# The Disputed Federalist Papers: SVM Feature Selection via Concave Minimization

Glenn Fung
Siemens Medical Solutions

In this paper, we use a method proposed by Bradley and Mangasarian “Feature Selection via Concave Minimization and Support Vector Machines” to solve the well-known disputed *Federalist Papers* classification problem. We find a separating plane that classifies correctly all the ”training set” papers of known authorship, based on the relative frequencies of only three words. Using the obtained separating hyperplane in three dimensions, all of the 12 disputed papers ended up on the Madison side of the separating plane. This result coincides with previous work on this problem using other classification techniques.

Categories and Subject Descriptors: I.2.7 [**Text Analysis**]: Natural Language Processing; G.1.6 [**Linear programming**]: Optimization; K.3.2 [**Concept Learning**]: learning

General Terms: Algorithms, Experimentation, Performance

Additional Key Words and Phrases: Text Classification, Support Vector Machines

## 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”* [1964] using statistical inference concluded: *“In summary, we can say with better foundation than ever before that Madison was the author of the 12*

Author’s address: Glenn Fung, Computer Aided Diagnosis & Therapy Solutions, Siemens Medical Solutions, 51 Valley Stream Parkway, Malvern PA 19355.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 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 I. Function Words and Their Code Numbers

*disputed papers".*

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

In 1998 Bosch and Smith [Bosch and Smith 1998] used a method proposed by Bennett and Mangasarian [Bennett and Mangasarian 1992], 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:

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

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

### 1.4 Description of the Data

The data we used in this project is identical to the data used by Bosch and Smith [Bosch and Smith 1998]. They first produced machine-readable texts of the papers with a scanner and then they used Macintosh software to compute relative frequencies for 70 function words that Mosteller and Wallace identified as good candidates for author-attribution studies.

The data was obtained 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 [Mosteller and Wallace 1964]) 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 occurrences per 1000 words of the text) of the 70 function words (See Table 1).

Based on the relative frequencies of the words, each paper can be represented as a vector with 70 real-valued components. This means our *training* dataset is now represented by a real-valued matrix $A \in R^{106 \times 70}$ where each row of the matrix represents a federalist paper which authorship is already known. We also define a diagonal label matrix $D$ containing the information of the labels of the training dataset. If a training datapoint $A_i$ belongs to the "Madison class" then $d_{ii} = 1$, if it belongs to the "Hamilton class" then $d_{ii} = -1$.

## 2. NOTATION

We now describe our notation and give some background material. All vectors will be column vectors unless transposed to a row vector by a prime $'$. For a vector $x$ in the $n$-dimensional real space $R^n$, $|x|$ will denote a vector in $R^n$ of absolute values of the components of $x$. For a vector $x \in R^n$, $x_*$ denotes the vector in $R^n$ with components $(x_*)_i = 1$ if $x_i > 0$ and 0 otherwise (i.e. $x_*$ is the result of applying the step function component-wise to $x$). The base of the natural logarithm will be denoted by $\varepsilon$, and for a vector $y \in R^m$, $\varepsilon^{-y}$ will denote a vector in $R^m$ with components $\varepsilon^{-y_i}$, $i = 1, \ldots, m$. For $x \in R^n$ and $1 \leq p < \infty$, the $p$-norm and the $\infty$-norm are defined as follows:

$$\|x\|_p = \left( \sum_{j=1}^{n} |x_j|^p \right)^{\frac{1}{p}}, \ \|x\|_\infty = \max_{1 \leq j \leq n} |x_j|.$$

The notation $A \in R^{m \times n}$ will signify a real $m \times n$ matrix. For such a matrix $A'$ will denote the transpose of $A$, and $A_i$ will denote the $i$-th row of $A$. A column vector of ones in a real space of arbitrary dimension will be denoted by $e$. Thus, for the column vectors $e$ and $y$ in $R^m$, the scalar product $e'y$ denotes the sum $\sum_{j=1}^{m} y_i$. A vector of zeros in a real space of arbitrary dimension will be denoted by 0. A separating plane, with respect to two given point sets $\mathcal{A}$ and $\mathcal{B}$ in $R^n$, is a plane that attempts to separate $R^n$ into two halfspaces such that each open halfspace contains points mostly of $\mathcal{A}$ or $\mathcal{B}$. A real valued function $f(x)$ on $R^n$ is concave ("mountain-like") if linear interpolation between two function values never overestimates the function.

