

525 Computing Project, Fall 2009*

The Disputed Federalist Papers

Linear and quadratic programming can be used to solve problems in many applications. In this project, we will apply quadratic programming to a machine learning formulation to determine 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, obtained from [1], is available in the file `federalData.mat` in the public directory `~cs525-1/public`. The file contains a data matrix with 118 lines of data, one line per paper. (A number of other papers with known authorship of either Hamilton or Madison were added to the Federalist Papers mentioned above, to provide extra data on the vocabularial habits of the two authors.) The first entry in each line contains the code number of the author, 1 for Hamilton (56 paper total), 2 for Madison (50 paper total), and 3 for the disputed papers (12 total). The remaining entries contain 70

*Due at 5:00pm on Dec 14, 2009

floating point numbers that correspond to the relative frequencies (number of occurrences per 1000 words of the text) of the 70 function words, which 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 hyperplane in this case) to determine if a disputed paper was authored by Hamilton or Madison. To do this, you will divide the papers with known authors into a “training set” and a “tuning set” to generate your separating plane. The disputed papers will form a “testing set”. Once we have calculated the separating plane, we’ll test to see which side of the plane the disputed papers lie on. An appropriate plane can often be determined by solving a linear or quadratic program, as we now describe.

A linear function f will be constructed to “usually” have the following property:

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

where $f(x) = w'x - \gamma$, with $w \in \mathbb{R}^{70}$ and $\gamma \in \mathbb{R}$ to be determined from the training data. In other words, the plane $w'x = \gamma$ separates (most of) Madison’s papers from Hamilton’s in the space \mathbb{R}^{70} of word frequencies. We determine $w \in \mathbb{R}^{70}$ and $\gamma \in \mathbb{R}$ by solving a quadratic program involving the training data.

If we 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{H} by a matrix $H \in \mathbb{R}^{k \times n}$, then the problem becomes one of choosing w and γ to solve the following minimization problem:

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

Here e_m and e_k are vectors of lengths m and k , respectively, whose entries are all 1, while $((z)_+)_i = \max\{z_i, 0\}$, $i = 1, 2, \dots, m$ 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 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:

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

If the data sets \mathcal{M} and \mathcal{H} are separable, then this linear program may have multiple solutions. The “best” of these solutions can be defined to be

the one that maximizes 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:

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

Here μ is a penalty parameter. As its value is increased, $\|w\|_2$ tends to decrease. This formulation is effective even when the two sets are *not* separable. **This quadratic programming formulation is the one you will work with in this project.**

On the CS department unix machines, a standard routine `cplexqp` written in MATLAB is provided for solving quadratic programs of the form

$$\min \left\{ c'x + \frac{1}{2}x'Qx \mid Ax \leq, =, \geq b, \bar{l} \leq x \leq \bar{u} \right\}$$

which can be called by using `[obj,x]=cplexqp(c,A,b,Q,lb,ub,le,ge)`, where `le` (`ge`) are indices of less-than (greater-than) inequalities. Type `help cplexqp` within MATLAB to get more information about this quadratic programming routine.

The MATLAB routine `getFederalistData.m` (available in `~cs525-1/public`) parses the data in `federalData.mat` to produce a training matrix `train` (consisting of 86 entries with known authorship), a tuning matrix `tune` (consisting of the other 20 papers of known authorship), and a testing matrix `test` (consisting of the 12 entries with unknown authorship).

Answer the following questions.

1. Write MATLAB code to solve the QP to find the classifying hyperplane, where the matrices M and H are extracted from the training matrix `train` described above. Write a loop to solve the QP for these values of μ : 0, .001, .01, .1, 1, 10, and 100. For each value of μ , calculate and print the following information:
 - the optimal QP objective,
 - the values of γ and $\|w\|_2$,

- the number of misclassified points (those that lie on the wrong side of the classifying hyperplane) from the training set, which is described in the matrix `train`.
 - the number of misclassified points from the tuning set, which is described in the matrix `tune`.
 - the predictions of authorship for the 12 disputed papers. (To calculate this prediction, take each of the rows in `test` and determine which side of the classifying hyperplane they lie on.)
2. From your results in Q.1, comment on the following questions: What is the effect of μ on $\|w\|_2$ and on the quality of the classifying hyperplane? What is the best value of μ ?
 3. Fix $\mu = .05$ for purposes of this question and the next. Suppose that you want to use only 2 of the 70 attributes (word frequencies) for your prediction of authorship. Determine which pair of attributes is most effective in determining a correct prediction as follows. For each of the $\binom{70}{2} = 2415$ pairs of possible attributes, determine a separating hyperplane in \mathbb{R}^2 . (Note that for these problems, $n = 2$ in our formulation above, in contrast to $n = 70$ when we use all the features.) For each plane, determine the number of misclassified cases from the *tuning* set. 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);
```

where `i` and `j` denote the attribute pair and `wrong` indicates the number of misclassified points from the `tune` matrix. *In your submitted output, print this line for only those pairs that include either 10 or 11 as one of the two attributes.*

Keep track of the best-performing two-attribute classifier (breaking ties using some procedure of your own design). Report the best classifier, and the two words that it uses, along with the number of misclassified tuning points for this classifier.

4. Use the optimal classifier from Q. 3, predict the authorship of each of the twelve disputed papers. Plot all the data points according to the “best” two attributes 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. Use MATLAB draw in the calculated line $w'x = \gamma$. Check to see if the number of misclassified papers shown in your plot agrees with your calculations. (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 hard copies of your results and m-files.

References

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