A Theoretical Characterization of Linear SVM-Based Feature Selection

Douglas Hardin DOUG.HARDIN@VANDERBILT.EDU
Department of Mathematics, Vanderbilt University, Stevenson Center, Nashville, TN 37240-0001

Ioannis Tsamardinos IOANNIS.TSAMARDINOS@VANDERBILT.EDU
Dept. of Biomedical Informatics, Vanderbilt University, 2209 Garland Ave, Nashville, TN 37232-8340, USA

Constantin F. Aliferis CONSTANTIN.ALIFERIS@VANDERBILT.EDU
Dept. of Biomedical Informatics, Vanderbilt University, 2209 Garland Ave, Nashville, TN 37232-8340, USA

Abstract

Most prevalent techniques in Support Vector Machine (SVM) feature selection are based on the intuition that the weights of features that are close to zero are not required for optimal classification. In this paper we show that indeed, in the sample limit, the irrelevant variables (in a theoretical and optimal sense) will be given zero weight by a linear SVM, both in the soft and the hard margin case. However, SVM-based methods have certain theoretical disadvantages too. We present examples where the linear SVM may assign zero weights to strongly relevant variables (i.e., variables required for optimal estimation of the distribution of the target variable) and where weakly relevant features (i.e., features that are superfluous for optimal feature selection given other features) may get non-zero weights. We contrast and theoretically compare with Markov-Blanket based feature selection algorithms that do not have such disadvantages in a broad class of distributions and could also be used for causal discovery.

1. Introduction

Feature selection (also called variable selection) is the problem of selecting a subset of variables of minimal size with maximum predictive, classification, or diagnostic power relative to a target variable of interest Y. Being able to identify this minimal size set is important for treating the curse of dimensionality, for reducing the cost of observing the required variables for prediction, and for gaining insight into the domain. The problem is far from solved, and more pressing than ever given the recent emergence of large datasets.

A recent breakthrough in feature selection research is the development of Support Vector Machine (SVM) based techniques, that are scalable to thousands of variables and typically exhibit excellent performance in reducing the number of variables while maintaining or improving classification accuracy (Guyon et al., 2002).

A binary SVM classifier is a function of the form $g(\mathbf{x}) = \text{sgn}\,(\mathbf{w}\cdot\phi(\mathbf{x}) + b)$, where $\mathbf{x}$ is the input vector, ϕ a function that maps $\mathbf{x}$ from the original feature space to a new feature space, $\mathbf{w}$ a weight vector in the projected feature space, and b a real constant. In the linear SVM case the function $\phi(\mathbf{x})$ is the identity function. Training consists of identifying the weight vector $\mathbf{w}$ that maximizes the margin between the convex sets of each class, or a trade-off between the margin and the misclassifications of training instances. In this paper we only consider linear SVMs; the term "linear" is dropped for simplicity in the rest of the paper.

Obviously, if a variable has a corresponding zero weight it does not contribute to classification and thus, is irrelevant to the output of the SVM classifier and can be dropped from the model. For example, the Recursive Feature Elimination algorithm (Guyon et al., 2002), a prototypical, widely used, and successful feature selection algorithm, recursively identifies small weights in magnitude and removes the corresponding variables.

Appearing in *Proceedings of the 21st International Conference on Machine Learning*, Banff, Canada, 2004.

Although any variable with zero weight in a specific SVM classifier is indeed irrelevant for that classifier, what is currently unknown is *(a) whether the variable is truly irrelevant, i.e., with respect to any optimal classifier.* Conversely, it is also unknown *(b) whether all truly irrelevant variables will be assigned a zero weight during SVM feature selection.* In this paper, we are exploring a characterization of conditions under which such SVM feature selection algorithms will output all and only relevant variables.

In a recent review of the field of feature selection, Guyon *et al.* (Guyon & Ellisseeff, 2003) reads "The approaches [in feature selection] are very diverse and motivated by various theoretical arguments but a unifying theoretical framework is lacking". With this paper, we hope to stimulate research in such a unifying framework, by exploring the connections between notions of relevancy, Markov-Blanket and SVM based feature selection.

2. Preliminaries: Classification Using SVMs

In this section we review the hard and soft margin linear SVM classifiers for finite training data. We define the sample limit formulations of these SVM classifiers and show that the large sample formulation defines a unique weight vector $\boldsymbol{\omega}$ that is indeed the limit of the respective finite sample SVM weight vectors $\mathbf{w}_m$, where m is the number of training sample instances.

Given training data $\mathbf{X}_m := \{\mathbf{x}_1, \ldots, \mathbf{x}_m\} \subset \mathbf{R}^N$ together with labels $Y_m := \{y_1, \ldots, y_m\} \subseteq \{-1, 1\}$ a linear SVM produces a decision function $g : \mathbf{R}^N \to \{-1, 1\}$ of the form

$$g(\mathbf{x}) := \mathrm{sgn}\,(\mathbf{w}\cdot\mathbf{x} + b) \tag{1}$$

where the *weight vector* $\mathbf{w} = (w^1, \ldots, w^N) \in \mathbf{R}^N$ and the *offset* $b \in \mathbf{R}$ are chosen according to one of the following constrained optimization problems.

