TÉCNICO LISBOA

Data Science 2018/19

Exam A & B

January 28th, 2019
Duration: 2 hours

- No consultation allowed. **Mobile phones and calculatorsare forbidden**.
- Guarantee correct **identification** in 1st sheet (and provide your student card on the table).
- **Only this sheet is delivered and assessed**!
- Withdrawals: 1 hour after starting time. Room entries: up to 30 minutes of starting time.

Student ID: _____ **Name:** _____

III. Calculus(write the results inside the provided space)

Classification	
1	C4.5 and CART, since ID3 does not stop with those two levels
2	3x2=6
3	N, P(A|P)=1, P(A|N)=1/3
4	No, iclassifies in the other class
5	No, there are several neighbors

Clustering	
1	1/4 and 1/2
2	3/4 and 1/2
3	{x1,x3,x4},{x2}
4	2.5
5	5/4

Pattern Mining	
1	4, there are no 5-candidates
2	A
3	7% (the support of one of the patterns)
4	Yes, it is a subsequence of (B**CE**)(A**C**DE)(ABD**E**)

I. True or false (choose the right column)

Classification		
	T	F
1		x
2		x
3	x	
4		x
5		x

Pattern mining		
	T	F
6		x
7		x
8		x
9		x
10		x

Clustering		
	T	F
11		x
12	x	
13	x	
14	x	
15		x

Data reduction		
	T	F
16	x	
17	x	
18		x
19	x	
20		x

Regression		
	T	F
1		x
2		x
3		x
4	x	
5	x	

Time series		
	T	F
6		x
7		x
8	x	
9	x	
10		x

Pre-proc		
	T	F
11	x	
12	x	
13		x
14	x	
15	x	

Complex min.		
	T	F
16	x	
17	x	
18		x
19		x
20		x

II. Multiple choice(choose the **only** right answer)

Classification				
	A	B	C	D
1	x			
2			x	
3			x	

Clustering				
	A	B	C	D
1			x	
2				x
3			x	

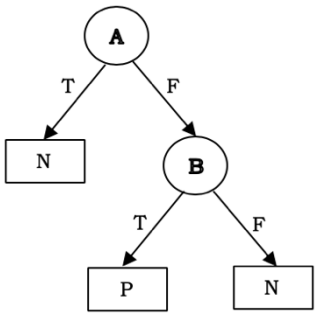
Others				
	A	B	C	D
1		x		
2	x			
3			x	

I. Calculus (14 questions = 8.85 v)

Group I. Classification [3.6v]

Consider the following dataset (note that the shadowed cells are contradictory among them).

A	F	F	F	F	F	F	F	F	T	T	T	T	T	T	T	T
B	F	F	F	F	T	T	T	T	F	F	F	F	T	T	T	T
C	F	F	T	T	F	F	T	T	F	F	T	T	F	F	T	T
Class	N	N	N	N	N	N	N	N	N	N	N	P	P	N	P	P



1. [0.5v] With which of the studied algorithm(s) is it possible to learn the decision tree in the figure?
2. [1.0v] Consider a random forest trainedon the given dataset and using C4.5 to train regular decision trees (trees with three nodes: the root and two siblings testing the same attribute). How many different classifiers would be trained?
3. [1.2v] How does naïve Bayes classify the instance (A=T,B=T,C=F)? Present the values for P(A=T|P) and P(A=T|N).
4. [0.5v] Does 1-NN classify the same instance correctly using the leave-one-out strategy?
5. [0.4v] And without removing the instance, is it possible to classify it?

