TÉCNICO LISBOA

Data Science
2019/20

Exam

Version A
January 27th, 2020
Duration: 2 hours

Rules:

- No consultation or calculator use is allowed.
- Delivery just the **last** sheet, with your identification and answers inside the grid.
- Withdrawals: 1 hour after starting time. Room entries: up to 30 minutes of starting time.

Consider a dataset composed by 768 records, described by 8 numeric variables, and classified according to the decision tree in Figure 1, except for the records signalled as errors in each leaf of the tree: there are 268 records classified as 'positive' (P) and 500 classified as 'negative' (N). The tree was trained with C4.5 algorithm, using the information gain criteria.
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.

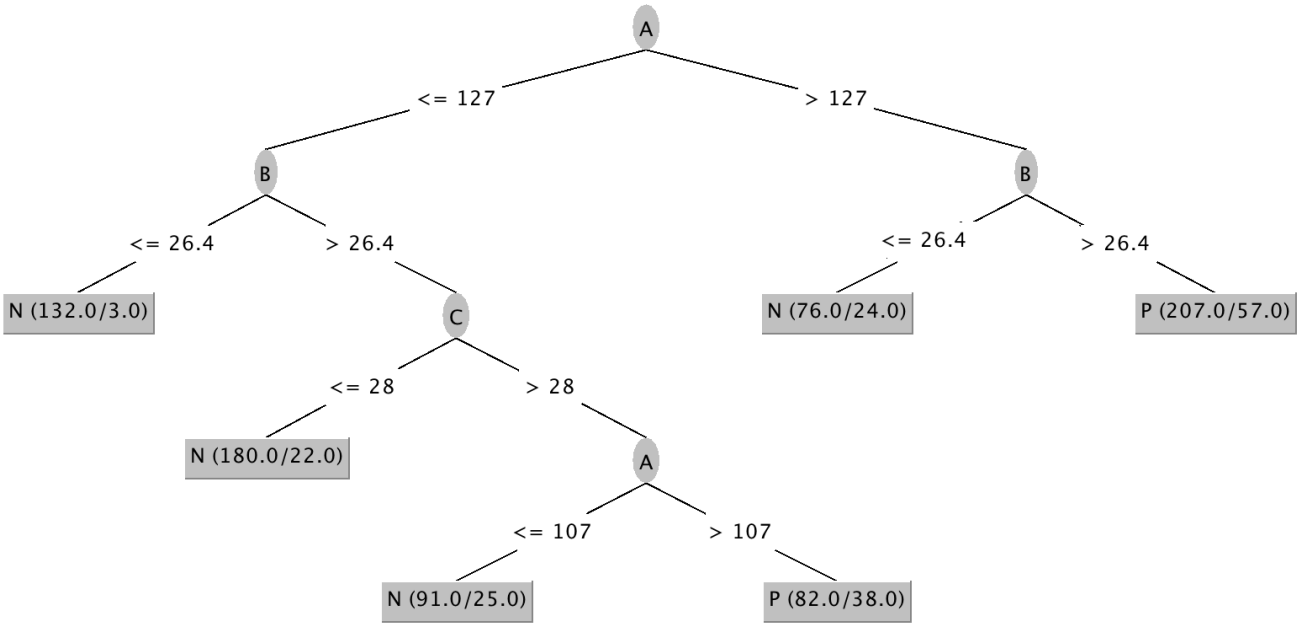


Figure 1 Decision tree trained from the dataset

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

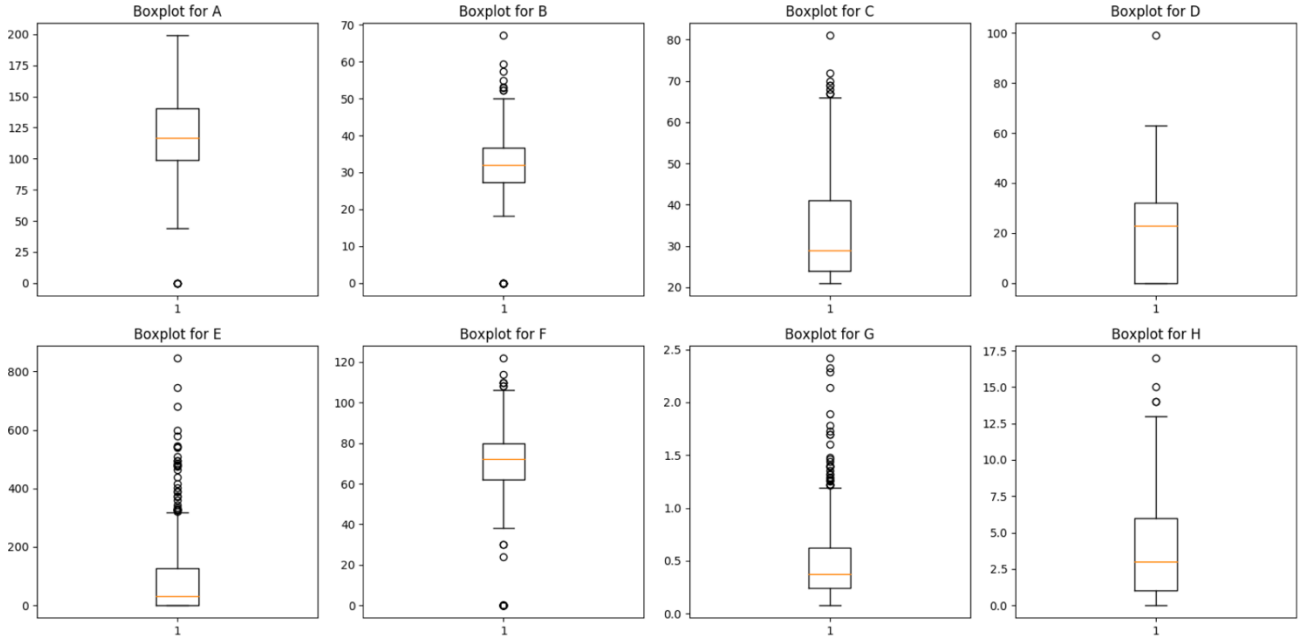


Figure 2 Boxplots for each variables

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. From the boxplots in Figure 2, we can conclude that variables D and H have similar statistical distributions.
2. Those boxplots show that the data is not normalized.
3. Variable C shows some outlier values.
4. Histograms are more adequate than boxplots to identify outlier values.
5. Boxplots are more adequate than histograms to assess if variables are balanced.

B. Data preparation (5 statements – 1.5 v)

1. Normalization of this dataset should have a high impact on naïve Bayes classifier.
2. Missing value imputation using the mean value per class improves the quality of discovered patterns.
3. Minkowski distances between time series disregard temporal dependencies.
4. Classic imputation methods over tabular data are generally adequate to impute missings in time series data.
5. Moving average can be applied to smooth a time series.

C. Complex and Big Data (5 statements – 1.5 v)

1. Spatial auto-correlation suggests that the observations within a given dataset lack independence.
2. Chords and phrases can be mined from univariate time series.