F.1. Finite Sample Hard-margin SVM

Minimize:

$$F_h(\mathbf{w}, b) := \mathbf{w}\cdot\mathbf{w} = \sum_{i=1}^{N} (w^i)^2 \tag{2}$$

Constraints:

$$y_k(\mathbf{w}\cdot\mathbf{x}_k + b) \geq 1 \text{ for } 1 \leq k \leq m. \tag{3}$$

The training data $(\mathbf{X}_m, Y_m)$ is *linearly separable* if there is at least one admissible $(\mathbf{w}, b)$ satisfying (3). Note that (2) and (3) form a strictly convex optimization problem and so there is a unique solution $(\mathbf{w}_{h,m}, b_{h,m})$ (e.g., see (Vapnik, 1998)). The distance (or "gap") between the planes $\mathbf{w}_{h,m}\cdot\mathbf{x} + b_{h,m} = 1$ and $\mathbf{w}_{h,m}\cdot\mathbf{x} + b_{h,m} = -1$ is given by $\Delta_m := 2/|\mathbf{w}_{h,m}|$. Hence, minimizing the objective function (2) is equivalent to maximizing the "gap" Δ_m. Of course, the hard-margin SVM is non-trivial only for linearly separable data.

F.2. Finite Sample Soft-margin (p-norm) SVM

Minimize:

$$F_s(\mathbf{w}, b, \{\xi_k\}_{k=1}^m) := \mathbf{w}\cdot\mathbf{w} + \frac{C}{m}\sum_{i=1}^{m} \xi_k^p \tag{4}$$

Constraints:

$$y_k(\mathbf{w}\cdot\mathbf{x}_k + b) \geq 1 - \xi_k \text{ and } \xi_k \geq 0,\ (1 \leq k \leq m) \tag{5}$$

If $p > 1$, then it is known that the soft-margin optimization problem (4) and (5) has a unique minimizer. When $p = 1$, there may be multiple minimizers $(\mathbf{w}, b, \{\xi_k\}_{k=1}^m)$. However all of these minimizers share a common weight vector $\mathbf{w}$ (see (Burges & Crisp, 2000)). We denote the soft-margin SVM weight vector for the training data set $(\mathbf{X}_m, Y_m)$ by $\mathbf{w}_{s,m}$.

For fixed $\mathbf{w}$ and b, the soft-margin SVM objective function $F_s(\mathbf{w}, b, \{\xi_k\}_{k=1}^m)$ in (4) is clearly minimized by $\xi_k = [1 - y_k(\mathbf{w}\cdot\mathbf{x}_k - b)]_+$, $k = 1, \ldots, m$, where $x_+ := \max(x, 0)$. Hence the soft-margin SVM may also be recast as an unconstrained optimization problem

F.3 Soft-margin SVM unconstrained form

Minimize:

$$F_s^*(\mathbf{w}, b) := \mathbf{w}\cdot\mathbf{w} + \frac{C}{m}\sum_{k=1}^{m} [1 - y_k(\mathbf{w}\cdot\mathbf{x}_k - b)]_+^p. \tag{6}$$

2.1. Sample limit SVMs

We now suppose that $(\mathbf{x}_k, y_k)_{k=1}^\infty$ is an infinite sequence of independent samples of a random variable $\mathbf{Z} = (\mathbf{X}, Y)$ that takes values in $B \times \{-1, +1\}$ for some bounded set $B \subset \mathbf{R}^N$ according to a (Borel) probability measure $p(\mathbf{x}, y)$. To avoid trivialities we assume $0 < P(Y = 1) = \int_{\mathbf{x}\in B}\ dp(\mathbf{x}, 1) < 1$.

We are interested in the limiting behavior of the weight vectors $\mathbf{w}_m$ associated with the training data $(\mathbf{x}_k, y_k)_{k=1}^m$. If A is an open set in $B \times \{\pm 1\}$ with

$P(A) > 0$, then almost surely (i.e. with probability 1) there will be some $(\mathbf{x}_k, y_k) \in A$. Hence, it is natural to consider the following "sample limit" version of the hard-margin SVM.

L.1. Sample limit hard-margin SVM

Minimize:

$$F_h(\mathbf{w}, b) := \mathbf{w}\cdot\mathbf{w} \tag{7}$$

Constraints:

$$Y(\mathbf{w}\cdot\mathbf{X} + b) \geq 1 \text{ almost surely.} \tag{8}$$

The random variable $(\mathbf{X}, Y)$ is *linearly separable* if there is some $(\mathbf{w}, b)$ such that (8) holds. If $(\mathbf{X}, Y)$ is linearly separable, then (as in the finite-sample case) there is a unique minimizer $(\boldsymbol{\omega}_h, b_h)$ of (7) satisfying (8). The "gap" Δ between the decision surfaces $\boldsymbol{\omega}\cdot\mathbf{x} + b = \pm 1$ is then given by $\Delta := 2/|\boldsymbol{\omega}|$.

Next we consider the sample limit for the soft-margin SVM. For fixed $(\mathbf{w}, b)$ the slack variable ξ_k is given by $\xi_k = [1 - y_k(\mathbf{w}\cdot\mathbf{x}_k + b)]_+$. Hence, $(\xi_k)_{k=1}^m$ is a sequence of independent samples of the random variable $\xi = [1 - Y(\mathbf{w}\cdot\mathbf{X} + b)]_+$. Thus we have $(1/m)\sum_{k=1}^m \xi_k^p \to E(\xi)$ as $m \to \infty$ almost surely where $E(\cdot)$ denotes the expectation with respect to $p(\mathbf{x}, y)$. This suggests the following "sample limit" soft-margin SVM constrained optimization problem:

L.2. Sample limit soft-margin SVM

Minimize:

$$\mathcal{F}_s(\mathbf{w}, b, \xi) := \mathbf{w}\cdot\mathbf{w} + CE(\xi^p) \tag{9}$$

Constraints:

$$Y(\mathbf{w}\cdot\mathbf{X} + b) \geq 1 - \xi \text{ and } \xi \geq 0 \text{ a. s.,} \tag{10}$$

where $\xi = \xi(\mathbf{X}, Y)$ is a random variable. In fact we clearly must have $\xi = [1 - Y(\mathbf{w}\cdot\mathbf{X} + b)]_+$ almost surely for any minimizer of (L.2) and so we can recast (L.2) in the equivalent unconstrained form:

L.3. Sample limit soft-margin SVM (unconstr.)

