DIMACS Series in Discrete Mathematics
and Theoretical Computer Science

# Breast Cancer Survival and Chemotherapy: A Support Vector Machine Analysis

Y.-J. Lee, O. L. Mangasarian, and W. H. Wolberg

Abstract. A linear support vector machine (SVM) is used to extract 6 features from a total of 31 features in a dataset of 253 breast cancer patients. Five features are nuclear features obtained during a non-invasive diagnostic procedure while one feature, tumor size, is obtained during surgery. The linear SVM selected the 6 features in the process of classifying the patients into node-positive (patients with some metastasized lymph nodes) and node-negative (patients with no metastasized lymph nodes). Node-positive patients are typically those who undergo chemotherapy. The 6 features were then used in a Gaussian kernel nonlinear SVM to classify the patients into three prognostic groups: good (node-negative), intermediate (1 to 4 metastasized nodes) and poor (more than 4 metastasized nodes). Very well separated Kaplan-Meier survival curves were constructed for the three groups with pairwise p-value of less than 0.009 based on the logrank statistic. Patients in the good prognostic group had the highest survival, while patients in the poor prognostic group had the lowest. The majority (72.8%) of the good group did not receive chemotherapy, while the majority (87.5%) of the poor group received chemotherapy. Just over half (56.7%) of the intermediate group received chemotherapy. New patients can be assigned to one of these three prognostic groups with its associated survival curve, based only on 6 features obtained before and during surgery, but without the potentially risky procedure of removing lymph nodes to determine how many of them have metastasized.

## 1. Introduction

Support vector machines [**Vap95, CM98**] as well as linear programming approaches [**Man65, Man68, BM92**] have been extensively used in machine learning and data mining applications [**Man97, BMS98, BM98a, BM98b, BFM99**] as well as in medical applications [**MSW95, WSM93, WSHM95b, WSHM95a, WSM97**]. In this work we attempt to classify breast cancer patients using a criterion that is closely related to the decision whether a patient is prescribed to have chemotherapy treatment or not. This criterion is the presence of metastasized lymph nodes under the patient's armpits (node-positive) or their absence

1991 *Mathematics Subject Classification.* Primary 68T10, 62N01; Secondary 90C90.

The research described in this Data Mining Institute Report 99-10, December 1999, was supported by National Science Foundation Grants CCR-9729842 and CDA-9623632, by Air Force Office of Scientific Research Grant F49620-97-1-0326 and by the Microsoft Corporation.

(node-negative). Lymph nodes are removed during surgery in conjunction with the removal of the malignant tumor from the breast. This potentially risky procedure which can cause arm swelling and increased susceptibility to infection can possibly be eliminated by using the classification procedures proposed here. By using a linear support vector machine we first select 6 out of 31 available features to classify patients into node-positive and node-negative patients. We then use these six features in a nonlinear support vector machine to classify 253 breast cancer patients into three prognosis groups: a good prognosis group (GPG) which is node-negative, an intermediate prognosis group (IPG) which has 1-4 metastasized lymph nodes, and a poor prognosis group (PPG) which has more than 4 metastasized lymph nodes. It turns out that these three groups have well separated survival curves [**KM58, Kle96**] with very small inter-curve p-values (Figure 3). In addition, 72.8% of the patients in GPG did not take chemotherapy, while 87.5% of the patients in PPG took chemotherapy. Only 56.7% of patients in IPG took chemotherapy. Thus our classification procedure can be utilized to assign new patients to one of three prognostic groups with an associated survival curve and a possible indication of the utilization of chemotherapy or not.

The paper is organized as follows. Section 2 gives a basic description of the essentials of support vector machines. Section 3 applies the ideas described in Section 2 to the problem of classifying patients using lymph node status as a classification criterion. Section 4 gives some brief conclusions.

