11.128/14/49 Frank Levy
Spring 2005

Solutions to Problem Set 2

1) (a) Here is one way to proceed: look at the means and ranges of these variables for the three species. We see that setosa is the smallest of these species; in fact, the maximum petal length and petal width for setosa are smaller than the minimum for the other two species. So one rule is as follows:

If length of petal is less than 2.5cm and width of petal is less than 0.7, then species of the flower is *setosa*. (RULE 1)

Drawing the graph of sepal width vs. petal width, we find that for species 3, most of the petal widths are above 1.75cms, so we can write some more rules as:

If petal width is greater than 1cm and petal width is less than 1.75cm, then species is *versicolor*. (RULE 2)

If petal length is greater than 3cm and petal length is less than 5.5cm and petal width is less than 1.75cm, then species is *versicolor*. (RULE 3)

Graphing petal width against petal length gives a rule for virginica:

If petal length is greater than 4.5cm and petal width is greater than 1.75cm, then species is *virginica*.(RULE 4)

Lastly, I put in a fairly arbitrary rule to take care of any missing classifications:

If still unclassified, then species is *setosa* (RULE 5).

(b) Using these rules, we get the following table:

	Species			
Classified	**1**	**2**	**3**	**Total**
1	50	0	0	50
2	0	49	5	54
3	0	1	45	46
Total	50	50	50	150

(c) The evaluation criterion is: C = 6/150 + 0.15*5 = 0.79

(d) Now, if we delete Rule 3, the table becomes:

Species				
Classified	**1**	**2**	**3**	**Total**
1	50	7	0	57
2	0	42	5	47
3	0	1	45	46
Total	50	50	50	150

C = 13/150 +0.15*4 = 0.69

So, even though the misclassification has gone up, deleting this rule still does better in terms of this particular evaluation criterion.

2) a) In the initial processing, the software has to divide the grid into objects – in this case, cell clusters and background material. It does this in the same way the human eye would do it, grouping pixels of the same gray and looking for "edges" where adjacent pixels have sharply different intensities.

A second part of the initial processing is the selection features of the cell cluster that will be used as the input variables for the neural net. Presumably, these features are the length of the circumference and several different measures of the diameter (since the cluster is probably not totally circular. (This part of the answer might be presented in part b rather than in this part.

b) The software would have to first have to be "trained" on a sample which included both cell clusters and non-cell clusters. The input variables would be the "features" described in the previous answer. The estimation would be iterated numerous times, with weights and thresholds being modified each iteration, until the neural net could distinguish between potentially cancerous clusters and non-cancerous clusters. Once the weights were identified, the estimated equations would be included as part of the software to predict whether each cluster looked cancerous or non-cancerous.

c) Begin with a null hypothesis that the woman does not have cancer. In this case, a Type I error occurs when the software incorrectly identifies a cancerous cell cluster – i.e. when the software rejects a true null hypothesis. In this case, the loss consists of unnecessary biopsies or other procedures together with substantial unwarranted anxiety. A Type II error occurs when the software sees no cancerous cell cluster but the hypothesis is false – i.e. the software incorrectly accepts a false null hypothesis. In this case, the loss is the possibility that the woman dies of undiagnosed breast cancer or, at minimum, that the cancer is discovered far later than it should have been.

While neither loss is trivial, the relative losses suggest the software should be "tuned" toward false rejection rather than false acceptance.

d) The software is designed to process a very particular kind of information and to identify very specific shapes – i.e. the software "knows" that only one kind of problem is coming. By contrast, a human radiologist is trained to read mammograms, x-rays, MRI's of abdomens, CT-Scans of brains – i.e. a wide variety of information in which he/she is expected to identify a wide variety of possible abnormalities. To make an analogy to speech recognition software, it is difference between writing a pattern recognition routine where you know the only spoken words will be names of big cities (e.g. for AMTRAK) versus writing a pattern recognition that is supposed to recognize general conversation about any subject**.**

3) The psychologist's argument largely contradicts Blois' funnel diagram. The Blois diagram would argue that it takes human pattern recognition—not a rule—to evaluate many different characteristics that each prisoner brings to a parole hearing. The psychologist is arguing the opposite – that these characteristics can be combined into a rule (in this case, the rule involves plugging characteristics into a previously estimated logit or probit equation) and the resulting prediction of the probability of recidivism will be more accurate than the pattern recognition and deliberations of the parole board members.