3. An atomset is a frequent pattern in a relational data structure.
4. Association rule mining can be adequately parallelized using horizontal data partitioning.
5. Incremental data mining algorithms generally revisit previous data observations to learn models once new data observations are given.

D. Pattern Mining (5 statements - 1.5 v)

Consider the dataset above, where each variable was binarized, with A≤127, B≤26.4, C≤28, D≤20, E≤75, F≤70, G≤0.5 and H≤50. In this context we consider that item A occurs in each transaction where A assumes a value less or equal than 127, and similar for all the other variables. Consider a minimum support threshold of 10%, and that ABCD, ABCE, BCDE, CEF and CEG are the only patterns found up to the 3rd apriori iteration. Remember that a *pattern* is a maximal frequent itemset.

1. The support for A and B is larger than 25%.
2. BCE is a frequent 3-itemset.
3. The apriori algorithm counts the support for the candidate CEFG.
4. It is possible to know that ABCDE is frequent without counting its support.
5. The rule B=>A is better than the rule A=>B.

E. Social Network Analysis and Ethical Concerns (5 statements – 1.5 v)

1. The centrality of a node in a social network is given as a function of its outdegree.
2. The prestige of a node in a social network is easily fooled / influenced be the node owner.
3. The transition probability from a node *i* to any other node following a random surfing is given by $1/O_i$, with O_i the out-degree of node *i*.
4. The central entities on the GDPR are the companies and the people.
5. According to GDPR, data shall be collected for specified, explicit and legitimate purposes, but it can be further processed under public interest or historical research purposes.

Part II

F. Classification (4 v)

Consider the dataset described, but with binarized variables as before: A≤127, B≤26.4, C≤28, D≤20, E≤75, F≤70, G≤0.5 and H≤50, and selecting the k-best features with k=2 and information gain.
Consider the instance Z=(A≤127, B≤26.4, C≤28, D≤20, E≤75, F≤70, G≤0.5 and H≤50).

