CMPT-740 Fall 2003
Foundations of Data Mining
Martin Ester

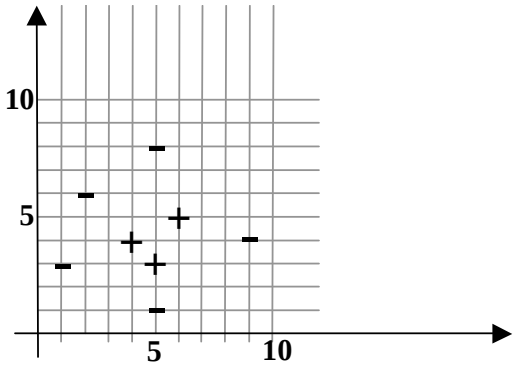
Assignment 1

Total marks: 60
Due date: September 24, 2003

Assignment 1.1

Consider the instance space consisting of integer points in the two-dimensional plane and the space of hypotheses consisting of integer rectangles. Thus, hypotheses are of the form $(a \leq x \leq b) \wedge (c \leq y \leq d)$ with integers a, b, c, d.

Let positive (+) and negative (-) training examples be given according to the following diagram:



(a) What is the specific boundary *S* of the resulting version space? What is the general boundary *G* of the version space? For S and G, write down the corresponding hypotheses and draw them in on the diagram.

(b) Suppose the learning algorithm may get back to some supervisor and ask him for the proper label of further training examples suggested by the learner. Suggest one training example which is guaranteed to reduce the size of the version space (independent from its class label) and another one that will not reduce its size.

(c) What is the minimum number of training examples which will allow the Candidate-Elimination algorithm to perfectly learn any given concept (rectangle), i.e. to return a version space containing only the target concept? Explain your answer and provide an example of such a minimal set of training examples.

Assignment 1.2

Assume the following training dataset of *Employee* records for learning the concept "Manager":

Name	**Birthdate**	**Sex**	**Citizenship**	**Department**	**Salary**	***Manager***
Enns	Jan 24, 1971	Female	Canadian	Sales	-	No
Walter	Mar 4, 1960	Male	Canadian	Finances	100430	Yes
Ferguson	Sep 18, 1965	Male	Canadian	Marketing	164320	Yes
Zhang	Dec 2, 1983	Male	Chinese	Sales	32100	No
Muller	Oct 3, 1985	Male	German	Production	24800	No
Wright	Jan 20, 1975	-	Canadian	Production	31245	No
Li	Feb 25, 1960	Male	Chinese	Production	45376	Yes
Capolla	Sep 30, 1972	Female	Italian	Marketing	64800	No

We want to use the Candidate-Elimination algorithm to learn the target concept. How would you preprocess the data for this data mining task? In particular, what kind of

(a) data cleaning
(b) data integration
(c) data transformation
(d) data reduction

would you perform? Explain your choices.