Data Science

2020/21

Exam 1

January 19th, 2021

Duration: 2 hours

Student ID: __________ Name: ______________________________

Rules:

- No consultation or calculator use is allowed.
- Delivery just the **this** sheet, with your identification and answers inside the grid.
- Withdrawals: 1 hour after starting time. Room entries: up to 30 minutes of starting time.
- Each group counts at most 2 and at least 0 points. Each correct answer counts 0.4 points and each wrong one counts -0.2.

Solution

Data Profiling

	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

Classification

	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	

Clustering

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

Time Series

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

SNA

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

Ethics

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

Deloitte Case Study

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

Data Description

Consider the problem of diagnosing arrhythmia in patients, through the use of a dataset with 452 medical records, described by 250 variables. One of these variables, call it `Z`, contains the type of arrhythmia detected in each positive patient, and 0 if the problem was not diagnosed. From it, the variable `class` was derived assuming the value `regular` whenever `Z=0` (245) and `abnormal` (207) otherwise.

The tree on the left was learned through the C4.5 algorithm and the information gain criteria, when applied over 302 of the 452 records available, and the target variable `class`, after applying some preparation techniques.

The tree was printed through `sklearn.tree` package. Each node in the tree shows the variable tested, the number of records satisfying the branch conditions, the number of records from `regular` and `abnormal` classes, respectively, and the label predicted by the tree. For example, the leaf on the left covers 58 records, where 10 are `regular` and 48 are `abnormal`, and the tree classifies records in the branch as `abnormal`.

