

525 Computing Project 2, Spring 2012*

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 short 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 [?], is available on the course website (`federalData.mat`). 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 papers in total), 2 for Madison (50 papers in total), and 3 for the disputed papers (12 in total). The remaining entries contain 70 floating point numbers that

*Due in class on Dec 5, 2011

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 compute a linear support vector machine (as we have discussed in class and in Chapter 9 in the text) to determine if an 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” for separating plane. The disputed papers form a “testing set”. Once we have calculated the separating plane, we will test to see which side of the plane the disputed papers lie on.

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). Each row of `train` and `tune` consists of a label followed by a feature vector in \mathbb{R}^{70} .

Use the following quadratic programming formulation of the SVM for this problem:

$$\begin{aligned} & \text{minimize} && \frac{1}{N} \sum_{k=1}^N s_k + \frac{\mu}{2} w^T w \\ & \text{subject to} && y_i(w^T x_i + b) \geq 1 - s_i \quad \text{for } i = 1, \dots, N \\ & && s \geq 0 \end{aligned}$$

where (w, b) are the decision variables we’d like to learn and N is the total number of training examples. Set $y_i = 1$ for Hamilton and $y_i = -1$ for Madison. Each x_i is an example vector in \mathbb{R}^{70} . The output (w, b) will parametrize a linear classifier f such that

$$f(x_i) = w^T x_i + b$$

will be mostly greater than zero on the Hamilton documents and mostly less than zero on the Madison documents. μ is the regularization parameter. Large values of μ result in small values of $\|w\|_2$.

You can use any solver you would like. I would recommend using the interpreter `cvx` (<http://cvxr.com/cvx/>) with a free solver like SDPT3 (<http://www.math.nus.edu.sg/~mattohkc/>). You can also use CPLEX in the CS labs (see the course webpage for more details) or you can use Matlab’s `quadprog`.

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 b 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.) Express your result with one line of output for each of the 12 papers, indicating the paper number, predicted author, and margin (the value of $w'x + b$ for the x corresponding to this paper).
2. From your results in Q.1, answer the following questions, and give reasons for your answers. 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 $\mu = .1$ 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. 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 $w'x = -b$. 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.