Minimize:

$$\mathcal{F}_s^*(\mathbf{w}, b) = \mathbf{w}\cdot\mathbf{w} + CE([1 - Y(\mathbf{w}\cdot\mathbf{X} + b)]_+^p). \tag{11}$$

Sample limit SVMs have been considered in (Steinwart, 2003). There it is shown that the finite sample SVM decision functions converge in probability to a limiting set valued decision function. The following lemma establishes that in both the hard and soft-margin cases the finite sample weight vectors converge almost surely to the respective hard or soft-margin sample limit weight vector as the number of samples $m \to \infty$.

Lemma 1. *(a) Suppose $(\mathbf{X}, Y)$ is linearly separable. Then the sample limit hard-margin SVM problem (L.1) has a unique minimizer $(\boldsymbol{\omega}_h, \beta_h)$. Let $(\mathbf{w}_{h,m}, b_{h,m})$ denote the unique minimizer for the finite sample hard-margin SVM problem (F.1) with m samples. Then*

$$\lim_{m\to\infty} \mathbf{w}_{h,m} = \boldsymbol{\omega}_h \qquad a.s.$$

(b) The sample limit soft-margin (unconstrained) SVM problem (L.3) has a (global) minimizer $(\boldsymbol{\omega}_s, \beta_s)$. The weight vector $\boldsymbol{\omega}_s$ is unique, i.e. if $\mathcal{F}_s^*(\boldsymbol{\omega}, \beta) = \mathcal{F}_s^*(\boldsymbol{\omega}_s, \beta_s)$ then $\boldsymbol{\omega} = \boldsymbol{\omega}_s$. Furthermore,

$$\lim_{m\to\infty} \mathbf{w}_{s,m} = \boldsymbol{\omega}_s \qquad a.s.$$

For $p > 1$, we also have $\beta = \beta_s$.

3. Treatment of Irrelevant Variables by Linear SVMs

In this section, we show that irrelevant variables will get zero weights in the sample limit, both in the hard and the soft margin case. In addition, we provide bounds on the size of the weights of the irrelevant variables in the finite sample case.

If $J \subseteq \{1, \ldots, N\}$ and $\mathbf{x} \in \mathbf{R}^N$, we let $\mathbf{x}^J$ denote the vector whose ith component is x^i if $i \in J$ and is 0 otherwise.

We define as *irrelevant* a set of variables $I \subseteq \{1, \ldots, N\}$ if $\mathbf{X}^I$ is independent of $(\mathbf{X}^R, Y)$, i.e., $p(\mathbf{X}^I, \mathbf{X}^R, Y) = p(\mathbf{X}^I)\cdot p(\mathbf{X}^R, Y)$, where R consists of the rest of the variables, for any value of $\mathbf{X}^I, \mathbf{X}^R$, and Y. Intuitively, this definition of irrelevancy means that the set of irrelevant variables provides no information for the target and additionally, no information for the relevant variables. It is introduced here because it facilitates mathematical analysis while being highly intuitive. In addition, this definition ties with the Kohavi and John definition or irrelevancy in many distributions (see Section 4).

Theorem 1. *Suppose $I \subseteq \{1, \ldots, N\}$ is an irrelevant set of variables for $(\mathbf{X}, Y)$. Then:*

(a) If $(\mathbf{X}, Y)$ is linearly separable and $\boldsymbol{\omega}_h$ denotes the large-sample limit hard -margin SVM weight vector then $\boldsymbol{\omega}_h^i = 0$ for $i \in I$.

(b) If $\boldsymbol{\omega}_s$ denotes the large-sample limit soft-margin SVM weight vector, then $\boldsymbol{\omega}_s^i = 0$ for $i \in I$.

While the above theorem guarantees that all irrelevant variables will get zero weights in the sample limit

for practical purposes, it would be useful to know how large the weights of the irrelevant variables may be with finite sample. The following lemma provides bounds for $|\boldsymbol{\omega} - \mathbf{w}_m|$ in terms of $\Delta - \Delta_m$.

Lemma 2. *Suppose $(\mathbf{X}, Y)$ is linearly separable. Let $\boldsymbol{\omega}_h$ (respectively, $\mathbf{w}_{h,m}$) denote the hard-margin SVM weight vector for $(\mathbf{X}, Y)$ (respectively, for the training set $(\mathbf{x}_k, y_k)_{k=1}^m$) Then*

$$|\boldsymbol{\omega}_h - \mathbf{w}_{h,m}| \leq 2\frac{\sqrt{\Delta_m + \Delta}}{\Delta_m \Delta}\sqrt{\Delta_m - \Delta}. \qquad (12)$$

Combining Lemma 2 and part (a) of Theorem 1 then gives:

Corollary 1. *Suppose $(\mathbf{X}, Y)$ is linearly separable and that $I \subseteq \{1, \ldots, N\}$ is an irrelevant set of variables for $(\mathbf{X}, Y)$. Then*

$$\sum_{i \in I} (w_{h,m}^i)^2 \leq 4\left(\frac{\Delta_m + \Delta}{\Delta_m^2 \Delta^2}\right)(\Delta_m - \Delta),$$

where $w_{h,m}^i$ is the ith component of $\mathbf{w}_{h,m}$.