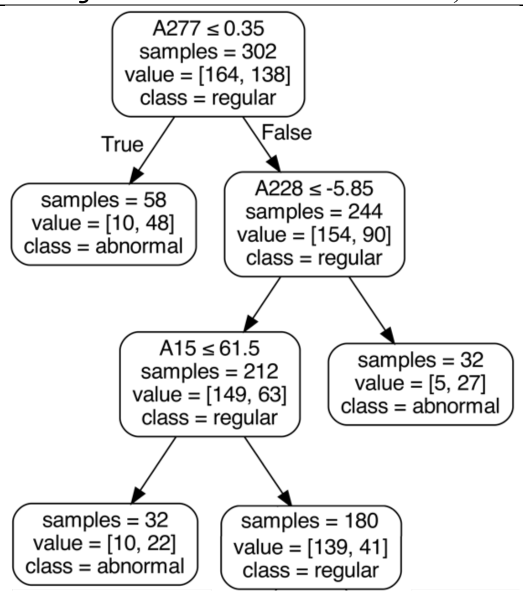

Figure 1 Decision Tree trained over 302 records

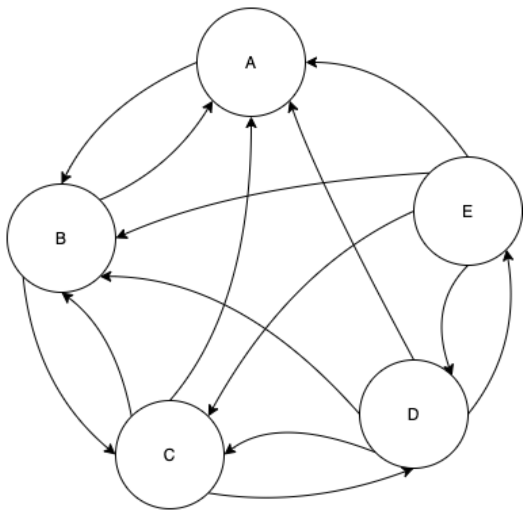

Figure 2 Social network

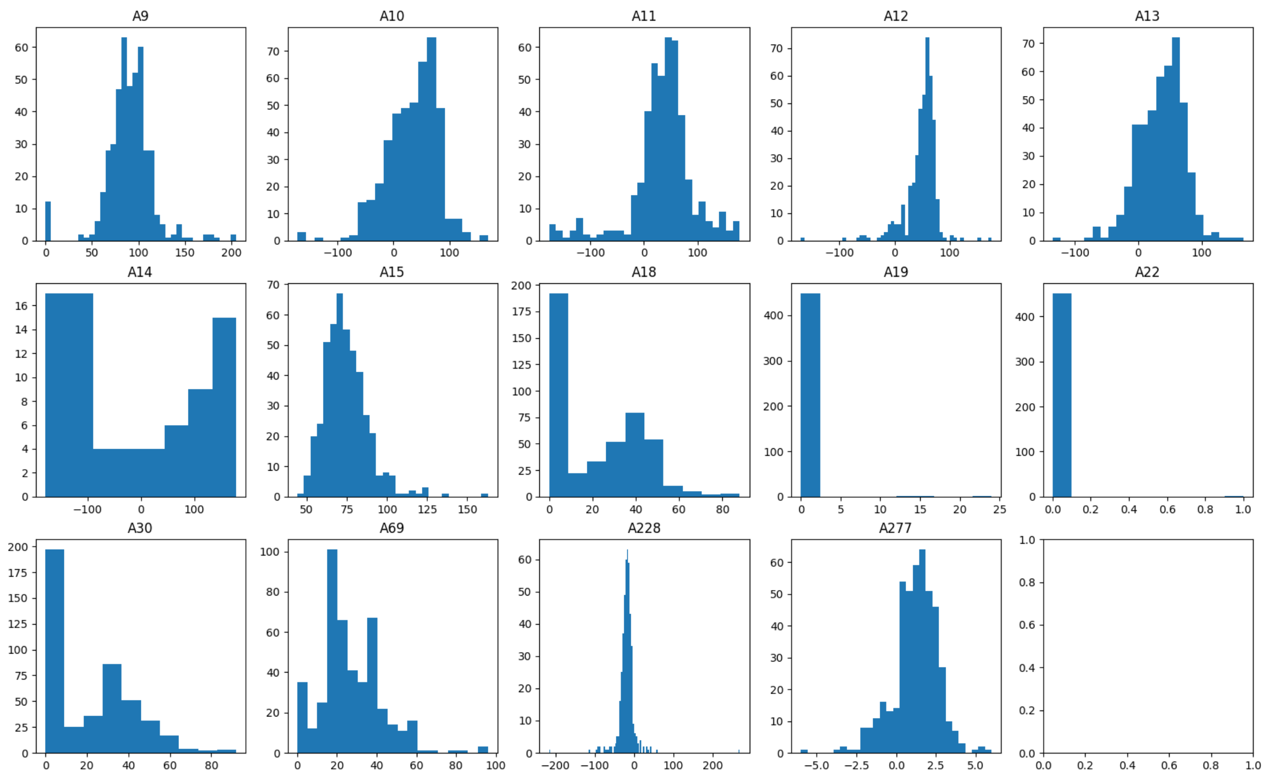

Figure 3 Histograms for some variables

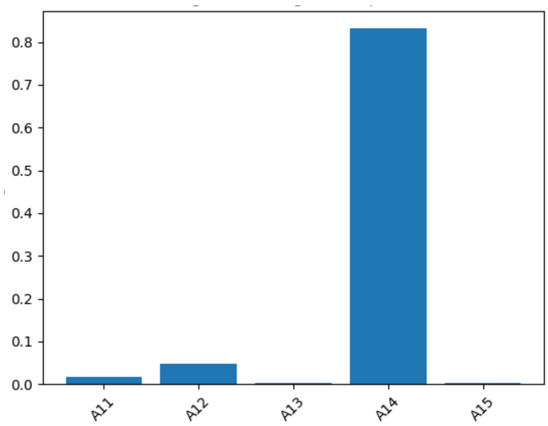

Figure 4 Variables with Missing Values

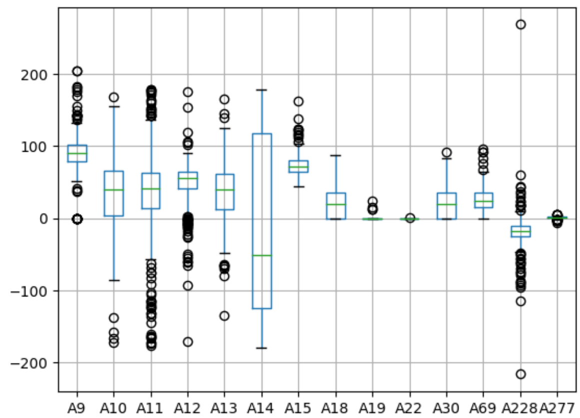

Figure 5 Boxplots for some variables

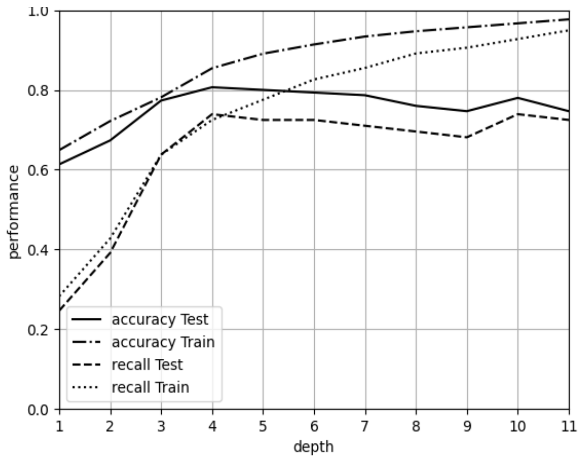

Figure 6 Performance of different decision trees specializations

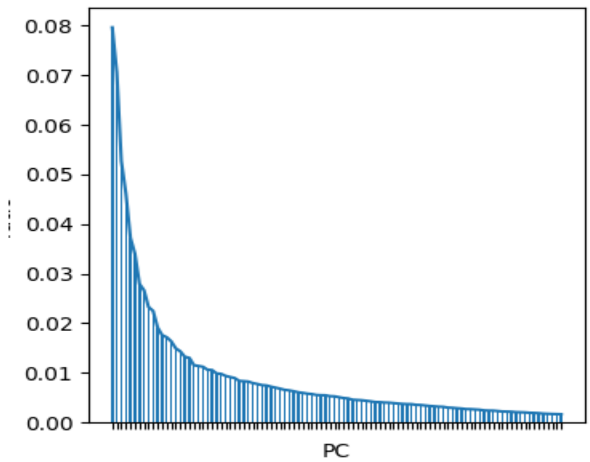

Figure 7 Explained variance for the first 100 principal components

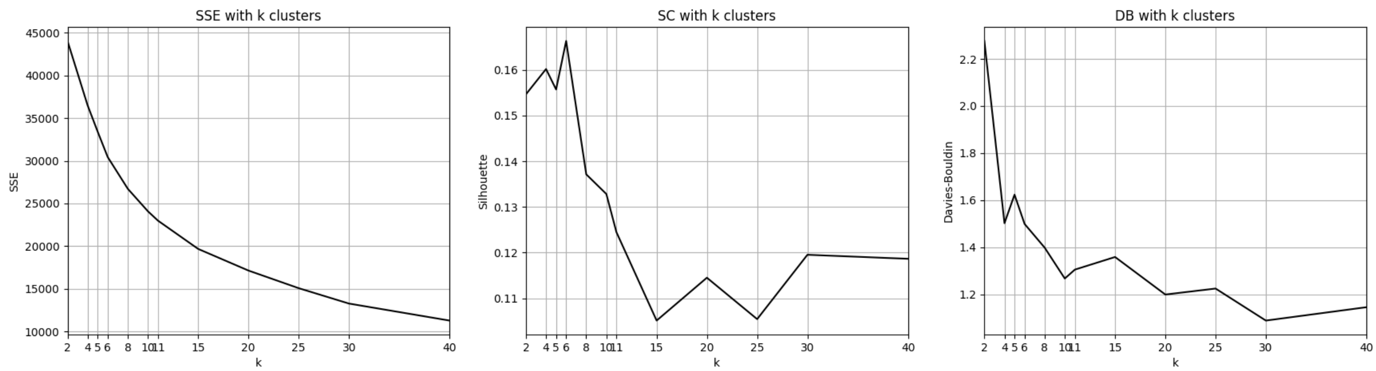

Figure 8 Algorithm performance for different number of clusters

