5. Scatterplots are more adequate to identify correlation than boxplots.

B. Data preparation (5 statements – 1.5 v)

1. Normalization of this dataset could **not** have impact on a KNN classifier.
2. Discretization, independently of the technique applied, could change the performance of a decision tree trained over this dataset.
3. There is evidence in favour for sequential backward selection to select variable F previously than variable A.
4. The application of PCA generates at most six principal components where each eigenvector is necessarily given by a vector of length six.
5. Knowing that C and F are strongly correlated (correlation=1), we can say that removing one of those variables, would not have any impact on the performance of a KNN classifier.

C. Classifiers Evaluation (5 statements – 1.5 v)

Knowing that we call *positives* to the records belonging to the minority class, and with respect to the decision tree presented in Figure 1.

1. The confusion matrix is

	predicted	
real	108	18
	6	218

2. Sensitivity (also known as recall) is higher than the specificity.
3. According to Occam's razor, any other model with the same accuracy will be considered as good as this one.

Suppose that we trained some trees with different sizes over the same dataset (each is a specialization of the smaller ones), and collected their performance over an independent dataset (test set), illustrated in Figure 3: accuracy on the left and ROC charts on the right.

4. we can say that the tree goes in overfitting for depths above 2.
5. according to the ROC chart the tree that outperforms the others is the one with depth=2.

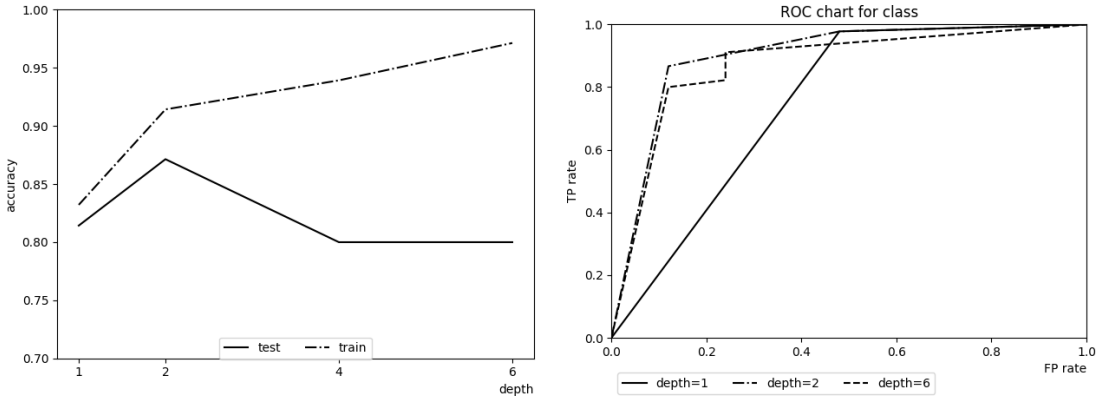

Figure 3 Accuracy and ROC chart for decision trees specializations

D. Biclustering (5 statements - 1.5 v)

Consider a dataset with n=4 observations and m=3 variables and the following biclusters $B_1=(I_1=\{x_2,x_3\},J_1=\{y_1,y_2,y_3\})$ and $B_2=(I_2=\{x_1,x_4\},J_2=\{y_2,y_3\})$, where B_1 is

	y_1	y_2	y_3
x_2	1	2	2
x_3	2	3	3

1. B1 is an additive bicluster with shifting factors $\gamma 2=0$ and $\gamma 3=1$
2. B1 is an order-preserving bicluster
3. B1 has perfect quality (no noise)
4. The biclustering solution {B1, B2} is exclusive on rows and exhaustive on columns
5. Biclusters B1 and B2 have overlapping entries

E. Other topics (5 statements – 1.5 v)

1. A global outlier is an observation inconsistent with its neighbours
2. Given a set of itemset sequences, sequential pattern mining is able to discover both frequent precedencies and co-occurrences
3. A motif is a recurring pattern within a single time series
4. The learning of decision trees can be adequately parallelized using horizontal data partitioning
5. Data stream mining algorithms generally rely on continuously updatable statistics/models

Part II

F. Classification (6 v)

Consider the dataset described before.

1. [0.5 v] What would be the accuracy and sensitivity of the pruned tree, if having just two nodes?