A word about our notation now. All vectors will be column vectors unless transposed to a row vector by a prime superscript $'$. The scalar (inner) product of two vectors $x$ and $y$ in the $n$-dimensional real space $R^n$ will be denoted by $x'y$. For a general norm $\|\cdot\|$ on $R^n$, the dual norm $\|\cdot\|'$ on $R^n$ is defined as

$$\|x\|' := \max_{\|y\|=1} x'y. \tag{1.1}$$

For $p, q \in [1, \infty]$, $\frac{1}{p} + \frac{1}{q} = 1$, the $p$-norm $\|\cdot\|_p$ and $q$-norm $\|\cdot\|_q$ are dual norms. Thus the 1-norm $\|\cdot\|_1$ and $\infty$-norm $\|\cdot\|_\infty$ are dual norms while the 2-norm is self-dual. For an $\ell \times n$ matrix $A$, $A_i$ will denote the $i$th row of $A$. The identity matrix in a real space of arbitrary dimension will be denoted by $I$, while a column vector of ones of arbitrary dimension will be denoted by $e$. We shall employ the MATLAB "dot" notation [**MAT92**] to signify application of a function to all components of a matrix or a vector. For example if $A \in R^{\ell \times n}$, then $A^2_\bullet \in R^{\ell \times n}$ will denote the matrix obtained by squaring each element of $A$.

## 2. Support Vector Machines

We begin with the simple linear support vector machine formulation as follows:

$$\begin{array}{rlcl} \min\limits_{w,\gamma,y} & \nu e'y + \|w\| & & \\ \text{s.t.} & D(Aw - e\gamma) + y & \geq & e \\ & y & \geq & 0. \end{array} \tag{2.1}$$

Here, $\nu$ is a positive weight, $\|\cdot\|$ is an arbitrary norm and the $m \times n$ matrix $A$ represents $m$ given points in $R^n$ which belong to class 1 or -1 depending on whether the corresponding elements of the given $m \times m$ diagonal matrix $D$ are 1 or -1 respectively. The plane in $R^n$, with normal $w$ and distance from the origin $\frac{|\gamma|}{\|w\|}$

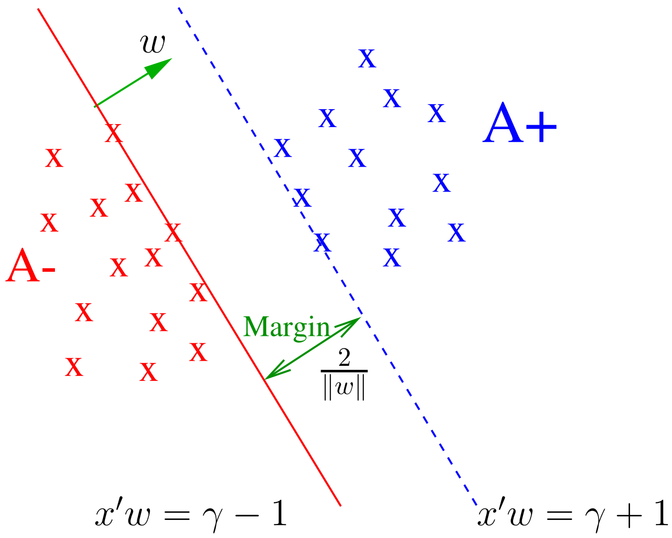


FIGURE 1. A separating plane $x'w = \gamma$, midway and parallel the bounding parallel planes $x'w = \gamma \pm 1$ with maximal separation margin $\frac{2}{\|w\|}$, is obtained by solving the mathematical program (2.1).

from the origin, generated by a solution $(w, \gamma)$ of the above mathematical program:

$$x'w = \gamma, \tag{2.2}$$

strictly separates the set $A+ := \{A_i \mid D_{ii} = 1\}$ from $A- := \{A_i \mid D_{ii} = -1\}$, if and only if the slack variable $y \in R^m$ is zero as shown in Figure 1. The margin or distance (measured by a norm $\|\cdot\|'$ that is dual to the norm $\|\cdot\|$ of (2.1)) between the bounding planes $x'w = \gamma \pm 1$ is given by $\frac{2}{\|w\|}$ [**Man99**]. These two planes are parallel to the plane $x'w = \gamma$ and bound the sets $A+$ and $A-$. For $\nu \geq \bar{\nu}$ for some $\bar{\nu} > 0$ [**MM79**, Theorem 4], the margin is maximized by a solution of (2.1) that suppresses the slack variable $y$ first. When the sets $A+$ and $A-$ are not strictly linearly separable, some components $y_i$ of the slack variable $y$ will be positive. The points $A_i$ in either $A+$ or $A-$ corresponding to positive $y_i$, as well as those corresponding to zero $y_i$ but with positive Lagrange multipliers, are termed *support vectors*. Points $A_i$ with positive $y_i$ lie on the wrong sides of the bounding planes $x'w = \gamma \pm 1$, while points with positive multipliers and zero slacks $y_i$ lie on the bounding planes.

