Challenges in Statistical Machine Learning

John Lafferty
Computer Science Department
Machine Learning Department
Carnegie Mellon University

Larry Wasserman
Department of Statistics
Machine Learning Department
Carnegie Mellon University

March 23, 2006

1. Introduction

Machine learning and statistics are one and the same discipline, with different communities of researchers attacking essentially the same fundamental problems from different perspectives. In this note we briefly describe some current challenges in the field of statistical machine learning that cut across the communities. We focus on areas where active development of learning techniques demonstrates promising performance, but with significant gaps in the theoretical foundations; filling the gaps will help to explain and improve upon this performance. The themes are high dimensional data, sparsity, semi-supervised learning, the relation between computation and risk, and structured prediction. Our selection of these themes is highly biased (and therefore has high risk), but we believe that these challenges can benefit from a combination of the statistics and computer science perspectives on learning from data.

2. Sparse Learning in High Dimensions

Most statistical theory is based on asymptotic approximations that allow the sample size n to grow large. When the number of variables d in the model is large, however, this theory can be misleading. One important challenge in statistical machine learning is to develop relevant theory and methods when the dimension of the data grows with the number of data points. Such a theory should yield insights for real data sets with moderate sample sizes but large dimensions. Sparsity clearly has to play a central role in this emerging theory.

In the standard statistical prediction problem, we observe n pairs of data $(X_1, Y_1), \ldots, (X_n, Y_n)$ where $X_i = (X_{i1}, \ldots, X_{id})^T$ is a d-dimensional vector of covariates and Y_i is a response. The goal is to predict Y from X. The usual regression model is $Y_i = m(X_i) + \epsilon_i,\ i = 1, \ldots, n$, where $m : \mathbb{R}^d \to \mathbb{R}$ is the unknown regression function and ϵ_i is a mean 0 noise variable. Estimating m nonparametrically is hopeless if d is large, unless we add extra assumptions. For example, it is well known that

$$\liminf_{n\to\infty} n^{4/(4+d)} \inf_{\widehat{m}_n} \sup_{m\in\mathcal{M}} R(\widehat{m}_n, m) \ > \ 0$$

where $R(\widehat{m}_n, m) = \mathbb{E}_m \int (\widehat{m}_n(x) - m(x))^2 \, dx$ is the *risk* of the estimate $\widehat{m}_n$ constructed on a sample of size n and $\mathcal{M}$ is the Sobolev space of order two. This implies that the best rate of convergence is $n^{-4/(4+d)}$, which in turn implies that the sample size n needs to grow exponentially with dimension d to keep the risk small. This is the statistical *curse of dimensionality.* The computational burden also increases exponentially with dimension. This is the computational curse of dimensionality. It is worth pointing out that even the parametric, linear model is difficult both statistically and computationally if d is very large.

For some applications it is reasonable to expect that m is *sparse* in some sense. In such cases it may be possible to "beat" the computational and statistical curses of dimensionality using various greedy algorithms, but little theoretical support is currently available for such techniques. In the linear case, it might be reasonable to assume that $\|\beta\|_1 \equiv \sum_j |\beta_j|$ is small, which implies that many of the β_j's are close to zero. Alternatively, one might assume that $m(x)$ actually only depends on a small number r of the covariates so that $m(x) = \sum_{j \in R} \beta_j x_j$ where R has cardinality r. Such sparsity assumptions play a critical role in many new methods for high-dimensional problems. The lasso estimator (Tibshirani, 1996) of β in the linear model $m(x) = x^T \beta$ is

$$\widehat{\beta} = \operatorname{argmin}_\beta \left(\sum_{i=1}^n (Y_i - \beta^T X_i)^2 + \lambda \|\beta\|_1 \right).$$

The same estimator was proposed in signal processing under the name basis pursuit (Chen et al., 1998). Since the optimization problem is convex, the estimator can be found efficiently, in principle. Justification for the estimator, recently provided by Donoho (2004) in the signal processing context, hinges on sparsity; related recent work is Fu and Knight (2000) and Fan and Peng (2004). For nonparametric problems Zhang et al. (2005) use likelihood basis pursuit, essentially the lasso adapted to the spline setting.