While the quantity Δ is unknown, the above result is a first step towards providing practical bounds. Such bounds are important for algorithms as Recursive Feature Elimination in determining the threshold of weights below which any variable should be labeled as irrelevant and be filtered out. For example, one could fit a model of Δ_m as a function of sample size and use it to estimate Δ; subsequently, order the variables by ascending weights and select to remove the first k for which the above bound is satisfied.

4. Treatment of Relevant Variables by SVMs

In this section we define strongly and weakly relevant features and show that a linear SVM may assign zero weight to strongly relevant variables, and non-zero weights to weakly relevant variables. This implies that under certain conditions (see section 5) the SVM output is neither sound nor minimal.

A number of researchers have attempted to provide "reasonable" definitions of relevancy. Kohavi and John (Kohavi & John, 1997) review several such definitions and conclude with the following ones.

A feature X^i is *strongly relevant to* Y if

$$p(Y = y|X^i = x^i, S^i = s^i) \neq p(Y = y|S^i = s^i).$$

for *some* values y, x^i, s^i of Y, X^i, S^i, for which $p(X^i = x^i, S^i = s^i) > 0$, where S^i is the set of remaining variables $S^i = V \setminus \{Y, X^i\}$.

A feature X^i is *weakly relevant to* Y if it is not strongly relevant, and there exists a subset of features U^i of S^i, and a set of values u^i, x^i, y for U^i, X^i, and Y for which $p(X^i = x^i, U^i = u^i) > 0$ and

$$p(Y = y|X^i = x^i, U^i = u^i) \neq p(Y = y|U^i = u^i).$$

A feature is *relevant to* Y if it is weakly or strongly relevant to Y, otherwise a feature is *KJ-irrelevant to* Y (KJ stands for Kohavi and John). In other words, a feature X^i is *KJ-irrelevant* if for any $S \subseteq V \setminus \{Y, X^i\}$, and any values s, y, x^i of S, Y, X^i for which $p(S = s, X^i = x^i) > 0$, X^i is independent of Y conditioned on S.

Intuitively, the KJ-irrelevant variables provide no information for the distribution of Y. The irrelevant variables, as defined in this paper, provide no information neither for Y nor for the relevant (weakly or strongly) variables. Thus,

Lemma 3. *If X^I is a set of irrelevant features, then $\forall X^i \in X^I, X^i$ is KJ-irrelevant.*

The converse also holds in faithful distributions. The latter is a broad class of distributions for which there is a Bayesian Network that entails all probabilistic dependencies and independencies observed in the distribution (Meek, 1995).

Lemma 4. *The set $\{X^i | X^i$ is KJ-irrelevant to $Y\}$ is the set of irrelevant variables X^I in any distribution faithful to a Bayesian Network.*

An SVMs may assign zero weights to both weakly and strongly relevant variables. For example consider the parity function where Y is the exclusive OR of X^1 and X^2. That is, $p(Y = 1|X^1 = 1, X^2 = 0) = p(Y = 1|X^1 = 0, X^2 = 1) = 1$ and $p(Y = -1|X^1 = 1, X^2 = 1) = p(Y = -1|X^1 = 0, X^2 = 0) = 1$ while $P(X^1 = 1) = P(X^1 = 0) = P(X^2 = 1) = P(X^2 = 0) = 1/4$. The distribution is non-linearly separable and the soft-margin linear SVM returns a zero weight vector.

An example of assigning non-zero weights to weakly relevant variables is given by the following distribution: $p(Y = 1|X^1 = 0, X^2 = 0) = p(Y = 1|X^1 = 1, X^2 = 1) = p(Y = -1|X^1 = 1 + \epsilon, X^2 = 0) = 1$, $p(X^1 = 0, X^2 = 0) = p(X^1 = 1, X^2 = 1) = p(X^1 = 1 + \epsilon, X^2 = 0) = 1/3$. The distribution is shown pictorially in Figure 1.

Variable X^2 is not a strongly relevant one, because $p(Y|X^1, X^2) = p(Y|X^1)$, for all values of the variables. But, X^2 is weakly relevant, because it is not strongly relevant and for $S = \emptyset, Y = 1, X^2 = 1$, $p(Y = 1|X^2 = 1, S) \neq p(Y = 1|S)$.

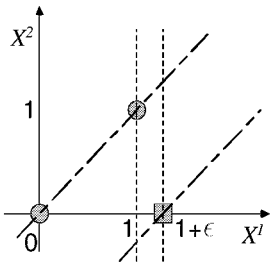

Figure 1. An example where the weakly relevant variable X^2 receives a non-zero weight by the maximum margin classifier (dashed diagonal lines). The gap corresponding to the classifier that assigns a zero weight to X^2 (dashed vertical lines) can have an arbitrarily smaller gap.

5. Relevancy, Markov-Blankets, and Optimal Feature Selection

Researchers are interested in the concept of relevancy in the context of feature selection because they follow the intuition that a relevant variable should be included in the selected variables and all irrelevant variables should not. Moreover, an implicit consensus in the feature selection research so far has been that relevancy can and should be defined independently of both the classifier to be used and the evaluation metric; in other words, the relevancy of a variable should depend on the probability distribution of the data, not whether SVMs or any other specific classifier inducer will be used to build the final model. It would then be a matter of designing efficient algorithms that identify the relevant features to solve the feature selection problem.