We shall employ the 1-norm in the SVM formulation which leads to the following linear programming formulation which has been shown [**BM98a**] to have one of the best feature selection properties among all norms including the conventional 2-norm-squared formulation [**Vap95, CM98**]:

$$\begin{array}{rrcll} \min\limits_{w,\gamma,s,y} & & \nu e'y + e's & & \\ \text{s.t.} & & D(Aw - e\gamma) + y & \geq & e \\ & -s \leq & w & \leq & s \\ & & y & \geq & 0. \end{array} \tag{2.3}$$

This linear programming formulation will be our *SVM1* formulation that will be used for feature selection in the next section of the paper.

In order to obtain a nonlinear support vector machine with a nonlinear separating surface, we make use of the following transformation based on the dual formulation for the 2-norm-squared support vector machine [**Man00**]:

$$w = A'Du \tag{2.4}$$

Substitution in (2.1) gives:

$$\begin{array}{lrcl} \min\limits_{u,\gamma,y} & \nu e'y + \|A'Du\| & & \\ \text{s.t.} & D(AA'Du - e\gamma) + y & \geq & e \\ & y & \geq & 0. \end{array} \tag{2.5}$$

Replacing $AA'$ by a nonlinear kernel $K(A, A') : R^{m\times n} \times R^{n\times m} \longrightarrow R^{m\times m}$, and $\|A'Du\|$ by a convex function of $u$, typically a norm, results in the following nonlinear generalized support vector machine (GSVM) [**Man00**]:

$$\begin{array}{lrcl} \min\limits_{u,\gamma,y} & \nu e'y + \|u\|_p & & \\ \text{s.t.} & D(K(A, A')Du - e\gamma) + y & \geq & e \\ & y & \geq & 0, \end{array} \tag{2.6}$$

with the following nonlinear separating surface instead of a plane:

$$K(x', A')Du = \gamma. \tag{2.7}$$

Typical kernel functions are the following:

- **Polynomial Kernel** $(AA' + \mu aa')^d_\bullet$,
  where $[\cdot]^d_\bullet$ denotes component-wise exponentiation as in MATLAB [**MAT92**].
- **Radial Basis (Gaussian) Kernel** $\varepsilon^{-\mu\|A_i - A_j\|^2}$, $i, j = 1, \ldots, m$.
- **Neural Network Kernel** $(AA' + \mu aa')_{\bullet *}$,
  where $[\cdot]_{\bullet *}$ denotes the step function: $R \longrightarrow \{0, 1\}$ component-wise.

Using the 1-norm in (2.6) leads to a linear programming formulation of the nonlinear GSVM that we will be using in this work:

$$\begin{array}{llcl} \min\limits_{u,\gamma,s,y} & & \nu e'y + e's & \\ \text{s.t.} & & D(K(A, A')Du - e\gamma) + y & \geq & e \\ & -s \leq & u & \leq & s \\ & & y & \geq & 0. \end{array} \tag{2.8}$$

This nonlinear GSVM, which is capable of discriminating between elements of fairly complex sets, such as the black and white squares of a checkerboard [**MM99, LM99**], will be used in the next section. We note that even though the nonlinear GSVM (2.8) is capable of achieving fairly complex separation, it is not suited for feature selection, because the kernel $K(A, A')$ creates interactions between all the features through the variable $u$ which is not present in SVM1 (2.3). That is the reason we used the linear SVM1 for feature selection and the nonlinear GSVM for the final separation.