## 3. THE LINEAR SUPPORT VECTOR MACHINE

We consider the problem, depicted in Figures 1 and 2, of classifying $m$ points in the $n$-dimensional real space $R^n$, represented by the $m \times n$ matrix $A$, according to membership of each point $A_i$ in the class $A+$ or $A-$ as specified by a given $m \times m$ diagonal matrix $D$ with plus ones or minus ones along its diagonal. For this problem the standard support vector machine with a linear kernel [Vapnik 1995; Cherkassky and Mulier 1998] is given by the following quadratic program with parameter $\nu > 0$:

$$\begin{array}{rrcl} \min\limits_{(w,\gamma,y) \in R^{n+1+m}} & \nu e'y + \frac{1}{2}w'w & & \\ \text{s.t.} & D(Aw - e\gamma) + y & \geq & e \\ & y & \geq & 0. \end{array} \qquad (1)$$

Written in individual component notation, and taking into account that $D$ is a diagonal matrix of ± 1, this problem becomes:

$$\begin{array}{rrcl}
\min\limits_{(w,\gamma,y)\in R^{n+1+m}} & \nu\sum\limits_{i=1}^{m} y_i + \frac{1}{2}\sum\limits_{j=1}^{n} w_j^2 & & \\
\text{s.t.} & A_i w + y_i & \geq & \gamma + 1, \ \text{for} \ \ D_{ii} = 1 \\
 & A_i w - y_i & \leq & \gamma - 1, \ \text{for} \ \ D_{ii} = -1 \\
 & y_i & \geq & 0 \\
 & i & = & 1 \ldots = m.
\end{array} \tag{2}$$

Here, $w$ is the normal to the bounding planes:

$$\begin{array}{rcl}
x'w & = & \gamma + 1 \\
x'w & = & \gamma - 1,
\end{array} \tag{3}$$

and $\gamma$ determines their location relative to the origin. See Figure 1. The two classes are strictly linearly separable when the error variable $y = 0$ in (1)-(2), as in the case of Figure 1. If the two classes are strictly linearly separable the plane $x'w = \gamma + 1$ bounds the class $A+$ points, while the plane $x'w = \gamma - 1$ bounds the class $A-$ points as follows:

$$\begin{array}{rcl}
A_i w & \geq & \gamma + 1, \ \text{for} \ \ D_{ii} = 1 \\
A_i w & \leq & \gamma - 1, \ \text{for} \ \ D_{ii} = -1.
\end{array} \tag{4}$$

The linear separating surface is the plane:

$$x'w = \gamma, \tag{5}$$

midway between the bounding planes (3). The quadratic term in (1), which is twice the reciprocal of the square of the 2-norm distance $\frac{2}{\|w\|_2}$ between the two bounding planes of (3) (see Figure 1), maximizes this distance, often called the "margin". Maximizing the margin enhances the generalization capability of a support vector machine [Vapnik 1995; Cherkassky and Mulier 1998].

If the classes are linearly inseparable then the two planes bound the two classes with a "soft margin" (i.e. bound approximately with some error) determined by the nonnegative error variable $y$, that is:

$$\begin{array}{rcl}
A_i w \ + \ y_i & \geq & \gamma + 1, \ \text{for} \ D_{ii} = 1 \\
A_i w \ - \ y_i & \leq & \gamma - 1, \ \text{for} \ D_{ii} = -1.
\end{array} \tag{6}$$

The 1-norm of the error variable $y$ is minimized parametrically with weight $\nu$ in (1) resulting in an approximate separation as depicted in Figure 2, for example. Points of $A+$ that lie in the halfspace $\{x \mid x'w \leq \gamma + 1\}$ (i.e. on the plane $x'w = \gamma + 1$ and on the wrong side of the plane) as well as points of $A-$ that lie in the halfspace $\{x \mid x'w \geq \gamma - 1\}$ are called *support vectors.*

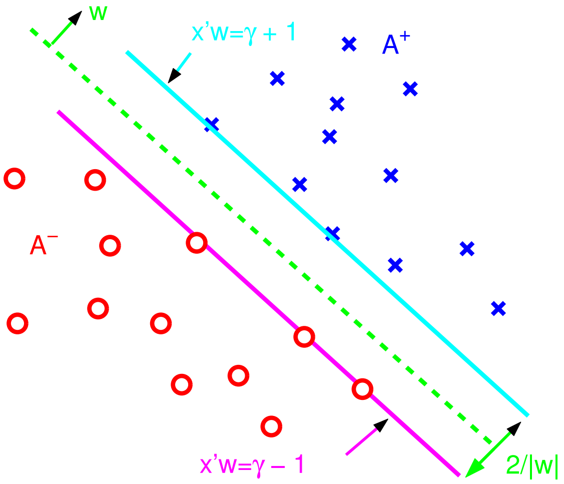

Fig. 1. **The Linearly Separable Case: The bounding planes of equation (3) with margin $\frac{2}{\|w\|_2}$, and the plane of equation (5) separating $A+$, the points represented by rows of $A$ with $D_{ii} = +1$, from $A-$, the points represented by rows of $A$ with $D_{ii} = -1$.**