In (Tsamardinos & Aliferis, 2003) we showed that *the above intuition is false* in the following sense:

Proposition 1: There is no definition of relevancy independent of the learner used to build the classification model and the performance metric used (e.g., accuracy) such that, for all distributions the relevant features are the solution to the feature selection problem.

Thus, one needs to specify a (class of) learning algorithm(s), a (set of) performance metric(s), and a class of distributions on which, a definition of relevancy can be given such that there is a correspondence between the set of features that are relevant and the solution to the feature selection problem. In (Tsamardinos & Aliferis, 2003) we also showed that:

Proposition 2: When the learning algorithm can learn any function (e.g., neural networks, k-nearest neighbors, etc.) and the performance metric is *calibrated accuracy with a preference for smaller models* then the solution to the feature selection problem is the *Markov Blanket.*

The Markov Blanket is defined as the smallest subset of variables $MB(Y)$ such that all remaining variables are independent of Y given $MB(Y)$.An important connection between relevancy and the Markov Blanket was shown (Tsamardinos & Aliferis, 2003) to be the following:

Proposition 3: The Markov Blanket of Y in faithful distributions coincides with the set of *strongly relevant* features as defined by Kohavi and John. Furthermore, in such distributions the $MB(Y)$ is unique and has a graphical interpretation too: it is the set of parents, children, and parents of children in any Bayesian Network that is faithful to the joint distribution.

Finally, when the joint distribution of the data can be faithfully represented by some Causal Bayesian Network and all confounders of each pair of variables are observed (Causal Sufficiency) then the following relationship exists between the Markov Blanket variables and the local causal structure around the variable Y:

Proposition 4: The Markov Blanket of Y is the set of direct causes, direct effects, and direct causes of direct effects of Y.

There are currently several algorithms available (Margaritis & Thrun, 1999; Tsamardinos et al., 2003; Aliferis et al., 2003b) that can identify the Markov Blanket in faithful distributions (i.e., the strongly relevant features) and thus solve the feature selection problem, under the conditions in Proposition 2 both theoretically and empirically. Additionally, the selected feature subset has a causal interpretation according to Proposition 4.

6. Conclusions and Open Problems

We provided an initial characterization of the behavior of weight-vector based linear SVM feature selection, both for the hard and the soft margin formulation. We show that for reasonable definitions of irrelevancy, in the sample limit an SVM will remove the (theoretically) irrelevant features. This partly explains the empirical success of these methods. However, the selected features will not in the general case contain the smallest feature subset, neither the correct one, nor is there a specific causal interpretation of the selected features. We emphasize the latter because in several domains feature selection is performed precisely to gain an understanding of the causal underlying mechanisms of the domain, e.g., biomarker selection in array gene expression data analysis.

In addition, we would like to point out that empirical evaluation of SVM-based feature selection is still an open area of research with no conclusive results as of yet. For example, several SVM feature selection methods have been conducted in the array gene expression domain (see special issue (Guyon & Ellisseeff, 2003) and the references therein). However gene expression values in microarrays are highly interrelated so that if a non-optimal feature selection algorithm routinely misses strongly relevant features but also does not remove all weakly relevant ones, (like SVM feature selection) the algorithm's performance may be only mildly affected by these theoretical weaknesses. This intuition is strengthened by the fact that it has been shown that in gene expression data even random gene subsets give good classification performance (Aliferis et al., 2003a).

The statistics community has also extensively examined the problem of whether a variable X^i should be included in a model or not. There are several tests that determine whether the weight of a variable is statistically significantly different than zero (e.g., see (Draper & Smith, 1981) in the context of multivariate linear regression model selection). We note that all SVM-related results in the present work apply to the linear (soft as well as hard-margin) classifier. However the notions of relevancy employed are model-independent. In other words, the definitions of relevancy only involve properties of the joint distribution of the data and not the inductive bias of some classifier. While linear SVMs in the sample limit will identify the above model-independent irrelevant variables, the same behaviour is not obtained unfortunately with weakly and strongly relevant features by the linear SVM. It is not clear at this time whether or how a sufficiently powerful kernel function combined with appropriate identification of irrelevant features would circumvent this limitation in the non-linear SVM case. We also did not address how the present results relate to the selection of features in the statistical literature of linear regression model selection. These connections could all be interesting topics for future work.

In the present paper we consider "hard" feature selection in which features are either accepted in the classifier model or considered irrelevant and thrown away. In addition our analysis of feature relevance is query independent. Other authors (Peng et al., 2003; Domeniconi & Gunopulos, 2001) have presented "soft" feature selection schemes in which features are weighted according to their SVM-based class discriminatory contribution with respect to specific queries. These feature weighting methods were shown to augment the performance of the KNN classifier in high-dimensional spaces.

Acknowledgments

We would like to thank the anonymous reviewers for their comments. This research was supported in part by NIH grants RO1 LM007948-01 and P20 LM 007613-01.

References

Aliferis, C. F., Tsamardinos, I., Massion, P., Statnikov, A., & Fananapazir, N. (2003a). Why classification models using array gene expression data perform so well. *International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences (METMBS).*

Aliferis, C. F., Tsamardinos, I., & Statnikov, A. (2003b). HITON, a novel markov blanket algorithm for optimal variable selection. *American Medical Informatics Association (AMIA)* (pp. 21–25).

Burges, C. J. C., & Crisp, D. J. (2000). Uniqueness of the SVM solution. *NIPS* (pp. 223–229).