## 3. Breast Cancer Prognosis and Chemotherapy

In this section of the paper we shall use the support vector machines described in the previous section to discriminate between 253 breast cancer patients. The dataset for these patients is publicly available [**WLM99**]. Our criterion for discrimination will be the absence or presence of metastasized lymph nodes (1 to 30 nodes) removed from the patient's armpit during surgery. Lymph node metastasis is a strong indicator for chemotherapy. The total number of features used to constitute the $n$-dimensional space in which the separation is accomplished is made up of the mean, standard deviation, and maximum (worst) value of ten cytological nuclear measurements of size, shape and texture taken from the patients breast by a non-surgical fine needle aspirate procedure [**Str94, MSW95**] together with the tumor size excised from the patient's breast during surgery. Taken together these constitute the 31-dimensional feature space over which classification according to lymph node metastasis will be achieved.

Before proceeding with our classification process we exhibit Kaplan-Meier survival curves [**Kle96**] for 140 of the 253 patients who had chemotherapy and for the remaining 113 patients that did not have chemotherapy in Figure 2. The solid curve gives the percent of surviving patients as a function of time for the group that received chemotherapy, while the dashed curve gives the corresponding survival curve for the group that did not receive chemotherapy. One possible interpretation of this curve is that the patients not receiving chemotherapy and who have a better survival curve than those who are receiving chemotherapy, are not being compromised by not receiving chemotherapy. Another interpretation is that most patients who had chemotherapy were preselected for having lymph node metastasis and hence had inherently worse cancers. The p-value based on the logrank statistic [**Kle96**] relating these two curves is 0.0018. Hence these two curves are statistically distinct. (The logrank statistic is a comparative measure of the survival times of two groups which calculates a weighted sum of the difference, at each time interval, of the mortality ratio between the two groups: $\frac{m_1}{m_2}$, and the population ratio between the two groups: $\frac{n_1}{n_2}$. Here, $m_i$ is the number of deaths in group $i$ and $n_i$ is the number of people in group $i$. The logrank statistic for two groups has a Chi-squared distribution with degree of freedom one [**Kle96**]. Larger logrank statistic indicates significantly different survival curves. That corresponds to a small p-value. The p-value is the probability that the two survival curves are statistically the same.)

Our next result is the classification of the 253 patients into three prognostic groups as follows:

1. **Good Prognosis** "Node-negative patients": Patients with no metastasized lymph nodes.
2. **Intermediate Prognosis** "Node-positive patients" with 1 to 4 metastasized lymph nodes.
3. **Poor Prognosis** "Node-positive patients" with more than 4 metastasized lymph nodes.

To achieve this classification we first used SVM1 (2.3) to separate node-negative patients from node-positive patients. By using a sufficiently small value of $\nu$ in (2.3), $\nu = 0.02$, the SVM1 formulation (2.3) suppressed 25 of the 31 features and selected a total of 6 features as follows:

1. **Mean Area**
2. **Standard Deviation of Area**
3. **Worst Area**
4. **Worst Texture**
5. **Worst Perimeter**
6. **Tumor Size** (from surgery)

Note that the first five are cytological features obtained from the fine needle aspirate taken during diagnosis.

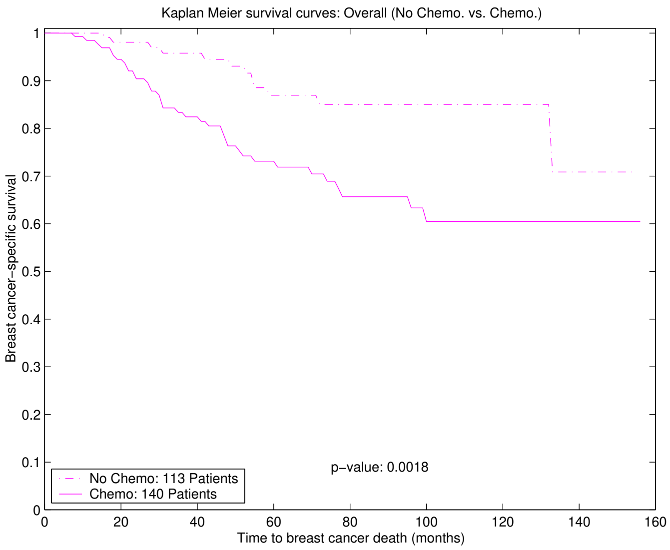


FIGURE 2. Kaplan-Meier survival curve (solid line) for 140 of the 253 patients who received chemotherapy and the corresponding survival curve (dashed line) for the remaining 113 patients that did not receive chemotherapy. p-value=0.0018.

