11.128/14.49 Frank Levy
Spring 2005 Distributed: Session 9

Problem Set 2
Due: Session 11 (in class)

1) The course website has a dataset containing the characteristics of 150 flowers belonging to three different species of iris: *Iris setosa* (species=1), *Iris versicolor* (species=2) and *Iris virginica* (species=3). The dataset has information on four attributes of the iris flowers: sepal length, sepal width, petal length and petal width, all in centimeters (see website for pictures of an iris flower). The pattern recognition problem here is to classify the flowers into correct species based on these attributes.

(a) Develop a set of simple rules to help you do this. Your rules should be of the attribute-object-value format. For example:

If LENGTH (attribute) of SEPAL (object) is less than 5cm (value), then SPECIES (attribute) of FLOWER (object) is VERSICOLOR (value).

You are in essence building an "expert system" to do this classification. Your rules can have more statements in the IF part (i.e. 2 or more conditions to be satisfied) and more than one value in the THEN part as well. Explain why you make each rule.

(Hint: to start building up such rules, use a software package of your choice (e.g. Excel, Stata, SAS, etc.) to look at the data, which is in a text file on the course website. You would need to do things like look at the average or draw histograms of the attributes in the three categories. Another thing to do would be to graph these attributes against each other, marking out the species (in effect, looking at combinations of attributes).)

(b) Once you have made a set of rules (typically, 3-8 rules), apply your rules to the dataset, i.e. classify each flower as a certain species based on your rules. Make a table of the following kind to evaluate how well your rules work (put in the number of flowers in each cell):

		True Species		
Classified As		**Setosa**	**Versicolor**	**Virginica**
	Setosa			
	Versicolor			
	Virginica			

(c) For your "expert" system, calculate the following evaluation criterion:

C = Misclassification probability + 0.15*Number of rules used

where the misclassification probability = # flowers misclassified/total flowers

(d) (Extra Credit) From your set of rules, delete any one rule. Redo parts (b) and (c). Is your measure of C greater or less than before?

2) Mammograms are specialized x-ray or ultrasound images used to detect potentially cancerous cell clusters in a woman's breast. For a number of years, it has been possible to digitize these images. Digitization allows the information in an image to be processed by a computer. About three years ago, the U.S. Food and Drug administration approved software from several companies that provides a 'digital second opinion' on a mammogram. The software scans the digitized image and flags "regions of interest" that may represent cancerous cell clusters and so should be examined with particular care by a human radiologist. These programs are known as Computer Assisted Diagnosis software (CAD).

a) From the programmer's perspective, we can think of the x-ray or ultrasound as producing a matrix or grid of grayscale pixels – digital representations of points of light – with each point having an intensity that runs from totally black to pure white. Except for the absence of color, this grayscale grid is similar to the grid of photons that hits your retina when you look at, say, a picture in a magazine. Using your knowledge of human vision processing, discuss the perception problems the CAD software must first solve before it can start looking for cancerous cell clusters. (Relevant reading = first 16 pages of the Anderson chapter and Duda et. al).

b) After the initial processing has been done, describe how neural net software might be used to determine whether a given cluster of pixels in the information represents a cancerous cell cluster. Where does "training" fit into your explanation? (Relevant reading = Duda et. all, Quinn and Johnson, and the optional reading by Chidley).

c) From both the readings and class discussion, we know that pattern recognition software, including neural nets, makes both Type I and Type II classification errors – e.g. incorrectly classifying a sea bass as a salmon and incorrectly classifying a salmon as a sea bass. The relative frequency of Type I and Type II errors depends on how the programmer "fine-tunes" the software. This fine tuning should reflect the problem's loss-function. Describe the loss function associated with errors in reading a woman's mammogram and the problems this loss function creates for the fine-tuning decision.

d) In the chapter we read earlier this year, Herbert Simon argued that it was very hard to write software that was very flexible in the problems it could solve. He argued that it was possible to work around this obstacle to an extent by routinizing the problem which would allow it to be solved by simpler software. For example, the computerized robot that inserts a windshield into a pick-up truck frame relies on the fact that the truck frame

will always occupy a certain position, the pallet holding the windshield will always occupy a certain position, etc. Briefly explain how Simon's reasoning applies to software that detects potential cancerous cell clusters in mammograms.

3) Consider the problem of a parole board determining whether a prisoner is likely to commit more crimes if he/she is put on parole – in criminologists' terms, whether the prisoner will be a recidivist. At least one psychologist has argued strongly that a probit or logit regression based on age, previous crimes, marital status, etc. can predict recidivism more accurately than the careful deliberations of a half-dozen parole board members who interview the prisoner and review his/her files. (Recall that we discussed such regressions as one way of determining whether a mortgage application should be approved.) How does this argument fit or contradict the arguments of Blois in "Clinical Judgment and Computers"?