An alternative method is L_2 boosting, closely related to forward stepwise regression and matching pursuit. For standardized variables (zero mean, unit variance), the L_2 boosting algorithm for estimating $m(x) = x^T \beta$ can be expressed as the iteration $\widehat{m} \longleftarrow \widehat{m} + \nu \widehat{\beta}_j x_j$ where x_j is the variable most correlated with the current residuals, $\widehat{\beta}_j$ is the corresponding least squares estimate and $\nu > 0$. Bühlmann (2006) showed the following remarkable result. Let the number of variables $d(n)$ increase as $d = O(\exp(Cn^\gamma))$ for some $\gamma > 0$ and $C > 0$. Assuming the sparsity condition $\limsup_{n\to\infty} \sum_{j=1}^{d(n)} |\beta_j| < \infty$ as well as some mild technical conditions, L_2 boosting is consistent: $\mathbb{E}|\widehat{m}(X) - m(X)|^2 = o(1)$ as $n \to \infty$.

In the nonparametric setting, recent work of Lafferty and Wasserman (2005) has developed the *rodeo*, which stands for "regularization of derivative expectation operator." It is based on fitting a local linear regression with large bandwidths, and then incrementally reducing the bandwidths in small, greedy steps. The decision of whether or not to change a bandwidth is based on a statistical test on the size of the derivative. Assuming sparsity, Lafferty and Wasserman (2005) show that the resulting estimator has nearly optimal rates of convergence for Sobolev classes in high dimensions, as if the relevant variables were isolated and known in advance.

But a convergence rate such as $n^{-4/(4+r)}$ for a function depending on r relevant variables in d dimensions may still not be strong enough to explain the impressive performance seen by many heuristic machine learning algorithms empirically. Recently, Audibert and Tysbakov (2005) have proposed a framework under which *superlinear* rates of convergence can be obtained for plug-in classifiers, that is, classification rules of the form

$$\widehat{m}(x) = \mathbb{I}\left(\widehat{\eta}_n(x) > \tfrac{1}{2}\right)$$

where $\widehat{\eta}_n$ is an estimate of the regression function $\eta(X) = \mathbb{P}(Y = 1 \mid X)$ for a binary classification problem $Y \in \{0, 1\}$. In particular, it is shown that the optimal rate satisfies

$$\sup_{\mathbb{P}} \left\{ \mathbb{E}\,\mathbb{P}(\widehat{m}_n \neq Y) - \inf_m \mathbb{P}(m(X) \neq Y) \right\} = O\left(n^{-\frac{\beta(1+\alpha)}{2\beta+d}} \right)$$

where $\mathbb{P}$ is a probability distribution on (X, Y), $\widehat{m}_n$ is the plug-in classification rule for a local polynomial estimator on a sample of size n, β is the Hölder exponent of the regression function $\mathbb{P}(Y \mid X)$, and α is the exponent in the *margin assumption*

$$\mathbb{P}\left(0 < \left|\eta(X) - \tfrac{1}{2}\right| \leq t\right) \leq Ct^{\alpha}, \quad \forall t > 0$$

(The relation between the margin condition and variance-mean bounds is discussed in Shen and Wang (2006).) However, to beat the statistical curse, such analysis requires that $\beta\alpha = O(d)$. With the dimension d growing with sample size n, this may not be realistic, since it may require the margin to be too large, or the decision boundary to be too smooth. Moreover, the method assumes that the bandwidth is selected as $h = n^{-1/(2\beta+d)}$, which, apart from not being data-dependent, does not allow for sparsity and does not address the computational curse. But these impressive results suggest that significant advances are being made toward a realistic theory for learning in high dimensions.

3. Semi-Supervised Learning

In a typical machine learning problem, labeled examples are time consuming and expensive to obtain relative to raw data, since the labeling may require expensive experiments, clinical trials, or the efforts of human experts who must often be quite skilled. Indeed, if it were otherwise, the machine learning problem would probably not be of significant interest in the first place. For example, it is easy to collect acoustic speech by pointing a microphone at a TV or radio, but accurately transcribing the speech requires significant time and effort. The challenge of *semi-supervised learning* is to somehow leverage large amounts of unlabeled data in order to improve upon a learning algorithm that uses only labeled data. Interestingly, while this problem has attracted significant attention recently in the machine learning community, there is little work in this direction in the more traditional statistics literature.

Several novel approaches to this problem have been proposed recently, with results that suggest significant improvements may be obtainable by combining labeled with unlabeled data. However, from a theoretical standpoint the problem is wide open. Among these recent methods is a promising family of techniques that exploit the "manifold structure" of the data. Such methods are generally based upon an assumption that similar unlabeled examples should be given the same classification. The learning methods have intimate connections with random walks, electric networks and spectral graph theory, heat kernels and normalized cuts used in image processing.