Group II. Regression and Clustering [3.15v]

Consider the dataset below and answer the questions:

	y1	y2	*cluster*	*class*	*z*
x1	1	2	C1	A	5.5
x2	3	4	C1	A	9.0
x3	0	1	C1	B	5.0
x4	1	0	C2	B	6.5

1. [0.8v] Given the clusters C1 and C2 in the dataset, what is the silhouette index for x1 using Chebyshev distance? And Manhattan distance?
2. [0.85v] Given the ground truth (*class*), how much is the purity of the clustering solution (*cluster*)? And the rand index?
3. [0.6v] Considering uniquely attribute y1, separate observations in two clusters using agglomerative clustering with maximum link.
4. [0.6v] What is the *residual* value associated with observation x1 assuming amultiple linear regression with $\hat{\boldsymbol{\beta}}=[\hat{\beta}_0=3, \hat{\beta}_1=1, \hat{\beta}_2=2]$.
5. [0.4v] Assuming a constant regression model $\hat{z}=6$, calculate the MAE.

Group III. Pattern Mining [2.1v]

Given a transaction dataset, where the only patterns are (ABDE) s=5%, (BCE) s=7% and (ACDE) s=9%.

1. [0.3v] How many times does Apriori algorithm scans the dataset?
2. [0.7v]For which 3-candidates, the Apriori algorithm counts the support for...

A. All generated candidates	B. Only for the proper maximal subsets of each 4-pattern	C. It's not possible to answer

3. [0.8v] Consider that the three patterns are frequent itemsets found from a set of sequences, with the corresponding supports. What is the possible maximum support for the sequence **(AB)(BCE)D**?
4. [0.3v] In the same conditions, can **(CE)CE** be a frequent sequence?

II. True or False(40 statements = 8v)

Please mark the following statements as **T**rue or **F**alse (+0.2 correct, –0.1 wrong):

Group I. Classification

1. Given an unbalanced dataset, a classifier with 90% testing accuracy is always more useful than a classifier with 80% testing accuracy.
2. Naive Bayes is a linear classifier (linear boundaries to separate observations).
3. All attributes are equallyimportant in Naïve Bayes.
4. Normalizing attributes improve the performance of any classifier.
5. Feature selection does not usually improve the performance of KNN.

Group II. Patternmining and Biclustering

1. The anti-monotonic property states that supersets of a frequent itemset are infrequent.
2. The lift measure of an association rule A⇒B does not change if we add a new transaction that does not contain either A or B.
3. The pattern AB(A,B)AC cannot be discovered by PrefixSpan.
4. Given a dataset and a bicluster in it, a false positive bicluster is a statistically significant one that was not found.
5. A biclustering solution with 2 biclusters with overlapping cells is always non-exhaustive on rows and columns.

Group III. Clustering

1. The sum of diagonal entries in a pairwise distance matrix equals one.
2. When pairs of observations are known to belong to the same cluster, we face a semi-supervised clustering task.
3. In k-medoids, the centroid is given by the mode of categorical attributes and median of numerical attributes.
4. Agglomerative clustering algorithms allow to decide the number of clusters after clustering is done.
5. k-means does not adequately identify spherical groups of observations.

Group IV. Data reduction

1. Feature selection can be applied supervisedly with a numeric output variable.
2. Principal component analysis is a centered singular value decomposition.
3. The largest eigenvector of the covariance matrix is the direction of minimum variance in the data.
4. The generalized forward subset selection greedily adds a fixed number of features per iteration that most improves cross-validation accuracy.
5. Given a *m*-dimensional dataset, PCA can reconstruct any data point using m-1 components of PCA with zero reconstruction error.

Group V. Regression

1. The higher the covariance (in absolute value) between two attributes, the higher their correlation.
2. A linear regression with more than one dependent variable is called multiple linear regression.
3. Cross-validation can be applied to minimize the overfitting propensity of a regression model.
4. A logistic regression model is a classification model.
5. By minimizing regression coefficients, Lasso estimation is useful to discard attributes that do not help to estimate the output variable.

Group VI. Time series data

1. When applying the DFT, the sampling rate at which a signal is measured needs to be high enough to adequately model low frequencies.
2. While DWT offers a temporal-based decomposition of a signal, a short DFT (SDFT) strictly offers a frequency-based decomposition.
3. Contrasting with SAX, codebooks encode motifs discovered using multiple temporal-resolutions.
4. In addition to seasonal variation, time series can be described by a cyclical variation component.
5. A time series has a linear trend if the *p*-value of Dickey-Fuller test is low (typically less than 0.01).

