CMPT-740 Fall 2003
Foundations of Data Mining
Martin Ester

Midterm Exam with Solutions

Total marks: 200
Date: October 29, 2003

Question 1

Consider the task of learning the concept "EnjoySport". Let H be the hypothesis space consisting of conjunctions of attribute constraints and H' be the hypothesis space consisting of pairwise disjunctions of the hypotheses in H. For example, a typical hypothesis in H' is

$$< ?, \text{Cold}, \text{High}, ?, ?, ? > \vee < \text{Sunny}, ?, \text{High}, ?, ?, \text{Same} > .$$

Trace the Candidate-Elimination algorithm for the hypothesis space H' given the following sequence of training examples:

Example	Sky	AirTemp	Humidity	Wind	Water	Forecast	EnjoySport
1	Sunny	Warm	Normal	Strong	Warm	Same	Yes
2	Sunny	Warm	High	Strong	Warm	Same	Yes
3	Rainy	Cold	High	Strong	Warm	Change	No
4	Sunny	Warm	High	Strong	Cool	Change	Yes

After each new training example, show the contents of the specific boundary *S* and the general boundary *G* of the resulting version space.

Initially:

S = $\{(<\emptyset,\emptyset,\emptyset,\emptyset,\emptyset,\emptyset> \vee <\emptyset,\emptyset,\emptyset,\emptyset,\emptyset,\emptyset>)\}$
G = $\{(<?,?,?,?,?,?> \vee <?,?,?,?,?,?>)\}$

After training example 1:

S = $\{(< Sunny, Warm, Normal, Strong, Warm, Same > \vee <\emptyset,\emptyset,\emptyset,\emptyset,\emptyset,\emptyset>)\}$
G = $\{(<?,?,?,?,?,?> \vee <?,?,?,?,?,?>)\}$

After training example 2:

S=
$\{(< Sunny, Warm, Normal, Strong, Warm, Same > \vee < Sunny, Warm, High, Strong, Warm, Same >)\}$
G = $\{(<?,?,?,?,?,?> \vee <?,?,?,?,?,?>)\}$

After training example 3:

S=

$$\{(< \mathit{Sunny}, \mathit{Warm}, \mathit{Normal}, \mathit{Strong}, \mathit{Warm}, \mathit{Same} > \vee < \mathit{Sunny}, \mathit{Warm}, \mathit{High}, \mathit{Strong}, \mathit{Warm}, \mathit{Same} >)\}$$

$$\begin{array}{ll} & \{(< \mathit{Sunny}, ?, ?, ?, ?, ? > \vee < ?, \mathit{Warm}, ?, ?, ?, ? >), (< \mathit{Sunny}, ?, ?, ?, ?, ? > \vee < ?, ?, ?, ?, ?, \mathit{Same} >), \\ G = & (< \mathit{Sunny}, ?, ?, ?, ?, ? > \vee < ?, ?, \mathit{Normal}, ?, ?, ? >), (< ?, ?, ?, ?, ?, \mathit{Same} > \vee < ?, \mathit{Warm}, ?, ?, ?, ? >), \\ & (< ?, \mathit{Warm}, ?, ?, ?, ? > \vee < ?, ?, \mathit{Normal}, ?, ?, ? >), (< ?, ?, ?, ?, ?, \mathit{Same} > \vee < ?, ?, \mathit{Normal}, ?, ?, ? >)\} \end{array}$$