We next used the nonlinear support vector machine GSVM (2.8) with a Gaussian kernel with $\nu = 0.096$ to achieve our three-class separation as follows:

1. **GPG**: Good prognosis group (0 lymph node metastasis) was separated from all the rest for which:
2. **IPG**: Intermediate prognosis group (1-4 lymph node metastasis) was separated from:
   **PPG**: Poor prognosis group (lymph node metastasis $> 4$).

Note that the separation in Step 2 above included some node-negative patients in IPG because the separation of Step 1 is not perfect and GPG does not include *all* node-negative patients. Tenfold cross validation correctness [**Sto74**] for the above separation was 72.8%.

Based on the above separation we constructed three Kaplan-Meier survival curves for each of the three prognostic groups: GPG, IPG and PPG as shown in Figure 3. Note that the three curves are very well separated and the p-values between the three groups are as follows:

1. GPG-IPG p-value=0.0090
2. GPG-PPG p-value=9.5769e-06
3. IPG-PPG p-value=0.0037

These p-values are substantially smaller than the conventional 0.05 cutoff used to reject the null-hypothesis (survival curves are the same). Hence these curves are indeed statistically significantly different from each other.

It is worth noting within each of the three prognosis groups how many patients received and did not receive chemotherapy:

1. **GPG**: 19 patients received chemotherapy; 51 did not (27.2%;72.8%)
2. **IPG**: 72 patients received chemotherapy; 55 did not (56.7%;43.3%)

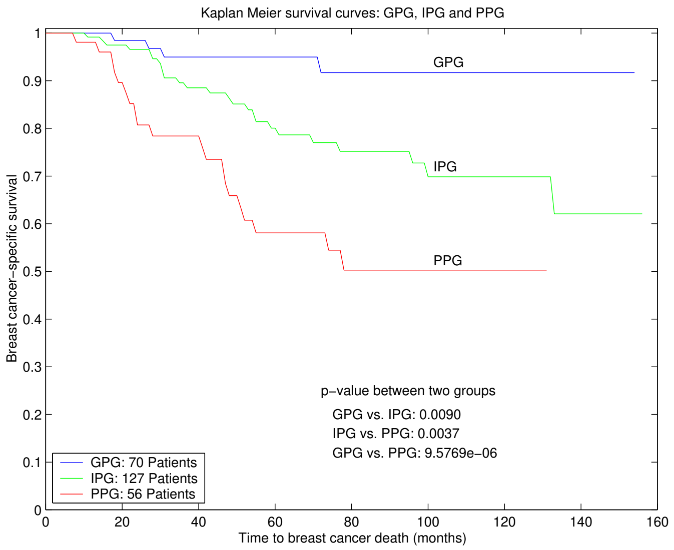


FIGURE 3. Kaplan-Meier survival curves for GPG, IPG and PPG: the good, intermediate and poor prognosis groups. The classification criterion is based on the number of metastasized lymph nodes. The features used for the classification are: mean area, standard deviation of area, worst texture, worst perimeter and tumor size. The pairwise p-values are: GPG-IPG= 0.0090, GPG-PPG= 9.5769e-06, IPG-PPG= 0.0037.

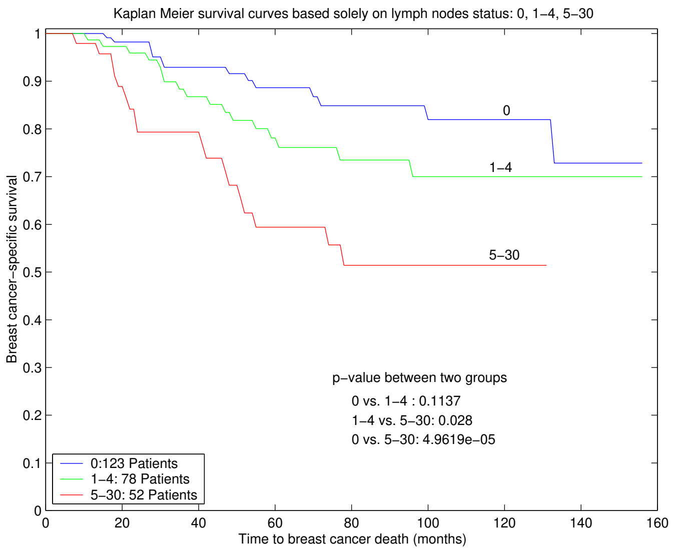


FIGURE 4. Kaplan-Meier survival curves for 3 groups based solely on lymph node status: 0, [1-4] and [5-30]. The pairwise p-values are: [0]-[1-4]:0.1137, [1-4]-[5-30]:0.028, [0]-[5-30]:4.9619e-05.