Group VII. Pre-processing

1. An outlier can be either inconsistent with the remaining data or with its neighbourhood.
2. Clustering can be applied to perform outlier analysis and subsampling.
3. When assessing a classifier, the imputation of missing values should be always performed prior to cross-fold validation as long as imputation does not depend on the class.
4. In proximity-based approaches for outlier analysis, either outliers have distant nearest neighbours or density around outliers differs from density around neighbours.
5. Normalizing attributes **can** affect the outcome of an equal-width/range discretization.

Group VIII. Complex data mining

1. Minkowski distances are more adequate than DTW for time series clustering if we do not want to tolerate temporal misalignments.
2. Associative classifiers for spatiotemporal data rely on spatiotemporal pattern mining.
3. Chords and phrases are temporal patterns for univariate time series data.
4. Horizontal data partitioning principles can be used to distribute the learning of decision trees since the information gain of each data attribute is independently tested.
5. *k*NN is always able to efficiently and incrementally learn from data streams as long as *k* is less than the number of simultaneously arriving observations.

III. Multiple Choice (9 questions 0.35 each = 3.15v)

Select the **only true answer** (just one). Wrong answers discount half the grade.

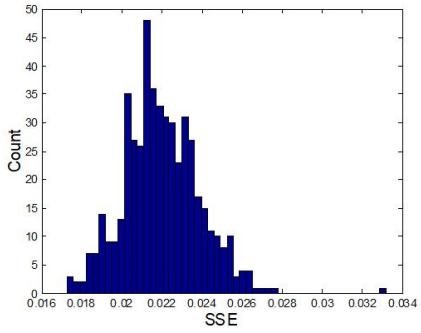
Group I. Classification

1. In which of the following scenario a gain ratio is preferred over Information Gain?
   a. When a categorical variable has high cardinality
   b. When a categorical variable has low cardinality
   c. Numeric variables
   d. None of above

2. A random forest with low training error is getting abnormally bad performance on the validation set. What could be causing the problem?
   a. Decision trees are too weak
   b. Randomly sampling too few features when choosing a split
   c. Too few trees in the ensemble
   d. None of above

3. What does it mean to perform a data bootstrap?
   a. To sample m features with replacement from the total m
   b. To sample m features without replacement from the total m
   c. To sample n examples with replacement from the total n
   d. To sample n examples without replacement from the total n

Group II. Clustering



1. Consider the following analysis of sum squared errors gathered from a thousand of randomized datasets using *k*-means:
   a. A SSE in [0.02,0.023] is statistically significant
   b. A SSE above 0.34 is statistically significant
   c. A SSE below 0.017 is statistically significant
   d. None of above

2. Which of the following is **not** applicable to the *k*-Means algorithm:
   a. Dependent on good initialization/seeding
   b. Sensitive to outliers and noisy data
   c. Not suitable to discover clusters with non-convex shapes
   d. Not able to separate clusters when their variance is small in all directions

3. When we are interested in generating overlapping cluster membership, we should use:
   a. *k*-medoids clustering
   b. Hierarchical clustering
   c. Model-based clustering
   d. Density-based clustering

Group III. Others

1. Which of the following factors does **not** contribute to increase the average size of biclusters:
   a. Increasing tolerance to noise
   b. Given perfect quality, increasing the cardinality of attributes in discrete data
   c. Looser coherence strength (higher deviations allowed) in real-valued data
   d. More flexible coherence assumptions (e.g. choosing additive instead of constant assumption)

2. Which of the following is **not** a typical property of smoothed regression models (such as a multiple linear regression model):
   a. Small training error
   b. Low overfitting risk
   c. Low complexity
   d. Interpretability

3. Given the already normalized time series x1=<-2,0,2> and x2=<2,1>, select the correct answer:
   a. PAA representation of x1 with length 2 is <-1,1>
   b. SAX representation of x2 using two symbols is <b, a>
   c. DTW with Manhattan loss between x1 and x2 is 6
   d. The number of DTW alignments between x1 and x2 is 4