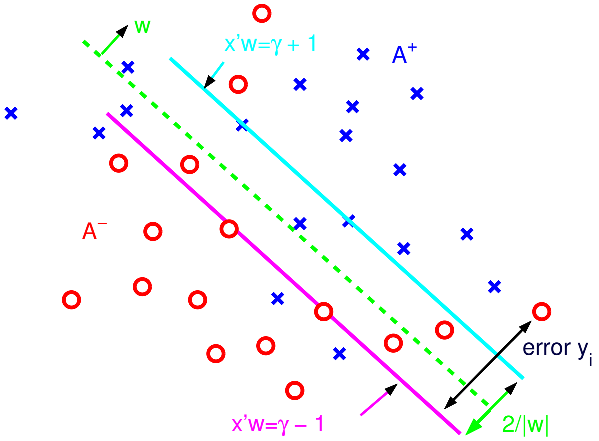

Fig. 2. **The Linearly Inseparable Case: The approximately bounding planes of equation (3) with a soft (i.e. with some error) margin $\frac{2}{\|w\|_2}$, and the plane of equation (5) approximately separating $A+$ from $A-$.**

Support vectors, which constitute the complement of the strictly separated points by the bounding planes (3), completely determine the separating surface. Minimizing the number of such exceptional points can lead to a minimum length description model [Mitchell 1997, p. 66],[Blumer et al. 1987] that depends on much fewer data points. Computational results indicate that such lean models generalize as well as or better than models that depend on many more data points. In the next section of the paper an algorithm that minimizes the norm of the error for the misclassified points as well as the number of input space features.

## 4. FEATURE SELECTION VIA CONCAVE MINIMIZATION

In order to make use of a faster linear programming based approach, instead of the standard quadratic programming formulation (1), we reformulate (1) by replacing the 2-norm by a 1-norm as follows [Mangasarian 2000; Bradley and Mangasarian 1998]:

$$\begin{array}{rl} \displaystyle\min_{(w,\gamma,y)\in R^{n+1+m}} & \nu e'y + \|w\|_1 = \nu\displaystyle\sum_{i=1}^{m} y_i + \sum_{j=1}^{n} |w_j| \\ \text{s.t.} & D(Aw - e\gamma) + y \geq e \\ & y \geq 0. \end{array} \tag{7}$$

This SVM$\|\cdot\|_1$ reformulation in effect maximizes the margin, the distance between the two bounding planes of Figures 1 and 2, using a different norm, the $\infty$-norm, and results with a margin in terms of the 1-norm, $\frac{2}{\|w\|_1}$, instead of $\frac{2}{\|w\|_2}$ [Mangasarian 1999]. The mathematical program (7) is easily converted to a linear program as follows:

$$\begin{array}{rl} \displaystyle\min_{(w,\gamma,y,v)\in R^{n+1+m+n}} & \nu e'y + e'v = \nu\displaystyle\sum_{i=1}^{m} y_i + \sum_{j=1}^{n} v_j \\ \text{s.t.} & D(Aw - e\gamma) + y \geq e \\ & v \geq w \geq -v \\ & y \geq 0, \end{array} \tag{8}$$

where, at a solution, $v$ is the absolute value $|w|$ of $w$. This fact follows from the constraints $v \geq w \geq -v$ which imply that $v_i \geq |w_i|,\ i = 1\ldots,n$. Hence at optimality, $v = |w|$, otherwise the objective function can be strictly decreased without changing any variable except $v$. We will modify this linear program so as to generate an SVM with as few nonzero components of $w$ as possible by adding an error term $e'|w|_*$ to the objective function, where $_*$ denotes the step function as defined in the Introduction. The term $e'|w|_*$ suppresses nonzero component of the vector $w$ and results in separating hyperplanes that depend on fewer features.

$$\begin{array}{rrcl} \displaystyle\min_{(w,\gamma,y,v)\in R^{n+1+m+n}} & \nu e'y + e'v_* & & \\ \text{s.t.} & D(Aw - e\gamma) + y & \geq & e \\ & v \geq\ w & \geq & -v \\ & y & \geq & 0. \end{array} \tag{9}$$

