Data Science 2018/19

Exam

January 14th, 2019
Duration: 2 hours

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

Student ID: _____ **Name:** _____

I. True or false (write T or F on cells below)

Classification	
1	F
2	T
3	F
4	T
5	T

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

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

BiClustering	
16	T
17	T
18	T
19	T
20	F

Data reduction	
21	F
22	T
23	T
24	F
25	F

Regression	
26	F
27	T
28	T
29	T
30	F

Time series	
31	T
32	T
33	F
34	T
35	F

Preprocessing	
36	F
37	T
38	T
39	T
40	T

Complex data	
41	T
42	F
43	T
44	F
45	F

II. Multiple choice(e.g. “a b c” or “none”)

Classification	
1	BC
2	AC
3	BC
4	ABD

Clustering	
5	AB
6	ABCD
7	BD
8	AB

Regression	
9	A
10	BCD
11	C
12	BD

III. Calculus

Classification	
1	A=X2<=1 e X1<=2.5
2	19/20 e 14/15
3	13/20
4	2
5	Positive

Time Series	
6	<0,-1,-2>
7	6 and 4
8	4
9	5
10	<4,3,5,4,4> and √(82/5)

Data reduction	
11	2
12	80%
13	[1 -1]
14	[1 3 -2]

I. True or False(45 statements = 9.0v)

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

Group I. Classification

1. A 1-NN classifier has always zero training error.
2. Increasing the depth of a decision tree cannot increase its training error.
3. A negative observation that is wrongly labelled is termed false negative.
4. In theory, a decision tree learned from data with m binary attributes can represent any Boolean function over those m attributes if enough observations are provided.
5. In cross-validation, the higher the number of folds, the higher the number of training observations per fold.

Group II. Pattern mining

6. The monotonic property states that the subsets of a frequent itemset are also frequent.
7. Considering the use of equal-frequency/depth discretization, the prior normalization of an attribute affects the produced bins.
8. The assessment of lift is particularly relevant since the rule's consequent can appear on transactions without the rule's antecedent.
9. A closed itemset is always a maximal itemset.
10. Assuming the lengthiest pattern is p, then Apriori performs at most O(p) database scans.

Group III. Clustering

11. Hamming distance is adequate to handle ordinal data with high cardinality.
12. A rand index close to zero suggests that the clustering algorithm was not able to guarantee high cluster dissimilarity.
13. Purity is biased when the number of found clusters approaches the total number of observations.
14. Both k-median and k-medoids are more robust to outliers than k-means.
15. Complete link criterion tends to break large clusters and is biased towards globular clusters.

Group IV. Biclustering

Consider the following dataset and biclusters:

	y_1	y_2	y_3	y_4
x_1	1	1	0	2
x_2	1	2	1	3
x_3	1	2	1	3
x_4	4	1	1	0

16. B=(I=$\{x_2,x_3\}$,J=$\{y_2,y_3,y_4\}$) is an order-preserving bicluster.
17. The largest constant bicluster with I=$\{x_2,x_3\}$ has pattern of length $|\varphi_B|=4$.
18. B=(I=$\{x_1,x_2,x_3\}$,J=$\{y_2,y_3,y_4\}$) is an additive bicluster with shifting factors $\gamma_B=\{\gamma_1=0,\gamma_2=1,\gamma_3=1\}$.
19. B_1=(I1=$\{x_1,x_2,x_3\}$,J1=$\{y_1,y_2\}$) and B_2=(I2=$\{x_2,x_3,x_4\}$,J2=$\{y_2,y_3\}$) are plaid biclusters with pattern $\varphi_{B1}=\{c_1=1,c_2=1\}$ and $\varphi_{B2}=\{c_1=1,c_2=1\}$.
20. Constant bicluster B=(I=$\{x_1,x_2,x_3\}$,J=$\{y_1,y_2,y_3\}$) with $\eta_{ij} = 0$ has a quality of 0.75.

Group V. Data reduction

21. Filter procedures for feature selection are commonly applied with accuracy measures.
22. Learning curves can be considered to estimate the best number of features to select.
23. Spearman correlation is preferred over Pearson correlation if the order of quantities is more relevant than their absolute value.
24. Generally, backward subset selection is more computationally expensive than forward subset selection if the majority of features are non-redundant and discriminative.
25. Linear discriminant analysis (LDA) finds the axes that show greatest variation for the observations of each class.

Group VI. Regression

26. Given attributes y1, y2 and y3, if covariance (in absolute value) between y1 and y2 is higher than covariance between y1 and y3, then y1 and y2 have higher correlation.
27. Given a linear regression model, if a scatter plot shows that the residuals are highly correlated (Pearson correlation close to 1), then the learned model is not good.
28. Multiple linear regression is sensitive to outliers.
29. Decision tree regressors can only estimate as many quantities as the number of leaves.
30. AUC can be also considered to evaluate regression models.

