CS 525 Class Project
Breast Cancer Diagnosis via Quadratic Programming*
Fall, 2015
Due 15 December 2015, 5:00pm

In this project, we apply quadratic programming to breast cancer diagnosis.

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

The database is available thorugh the class web site, as the files `wdbc.data` and `wdbc.names`. You should read the file `wdbc.names`, which explains some background information and a description of the structure of the data file.

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

Attributes 3 to 32 of each piece of data form a 30-dimensional vector— a point in 30-dimensional real space $\mathbb{R}^{30}$. A training set, consisting of two disjoint point sets in $\mathbb{R}^{30}$ representing confirmed benign and malignant fine needle aspirates (FNAs), is used to generate the separating plane. (Figure 1 shows the different appearance of malignant and benign samples.) Each

*November 4, 2015

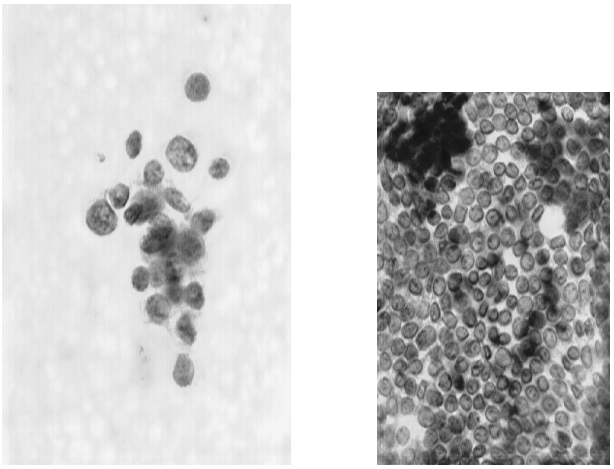

Figure 1: Nuclei of cells of malignant(left) and benign(right) fine needle aspirates taken from patients' breasts

sample point is labeled either "B" of "M", to indicate whether it is benign or malignant. We can therefore form two sets of points in the space $\mathbb{R}^{30}$: a set $\mathcal{B}$ of points corresponding to the benign tumors, and a set $\mathcal{M}$ of malignant tumors. Our aim is to compute a plane in $\mathbb{R}^{30}$ which separates these two sets, as much as possible. When a new sample point is obtained, we diagnose it as "benign" or "malignant" depending on whether it lies on the "$\mathcal{B}$ side" of the separating plane, or on the "$\mathcal{M}$ side".

We say "as much as possible" because it may be that the data in $\mathcal{B}$ and $\mathcal{M}$ is interspersed in a way that makes it impossible to find a plane that performs the separation cleanly. In this case, we seek a plane that minimizes some measure of the classification error. This classification error for a point is strictly positive if it lies on the wrong side of the plane (for example, the point represents a benign sample but lies on the "$\mathcal{M}$ side" of the plane), and zero otherwise.

Algebraically, the separating plane is defined as a linear function f with the following desired property:

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

This function is given by $f(x) = w'x - \gamma$, where $w \in \mathbb{R}^{30}$ and $\gamma \in \mathbb{R}$ are to be determined from the training data. The quantities w and γ are determined

by solving a quadratic program in MATLAB. The form of this quadratic program is similar to one proposed in [1, 4]; see also [3].

If represent the sets of m points $\mathcal{M}$ by a matrix $M \in \mathbb{R}^{m\times n}$ and the set of k points $\mathcal{B}$ by a matrix $B \in \mathbb{R}^{k\times n}$, then the problem becomes one of choosing w and γ to solve the following minimization problem:

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

Here e_m and e_k are vectors of lengths m and k, respectively, whose entries are all 1, while $((z)_+)_i = \max\{z_i, 0\}$, $i = 1, 2, \ldots, m$ and $\|z\|_1 = \sum_{i=1}^m |z_i|$ for $z \in \mathbb{R}^m$. 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 $\|\cdot\|_1$ functions can be eliminated to yield the following linear programming reformulation:

$$\begin{array}{c} \min_{w,\gamma,y,z} \left(\frac{1}{m}e'y + \frac{1}{k}e'z\right) \quad \text{subject to} \\ \begin{array}{rcl} Mw - e\gamma + y & \geq & e, \\ -Bw + e\gamma + z & \geq & e, \\ y \geq 0,\ z \geq 0. & & \end{array} \end{array}$$

