CS 525, Semester II, 1993–94

Project[1]

Friday 24th March, 1994

Linear programming is used in many applications. In this project, we will investigate the use of linear programming for breast cancer diagnosis.

The project will use the Wisconsin Breast Cancer Database made publicly available by Dr. William H. Wolberg of the University of Wisconsin Hospitals. The database consists of samples that arrive periodically as Dr. Wolberg reports his clinical cases and therefore reflects this chronological grouping of the data.

The data provided for this project consists of files "training.dat" which are the first 367 examples which should be used for generating a discriminant function, and "testing.dat" that should be used to test the discriminant function. There are 11 attributes per data point, with one data point per line. Attributes are separated by spaces. The attributes are as follows:

Field	Attribute
1	Sample code number
2	Class: 2 for benign, 4 for malignant
3	Clump Thickness
4	Uniformity of Cell Size
5	Uniformity of Cell Shape
6	Marginal Adhesion
7	Single Epithelial Cell Size
8	Bare Nuclei
9	Bland Chromatin
10	Normal Nucleoli
11	Mitoses

You should use attributes 3 to 11 (each of which take on values between 0 and 10) to form a 9-dimensional vector representing each case as a point in 9-dimensional real space. In generating the diagnostic program, you should begin with the training set consisting of two disjoint point sets in the nine-dimensional real space $\mathbf{R}^9$ representing confirmed benign and malignant fine needle aspirates (FNA's). An example of a benign FNA and a malignant FNA are give in Figure 1, from which the oncologist has determined the corresponding attributes. You should generate a discriminant function that distinguishes between the two sets – called $\mathcal{B}$ (benign) and $\mathcal{M}$ (malignant). This discriminant function is subsequently used to determine whether new aspirates are benign or malignant.

The discriminant function that you should use is determined by solving a single linear program in MATLAB and is based on a formulation proposed in [**?**]. The details of how to formulate the correct linear program now follow.

[1]Project due in class Monday 2nd May, 1994. No late projects accepted.

