Data Mining and Machine Learning
Final examination – 2h

Lessons notes and slides are permitted.

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

Exercise 1 : Clustering **(10 pts)**

1) Use the dissimilarity matrix in the table below to perform **single link** hierarchical clustering. Show step by step the modifications in the dissimilarity matrix. (4 pts)

	P1	P2	P3	P4	P5
P1	0.00				
P2	0.90	0.00			
P3	0.59	0.36	0.00		
P4	0.45	0.53	0.56	0.00	
P5	0.65	0.02	0.15	0.24	0.00

2) Show your results by drawing a dendrogram. The dendrogram should clearly show the order in which the points are merged. (1 pt)
3) How many sets of cluster can you deduced from the dendrogram ? (1 pt)
4) Given a dendrogram, how would you proceed to provide a set of clusters to a user. (1 pt)
5) Illustrate the strength and weakness the two hierarchical method principles : single link and complete link. (1 pt)
6) Could you apply k-means on this dissimilarity matrix to build cluster ? Explain your answer. (2 pts)

Exercise 2 : Pattern mining **(6 pts)**

Let I = {a, b, c, d, e} be a set of items ; Y is a frequent itemset ; X is a proper subset of Y ; and X is not empty.

Theorem :

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

1. Prove the above theorem. (2 pts)
2. Given the positive border (PB), how could you extract the list of frequent itemsets ? Illustrate your answer with the following PB = {ad, ace, bcd} (1 pt)
3. Given the negative border (NB), how could you extract the list of infrequent itemsets ? Illustrate your answer with the following NB = {ab, acd, be, de} (1 pt)
4. Given the positive border, how could you extract the list of frequent closed itemsets (FCI) ? Provide details. (2 pts)

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

Given the following dataset (see Appendix) to illustrate your answer. Justify your answer.

1. Discuss the importance of the training set within the classification process. (1 pt)
2. Why k-nearest-neighbor (k-NN) is a lazy learning ? (1 pt)
3. What is the difference between decision tree classification and rule-based classification ? (2 pts)

APPENDIX

```
@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
pre-presbyopic,myope,no,normal,soft
young,myope,yes,reduced,none
young,myope,yes,normal,hard
young,hypermetrope,no,normal,soft
presbyopic,myope,yes,normal,hard
young,hypermetrope,no,reduced,none
young,hypermetrope,yes,reduced,none
young,hypermetrope,yes,normal,hard
young,myope,no,normal,soft
pre-presbyopic,myope,no,reduced,none
pre-presbyopic,myope,yes,reduced,none
pre-presbyopic,myope,yes,normal,hard
pre-presbyopic,hypermetrope,no,reduced,none
pre-presbyopic,hypermetrope,no,normal,soft
pre-presbyopic,hypermetrope,yes,reduced,none
pre-presbyopic,hypermetrope,yes,normal,none
presbyopic,myope,no,reduced,none
presbyopic,myope,no,normal,none
presbyopic,myope,yes,reduced,none
presbyopic,hypermetrope,no,reduced,none
presbyopic,hypermetrope,no,normal,soft
presbyopic,hypermetrope,yes,reduced,none
presbyopic,hypermetrope,yes,normal,none
```