CS 635 Project

Spring Semester 1999

Glenn Fung

1 Introduction

1.1 *The 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.

1.2 Mosteller and Wallace (1964)

In 1964 Mosteller and Wallace in the book *"Inference and Disputed Authorship: The Federalist"* [5] using statistical inference concluded: *"In summary, we can say with better foundation than ever before that Madison was the author of the 12 disputed papers".*

1.3 Robert A. Bosch and Jason A. Smith (1998)

In 1998 Bosch and Smith [2] used a method proposed by Bennett and Mangasarian [1], that utilize linear programming techniques to find a separating hyperplane. *Cross-validation* was used to evaluate every possible set comprised of one, two or three of the 70 function words. They obtained the following hyperplane:

$$-0.5242as + 0.8895our + 4.9235upon = 4.7368.$$

using this hyperplane they found that all the 12 of the disputed papers ended up on the Madison side of the hyperplane, that is (> 4.7368).

2 My Previous work in this Subject

Using the method proposed by Bradley and Mangasarian , *"Feature Selection via Concave Minimization and Support Vector Machine"* [3] I obtained the following separating hyperplane in three dimensions:

$$-0.5368to - 24.6634upon - 2.9532would = -66.6159,$$

1	*a*	15	*do*	29	*is*	43	*or*	57	*this*
2	*all*	16	*down*	30	*it*	44	*our*	58	*to*
3	*also*	17	*even*	31	*its*	45	*shall*	59	*up*
4	*an*	18	*every*	32	*may*	46	*should*	60	*upon*
5	*and*	19	*for*	33	*more*	47	*so*	61	*was*
6	*any*	20	*from*	34	*must*	48	*some*	62	*were*
7	*are*	21	*had*	35	*my*	49	*such*	63	*what*
8	*as*	22	*has*	36	*no*	50	*than*	64	*when*
9	*at*	23	*have*	37	*not*	51	*that*	65	*which*
10	*be*	24	*her*	38	*now*	52	*the*	66	*who*
11	*been*	25	*his*	39	*of*	53	*their*	67	*will*
12	*but*	26	*if*	40	*on*	54	*then*	68	*with*
13	*by*	27	*in*	41	*one*	55	*there*	69	*would*
14	*can*	28	*into*	42	*only*	56	*things*	70	*your*

Table 1: Function Words and Their Code Numbers

which was obtained with $\alpha = 18$ and $\lambda = 0.03$ starting from a random point. This hyperplane classified all the training data correctly and all the 12 of the disputed papers ended up on the Madison side of the hyperplane (> -66.6159).

3 Finding a non-linear Separating Surface Using only two Features

In this project I used the theory proposed by Mangasarian in *"Generalized Support Vector Machines"* [4] in order to find a separating non linear function that classifies correctly all the training set papers using the minimum possible number of features.

3.1 The data

The data used in this project is identical to the data used by Bosch and Jason. They first produced machine-readable texts of the papers with a scanner and after that they used Macintosh software to compute relative frequencies for 70 function words that Mosteller and Wallace identified as good candidates for author-attribution studies.

I got the data from Bosch in a text file. The file contains 118 pairs of lines of data, one pair per paper. The first line in each pair contains two numbers: the code of the paper (see pages 269 and 270 of [5]) and 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 second line contains 70 floating point numbers that correspond to the relative frequencies (number of ocurrences per 1000 words of the text) of the 70 function words (See Table 1). Since all the calculations and the programming was made with MATLAB, I put all the data in MATLAB format in different arrays containing all the data.

3.2 Brief Description of the method used

In order to find a nonlinear separating surface I used the formulation (6) proposed in [4]:

$$\min_{u,\gamma,y} ve^T y + \frac{1}{2}u^T Hu$$

subject to

$$D(K(A, A^T)Du - e\gamma) + y \geq e$$

$$y \geq 0.$$

Once this problem quadratic problem is solved the separating surface induced by the kernel K is given by:

$$K(x^T, A^T)Du = \gamma$$

Here A is the data matrix, in our case each row of A represent a paper (106) and each column of A represents a feature (70), so A is a 106×70 matrix. In other words A is given by:

$$A = \left(\frac{H}{M}\right)$$

Where, $M \in \Re^{50\times 70}, H \in \Re^{56\times 70}$are matrices containing the corresponding relative frequencies for all the 70 function words for the Madison (M) and the Hamilton (H) attributed papers, $m = 50, k = 56$ and e is a vector of ones of suitable dimension.

