Final examination
Machine Learning and Data Mining

ONLY Slides are permitted.

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

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

Part I

Exercise 1 : Classification **(6 pts)**

1) Briefly outline the topic of your homework (1,5 pt)
2) Outline the major steps of naive bayes classification. (1 pt)
3) Is post-pruning preferable to prepruning in decision tree classification ? Explain. (1 pt)
4) Given the dataset below, explain how to build and evaluate a classifier. (2,5 pts)

```
@relation contact-lenses
@attribute age                  {young, pre-presbyopic, presbyopic}
@attribute spectacle-prescrip   {myope, hypermetrope}
@attribute astigmatism          {no, yes}
@attribute tear-prod-rate       {reduced, normal}
@attribute contact-lenses       {soft, hard, none}

@data
% 24 instances
young,myope,no,reduced,none
young,myope,no,normal,soft
….
pre-presbyopic,myope,no,reduced,none
…
presbyopic,hypermetrope,yes,reduced,none
presbyopic,hypermetrope,yes,normal,none
```

Exercise 2 : Association rules **(10 pts)**

Let {a, b, c, d, e} be a set of items ; {t_1, t_2, t_3, t_4, t_5, t_6} be a set of transactions described as :

t_1 = b, c, d t_2 = a, b, c, d, e t_3 = a, b, c, e t_4 = a, b, d, e t_5 = b, c, e t_6 = a, b, d, e

The minimum support threshold is to set to 3/6. Let U is a frequent itemset ; V is a proper subset of U ; and V is not empty.

Theorem : If a rule V → U–V does not satisfy the confidence threshold, then any rule V' → U-V', where V' is a subset of V, must not satisfy the confidence threshold as well.

1. Prove the theorem above . (1 pt)
2. Find all frequent itemsets using Apriori algorithm. Provide details. (2 pts)
3. Draw the lattice of frequent itemsets. (1 pt)
4. Find all frequent closed itemsets. Explain. (1,5 pt)
5. Find one frequent closed itemset that has two minimal generators. Provide details. (1,5 pt)
6. Find all frequent maximal itemsets. (1 pt)
7. Give the negative border of frequent itemsets. (1 pt)
8. Let the minimal confidence equal to 66%, list three valid rules. Provide details. (1 pt)

Part II MANDATORY : Use another copie

Exercise 3 : Clustering **(4 pts)**

1. Use the complete-link agglomerative clustering with the Euclidean distance to group the data described as follow :

x1	x2	x3	x4	x5
0	1	4	7	9

Show, for each epoch, the dissimilarity matrix and the dendrogram.

2. Cut the final dendrogram in order to get two clusters and give the set of objects for each group.

3. Name two algorithms enable to find correctly the groups for the dataset below. Explain briefly your choices.