3. **PPG**: 49 patients received chemotherapy; 7 did not (87.5%;12.5%)

Within each prognosis group, there was no statistically significant difference based on p-values, between the survival curves of patients who took and did not take chemotherapy.

In contrast to the three survival curves based on the GPG, IPG and PPG groups, we constructed three other survival curves based solely on the number of metastasized lymph nodes, the conventional indicator for chemotherapy prescription, as shown in Figure 4. It is interesting to observe the large p-values between the node-negative survival curve and the 1-4 lymph node survival curve (p-value=0.1137), and between the latter curve and the 5-30 lymph node survival curve (p-value=0.028). This indicates that our criteria for GPG-IPG-PPG classification based on the six features selected by SVM1, which do not include lymph node status, yield statistically more distinct survival curves than those based on lymph node status by itself.

## 4. Conclusions

Based on the results presented in this work, the following conclusions can be made:

- Support vector machines can be used to:
  - Select features that are important for prognosis.
  - Classify breast cancer patients into three groups with well separated survival curves.
- Most (72.8%) patients in good group received no chemotherapy.
- Most (87.5%) patients in poor group received chemotherapy.
- Just over half the patients (56.7%) in intermediate group received chemotherapy.
- Survival curves and possible chemotherapy decision for new patients can be determined by assigning them to one of the three groups based on five nuclear features and tumor size without the need for lymph node status.

## References

[BFM99] P. S. Bradley, Usama M. Fayyad, and O. L. Mangasarian, *Data mining: Overview and optimization opportunities*, INFORMS Journal on Computing **11** (1999), 217–238, ftp://ftp.cs.wisc.edu/math-prog/tech-reports/98-01.ps.

[BM92] K. P. Bennett and O. L. Mangasarian, *Robust linear programming discrimination of two linearly inseparable sets*, Optimization Methods and Software **1** (1992), 23–34.

[BM98a] P. S. Bradley and O. L. Mangasarian, *Feature selection via concave minimization and support vector machines*, Machine Learning Proceedings of the Fifteenth International Conference(ICML '98) (San Francisco, California) (J. Shavlik, ed.), Morgan Kaufmann, 1998, ftp://ftp.cs.wisc.edu/math-prog/tech-reports/98-03.ps, pp. 82–90.

[BM98b] P. S. Bradley and O. L. Mangasarian, *Massive data discrimination via linear support vector machines*, Tech. Report 98-05, Computer Sciences Department, University of Wisconsin, Madison, Wisconsin, May 1998, *Optimization Methods and Software*, to appear. ftp://ftp.cs.wisc.edu/math-prog/tech-reports/98-03.ps.

[BMS98] P. S. Bradley, O. L. Mangasarian, and W. N. Street, *Feature selection via mathematical programming*, INFORMS Journal on Computing **10** (1998), no. 2, 209–217, ftp://ftp.cs.wisc.edu/math-prog/tech-reports/95-21.ps.

[CM98] V. Cherkassky and F. Mulier, *Learning from data - concepts, theory and methods*, John Wiley & Sons, New York, 1998.

[Kle96] David G. Kleinbaum, *Survival analysis*, Springer-Verlag, New York, 1996.

[KM58] E. L. Kaplan and P. Meier, *Nonparametric estimation from incomplete observations*, Journal of the American Statistical Association **53** (1958), 457–481.

[LM99] Yuh-Jye Lee and O. L. Mangasarian, *SSVM: A smooth support vector machine*, Tech. Report 99-03, Data Mining Institute, Computer Sciences Department, University of Wisconsin, Madison, Wisconsin, September 1999, ftp://ftp.cs.wisc.edu/pub/dmi/tech-reports/99-03.ps.

[Man65] O. L. Mangasarian, *Linear and nonlinear separation of patterns by linear programming*, Operations Research **13** (1965), 444–452.