D is a given diagonal matrix of plus or minus 1, depending on whether A_i belongs to class 1 or -1. H can be any 106×106 positive definite matrix, in our case $H = I$. I tried different values v ranging from 0.5 to 100 and finally $v = 60$ was chosen. I also choose a simple quadratic kernel:

$$K(A, B) = (AB + ee^t)^2 \tag{1}$$

After that, I used the technique known as *ten fold cross validation* in order to choose the best two features to work with, so I divided my original set of 106 papers in ten smaller subsets of sizes 10 and 11, and for each pair of features I solve the problem 10 times taking every time one of the smaller subsets as testing set and the reminding ones as a training set, it means I solve about $2415 \times 10 = 24150$ problems.

3.3 the Results

I used MATLAB to manipulate all the matrix structures, calculate all the K values and to plot the graphics showing the obtained results. In order to solve the required quadratic problem once all the computations where made in MATLAB I called GAMS from the MATLAB prompt using the interface described in *MATLAB and GAMS : Interfacing Optimization and Visualization Software* by Michael Ferris. In GAMS I used the solver MINOSNEW which gave me the faster times in previous tests.

3.3.1 Using the Quadratic kernel

After the ten cross validation procedure I selected the ten pair of features with the best overall performance in the testing set, However, in some of the obtained surfaces the results did not coincide with previous results (all the disputed papers not in the Madison side).Figure 1 shows one of the obtained surfaces using the quadratic kernel described above. I think this can be improved with more experimentation with the parameter v and different scalings of the data.

3.3.2 Using a 5 Degree Polynomial kernel

Then, I tried with a similar polynomial kernel of degree 5, this time I got one that separates all the data in the training set correctly and all the disputed papers were in the Madison side of the non linear surface (See figure 2 and 3).