To illustrate, we briefly mention the approach of Zhu et al. (2003) based on Gaussian random fields and harmonic functions defined with respect to discrete Laplace operators. Standard kernel regression corresponds to the locally constant estimator

$$\begin{aligned} \widehat{m}_n(x) &= \arg\min_{m(x)} \sum_{i=1}^n K_h(X_i, x)(Y_i - m(x))^2 \\ &= \frac{\sum_{i=1}^n K_h(X_i, x)\, Y_i}{\sum_{i=1}^n K_h(X_i, x)} \end{aligned}$$

where K_h is a symmetric kernel depending on bandwidth parameters h. In the semi-supervised approach of Zhu et al. (2003), the locally constant estimate $\widehat{m}(x)$ is formed using not only the labeled data, but also using the estimates at the *other* unlabeled points. Suppose that the first ℓ data points $(X_1, Y_1), \ldots, (X_\ell, Y_\ell)$ are labeled, and the next u points are unlabeled, $X_{\ell+1}, \ldots, X_{\ell+u}$. The semi-supervised regression estimate is then

$$\widehat{m}_n = \arg\min_m \sum_{i=1}^n \sum_{j=1}^n K_h(X_i, X_j)\,(m(X_i) - m(X_j))^2$$

where the minimization is carried out subject to the constraint $m(X_i) = Y_i,\ i = 1, \ldots, \ell$. Thus, the estimates are coupled. The local linear version would solve the least squares problem

$$\widehat{m}_n = \arg\min_m \sum_{i=1}^n \sum_{j=1}^n K_h(X_i, X_j)\,(\beta_0(X_i) - (X_i - X_j)^T \beta(X_j))^2$$

with the estimator $\widehat{m}_n(X_j) = \beta_0(X_j)$ for $j = \ell + 1, \ldots, \ell + u$.

This estimator can be viewed in several different ways. For example, it is the posterior of a Gaussian random field, corresponding to the configuration of the field with smallest total energy, subject to boundary conditions specified by the labeled points, thus solving a graphical Dirichlet problem (Doyle and Snell, 1984). In contrast, for multi-label *discrete* random fields, computing the lowest energy configuration is typically NP-hard, and approximation algorithms or other heuristics must be used, as have been extensively developed in the computer vision literature (Boykov et al., 2001). Another view is to note that the estimator can be written in closed form as

$$\widehat{m} = \Delta_{uu}^{-1}\,\Delta_{ul}\,Y = GY$$

where Δ_{uu} and Δ_{ul} denote appropriate blocks of the combinatorial Laplacian on the data graph,

$$\widehat{m} = (\widehat{m}(X_{\ell+1}), \ldots, m(X_{\ell+u}))^T$$

is the vector of estimates over the unlabeled test points, and $Y = (Y_1, \ldots, Y_\ell)^T$ is vector of labeled values. This expresses the *effective kernel* G in terms of the "data manifold," which can be thought of in terms of heat kernels for the discrete diffusion equations (Smola and Kondor, 2003). Related work in semi-supervised learning by Chapelle et al. (2002) uses eigenvalues of the Laplacian to create various kernels, and an approach of Belkin and Niyogi (2002) regularizes functions on the data graph by selecting the top p normalized eigenvectors of the Laplacian corresponding to the smallest eigenvalues.

While this preliminary work has been promising, with semi-supervised learning often dramatically outperforming conventional approaches, many important questions remain unresolved. For example, it is unknown how to handle noise in these methods, and how to construct the underlying graphs automatically from data, which encodes the data manifold. This latter problem can be viewed as equivalent to bandwidth selection, for which methods based on the rodeo may be applicable. Virtually nothing is known about minimax theory for such problems.

In the analysis of traditional approaches to kernel regression, it is well known that the actual kernel used is not as important as the choice of bandwidths. In particular, in one dimension the risk of the locally-constant (Nadaraya-Watson) estimator is

$$\begin{aligned} R(\widehat{m}_n, m) &= \frac{h^4}{4}\left(\int x^2 K^2(x)\,dx\right)\int\left(m''(x) + 2m'(x)\frac{f'(x)}{f(x)}\right)^2 dx \\ &\quad + \frac{\sigma^2\int K^2(x)dx}{nh}\int\frac{1}{f(x)}\,dx + o(nh^{-1}) + o(h^4) \end{aligned}$$

where $h \to 0$ and $nh \to \infty$. The multi-dimensional version is part of the analysis given by Ruppert and Wand (1994). The term $2m'(x)\frac{f'(x)}{f(x)}$ involving the first derivative of the regression function and the derivative of the logarithm of the sampling density f is called the *design bias*; it involves the distribution of the covariates. When using large amounts of unlabeled data, it can be assumed that the design bias is known. An analysis of semi-supervised regression and classification must somehow incorporate more global information about the sampling density—that is, the data manifold—and the smoothness of the regression function with respect to this manifold. It should be possible to establish rates that are faster than those obtained using only labeled data, under appropriate assumptions.

