Comp Sci 525 Computing Project
Breast Cancer Diagnosis via Linear Programming

Linear programming has many applications. In this project, we will use linear programming for breast cancer diagnosis.

The project will use the Wisconsin Diagnosis Breast Cancer Database (WDBC) made publicly available by Dr. William H. Wolberg of the Department of Surgery of the University of Wisconsin Medical School, Professor W. Nick Street of the Management Sciences Department of the University of Iowa and Prof. O.L. Mangasarian of the Computer Sciences Department of the University of Wisconsin.

The database is available at various ftp sites as well as in the public directory `~cs525-1/public` as the files wdbc.data and wdbc.names. You should read the file wdbc.names.

The idea of the project is to come up with a discriminant function (a separating plane in this case) to determine if an unknown sample is benign or malignant. In order to do this, you will use part of the data in the above database as a "training set" to generate your separating plane and the remaining part as a "testing" set to test your separating plane.

Attributes 3 to 32 form a 30-dimensional vector representing each case as a point in 30-dimensional real space R^{30}. To generate the separating plane, a training set, consisting of two disjoint point sets in R^{30} representing confirmed benign and malignant fine needle aspirates (FNAs), should be used. An example of a benign FNA and a malignant FNA are given in Figure 1, from which a computer program has determined the corresponding 30 attributes. A separating plane should be generated that distinguishes between the two sets called $\mathcal{B}$ (benign) and $\mathcal{M}$ (malignant). This separating plane is subsequently used to determine whether new aspirates are benign or malignant.

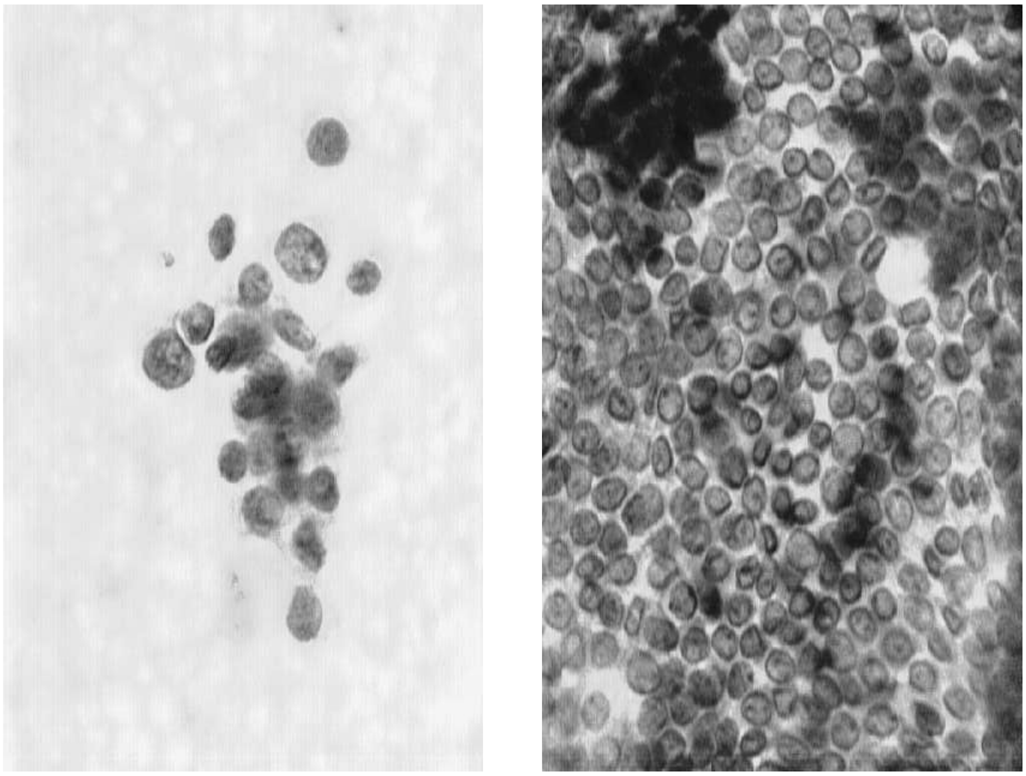

Figure 1: Nuclei of cells of malignant(left) and benign(right) fine needle aspirates taken from patients' breasts

The separating plane, to be determined by solving a single linear program in MATLAB is based on a formulation proposed in [1, 3]. Description of this project can also be found in [2]. Description of the linear program that generates the separating plane follows.

A linear function f will be constructed which has the property:

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

to the extent possible. This function is given by $f(x) = w'x - \gamma$, and determines a plane $w'x = \gamma$ that separates to the extent possible malignant points from benign ones in R^{30}. It remains to show how to determine $w \in \mathrm{R}^{30}$ and $\gamma \in \mathrm{R}$ from the training data.

It we let the sets of m points $\mathcal{M}$ be represented by a matrix $M \in \mathrm{R}^{m \times n}$ and the set of k points $\mathcal{B}$ be represented by a matrix $B \in \mathrm{R}^{k \times n}$, then the problem becomes one of choosing 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 ones, $((z)_+)_i = \max\{z_i, 0\}$, $i = 1, 2, \ldots, m$ and $\|z\|_1 = \sum_{i=1}^{m} |z_i|$ for $z \in \mathrm{R}^m$. This problem approximately minimizes the number of points that are misclassified by choosing w and γ to minimize the sum of the distances (times $\|w\|_2$) to the separating 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 programming reformulation;

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

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

$$\min \{c'x \mid Ax \leq b, L \leq x \leq U\}$$

which can be called by using `x=lp(c,A,b,L,U)`. Type `help lp` within MATLABto get more information about this linear programming routine.

The project consists of the following four parts:

1. Formulate the problem as a linear program. Solve the problem using the matrices M and B as a training set from the first 500 cases of the

wdbc.data file (see field 2). The last 69 will be used as a testing set as indicated below. An example of an m-file that extracts the training data and testing data as "matrices" from the original data file can be found in the public directory as `getdata.m`. **Make sure you print out w, γ and the minimum value of the LP.**

2. Test the separating plane on the 69 cases of the testing set. Report the number of misclassified points on the testing set. It is probably a good idea if you create an m-file to do this.

3. Suppose that the oncologist wants to use only 2 of the 30 attributes in his diagnosis. Determine which pair of attributes is most effective in determining a correct diagnosis as follows. Use each of $\binom{30}{2} = 435$ pairs of possible attributes and for each pair determine from the training set a separating plane in R^2. For each plane use the testing set with the corresponding pair of attributes to determine the number of misclassified cases. Make sure you print out the number of misclassified points in the testing set for each pair of attributes.

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

References

[1] K. P. Bennett and O. L. Mangasarian. Robust linear programming discrimination of two linearly inseparable sets. *Optimization Methods and Software*, 1:23–34, 1992.

[2] M. C. Ferris and O. L. Mangasarian. Breast cancer diagnosis via linear programming. *IEEE Computational Science and Engineering*, 2:70–71, 1995.

[3] O. L. Mangasarian, W. N. Street, and W. H. Wolberg. Breast cancer diagnosis and prognosis via linear programming. *Operations Research*, 43:570–577, 1995.