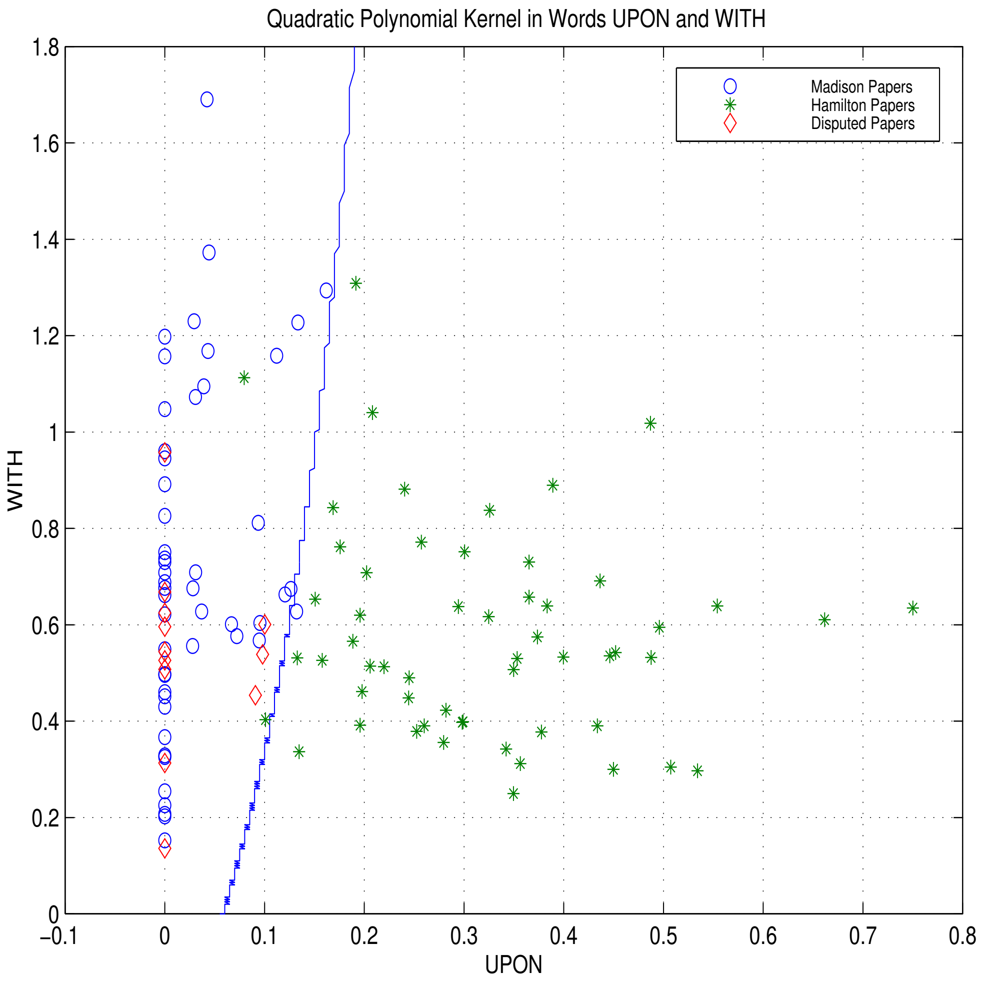

Figure 1: Separation using a Quadratic Kernel

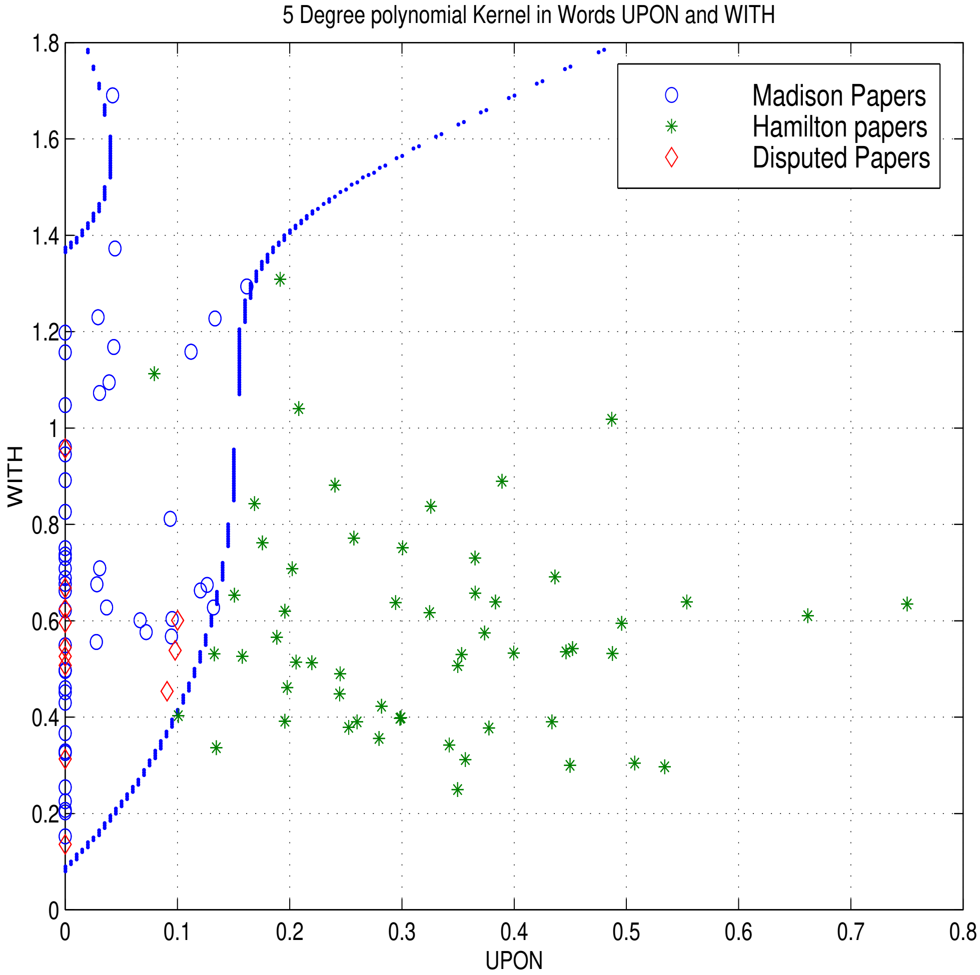

Figure 2: Separation using a 5 degree Kernel

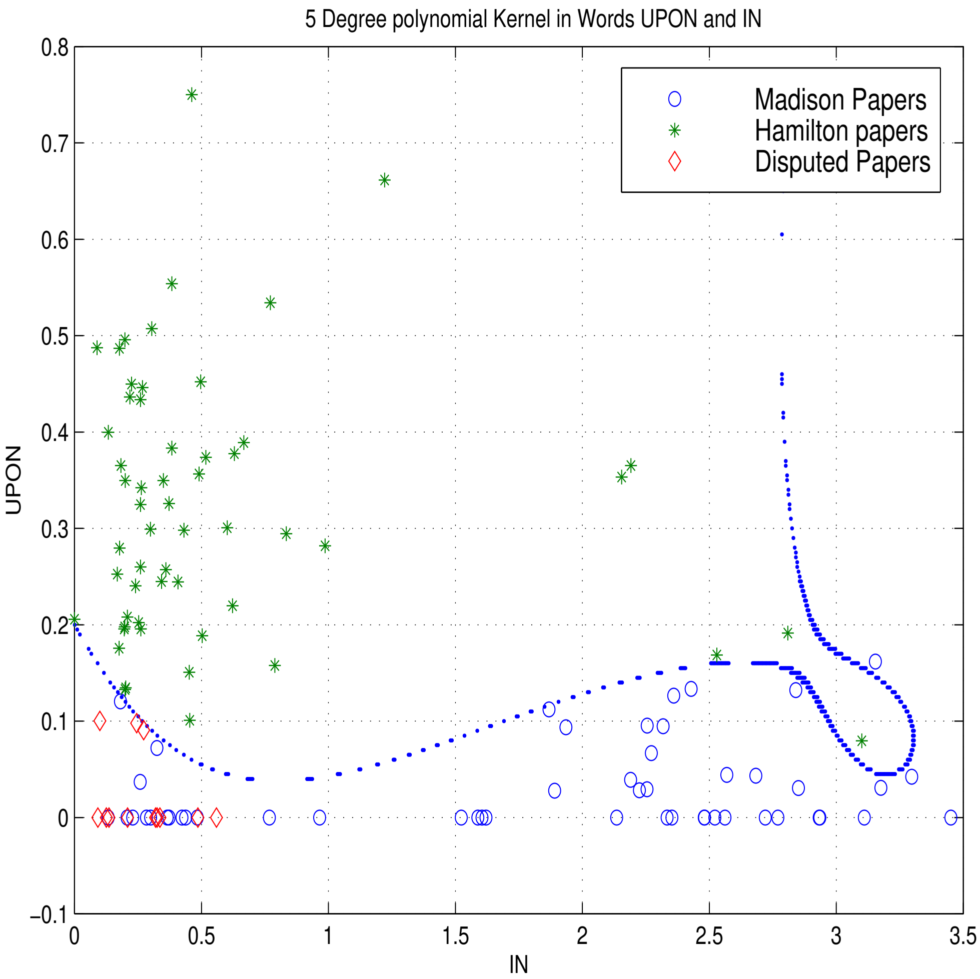

Figure 3: Complete Separation using a 5 degree Kernel

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. A. Bosch and J. A. Smith. Separating hyperplanes and the authorship of the disputed fedaralist papers. *American Mathematical Monthly*, 105(7):601–608, August-September 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] O. L. Mangasarian. Generalized support vector machines. In A. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 135–146, Cambridge, MA, 2000. MIT Press. ftp://ftp.cs.wisc.edu/math-prog/tech-reports/98-14.ps.

[5] F. Mosteller and D. L. Wallace. *Inference and Disputed Authorship: The Federalist*. Addison-Wesley, Massachusetts, series in behavioral science:quantitative methods edition, 1964.