Machine learning
MSc BioInfo 2013

G. R.

October 15, 2013

Send the answers as a .pdf file + the Progol .pl file to richard@irit.fr. All documents and notes are allowed.

1. **(4marks)** Let us consider the table below with simple interpretation: each line is a web surfer and 1 2 3 means that, during a session, the corresponding user clicked on page1, page2 and page3.

1	2	3	
1	4	5	
2	3	4	
1	2	3	4
2	3		
1	2	4	
4	5		
1	2	3	4
3	4	5	
1	2	3	

   Compute support, confidence and lift for the following association rules:

$$3 \leftarrow 2 \quad 1$$

$$2 \leftarrow 1 \quad 3$$

$$1 \leftarrow 2 \quad 3$$

2. **(4marks)** In $\mathbb{R}^2$, we consider the following set $\mathcal{C}$ of concepts: the disks i.e. the circles and their interior. Show that $VC_{dim}(\mathcal{C}) \geq 3$. Could you say more and compute the exact value of $VC_{dim}(\mathcal{C})$?

3. **(4marks)** Back to the algorithm studied during the lecture to learn rectangles in $\mathbb{R}^2$. Compute the minimal number of examples needed to get a precision of 10^{-2} with a confidence of 10^{-2}. Justify your answer.

4. **(4marks)** Write a Progol training set to learn a list permutation programme. We need positive examples, negative examples, mode declaration, type declaration and background knowledge.

5. **(4marks)** Binary Decision Trees (BDT) are powerful and effective tools when it comes to classify binary vectors. We want to estimate the VC_{dim} of the set of binary decision trees of a given height k. We assume that:

   - The representation space X is a cartesian product of dimension n, i.e. we have n attributes to represent our examples.
   - We consider only binary decision trees. i.e. a node has 2 children or is a leave (no child).
   - We consider 2 classes denoted $+$ et $-$. Our binary decision trees classify an element of X as $+$ or $-$.

   (a) We denote D_k the set of BDT over X of height k and C_k its cardinality. What is the value of C_0? Show that

   $$C_{k+1} = n \times C_k \times C_k$$

   .

   (b) We denote $L_k = log_2(C_k)$. Extract from the previous question the relationship between L_{k+1} and L_k. Why are we interested in L_k?

   (c) Show by induction that $L_k = (2^k - 1)(1 + log_2(n)) + 1$. Provide a conclusion.