Domeniconi, C., & Gunopulos, D. (2001). Adaptive nearest neighbor classification using support vector machines. *15th Conference in Neural Information Processing Systems (NIPS-01).*

Draper, N. R., & Smith, H. (1981). *Applied regression analysis.* John Wiley & Sons Inc. 2nd edition.

Guyon, I., & Ellisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research, 3,* 1157–1182.

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning, 46,* 389–422.

Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence, 97,* 273–324.

Kowalczyk, A. (2000). Maximal margin perceptron. *Advances in Large Margin Classifiers.* The MIT Press, Cambridge Massachusetts.

Margaritis, D., & Thrun, S. (1999). Bayesian network induction via local neighborhoods. *Advances in Neural Information Processing Systems 12 (NIPS).*

Meek, C. (1995). Strong completeness and faithfulness in bayesian networks. *Conference on Uncertainty in Artificial Intelligence* (pp. 411–418).

Peng, J., Heisterkamp, D., & Dai, H. K. (2003). SVM/LDA driven nearest neighbor classification. *IEEE Transactions on Neural Networks, 14*, 940–942.

Steinwart, I. (2003). Sparseness of support vector machines. *Journal of Machine Learning Research, 4*, 1071–1105.

Tsamardinos, I., & Aliferis, C. F. (2003). Towards principled feature selection: Relevancy, filters and wrappers. *Ninth International Workshop on Artificial Intelligence and Statistics (AI&Stats 2003).*

Tsamardinos, I., Aliferis, C. F., & Statnikov, A. (2003). Time and sample efficient discovery of Markov blankets and direct causal relations. *The Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 673–678).

Vapnik, V. (1998). *Statistical learning theory.* Johan Wiley & Sons, Inc. New York.

A. Proofs

We omit the Proof of Lemma 1 due to lack of space, which can be found in http://discover1.mc.vanderbilt.edu/ discover/ public/ supplements/ ICML2004/

Proof of Theorem 1 part (a). Suppose $(\mathbf{w}, b)$ separates $(\mathbf{X}, Y)$, that is, suppose (8) holds. Let c be such that both $P(\mathbf{w}^I \cdot \mathbf{X}^I \geq c)$ and $P(\mathbf{w}^I \cdot \mathbf{X}^I \leq c)$ are positive (for example, let $c = E(\mathbf{w}^I \cdot \mathbf{X}^I)$). We shall show that $(\mathbf{w}^R, b + c)$ also separates $(\mathbf{X}, Y)$. Since $\mathbf{w} \cdot \mathbf{w} = \mathbf{w}^R \cdot \mathbf{w}^R + \mathbf{w}^I \cdot \mathbf{w}^I \geq \mathbf{w}^R \cdot \mathbf{w}^R$ with equality only if $\mathbf{w}^I = \mathbf{0}$ it follows that $(\boldsymbol{\omega}_h)^I = 0$ which proves part (a) of Theorem 1. And so it remains to show that $(\mathbf{w}^R, b + c)$ separates $(\mathbf{X}, Y)$.

First, since $(\mathbf{w}, b)$ separates $(\mathbf{X}, Y)$ and since $\mathbf{w} \cdot \mathbf{X} = \mathbf{w}^R \cdot \mathbf{X}^R + \mathbf{w}^I \cdot \mathbf{X}^I$ we have

$$Y(\mathbf{w}^R \cdot \mathbf{X}^R + b + c) + Y(\mathbf{w}^I \cdot \mathbf{X}^I - c) \geq 1 \quad \text{(a. s.)}$$

and so

$$Y(\mathbf{w}^R \cdot \mathbf{X}^R + b + c) \geq 1 - Y(\mathbf{w}^I \cdot \mathbf{X}^I - c) \quad \text{(a. s.)} \quad (13)$$

Next, the independence of $\mathbf{X}^I$ and $(\mathbf{X}^R, Y)$ shows in the case $Y = 1$ that

$$\begin{aligned}
&P((Y(\mathbf{w}^R \cdot \mathbf{X}^R + b + c) \geq 1) \cap (Y = 1)) \\
&= P(Y(\mathbf{w}^R \cdot \mathbf{X}^R + b + c) \geq 1) \cap (Y = 1) \mid \mathbf{w}^I \cdot \mathbf{X}^I \leq c) \\
&\geq P(Y(\mathbf{w}^R \cdot \mathbf{X}^R + b + c) \geq 1 - Y(\mathbf{w}^I \cdot \mathbf{X}^I - c)) \\
&\qquad \cap (Y = 1) \mid \mathbf{w}^I \cdot \mathbf{X}^I \leq c) \\
&= P(Y = 1 \mid \mathbf{w}^I \cdot \mathbf{X}^I \leq c) = P(Y = 1)
\end{aligned}$$

where we have used the fact that (13) holds with probability 1 and that $P(Y = 1 \mid \mathbf{w}^I \cdot \mathbf{X}^I \leq c) = P(Y = 1)$ by independence. Similarly, for $Y = -1$ we have

$$\begin{aligned}
&P((Y(\mathbf{w}^R \cdot \mathbf{X}^R + b + c) \geq 1) \cap (Y = -1)) \\
&\geq P(Y(\mathbf{w}^R \cdot \mathbf{X}^R + b + c) \geq 1 - Y(\mathbf{w}^I \cdot \mathbf{X}^I - c)) \\
&\qquad \cap (Y = -1) \mid \mathbf{w}^I \cdot \mathbf{X}^I \geq c) \\
&= P(Y = -1).
\end{aligned}$$