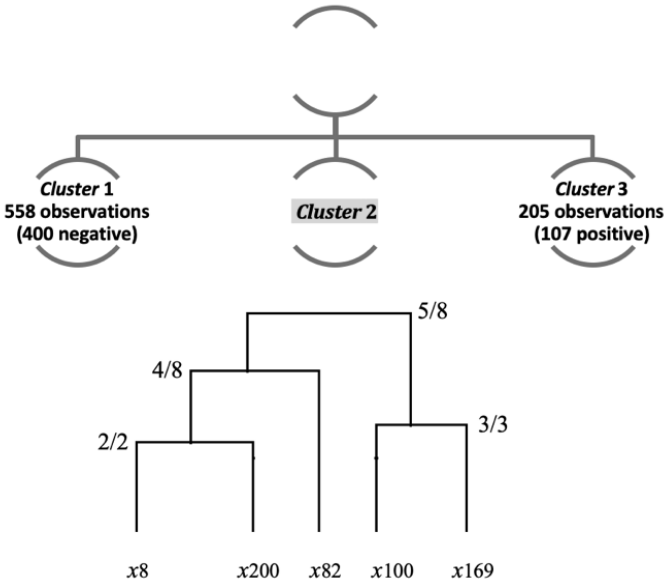
1. [1.0 v] What are the values for P(A≤127|positive) and P(B≤26.4|positive)?
2. [1.0 v] How would Naïve Bayes classify Z?
3. [0.5 v] How would KNN (K=132) classify Z? From those neighbours how many are positive?
4. [1.0 v] Suppose we train a Random Forest after applying the feature selection described. How many **different decision trees** can be trained over the original dataset (without sampling) and under the same conditions?
5. [0.5 v] How would that trained Random Forest classify Z? How many of the estimators would classify it as positive?

G. Classifiers Evaluation (4 v)

1. [1.0 v] Fill the confusion matrix for the original tree.
2. [0.75 v] What is its accuracy and sensitivity?
3. [0.75 v] How many nodes (not considering leaves) would have the best tree resulting from post-pruning the presented one if we accepted some accuracy reduction?
4. [0.75 v] How much is the accuracy loss for the new tree?
5. [0.75 v] Can we identify the existence of overfitting comparing the results from both trees? According to the Occam's razor, which one would be the best model (the original or the new tree)?

H. Clustering (4.5 v)

For the same dataset, A and B variables were selected and a distance metric fixed to produce the dendrogram in the right, with a minimum link criterion, annotated with single and complete linkage distances.



d	*x*8	*x*200	*x*82	*x*100	*x*169
*x*8		2	4	8	7
*x*200			?	6	8
*x*82				6	?
*x*100					3
*x*169					

1. [0.8v] Identify the missing entries in the pairwise distance matrix.
2. [1.2 v] Assuming k-medoids is applied to group observations within cluster 2 with x8 and x200 as starting medoids, m_1 and m_2 respectively. Assume d(x200, x82)=d(x82, x169)=6. Identify the sub clusters, sub_1 and sub_2, and the candidate medoids, m_1 and m_2, after one iteration.
3. [1.0 v] Consider the application of PCA, producing the eigenvectors μ_1=[0.2, 0.4] and μ_2=[0.3, 0.3] (after normalizing A and B, and ignoring the fact $\|\boldsymbol{\mu}_i\|$=1). Eigenvectors μ1 and μ2 were then applied to transform an observation x_{new} into x'_{new}=[2,-1]. Please recover the original values of x_{new}.

A new output variable *out* was identified, and kNN with k=1 (under a leave-one-out training strategy and the given distances) was applied to produce the estimates in the table below. Again, assume d(x200, x82)=d(x82, x169)=6.

	out	$\widehat{out}$
*x*8	?	4
*x*200	?	3
*x*82	6	?
*x*100	?	2
*x*169	?	5

4. [1.0 v] Identify the missing estimate $\widehat{out}82$ and the residuals for each one of the observations.
5. [0.5v] Compute the RMSE.

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	

Complex and Big Data		
	T	F
1	X	
2		X
3	X	
4	X	
5		X

Pattern Mining		
	T	F
1	X	
2	X	
3		X
4		X
5	X	

SNA and Ethics		
	T	F
1	X	
2		X
3	X	
4		X
5	X	

Part II. Calculus

	Classification		Classifiers Evaluation		Clustering and Regression
1	P(A≤127|P)= 94/268 P(B≤26.4|P)=27/268	1	predicted P N 194 74 P 95 405 N real	1	d(x_{200}, x_{82})=8 d(x_{82}, x_{169})=5
2	NB_Z→Negative	2	Accuracy= 599/768 = 78% Sensitivity= 194/268=72%	2	sub_1={x8,x82,x169} sub_2={x100,x200} m_1=x82 $m_{2=}$
3	KNN_Z→ Negative Nr.of P=3	3	Nr. of nodes= 2	3	x_{new}=[0.2*2+0.3*-1, 0.4*2+0.3*-1] = [0.1, 0.5]
4	Nr.of Trees=2	4	Loss= (599-593)/768 = 6/658 < 1%	4	$\widehat{out}82=3$ residuals=(-1,1,3,3,-3)
5	RF_Z→ Negative Nr Positive →0	5	Overfitting? →No Best → new tree	5	RMSE =SQRT(1/5*(1+1+3*3+3*3+3*3)) = SQRT(29/5)