If the data sets $\mathcal{M}$ and $\mathcal{B}$ are separable, then this linear program may have multiple solutions. In order to choose an appropriate solution from these, typically w is chosen to maximize the "separation margin" between the two datasets. It can be shown that the separation margin is given by the reciprocal of $\|w\|_2$, so we modify the above formulation to add a multiple of the two-norm of w to the objective:

$$\begin{array}{c} \min_{w,\gamma,y,z} \left(\frac{1}{m}e'y + \frac{1}{k}e'z\right) + \frac{\mu}{2}w'w \quad \text{subject to} \\ \begin{array}{rcl} Mw - e\gamma + y & \geq & e, \\ -Bw + e\gamma + z & \geq & e, \\ y \geq 0,\ z \geq 0. & & \end{array} \end{array}$$

Here μ is a penalty parameter. As its value is increased, the norm of the solution w tends to decrease. **This quadratic programming formulation is the one you will work with in this project.**

You can obtain a Matlab file `wdbcData.m` that contains a function that reads the data from `wdbc.data` and stores it in test and training matrices. See the comments at the start of this file for details.

The project consists of the following four parts.

1. Write code in Matlab and CVX to solve the QP formulation above.

   (a) Test your routine on the training data obtained by setting `fracTest=0.1` and `reord=0` in `help wdbcData`, using the value $\mu = 0.001$. (This will yield a training data set with 512 data points, consisting of records 1 through 512 from the file `wdbc.data`.)

   (b) Test your routine on the training data obtained by setting `fracTest=0.15` and `reord=1` in `help wdbcData`, using $\mu = 0.001$. (This will yield a training matrix of 484 records randomly selected from the 569 samples in the file `wdbc.data`.)

   Make sure you print out w, γ and the optimal objective of the quadratic program. Also, print out the number of misclassified points in the training set — the number of points in the training set that lie on the wrong side of the calculated plane.

2. By modifying your code for part 1, write a program to obtain the separating plane on the training set, and then determine the number of misclassified points on the corresponding *testing set*, for the following cases (use $\mu = 0.001$ for each):

   (a) `fracTest=0.1` and `reord=0`

   (b) `fracTest=0.15` and `reord=0`

   (c) `fracTest=0.20` and `reord=1`

   (d) `fracTest=0.05` and `reord=1`

   Print `fracTest`, `reord`, and the number of misclassified points in the testing set, for each case. (Do not print w, γ, or the optimal objective for these cases.)

3. Suppose that the oncologist wants to use only 2 of the 30 attributes to make the diagnosis. Determine which pair of attributes is most effective in obtaining a correct diagnosis as follows. First, obtain a training set by setting `fracTest=0.12` and `reord=0`. Considering each of the $\binom{30}{2} = 435$ pairs of possible attributes, use the training set and the formulation above to determine a separating plane in $\mathbb{R}^2$. *Use $\mu = 0.0008$ throughout.* For each plane use the training set with the corresponding pair of attributes to determine the number of misclassified cases.

Each time you find an attribute pair that gives the fewest number of training-set misclassifications encountered so far, print out a line of the form

```
fprintf('attributes %2d %2d: misclass %3d\n',i,j, wrong);
```

(Note: Do **not** print out this line for every one of the 435 pairs! Only do it when you find a pair that gives the best results so far.)

4. Apply the best performing answer from Part 3 above to the *testing* set. First, determine the number of incorrectly classified points in the testing set. Then, plot all the testing set points on a two dimensional figure using MATLAB's plotting routines. Use 'o' for benign points and '+' for malignant points in the plot. Use MATLABcommands to draw the calculated separating plane $w'x = \gamma$ on the plot. Check to see if the number of misclassified points agrees with the plot, and comment. (Note that some points may coalesce, so you may want to randomly perturb points by a small amount to visualize all the points.)

Hand in listings of your code and output, and a short (approximately one page) summary and discussion of your results.

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] P. S. Bradley and O. L. Mangasarian. Feature selection via concave minimization and support vector machines. In J. Shavlik, editor, *Machine Learning Proceedings of the Fifteenth International Conference(ICML '98)*, pages 82–90, San Francisco, California, 1998. Morgan Kaufmann. ftp://ftp.cs.wisc.edu/math-prog/tech-reports/98-03.ps.

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

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