Thus, $P(Y(\mathbf{w}^R \cdot \mathbf{X}^R + b + c) \geq 1)$

$$\begin{aligned}
&= P((Y(\mathbf{w}^R \cdot \mathbf{X}^R + b + c) \geq 1) \cap (Y = 1)) \\
&\quad + P((Y(\mathbf{w}^R \cdot \mathbf{X}^R + b + c) \geq 1) \cap (Y = -1)) \\
&\geq P(Y = 1) + P(Y = -1) = 1.
\end{aligned}$$

Since $\mathbf{w}^R \cdot \mathbf{X}^R = \mathbf{w}^R \cdot \mathbf{X}$ it immediately follows that $(\mathbf{w}^R, b + c)$ separates $(\mathbf{X}, Y)$ and the proof of part (a) is complete. □

We shall need the following lemma in the proof of part (b) of Theorem 1.

Lemma 5. *Suppose $p \geq 1$ and U, Y and V are real valued random variables such that $E(V) = 0$ and V is independent of (U, Y). Then $E([U + YV]_+^p) \geq E([U]_+^p)$.*

Proof. Fix $u, y \in \mathbf{R}$. If $u \geq 0$ then

$$[u + yV]_+ \geq u + yV = [u]_+ + yV$$

and so $E([u + yV]_+) \geq [u]_+ + yE(V) = [u]_+$ and if $u < 0$ the above inequality holds trivially since $E([u + yV]_+) \geq 0 = [u]_+$. Since the function $f(x) = x^p$ is convex for $p \geq 1$, Jensen's inequality implies

$$E([u + yV]_+^p) \geq E([u + yV]_+)^p \geq [u]_+^p.$$

$$\begin{aligned}
\text{so } E([U + YV]_+^p) &= E_{U,Y}(E_V([U + YV]_+^p \mid U, Y)) \\
&\geq E([U]_+^p).
\end{aligned}$$

□

Proof of Theorem 1 part (b). For $\mathbf{w} \in \mathbf{R}^N$ and $b \in \mathbf{R}$ let $e(\mathbf{w}, b) := E([1 - Y(\mathbf{w} \cdot \mathbf{X} + b)]_+^p)$ and let $c = E(\mathbf{w}^I \cdot \mathbf{X}^I)$.

Then Lemma 5 with $U := 1 - Y(\mathbf{w}^R \cdot \mathbf{X} + b + c)$ and $V := \mathbf{w}^I \cdot \mathbf{X} - c$ implies

$$\begin{aligned}
e(\mathbf{w},b) &= E([1 - Y(\mathbf{w}^R \cdot \mathbf{X} + b + c) - Y(\mathbf{w}^I \cdot \mathbf{X} - c)]_+^p) \\
&\geq E([1 - Y(\mathbf{w}^R \cdot \mathbf{X} + b + c)]_+^p) = e(\mathbf{w}^R, b + c).
\end{aligned}$$

Hence for any $\mathbf{w} \in \mathbf{R}^N$ there is some c such that

$$e(\mathbf{w}, b) \geq e(\mathbf{w}^R, b + c). \tag{14}$$

Using (14) we have

$$\begin{aligned}\mathcal{F}_s^*(\mathbf{w}^R, b + c) &= \mathbf{w}^R \cdot \mathbf{w}^R + Ce(\mathbf{w}^R, b + c) \\ &\leq \mathbf{w} \cdot \mathbf{w} + Ce(\mathbf{w}^R, b) = \mathcal{F}_s^*(\mathbf{w}, \beta)\end{aligned}$$

with strict inequality unless $\mathbf{w}^I = \mathbf{0}$. Hence we must have $\boldsymbol{\omega}^I = \mathbf{0}$ and part (b) of Theorem 1 is proved. □

Proof of Lemma 2. The SVM hard-margin classifier may also be determined geometrically from the training data as follows (e.g., see (Kowalczyk, 2000)). Recall that $X_+^{(m)} := \{\mathbf{x}_k | y_k = +1, 1 \leq k \leq m\}$ and $X_-^{(m)} := \{\mathbf{x}_k | y_k = -1, 1 \leq k \leq m\}$. Let $C_+^{(m)}$ and $C_-^{(m)}$ denote the closed convex hull of $X_+^{(m)}$ and $X_-^{(m)}$, respectively. Let $\mathbf{x}_+^{(m)} \in C_+^{(m)}$ and $\mathbf{x}_-^{(m)} \in C_-^{(m)}$ be a pair of points such that $|\mathbf{x}_+^{(m)} - \mathbf{x}_-^{(m)}| = \text{dist}\,(C_+^{(m)}, C_-^{(m)})$ where dist (A, B) denotes the minimum distance between sets A and B. Let $\mathbf{u}^{(m)} := (\mathbf{x}_+^{(m)} - \mathbf{x}_-^{(m)})$. Then

$$\mathbf{w}_m = \frac{2}{\Delta_m^2}\mathbf{u}^{(m)} \tag{15}$$

and $b_m = 1 - \mathbf{w}_m \cdot \mathbf{x}_+^{(m)}$ where $\Delta_m = |\mathbf{u}^{(m)}|$.