Group VII. Time series data

31. When applying the discrete Fourier transform (DFT) to analyze a signal with a fixed sampling rate, the time window (number of time points) needs to be high enough to model low frequencies.
32. A short discrete Fourier transform (SDFT) is a DFT on sliding segments of a signal.
33. SDFT is preferred over classic DFT when a time series is stationary.
34. While wavelets, such as Haar wavelets, are more appropriate to understand household electricity consumption signals, sinusoids are more appropriate to understand brain signals.
35. Codebook representations of time series are symbolic representations that produce time series with higher dimensionality than SAX representations.

Group VIII. Pre-processing

36. Outlier analysis aims to detect attributes with values deviating significantly from expectations.
37. In semi-supervised outlier analysis, it is more relevant to know non-outliers than outliers.
38. Statistical approaches to outlier detection assume data to be generated by a distribution to test outlier likelihood.
39. Clustering is an effective means to perform subsampling when $n_{new}/n_{original} << 1$ (where n_{new} and $n_{original}$ are respectively the number of observations in the final and original dataset).
40. Binning numeric variables and merging categoric values is a way of reducing domain cardinality.

Group IX. Mining complex data

41. Sequential pattern mining can be applied both on symbolic time series data and itemset sequence data
42. The orders between the items in a sequential pattern can be partially defined
43. Given a dataset with n time series, to learn a tabular data encoding using regression coefficients: there is the need to learn as many regressions as the number of observations
44. The spatial slicing principle for spatiotemporal data analysis is verified when spatial content can be separated from the remaining static and temporal content during the learning
45. Task partitioning procedures for distributed data mining aim to partition data into subsets and distribute their analysis across processors

II. Multiple Choice (12 questions 0.4v each = 4.8v)

Select **all the true answers** (none, one or more than one answer are allowed). The grade corresponds to k/n, with n the number of correct answers and k the number of correct options taken. Wrong answers discount 1/2n.

Group I. Classification

1. Consider the problem of building decision trees with k-ary splits (split one node into k nodes) with entropy impurity. Which of the following is/are true?
- **a.** The algorithm will always choose $k = 2$
- **b.** The algorithm will prefer high values of k
- **c.** There will be k-1 thresholds for a k-ary split
- **d.** This model is strictly more powerful than a binary decision tree

2. Which of the following are true about each individual tree in a random forest?
- **a.** Individual tree is built on a subset of the features
- **b.** Individual tree is built on all the features
- **c.** Individual tree is built on a subset of observations
- **d.** Individual tree is built on full set of observations

3. Why would we use a random forest instead of a decision tree?
- **a.** For lower training error
- **b.** To reduce underfitting propensity
- **c.** To reduce overfitting propensity
- **d.** To better approximate posterior probabilities
- **e.** To facilitate human interpretability

4. Consider the learning of a classifier from a dataset with 1000 attributes. 50 of them are discriminative. Another 50 features are direct copies of the first 50 attributes. The final 900 features are not informative. Assume there is enough data to reliably assess how useful features are:
- **a.** 100 features will be selected by mutual information filtering
- **b.** 100 features will be selected by a backward wrapper method
- **c.** 50 features will be selected by mutual information filtering
- **d.** 50 features will be selected by a forward wrapper method

Group II. Clustering

5. Given the following data, the true clusters are better described by:
- **a.** Model-based clustering than density-based clustering
- **b.** Soft clustering than deterministic clustering
- **c.** Classic partition-based clustering than fuzzy clustering
- **d.** Agglomerative-based clustering (single link) than model-based clustering

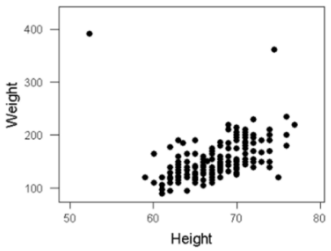
6. Partitioning-based clustering algorithms can be parameterized with specific:
- **a.** Number of clusters
- **b.** Seeding methods
- **c.** Similarity distances
- **d.** Centroid criteria
- **e.** Linkage criteria

7. Select the correct statements on the sum of squared errors (SSE):
- **a.** SSE is a measure of clustering separation
- **b.** SSE is not adequate to compare clustering solutions with a different number of clusters
- **c.** If SSE on a target dataset is ρ and the SSE expectations on randomized data is always lower than ρ, then the clustering solution is statistically significant regarding SSE
- **d.** If SSE on a target dataset is ρ and the SSE expectations follow a Gaussian distribution X and P(X<ρ)=1E-3, then the clustering solution is statistically significant regarding SSE

8. Select the advantages of clustering with minimum linkage (in contrast with maximum linkage)
- **a.** Ability to identify large clusters or different sizes
- **b.** Ability to identify clusters with non-elliptical shapes
- **c.** Robustness to outliers and noise
- **d.** Ability to model overlapping clusters

