Final examination
Machine Learning and Data Mining

Slides are permitted.

Nota Bene : Presentation will be taken into when given a mark. Marks are also subject to modification.

Two Parts that should be processed on different sheets of papers. Part I (Pr. Mephu Nguifo) and Part II (Dr. Antoine).

Part I

Exercise 1 : Classification **(12 pts)**

1) Briefly outline the major steps of decision tree classification. (1 pt)
2) Why is tree pruning useful in decision tree induction ? (1 pt)
3) What are the two common approaches to tree pruning ? Explain their principle. (1 pt)
4) What is a drawback of using a separate set of tuples to evaluate pruning ? (1 pt)
5) Compare the advantages and disadvantages of eager classification (e.g. decision tree, rules, Bayesian network) versus lazy classification (e.g., k-nearest neighbor, case-based reasoning). (1 pt)
6) Given the following relation « weather.symbolic » where the class attribute is « Play». What are the relevant attributes for classification ? Explain why. (1pt)
7) Considering that the training set contains the 8 first instances {O1 to O8}. Use the naive Bayes technique with Laplace formula and the relevant attributes, to predict the class of the 6 last instances {O9 to O14}. (3 pts)
8) Build the confusion matrix. (0,5 pt)
9) Give the error rate, the precision rate and the recall rate. (1,5 pt)
10) Which type of evaluation method is used here ? (1 pt)

```
@relation weather.symbolic
@attribute name          nominal
@attribute outlook       {sunny, overcast, rainy}
@attribute temperature   {hot, mild, cool}
@attribute humidity      {high, normal}
@attribute windy         {TRUE, FALSE}
@attribute play          {yes, no}
@data
O1,  overcast, hot,      high,     FALSE,    yes
O2,  rainy,    mild,     high,     FALSE,    yes
O3,  rainy,    cool,     normal,   FALSE,    yes
O4,  overcast, cool,     normal,   TRUE,     yes
O5,  sunny,    cool,     normal,   FALSE,    yes
O6,  sunny,    hot,      high,     FALSE,    no
O7,  sunny,    hot,      high,     TRUE,     no
O8,  rainy,    cool,     normal,   TRUE,     no
O9,  rainy,    mild,     normal,   FALSE,    yes
O10, sunny,    mild,     normal,   TRUE,     yes
O11, sunny,    mild,     high,     FALSE,    no
O12, overcast, mild,     high,     TRUE,     yes
O13, overcast, hot,      normal,   FALSE,    yes
O14, rainy,    mild,     high,     TRUE,     no
```

Exercise 2 : Association rules **(4 pts)**

Let {a, b, c, d} be a set of items ; {t_1, t_2, t_3, t_4, t_5, t_6, t_7, t_8, t_9, t_{10}} be a set of transactions ; {r_1, r_2, r_3, r_4, r_5, r_6, r_7, r_8} be a set of rules ; and {Frequency, Confidence, Pearl} be a set of measures.

The dominance relationship defined as follows : a rule r is said dominated by another one r', if for all used measures, r is less relevant than r'.

	a	b	c	d
t_1			×	×
t_2	×			
t_3	×			×
t_4			×	
t_5		×		×
t_6	×			×
t_7			×	
t_8				×
t_9		×	×	
t_{10}			×	×

(a) *A transaction dataset* $\mathcal{D}$

Rule	*Freq*	*Conf*	*Pearl*
r_1: $a \rightarrow d$	0.20	0.67	0.02
r_2: $b \rightarrow c$	0.10	0.50	0.00
r_3: $b \rightarrow d$	0.10	0.50	0.02
r_4: $c \rightarrow d$	0.20	0.40	0.10
r_5: $d \rightarrow a$	0.20	0.33	0.02
r_6: $d \rightarrow c$	0.20	0.33	0.10
r_7: $c \rightarrow b$	0.10	0.20	0.01
r_8: $d \rightarrow b$	0.10	0.17	0.02

(b) *A table relation* $\Omega(\mathcal{R}, \mathcal{M})$

Name	Definition	Domain
Frequency	$\frac{supp(X \cup Y)}{\lvert D \rvert}$	[0, 1]
Confidence	$\frac{supp(X \cup Y)}{supp(X)}$	[0, 1]
Pearl	$\frac{supp(X)}{\lvert D \rvert} \times \lvert \frac{supp(X \cup Y)}{supp(X)} - \frac{supp(Y)}{\lvert D \rvert} \rvert$	[0, 1]

(c) *Some measures of* $\mathcal{M}$

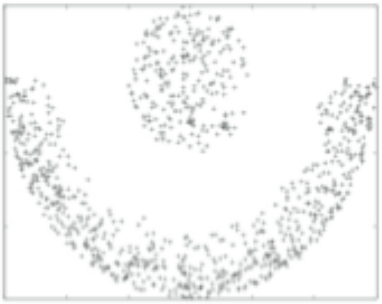
TABLE I
EXAMPLE OF A DATASET TRANSACTION AND MEASURES.

1. Give the definition of an undominated association rule. (0,5 pt)
2. List the set of undominated rules from table 1. Justify for each association rule why is it or not an undominated rules. (2 pts)
3. Given a list of all undominated association rules of a transactional table, is it possible to derived the set of all association rules. Justify your answer. (1 pt)
4. Consequently is there any relation between the number of association rules and the number of undominated rules. (0,5 pt)
5. **BONUS** : Provide an algorithm that allow to rank all the associations rules using the dominance relationship in an ascending order. (**3 pts**)

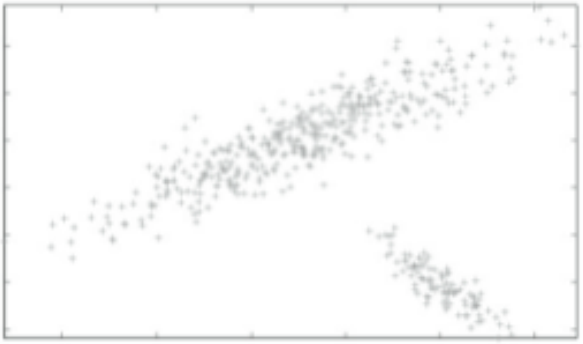
Part II (Mandatory : Use another copie)

Exercise 1 : Clustering **(4 pts)**

1. Which algorithms, between DBSCAN and k-means, should be used for the following dataset in order to give good results? Explain your answer.

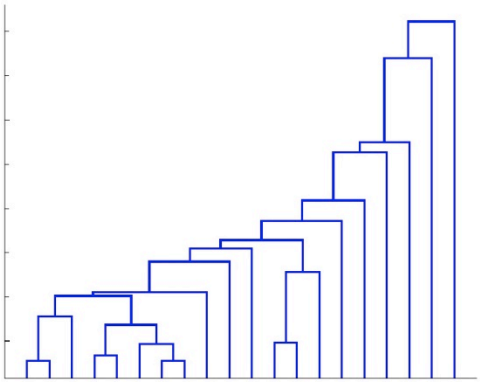


2. Which variants of k-means should be used for the following dataset? Note on the figure the final result of the algorithm using k=2 (i.e. the output of the algorithm, such as the centroids, ...)



3. For k-means, setting parameter k to 2 corresponds to background knowledge. Explain a way to define k automatically.

4. In the following dendrogram, explain and show the method to find 3 clusters.



5. The dendrogram corresponds to the result of a single-link method. From this figure, what can we say about the dataset?
6. Give two clustering algorithms (that are not from the same family) that take in entry a dissimilarity matrix.