Let C_+ (respectively, C_-) denote the closed convex hull of the support of the conditional probabilities $p(\mathbf{x}|y = 1)$ (respectively, $p(\mathbf{x}|y = -1)$). Then, as in the finite sample case, $\boldsymbol{\omega}$ and β may be calculated geometrically from any pair of closest points $\mathbf{x}_+ \in C_+$ and $\mathbf{x}_- \in C_-$. Let $\mathbf{u} = \mathbf{x}_+ - \mathbf{x}_-$ and then

$$\boldsymbol{\omega} = \frac{2}{\Delta^2}\mathbf{u} \tag{16}$$

where $\Delta = |\mathbf{u}| = |\mathbf{x}_+ - \mathbf{x}_-|$ is the minimum distance between the sets C_+ and C_-. Note that this distance is also the distance between the hyperplanes $\boldsymbol{\omega} \cdot \mathbf{x} + \beta = +1$ and $\boldsymbol{\omega} \cdot \mathbf{x} + \beta = -1$.

Clearly, we have $C_+^{(m)} \subset C_+$ and $C_-^{(m)} \subset C_-$ almost surely and so the finite sample $(\mathbf{x}_k, y_k)_{k=1}^m$ is almost surely linearly separable if the random variable $(\mathbf{X}, Y)$ is linearly separable and, hence, $\Delta \leq \Delta^m$ almost surely.

If $\mathbf{z}_+ \in C_+$ and $\mathbf{z}_- \in C_-$ then $\boldsymbol{\omega} \cdot \mathbf{z}_+ + \beta \geq 1$ and $\boldsymbol{\omega} \cdot \mathbf{z}_- + \beta \leq -1$ and so

$$\boldsymbol{\omega} \cdot (\mathbf{z}_+ - \mathbf{z}_-) \geq 2 \qquad \text{for } \mathbf{z}_+ \in C_+ \text{ and } \mathbf{z}_- \in C_-. \tag{17}$$

Let $\mathbf{u}$ and $\mathbf{u}^{(m)}$ be as above. Then (17) implies $\boldsymbol{\omega} \cdot \mathbf{u}^{(m)} \geq 2$ which by (16) is equivalent to $\mathbf{u} \cdot \mathbf{u}^{(m)} \geq |\mathbf{u}|^2$. Using (15) and (16) we then have

$$\begin{aligned}|\boldsymbol{\omega} - \mathbf{w}_m|^2 &= |\frac{2}{|\mathbf{u}|^2}\mathbf{u} - \frac{2}{|\mathbf{u}^{(m)}|^2}\mathbf{u}^{(m)}|^2 \\ &= 4\left(\frac{1}{|\mathbf{u}|^2} + \frac{1}{|\mathbf{u}^{(m)}|^2} - \frac{2\mathbf{u} \cdot \mathbf{u}^{(m)}}{|\mathbf{u}|^2|\mathbf{u}^{(m)}|^2}\right) \\ &\leq 4\left(\frac{1}{|\mathbf{u}|^2} + \frac{1}{|\mathbf{u}^{(m)}|^2} - \frac{2}{|\mathbf{u}^{(m)}|^2}\right) \\ &= 4\left(\frac{|\mathbf{u}^{(m)}| + |\mathbf{u}|}{|\mathbf{u}^{(m)}|^2|\mathbf{u}|^2}\right)\left(|\mathbf{u}^{(m)}| - |\mathbf{u}|\right)\end{aligned}$$

which, since $\Delta = |\mathbf{u}|$ and $\Delta_m = |\mathbf{u}^{(m)}|$, proves (12). □

Proof of Lemma 3. In the proof we use the property that if sets P and Q are independent, i.e., $p(P, Q) = p(P)p(Q)$, then all subsets $P' \subseteq P$, $Q' \subseteq Q$ are also independent: $p(P', Q') = p(P')p(Q')$.

Let $X^i \in X^I$, S be any subset of $V \setminus \{Y, X^i\}$. Furthermore, let $S^I = S \cap X^I$ and $S^R = S \cap X^R$. By the definition of irrelevancy we have that $p(X^I, Y, X^R) = p(X^I) \cdot p(X^R, Y)$. Since $\{X^i\} \cup S^I \subseteq X^I$ and $S^R \subseteq \{Y\} \cup X^R$ then, for all values of X^i and S for which $p(X^i, S) > 0$:

$$\begin{aligned}p(Y|X^i, S) = p(Y|X^i, S^I, S^R) &= \frac{p(Y, X^i, S^I, S^R)}{p(X^i, S^I, S^R)} \\ &= \frac{p(Y, S^R) \cdot p(X^i, S^I)}{p(X^i, S^I) \cdot p(S^R)} = \frac{p(Y, S^R)}{p(S^R)} \\ &= \frac{p(Y, S^R) \cdot p(S^I)}{p(S^R) \cdot p(S^I)} = \frac{p(Y, S^R, S^I)}{p(S^I, S^R)} = p(Y|S)\end{aligned}$$

Thus,$\forall X^i \in X^I, S \subseteq V \setminus \{Y, X^i\}$ where $p(X^i, S) > 0$, X^I is independent of Y conditioned on any S, and by definition X^I is KJ-irrelevant to Y. □

Proof of Lemma 4. Let X^i be any KJ-irrelevant variable to Y and X^r any relevant one. Faithfulness implies (by Theorems 5 and 6 (Tsamardinos & Aliferis, 2003)) that X^i has no (undirected) path to Y and that X^r does have a path to Y in any network faithfully representing the distribution. Thus, X^i has no path to X^r (or it would also have a path to Y through X^r). Thus, any subset S^I of the set of all X^i is independent of any subset S^R of the set of all X^r. We can now prove the lemma by following the steps of the previous proof in reverse order by noting the independence $p(S^I, S^R) = p(S^I) \cdot p(S^R)$ and that X^i is independent of Y conditioned on any S by KJ-irrelevance.

□