4. Computation and Risk

Statistical methods are usually aimed at finding procedures that make the mean prediction error, or risk, small. But these measures of risk ignore computation cost. It is important to develop new theoretical frameworks that combine statistical prediction error with computational complexity.

Computational learning theory has developed the PAC model of learning as a framework for studying the complexity of discrete classification problems (Valiant, 1984; Pitt and Valiant, 1988). Several significant advances have resulted directly from thinking about the computational and algorithmic aspects of machine learning within this framework. Examples include boosting (Freund and Schapire, 1996), exponentiated gradient algorithms for online learning (Kivinen and Warmuth, 1997; Kivinen et al., 1997), and Fourier based methods for Boolean problems (Kushilevitz and Mansour, 1993; Linial et al., 1993). More recent work has studied learning in the context of approximation algorithms (Alekhnovich et al., 2004). One important computational learning problem, which is closely related to the the sparse regression problem discussed above, is the task of learning a "k-junta," that is, a Boolean function that depends on only k of d Boolean variables, with $d \to \infty$. The brute force approach requires $O(d^k)$ examples to learn the function exactly. In the noise-free case an algorithm with time-complexity $O(d^{\frac{k\omega}{\omega+1}}) = O(d^{0.7k})$ was recently given by Elchanan et al. (2004), where $\omega < 2.37$ is the exponent of matrix multiplication. However, the problem becomes computationally intractable in the presence of noise, within the statistical query model (Kearns, 1998).

Overall, the PAC model's focus on the traditional complexity-theoretic dividing line of polynomial versus exponential time or space has resulted in the theory being largely built up around negative examples. This suggests that the underlying theoretical framework may be too rigid. It would be very interesting to develop new theoretical frameworks based on the *tradeoff* between computation and risk that is important in practice; this tradeoff appears to have largely been ignored in both statistical theory and computational learning theory.

The computation-risk tradeoff for learning is perhaps more akin to the classical theory of numerical optimization than it is to classical complexity theory and NP-completeness. In considering basic line search or trust region methods for unconstrained optimization, for example, one can consider a (locally) quadratically convergent Newton's algorithm, which may require $O(d^3)$ flops in each iteration, or a superlinearly convergent quasi-Newton algorithm that will require only $O(d^2)$ flops in each iteration, or a linearly convergent gradient descent algorithm that may cost $O(d)$ flops. Similarly, preconditioning methods for solving a linear system $Ax = y$ with conjugate gradient use a sparse matrix B to approximate A, and solve $B^{-1}Ax = B^{-1}b$. The time $T(A)$ required for an ϵ-approximate solution is then

$$T(A) = \sqrt{\kappa(A,B)}\,(m + T(B)) \log\left(\frac{1}{\epsilon}\right)$$

where m is the number of nonzero entries in A, $\kappa(A,B)$ is the condition number, measuring the quality of approximation, and $T(B)$ is the time required to solve $By = c$. In each case, one can trade off computation for the rate of numerical convergence to the solution.

Analogously, in nonparametric learning we expect to be able to trade off the amount of computation invested to search over a set of smoothing parameters against the rate of minimax convergence, or the richness of the function space that the method can learn at a given rate. A search procedure over subsets of size r requires $O(d^r)$ time, and as dimension grows, this cost for large r may be charged against the gains in statistical risk compared with a search over smaller sets of variables.

A classical minimax rate $\rho_{n,d}$ for n examples in d dimensions satisfies

$$\liminf_{n\to\infty} \inf_{\widehat{m}_n \in \mathcal{H}} \sup_{m\in\mathcal{F}} \rho_{n,d}\, \mathcal{R}(\widehat{m}_n, m) > 0$$

for a hypothesis class $\mathcal{H}$ and function space $\mathcal{F}$. It would be interesting to investigate new frameworks where the computational cost $\kappa_{n,d}$ for estimating a function on n examples in d dimensions is taken into account. In this setting, one could look for computational minimax rates satisfying

$$\liminf_{n\to\infty} \inf_{\widehat{m}_n \in \mathcal{H}} \sup_{m\in\mathcal{F}} \mathcal{U}(\kappa_{n,d}, \rho_{n,d})\, \mathcal{R}(\widehat{m}_n, m) > 0$$

where $\mathcal{U}$ plays the role of a utility function. If computational cost is not taken into account in the utility function, the classical minimax theory is recovered.

