Comp Sci 525 Computing Project
The Disputed Federalist Papers

Linear programming can be used to solve problems in many applications. In this project, we will use linear programming to detemine the authorship of the disputed federalist papers.

The Federalist Papers were written in 1787-1788 by Alexander Hamilton, John Jay and James Madison to persuade the citizens of the State of New York to ratify the U.S. Constitution. As was common in those days, these 77 shorts essays, about 900 to 3500 words in length, appeared in newspapers signed with a pseudonym, in this instance, "Publius". In 1778 these papers were collected along with eight additional articles in the same subject and were published in book form. Since then, the consensus has been that John Jay was the sole author of five of a total 85 papers, that Hamilton was the sole author of 51, that Madison was the sole author of 14, and that Madison and Hamilton collaborated on another three. The authorship of the remaining 12 papers has been in dispute; these papers are usually referred to as the disputed papers. It has been generally agreed that the disputed papers were written by either Madison or Hamilton, but there was no consensus about which were written by Hamilton and which by Madison.

The data is available in the file federalData.mat in the public directory `~cs525-1/public`. The data was obtained from [2]. The file contains a data matrix with 118 lines of data, one line per paper. The first entry in each line contains the code number of the author, 1 for Hamilton (56 total), 2 for Madison (50 total) and 3 for the disputed papers (12 total). The remaining entries contain 70 floating point numbers that correspond to the relative frequencies (number of occurrences per 1000 words of the text) of the 70 function words that are also available in the data file as an array of strings.

The idea of the project is to come up with a discriminant function (a separating plane in this case) to determine if an disputed paper was authored

by Hamilton or Madison. In order to do this, you will use the papers with known authors as a "training set" to generate your separating plane and the disputed papers as a "testing" set to test your separating plane. The separating plane, to be determined by solving a single linear program in MATLAB is based on a formulation proposed in [1, 5]. Further background of this project can also be found in [4]. A description of the linear program that generates the separating plane follows. The separating plane is subsequently used to determine the authorship of the disputed papers.

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

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

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 Madison's papers from Hamilton's in R^{70}. It remains to show how to determine $w \in \mathrm{R}^{70}$ 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 $\mathcal{H}$ be represented by a matrix $H \in \mathrm{R}^{k \times n}$, then the problem becomes one of choosing w and γ to

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

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 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;

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

If the datasets 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 $\|w\|$, so we modify the above formulation to add a multiple of the one-norm of w to the objective:

$$\begin{array}{ll} \min_{w,\gamma,y,z,s} & \left(\frac{1}{m} e'y + \frac{1}{k} e'z\right) + \mu e's \\ \text{s.t.} & Mw - e\gamma + y \geq e, -Hw + e\gamma + z \geq e, -s \leq w \leq s, y \geq 0, z \geq 0 \end{array}$$

This is the formulation [3] you should attempt to implement and solve.

On the departmental unix machines, a standard routine `cplexlp` written in MATLAB is provided for solving linear programs of the form

$$\min \left\{ c'x \mid Ax \leq b, Cx = d, \bar{l} \leq x \leq \bar{u} \right\}$$

which can be called by using `x=cplexlp(c,A,b,C,d,lb,ub)`, where `lb` $= \bar{l}$ and `ub` $= \bar{u}$. Type `help cplexlp` within MATLAB to get more information about this linear programming routine.

The project consists of the following four parts:

1. Formulate the problem as a linear program. Construct a tuning set based on extracting papers 31 to 50 from the training data, and removing these papers from the testing data. Solve the problem using the matrices M and H formed from the training set. The 20 (known authorship) papers in the tuning set will be used as indicated below. Try three values of μ, 0.1, 0.5 and 1.0. **Make sure you print out w, γ and the minimum value of the LP.**

2. Test the separating plane on the 20 papers of the tuning set. Report the number of misclassified papers on the tuning set. It is probably a good idea if you create an m-file to do this. What is the effect of μ? What is the best value of μ? Use this best value of μ for the remainder of the project. For this value of μ predict the authorship of the twelve disputed papers.

3. Suppose that you want to use only 2 of the 70 attributes (word counts) for your prediction. Determine which pair of attributes is most effective in determining a correct prediction as follows. Use each of $\binom{70}{2} = 2415$ pairs of possible attributes and for each pair determine from the training set a separating plane in R^2. For each plane use the tuning 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 tuning set for each pair of attributes using

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

   Also report the words that you used in your classifier.

4. Use the best performing answer from Part 3 above (break ties using a specified procedure), predict the authorship of each of the twelve disputed papers and plot all the data points on a two dimensional figure using MATLAB's built–in plotting routines. Use 'o' for Hamilton papers and '+' for Madison papers, and '*' for disputed papers in the plot. Then use MATLAB to draw in the calculated line $w'x = \gamma$. Check to see if the number of papers predicted for each author agrees with the plot and comment. (Note that some points may coalesce, so you may want to randomly perturb the points by a small amount to visualize all these points).

   Hand in your results and m-files.

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] R. Bosch and J. A. Smith. Separating hyperplanes and the authorship of the disputed federalist papers. *American Mathematical Monthly*, 105(7):601–608, 1998.

[3] 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.

[4] G. Fung. The disputed federalist papers: SVM feature selection using concave minimization. In *Proceedings of the 2003 Conference of Diversity in Computing*, pages 42–46, Atlanta, Georgia, 2003.

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