Intelligent Systems: Reasoning and Recognition

James L. Crowley

Conditions: You have the right to use any notes or written material. You may answer questions in English or in French. When appropriate, illustrate your answer with mathematics. Your written answers must be clear and legible. Illegible text will not be graded. Duration: 3 hours.

1) (4 points) You are asked to replace the confidence factor, CF, with Bayesian Probabilistic Reasoning in the MYCIN system. How does this change the way rules determine the confidence in facts? What is the new formula for the function COMBINE? Where would you get the probabilities for the facts and the rules?

2) (2 points) You are asked to program a planning system for Block World using the GRAPHSEARCH algorithm. What cost function would you use for planning? Is there an heuristic for your cost function that will provide an optimal search? If yes, what is the heuristic? If no, can you still use the GRAPHSEARCH algorithm? What kind of search does the algorithm perform?

3) (6 points) Bayesian Reasoning can be used to classify a body of text based on the frequency of occurrence of words. This can be used by a text editor to recognize the category of a document as a person is typing. Categories can represent types of documents (scientific, legal, journalistic, etc), different authors, or even different languages. Histograms (or bags) of words can be used to estimate the required probabilities for $P(E|H)$ and $P(E)$. For this task, assume that you have a large training corpus composed of examples of several categories of text.
a) (1 point) Explain how the training corpus can be used to construct a table for the frequency of occurrence for each word in the training data for each class of text. What it the probability that word, W, will occur in a sample from Class K? What is the probability that W will occur anywhere in the corpus?
b) (1 point) Propose a method to obtain an initial estimate for the probability that an unknown text (a probe) belongs to each category as the user types the first word.
c) (2 points) Explain how to use Bayesian reasoning to incrementally update the estimate for the probability of each category as the user types each new word in the unknown text.
d) (1 point) What happens if the unknown text contains a word that was not in the training corpus? What can you do to protect against this case?
e) (1 point) What Kernel function could you use to detect a particular category of text with a Support Vector Machine.

4) (8 points) The goal of this exercise is to program a system for probabilistic reasoning and recognition about text in CLIPS. For this you will need to determine the frequency of occurrence of words within different categories of text. You can assume that text is provided by a function named "read-paragraph" that creates a fact in working memory of the form:
        (Paragraph Category w1 w2 … wN)
where the <wn> are symbols representing the individual words.

Your system should use the following templates for Word and Category.

```
(deftemplate Word        ; structure for word count
     (slot CATEGORY (type SYMBOL))
     (slot WORD (type SYMBOL))
     (slot COUNT (type INTEGER))  ; Number of instances of word
)
```

```
(deftemplate Category
     (slot NAME) (type SYMBOL)) ; Category Name
     (slot M (type INTEGER)) ; Size of Training Set for Category
```

a) (1 point) Write a rule named MakeWords to generate a fact of type Word for each new word in a paragraph. Be sure to include the Category when you create each Word.

b) (1 point) Write a rule named CountWords that updates the count for each word in the paragraph. Make sure that this rule also updates the size of the training set, M, for the category.

c) (1 points) Assume that you have a very large corpus (training set) of text that includes several categories. Write a rule or set of rules to determine the sum of the word counts for each word in the training set. Store the result as facts of type Word with a category of "All". Name your rule(s) CountWordsForAllCategories.

d) (2 points) Assume the result of questions a, b and c. Given a word of an unknown category, write a rule or set of rules to printout the most likely category for the word.

e) (3 points) Given a paragraph of an unknown category, write a rule or set of rules named FindCategoryForParagraph to printout the most likely category for the paragraph.