Group III. Regression

9. The difficulties to learn a regression model to analyze the following dataset are:



- **a.** Presence of one or more outliers
- **b.** Differences in scale between variables
- **c.** Non-normalized data: height values higher than 50
- **d.** Curvilinear data
- **e.** Response variable is not quantitative

10. Which of the following criteria contribute to a smoothed regression model?

- **a.** Increasing depth of decision trees
- **b.** Increasing k of nearest neighbours
- **c.** Parameterizing nearest neighbours with uniform weights instead of distance-based weights
- **d.** Application of linear and kernel smoothers

11. Identify true statements on single-wise (versus local/piece-wise) regression models:

- **a.** *k*NN is a single-wise regressor
- **b.** Decision tree is a single-wise regressors
- **c.** Single-wise regressors typically have a higher generalization ability (low complexity term)
- **d.** Single-wise regressors typically suffer from underfitting risks

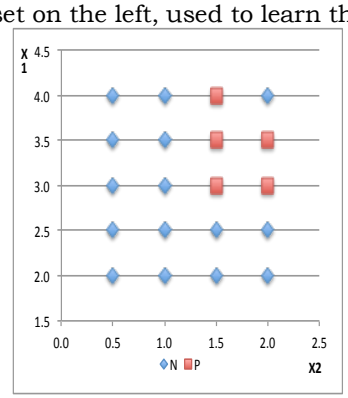
12. Why is PCA sometimes used as a preprocessing step before regression?

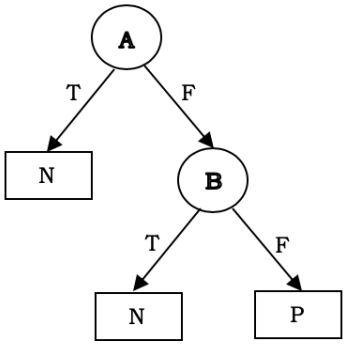
- **a.** To select predictors
- **b.** To minimize overfitting risks
- **c.** To expose information missing from the input data
- **d.** To make computation faster by reducing the dimensionality of the data

III. Calculus (14 questions = 6.2v)

Group I. Classification [2.5v]

Given the dataset on the left, used to learn the tree on the right with a pre-pruning strategy:





1. [0.5v] What are the tests performed in A and B when using information gain criteria to learn the tree?
2. [0.5v] What is the tree accuracy? And sensitivity for the negative class (N)?
3. [0.5v] How many instances is KNN (with k=3) able to correctly classify in the dataset, using the leave-one-out strategy?
4. [0.5v] Consider a random forest learnt using C4.5 to train decision stumps on the dataset. How many different classifiers would be trained?
5. [0.5v] How would that random forest classify the instance (4.0, 2.0)?

Group II. Time Series [2.4v]

Considering the following time series data
(where time series are assumed to be already normalized and DTW is applied with squared loss):

	time series	$\boldsymbol{z}$
x_1	<1, -1, 0>	5
x_2	<0, -2>	6
x_3	<-1, 1, 1>	7
x_4	<-1, 1, 0, 2, 3>	3

6. [0.2v] What is the PAA representation for x_2 with 3 segments?

7. [0.5v] What is the DTW cost between x_1 and x_3? How many alignments has this path?

8. [0.6v] Assuming a KNN with k=2 trained over $\{x_2,x_3,x_4\}$, and distances d(x_1, x_2)=2, d(x_1, x_3)=3, d(x_1, x_4)=1. What is the estimated quantity for x_1 when considering k=2 and distance weights?

9. [0.4v] Considering time series x_4 observed between t=1 and t=5. Assuming one differencing operation, what is the naïve forecast for t=7?

10. [0.7v] Considering time series x_4 observed between t=1 and t=5. Assuming a time series regression with $\widehat{\boldsymbol{\beta}}$=[$\widehat{\beta}_0$=2, $\widehat{\beta}_1$=1]: identify a) the residuals on the observed series, and b) the associated RMSE.

Group III. Data Reduction [1.3v]

Consider the following eigenvalues and eigenvectors (ignore the fact $\|v_i\|$=1) produced from a dataset with 3 attributes:

λ_1=2.5
λ_3=1
$\boldsymbol{v}_1$=[2 1 -2]
$\boldsymbol{v}_2$=[1 -2 0]

and a covariance matrix given by $\begin{bmatrix} 4 & 1 & 2 \\ 1 & 2.5 & 1 \\ 2 & 1 & 5 \end{bmatrix}$

11. [0.3v] What is the value for eigenvalue λ_2?

12. [0.3v] Assuming λ_2=1.5, what is the explained variability by the two first components?

13. [0.3v] Considering a (centred) observation with values x_1 = [1 1 1] and the application of PCA with the two first components, what are the component values for x_1?

14. [0.4v] In the same conditions, what are the recovered values of the first observation using the inverse PCA with two components?