Consider just variables A, B, C and D and that they have been binarized as in each node in the tree above (A≤0, B≤0, C≤-67, D≤0.2):

2. [1.0 v] how many different decision stumps could be trained by the Random Forest technique over the resulting dataset?
3. [1.5 v] how many of those trees would classify the instance (A=-1, B=-1, C=-100, D=-1) as bad? How would the random forest classify the instance if all and only those trees have been trained?

Suppose that we only consider the variables A and B to describe the data:

4. [1.5 v] How would KNN (k=50) classify (**A≤0, B=0**)? From those 50 neighbours, how many are *good*?
5. [1.5 v] How would Naïve Bayes classify a different instance (**A≤0, B=1**)? What are the values for P(A≤0|Bad) and P(A≤0|Good)?

G. Clustering and Time Series Similarity (3.5 v)

Consider the following four time series and the corresponding pairwise distance matrix, where DTW is given by the square root of the alignment cost under a squared loss

	time series
x_1	<0, 0, 2, 1>
x_2	<2, 2, 0>
x_3	<1, 2, 1, 1>
x_4	<0, 3, 1>

DTW	x_1	x_2	x_3	x_4
x_1	0	$\sqrt{9}$=3	$\sqrt{2}$=1.4	**?**
x_2	-	0	$\sqrt{3}$=1.7	$\sqrt{6}$=2.4
x_3	-	-	0	$\sqrt{2}$=1.4
x_4	-	-	-	0

1. [0.7 v] Identify the DTW distance between x_1 and x_4
2. [0.3 v] Identify the DTW alignment path between x_1 and x_4

Assuming $DTW(x_1, x_4)$=1.2, answer the following questions on *clustering*:

3. [1 v] Draw the dendogram under a complete (maximum) link criterion
4. [1 v] Identify the clusters produced using DBSCAN with p=2 and ε=2
5. [0.5v] Given $\{x_1,x_4\}$ and $\{x_2,x_3\}$ clusters, identify the silhouette of x_1

H. Forecasting and Regression (3v)

Consider the following time series with values measured between t_1 and t_5:

$$x =< 11, 9, 7, 6, 5 >$$

1. [0.4v] Considering differencing operations, identify the naïve forecasts on time points t_6 and t_7
2. [0.6v] Considering an ARIMA(p=2,d=1,q=0) where the auto-regressive component is given by
$$x(t)=0.2–x(t-1)+1.5x(t-2),$$
identify the ARIMA forecasts on time points t_6 and t_7

Considering now the decomposition of the time series x where the trend component is given by a linear regression with coefficients β=[14,-2] and no seasonal component:

3. [0.8 v] Identify the irregular /residual component of the time series
Assuming the irregular component to be given by <-4,-2,0,4,6>
4. [0.4 v] Is there some initial evidence for the linear regression forecaster to be biased?
5. [0.8 v] Compute the MAE

Good work!!!

Data Science
2019/20

Exam version A
January 13th, 2020

Student ID: _____ Name: _____

Part I. True or false

Data Exploration		
	T	F
1	X	
2		X
3		X
4		X
5	X	

Data Preparation		
	T	F
1		X
2	X	
3	X	
4	X	
5		X

Classifiers Evaluation		
	T	F
1	X	
2		X
3		X
4	X	
5	X	

Biclustering		
	T	F
1	X	
2	X	
3	X	
4	X	
5		X

Other Topics		
	T	F
1		X
2	X	
3	X	
4		X
5	X	

Part II. Calculus

	Classification		Clustering and time series		Forecasting and regression
1	ac=(66+19+224)/350=88% sen=85/126=67%	1	1.0	1	t_6=4, t_7=3
2	Nr trees = 6	2	$(x_{11},x_{41}),(x_{12},x_{41}),(x_{13},x_{42})(x_{14},x_{43})$	2	t_6=4.7, t_7=3.7
3	4 x bad, 2 x good RF → bad	3	$\{\{x_1,x_4\},x_3\},x_2\}$	3	-1,-1,-1,0,1 or 1,1,1,0,-1
4	KNN → bad Nr. of good = 0	4	$\{x_1,x_3,x_4\}$, $\{x_2\}$	4	Yes (uncaptured trend)
5	NB → bad P(A≤0|bad)=66/126 P(A≤0|Good)=0	5	$1-\frac{1.2}{2.2}$	5	16/5