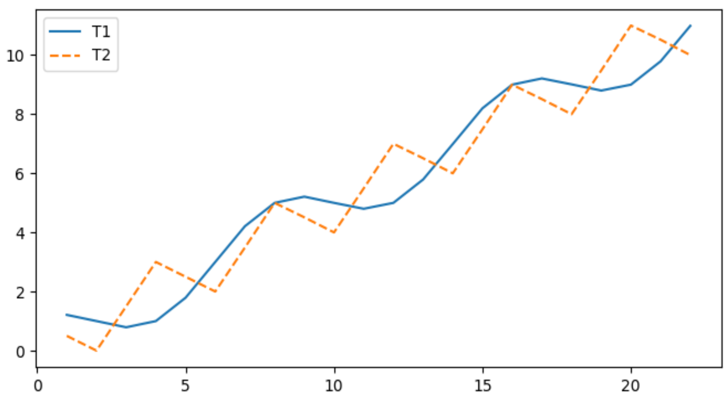

Figure 9 Time series

9.8	25.3	19.3	10.2	3.2	3.1
9	16.0	11.5	5.9	2.4	3.9
4.8	7.5	4.5	2.4	6.6	12.3
3	3.2	1.7	4.0	10.0	17.5
1.2	0.7	1.5	5.8	13.6	22.9
	0.5	2	5.5	9	10.5

Figure 10 Accumulated cost matrix between T1 and T2

A. Data Profiling

1. We face the curse of dimensionality when training a classifier with this dataset.
2. Variable Z is a false predictor.
3. Variables A19 and A22 are redundant, but we can't say the same for the pair A18 and A30.
4. Figure 4 doesn't show any missing values for A9, but these may be hidden as some non-pre-identified value.
5. Variables A14 and A228 seem to be useful for classification and clustering tasks.