Note that $w_i = 0$ implies that the separating hyperplane does not depends on feature $i$, this is:

$$wx - \gamma = \sum_{i=1}^{n} w_i x_i - \gamma = \sum_{i/w_i \neq 0} w_i x_i - \gamma$$

The discontinuity of the step function term $e'v_*$ is handled by using an smooth concave exponential approximation on the nonnegative real line [Mangasarian 1996] as is done in [Bradley and Mangasarian 1998]. For $v \geq 0$, the approximation of the step vector $v_*$ of (9) by the concave exponential, $v_{i*} \approx 1 - \varepsilon^{-\alpha v_i}$, $i = 1, \ldots, m$, that is:

$$v_* \approx e - \varepsilon^{-\alpha v},\ \alpha > 0, \tag{10}$$

where $\varepsilon$ is the base of natural logarithms, leads to the following smooth reformulation of problem (9):

$$\begin{array}{rl} \min\limits_{(w,\gamma,y,v) \in R^{n+1+m+n}} & \nu e'y + e'(e - \varepsilon^{-\alpha v}) \\ \text{s.t.} & D(Aw - e\gamma) + y \geq e \\ & v \geq w \geq -v \\ & y \geq 0. \end{array} \tag{11}$$

Note that:

$$e'(e - \varepsilon^{-\alpha v}) = n - \sum_{i=1}^{n} \varepsilon^{-\alpha v_i}. \tag{12}$$

It can be shown [Bradley et al. 1998, Theorem 2.1] that, for a finite value of the parameter $\alpha$ (appearing in the concave exponential), the smooth problem (11) generates an *exact* solution of the nonsmooth problem (9). We note that this problem is the minimization of a concave objective function over a polyhedral set. Even though it is difficult to find a global solution to this problem, a fast successive linear approximation (SLA) algorithm [Bradley et al. 1998, Algorithm 2.1] terminates finitely (usually in 4 to 7 steps) at a stationary point which satisfies the minimum principle necessary optimality condition for problem (11) [Bradley et al. 1998, Theorem 2.2] and leads to a locally minimal number of nonzero $w$, that is, a solution depending in fewer features.

ALGORIHTM 4.1. *Successive Linearization Algorithm (SLA) for (11). Choose* $\nu, \mu, \alpha > 0$. *Start with some* $(w^0, \gamma^0, y^0, v^0)$. *Having* $(w^i, \gamma^i, y^i, v^i)$ *determine the next iterate by solving the linear program:*

$$\begin{array}{rl} \min\limits_{(w,\gamma,y,v) \in R^{n+1+m+n}} & \alpha(\varepsilon^{-\alpha v^i})'(v - v^i) \\ s.t. & D(Aw - e\gamma) + y \geq e \\ & v \geq w \geq -v \\ & y \geq 0. \end{array} \tag{13}$$

*Stop when:*

$$\nu e'(y - y^i) + e'(v - v^i) + \alpha(\varepsilon^{-\alpha v^i})'(v - v^i) \leq tol \tag{14}$$

*Comment: The parameter $\alpha$ was set to 5. The parameters $\nu$ and $\mu$ were chosen with the help of a tuning set surrogate for a testing set to simultaneously minimize the number of support vectors, number of input space features and tuning set error. The tolerance for the stopping criteria tol was set to* $10^{-5}$.

We turn our attention now to numerical implementation and testing.

## 5. NUMERICAL RESULTS

A ten-fold cross validation procedure was applied to determine an optimal value for the parameter $\nu$ based on the training data. For each fold I defined a tuning set consisting of 10% of the data on that fold for tuning purposes. Based on tuning set results we picked the best value of $\nu$ in the set $\{2^i/i = -6, -5, \ldots, 0, \ldots, 5, 6\}$.

The initial starting solution $(w^0, \gamma^0)$ was generated randomly using a uniform distribution function that generates random numbers between 1 and 100. Note that a "good" but somewhat more expensive choice of an initial estimates for the algorithm 4.1is the solution to the linear programming formulation problem 8.

Using the approach described above We obtained a separating hyperplane only depending on three features in three dimensions:

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

which was obtained with $\nu = 2^{-5}$ starting from a random point.