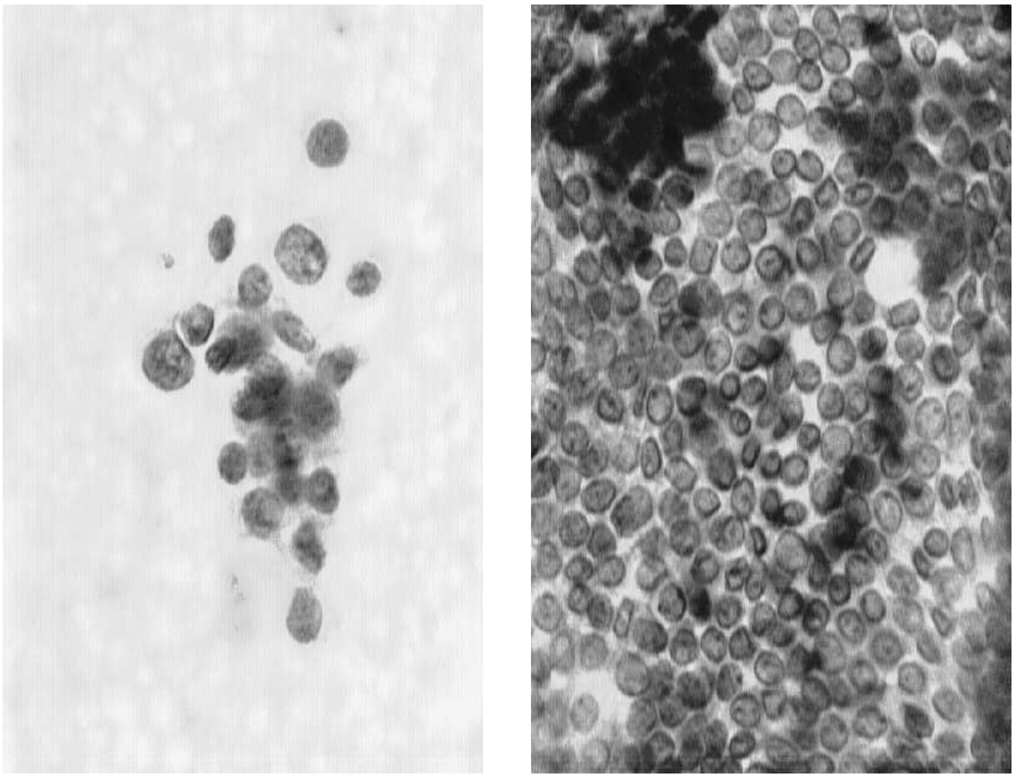

Figure 1: Nuclei of cells of malignant(left) and benign(right) fine needle aspirates taken from patients breasts

You need to construct a linear function f which has the property

$$f(x) > 0 \Longrightarrow x \in \mathcal{M}, \qquad f(x) \leq 0 \Longrightarrow x \in \mathcal{B}$$

The function you should generate is given as $f(x) = w^\top x - \gamma$, and determines a plane $w^\top x = \gamma$ that separates malignant points from benign ones in $\mathbf{R}^9$. It remains to show how to determine $w \in \mathbf{R}^9$ and $\gamma \in \mathbf{R}$ from the training data.

It we let the sets of m points $\mathcal{M}$ be represented by a matrix $M \in \mathbf{R}^{m \times n}$ and the set of k points $\mathcal{B}$ be represented by a matrix $B \in \mathbf{R}^{k \times n}$, then the problem becomes to choose w and γ to

$$\min_{w,\gamma} \frac{1}{m} \left\|(-Mw + e\gamma + e)_+\right\|_1 + \frac{1}{k} \left\|(Bw - e\gamma + e)_+\right\|_1$$

Here e is an appropriately dimensioned vector of all 1's, $((z)_+)_i = \max\{z_i, 0\}$, $i = 1, 2, \ldots, m$ and $\|z\|_1 = \sum_{i=1}^n |z_i|$, for $z \in \mathbf{R}^m$. This problem approximately minimizes the number of points that are misclassified by choosing w and γ to minimize the sum of the distances to the dividing plane whenever a point is on the incorrect side of the plane. The $(\cdot)_+$ and $\|\cdot\|$ functions can be eliminated by using the following linear program

$$\min_{w,\gamma,y,z} \left\{\frac{1}{m} e^\top y + \frac{1}{k} e^\top z \mid Mw - e\gamma + y \geq e, -Bw + e\gamma + z \geq e, y \geq 0, z \geq 0\right\}$$

A standard routine `lp` written in MATLAB is provided for solving linear programs of the form

$$\min\left\{c^\top x \mid Ax \leq b, L \leq x \leq U\right\}$$

which can be called by using `x=lp(c,A,b,L,U)`.

You are required to carry out the following steps.

1. Formulate the problem as a linear program. Solve the problem using the matrices M and B which can be determined from the training.dat file (see field 2). Determine how good the solution is by counting the number of misclassified points on the training set. Make sure you print w and γ out.

2. Test out your discriminant function on the testing set. Determine the number of misclassified points on the testing set.

3. Suppose that the oncologist wants to use only 2 of the 9 attributes in his diagnosis. Determine which pair of attributes is most effective in determining a correct diagnosis, again by using only the training set (with these 2 features) to determine the discriminant function and then using the testing set to determine the number of misclassifed points. Make sure you print out the number of misclassified points for each combination. Note that the scheme you have used above to determine which 2 attributes were the best is called chronological correctness.

4. Using the best answer from 3., plot all the testing points on a two dimensional figure using MATLAB's built–in plotting routines. Use 'o' for benign points and '+' for malignant points in your plot. Then use MATLAB to draw in the calculated line $w^\top x = \gamma$. Check to see if the number of misclassifed points agrees with your plot.