B. Data preparation

1. It is better to drop the variable A14 than removing all records with missing values.
2. Dummifying the variables will improve the mining results.
3. Removing the Z variable from the training will improve model performance over any non-observed records.
4. The first 10 principal components are enough for explaining half the data variance.
5. Feature generation based on both variables A9 and A19 seems to be promising.

C. Classifiers Evaluation

Consider the original dataset and the presented tree and the chart on Figure 6, reporting the accuracy and recall collected for different decision trees, trained with some algorithm with different pre-pruning requirements based on the maximum depth of the trees learned.

1. The number of True Positives is higher than the number of True Negatives for the presented tree.
2. The number of False Positives reported in the same tree is 25.
3. The recall for the same tree is less than 70%
4. We are able to identify the existence of overfitting for models with less than 4 nodes of depth.
5. The difference between recall and accuracy becomes smaller with the depth due to the overfitting phenomenon.

For the following two groups (D and E), consider the dataset above, now described by **all** variables **binarized** by computing each value that maximizes the **information gain** for each variable (as done with C4.5). For example, the item **A15** is supported in every record where **A15≤61.5**. (**~A15** represents the records where **A15>61.5**).

D. Classification

Suppose we apply a k-best feature selection, with k=3 and information gain.

1. We have enough information to say that **A15** was one of the selected variables.
2. The number of different decision trees trained over this dataset after applying the feature selection and using bootstrap resampling (usually used for training random forests) would be smaller than 100.
3. A random forest classifier trained under those conditions, and with some repeated models can show a better performance.
4. The binarization and feature selection reduced the diversity of the models in the ensemble, when compared to the diversity obtained in the original dataset.
5. Suppose A15, A228 and A277 are the selected features, KNN with K=1 classifies (**A15, A228, ~A277**) as abnormal.

E. Pattern Mining

Consider 10% as the minimum support threshold.

1. **A277** is frequent, but we can't be sure if the same happens to **A15** and **A228**.
2. **(A15, A228, ~A277)** is a frequent 3-itemset.
3. **~A277⇨A228** presents a confidence higher than 80%
4. **~A277⇨A228** presents a lift smaller then 0.05.
5. If any of **A15, A228** and **A277** were not frequent then **(A15, A228, A277)** would also not be frequent.

F. Clustering

Consider the given dataset again. A clustering algorithm was used to cluster the data, producing the results in Figure 8, reporting the sum of squared errors (SSE), silhouette coefficient (SC) and Davies-Bauldin (BD) metrics for each number of clusters.

1. The elbow-method should be used over all three charts to choose the best model.
2. According to SSE the best results are for 40 clusters.
3. According to DB the best results are for 30 clusters.
4. According to SC, the result with 4 clusters presents a good partition of the data.
5. One of the possible reasons to reach those results is the existence of several redundant variables.

G. Time Series

Consider the time series represented in Figure 9. (Consider only the following points for the required computations: T1=[1.2; 3; 4.8; 9; 9.8] and T2=[0.5; 2; 5.5; 9; 10.5]).

1. The time series T1 is stationary.
2. At a lower granularity (say twice) T2 would be stationary.
3. The dynamic time warping path between T1 and T2 is (1,1)(2,2)(3,3)(4,4)(5,5).
4. If T2 were the prediction of T1 according some regression model, its MAE would be between 0.6 and 0.7.
5. Applying a smoothing average to both T1 and T2 would reduce the distance between them.

H. Social Network Analysis

Consider the social network presented in Figure 2.

1. Node A is more central than node E.
2. Node A is more prestigious than node E.
3. The diameter of this network is 4 edges.
4. The smallest path from node A to node E is 4 edges.
5. If a new node only reachable from E and with no out link were added to the net, E's prestige rank would be higher than on the original network.

I. Ethical Concerns

Consider the original dataset, and suppose the described data controller is an hospital, following and treating people, who may suffer from arrhythmia. Additionally, consider a second and third institutions, M and R, that acquired the data for marketing and cancer research purposes, respectively.

1. In the GDPR context, the purpose limitation principle would be violated by institution M.
2. Institution R may legally process the data since it would be to protect the vital interests of natural people.
3. Anonymizing the data would be enough to make its processing legal by institution M.
4. 'Practice ethical data sharing' encloses privacy protection.
5. Any legal data processing would be ethical.

J. Deloitte Case Study (5 statements – 1.5 v)

Consider the data provided by Deloitte discussed in the Data Science classes.

1. The data available is described mostly by numerical variables.
2. The target variable was one of the provided variables.
3. The data temporality was explored in order to label the data.
4. The data was balanced not requiring the application of additional balancing techniques.
5. Using the data available is possible to define a social network, in order to study the influence of other subscribers in the churning process.

Good work!!!