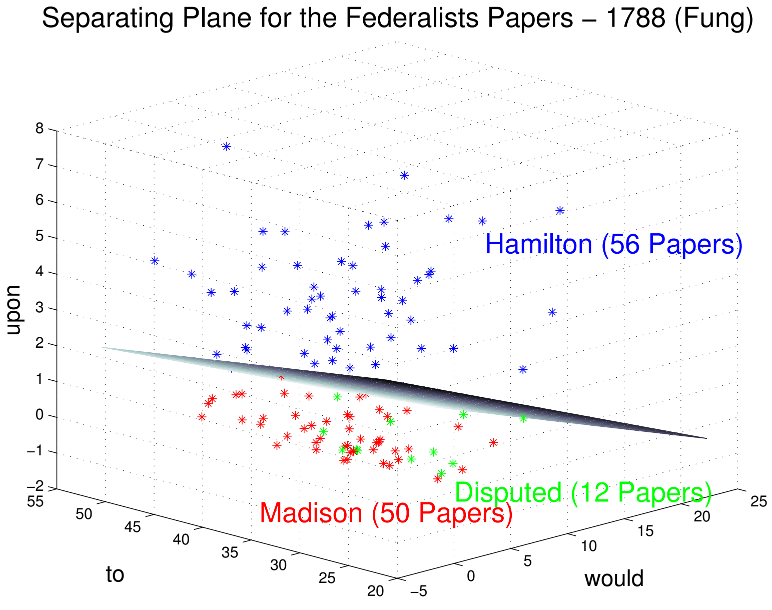

Fig. 3. Obtained Hyperplane in 3 dimensions

The final set of three features was chosen based on the number of times they appear through the ten-fold process. We obtained a hyperplane that classified all the training data correctly and all the 12 of the disputed papers ended up on the Madison side of the hyperplane ($> -66.6159$).

## 6. CONCLUSION

We have applied Support vector machine feature selection via concave minimization to solve the well-known disputed *Federalist Papers* classification problem. Our results are very similar with those obtained by Bosch and Smith [Bosch and Smith 1998]. Bosch and Smith tried to solve the problem using all the possibles combinations of 1,2 and 3 words out of 70. This method involves solving 57225 linear programming problems without considering the tuning phase. Instead using our approach we solve only 4 to 7 linear programming without including the tuning phase and around $5 \times 13 = 65$ including the tuning phase.

### ACKNOWLEDGMENTS

I wish to thank professor Olvi Mangasarian for all his guidance during the realization of this work and professor Robert Bosch for the data and the comments.

### REFERENCES

Bennett, K. P. and Mangasarian, O. L. 1992. Robust linear programming discrimination of two linearly inseparable sets. *Optimization Methods and Software 1*, 23–34.

Blumer, A., Ehrenfeucht, A., Haussler, D., and Warmuth, M. K. 1987. Occam's razor. *Information Processing Letters 24*, 377–380.

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

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

Bradley, P. S., Mangasarian, O. L., and Rosen, J. B. 1998. Parsimonious least norm approximation. *Computational Optimization and Applications 11,* 1 (October), 5–21. ftp://ftp.cs.wisc.edu/math-prog/tech-reports/97-03.ps.

Bradley, P. S., Mangasarian, O. L., and Street, W. N. 1998. Feature selection via mathematical programming. *INFORMS Journal on Computing 10,* 2, 209–217. ftp://ftp.cs.wisc.edu/math-prog/tech-reports/95-21.ps.

Cherkassky, V. and Mulier, F. 1998. *Learning from Data - Concepts, Theory and Methods.* John Wiley & Sons, New York.

Mangasarian, O. L. 1996. Machine learning via polyhedral concave minimization. In *Applied Mathematics and Parallel Computing - Festschrift for Klaus Ritter*, H. Fischer, B. Riedmueller, and S. Schaeffler, Eds. Physica-Verlag A Springer-Verlag Company, Heidelberg, 175–188. ftp://ftp.cs.wisc.edu/math-prog/tech-reports/95-20.ps.

Mangasarian, O. L. 1999. Arbitrary-norm separating plane. *Operations Research Letters 24*, 15–23. ftp://ftp.cs.wisc.edu/math-prog/tech-reports/97-07r.ps.

Mangasarian, O. L. 2000. Generalized support vector machines. In *Advances in Large Margin Classifiers*, A. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans, Eds. MIT Press, Cambridge, MA, 135–146. ftp://ftp.cs.wisc.edu/math-prog/tech-reports/98-14.ps.

Mitchell, T. M. 1997. *Machine Learning.* McGraw-Hill, Boston.

Mosteller, F. and Wallace, D. L. 1964. *Inference and Disputed Authorship: The Federalist*, Series in Behavioral Science:Quantitative Methods ed. Addison-Wesley, Massachusetts.

Vapnik, V. N. 1995. *The Nature of Statistical Learning Theory.* Springer, New York.