[Man68] O. L. Mangasarian, *Multi-surface method of pattern separation*, IEEE Transactions on Information Theory **IT-14** (1968), 801–807.

[Man97] O. L. Mangasarian, *Mathematical programming in data mining*, Data Mining and Knowledge Discovery **1** (1997), no. 2, 183–201, ftp://ftp.cs.wisc.edu/math-prog/tech-reports/96-05.ps.

[Man99] O. L. Mangasarian, *Arbitrary-norm separating plane*, Operations Research Letters **24** (1999), no. 1-2, ftp://ftp.cs.wisc.edu/math-prog/tech-reports/97-07r.ps.

[Man00] O. L. Mangasarian, *Generalized support vector machines*, Advances in Large Margin Classifiers (Cambridge, MA) (A. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans, eds.), MIT Press, 2000, ftp://ftp.cs.wisc.edu/math-prog/tech-reports/98-14.ps, pp. 135–146.

[MAT92] MATLAB, *User's guide*, The MathWorks, Inc., Natick, MA 01760, 1992.

[MM79] O. L. Mangasarian and R. R. Meyer, *Nonlinear perturbation of linear programs*, SIAM Journal on Control and Optimization **17** (1979), no. 6, 745–752.

[MM99] O. L. Mangasarian and David R. Musicant, *Data discrimination via nonlinear generalized support vector machines*, Tech. Report 99-03, Computer Sciences Department, University of Wisconsin, Madison, Wisconsin, March 1999, ftp://ftp.cs.wisc.edu/math-prog/tech-reports/99-03.ps.

[MSW95] O. L. Mangasarian, W. N. Street, and W. H. Wolberg, *Breast cancer diagnosis and prognosis via linear programming*, Operations Research **43** (1995), no. 4, 570–577.

[Sto74] M. Stone, *Cross-validatory choice and assessment of statistical predictions*, Journal of the Royal Statistical Society **36** (1974), 111–147.

[Str94] W. N. Street, *Cancer diagnosis and prognosis via linear-programming-based machine learning*, Computer Sciences Department, Mathematical Programming Technical Report and Ph.D. Dissertation 94-14, University of Wisconsin, Madison, Wisconsin, August 1994, http://www.cs.wisc.edu/tech-reports/.

[Vap95] V. N. Vapnik, *The nature of statistical learning theory*, Springer, New York, 1995.

[WLM99] W. H. Wolberg, Y.-J. Lee, and O. L. Mangasarian, *WPBCC: Wisconsin Prognostic Breast Cancer Chemotherapy Database*, Computer Sciences Department, University of Wisconsin, Madison, ftp://ftp.cs.wisc.edu/math-prog/cpo-dataset/machine-learn/WPBCC/, 1999.

[WSHM95a] W. H. Wolberg, W. N. Street, D. H. Heisey, and O. L. Mangasarian, *Computer-derived nuclear features distinguish malignant from benign breast cytology*, Human Pathology **26** (1995), 792–796.

[WSHM95b] W. H. Wolberg, W. N. Street, D. N. Heisey, and O. L. Mangasarian, *Computerized breast cancer diagnosis and prognosis from fine-needle aspirates*, Archives of Surgery **130** (1995), 511–516.

[WSM93] W. H. Wolberg, W. N. Street, and O. L. Mangasarian, *Breast cytology diagnosis via digital image analysis*, Analytical and Quantitative Cytology and Histology **15** (1993), no. 6, 396–404.

[WSM97] W. H. Wolberg, W. N. Street, and O. L. Mangasarian, *Computer-derived nuclear features compared with axillary lymph node status for breast carcinoma prognosis*, Cancer Cytopathology **81** (1997), 172–179.

Computer Sciences Department, University of Wisconsin, 1210 West Dayton St. Madison, WI 53706
*E-mail address*: yuh-jye@cs.wisc.edu

Computer Sciences Department, University of Wisconsin, 1210 West Dayton St. Madison, WI 53706
*E-mail address*: olvi@cs.wisc.edu

General Surgery, University of Wisconsin, H4/750 Clinical Sciences Center Madison, WI 53792
*E-mail address*: wolberg@surgery.wisc.edu