After training example 4:

$$\begin{array}{ll} S= & \{(< \mathit{Sunny}, \mathit{Warm}, \mathit{Normal}, \mathit{Strong}, \mathit{Warm}, \mathit{Same} > \vee < \mathit{Sunny}, \mathit{Warm}, \mathit{High}, \mathit{Strong}, ?, ? >), \\ & (< \mathit{Sunny}, \mathit{Warm}, ?, \mathit{Strong}, \mathit{Warm}, \mathit{Same} > \vee < \mathit{Sunny}, \mathit{Warm}, \mathit{High}, \mathit{Strong}, \mathit{Cool}, \mathit{Change} >)\} \end{array}$$

$$\begin{array}{ll} & \{(< \mathit{Sunny}, ?, ?, ?, ?, ? > \vee < ?, \mathit{Warm}, ?, ?, ?, ? >), (< \mathit{Sunny}, ?, ?, ?, ?, ? > \vee < ?, ?, ?, ?, ?, \mathit{Same} >), \\ G = & (< \mathit{Sunny}, ?, ?, ?, ?, ? > \vee < ?, ?, \mathit{Normal}, ?, ?, ? >), (< ?, ?, ?, ?, ?, \mathit{Same} > \vee < ?, \mathit{Warm}, ?, ?, ?, ? >), \\ & (< ?, \mathit{Warm}, ?, ?, ?, ? > \vee < ?, ?, \mathit{Normal}, ?, ?, ? >)\} \end{array}$$

Question 2

Consider the task of clustering a set of 100’000 protein sequences. Each protein is given by a sequence of symbols from the set of one-letter codes of the 20 different amino-acids, e.g.

AVFAMLCNFQDMAQSWKKKAVFAAGDE.

You have to satisfy the following two requirements:

- The resulting clustering should be hierarchical.
- The clustering method should be as efficient as possible.

You may make any additional assumptions on the application as long as they do not contradict these requirements.

(a) How would you represent the individual proteins? What distance function for pairs of proteins would you suggest? Explain your answers.

Protein representation as a string of amino-acids (represented using one-letter codes)

Appropriate distance functions:

(1) edit distance (using insertions, deletions and replacements of amino-acids as edit operations)
(2) distance based on scoring a pairwise alignment (using a method like BLAST)
(3) (weighted) number of common subsequences
(4) cosine similarity of frequency vectors (representing a protein by the numbers of occurrences of all frequent subsequences).

(b) What clustering algorithm would you apply? What is the format of the resulting clustering? What is the (approximate) runtime complexity of your method? Explain your answers.

Acceptable clustering formats and clustering algorithms:

(1) reachability plot: OPTICS with subsequent method for automatic cluster detection
(2) dendrogram: agglomerative or divisive hierarchical clustering.

All methods should be based on the distance function proposed in (a). Runtime complexities according to the chosen clustering method.

Question 3

Consider the task of classification of a transactional database, where each record consists of a set of items. For the training data set, we also know the corresponding class labels. A special form of association rules "$i_1 \wedge i_2 \wedge \ldots \wedge i_k \rightarrow c$ (support supp, confidence conf)", where $i_1, i_2, \ldots i_k$ are items and c denotes a class label, can be used for this classification task.

We restrict our discussion to two-class problems where we want to distinguish a target class from the contrasting class. The classifier should

- obtain 100 % recall w.r.t. the target class (on the training data) and
- have maximal precision w.r.t. the target class (on the training data).

We assume that an algorithm for determining all association rules of the above form is already given. We also assume that the minimum support was set such that each database record supports at least one of the frequent item sets, i.e. one of the rules. Your task is to design the model construction and model application based on the given set of all these association rules (for all possible items and for the two different class labels).

(a) Describe an algorithm that selects a subset of the set of all given association rules to construct a classification model satisfying the above requirements.

```
Sort all given association rules in descending order of confidence.
Start with an empty classifier.
While there are still some training records not covered by any chosen rule:
    (1) Select the association rule with the highest confidence and add it to the classifier.
    (2) Discard all training records covered by the new rule.
```

(b) How would your classifier be applied to classify an unseen object? Explain why your classifier satisfies the above requirements.

Model application:
Use the applicable association rule with the highest confidence.

This classifier achieves 100 % recall since each target class record is covered by at least one association rule (which predicts it as a target class element). We always apply the relevant rule with the highest confidence which maximizes the precision.