As an example, the rodeo method (Lafferty and Wasserman, 2005) is greedy in that it only tests the current fit against the next smallest bandwidth. A more sensitive test was used in Lepski et al. (1997), Lepski and Spokoiny (1997) and in the multivariate version in Kerkyacharian et al. (2001), the idea being to use the largest bandwidth h from a grid of bandwidths $\mathcal{H}_n$ such that $\widehat{m}_h$ is not significantly different from any $\widehat{m}_\eta$ where η varies over all bandwidths in $\mathcal{H}_n$ that are more refined than h. In contrast, the rodeo tests $\widehat{m}_h$ only against the set of bandwidths just smaller than h. The distinction is exhaustive search versus greedy search. The exhaustive method yields estimators that are adaptively minimax for L_r loss over a large scale of Besov spaces and losses, namely, $\mathcal{S} = \{B^s_{p,q} : 1 \leq p, q \leq \infty, s > (1/p - 1/q)_+\}$, while the rodeo achieves optimal rates only over $B^2_{2,2}$ and $r = 2$. But the rodeo involves much less computation. Thus, there is a *computation-adaptation tradeoff*. A compromise between these two extremes is to restrict the tests to a set of bandwidths $G(h)$ of varying polynomial size. Large G gives full adaptivity while small G saves computation. It should be possible to quantify the computation-adaptation tradeoff by finding the adaptivity scale $\mathcal{S}$ as a function of the size of the testing set G.

A related tradeoff is present in hierarchical regression and classification schemes such as dyadic decision trees, which were recently shown by Scott and Nowak (2006) to have nearly optimal rates of convergence, giving theoretical support to a family of techniques that have been popular for decades. Scott and Nowak (2006) prove that if one allows $M = O(\log n)$ dyadic splits in each of d covariates, and if the tree is chosen to minimize a penalized classification error, then the resulting classifier has adaptive minimax properties. For small dimensions the *optimal* tree, as determined by the penalized empirical risk, can be found using dynamic programming. But the search over all such trees is computationally intractable for large d; thus the statistical curse is in principle addressed, but only by ignoring the computational curse. It may be possible to quantify the tradeoff between adaptivity of the classifier and computational complexity as measured by the maximum depth of the search tree.

5. Structured Prediction

Structured prediction is a term used in the machine learning community for a classification or regression problem with non-iid data, where typically the dependencies are encoded in a graphical model. This topic has seen a good deal of activity in recent years, prompted by both technical advances and important applications. Problems such as speech recognition, image denoising, object recognition, natural language parsing, information extraction, handwriting recognition, gene prediction, machine translation and many others can be naturally cast as structured prediction problems. While many of the problems themselves are not new, the underlying methods are recent developments that have come on the heels of advances in kernel methods, approximate inference in graphical models, and large margin techniques for classification, including support vector machines. An incomplete sample of recent work on methods and applications of structured prediction includes (Lafferty et al., 2001; Collins, 2002; Pinto et al., 2003; McCallum, 2003; Kumar and Hebert, 2003; Sha and Pereira, 2003; Taskar et al., 2003; Altun et al., 2004; Tsochantaridis et al., 2005).

Formally, a structured prediction problem can be thought of as a multi-class problem with a large number of class labels, typically exponential in the number of variables. But in order to develop estimators and efficient algorithms, the structure of the problem must be taken into account. In the simplest case, the structure is a linear chain, as in a hidden Markov model. But when conditional models are used, complicated features of the entire input sequence can be incorporated. The number of parameters increases rapidly, so that regularization and sparsity become essential.

The maximum margin Markov network[1] framework of Taskar et al. (2003) is based on the use of loss functions that can be decomposed into a linear combination of losses associated with the cliques in a graphical model. Taskar et al. (2003) propose a generalization of the SVM hinge loss for structured problems, and show that the resulting optimization problem can be solved with the same techniques used for inference in undirected graphical models. The optimization algorithm is efficient if the underlying graph has low tree width. Closely related methods are developed by Tsochantaridis et al. (2005).

In some cases, covering number or support vector based generalization error bounds for structured prediction have been developed (Collins, 2002; Taskar et al., 2003); but little is currently known about convergence rates or minimax theory for these problems. The analysis will be complicated by the fact that the methods are often used in conjunction with approximate inference techniques for graphical models, for example, variational methods or relaxations of integer linear or quadratic programs. These methods themselves are not well understood statistically; for example, there is currently no reasonable analysis of the bias and variance properties of mean-field or structured variational approximations (Wainwright and Jordan, 2003), although such techniques are widely used in machine learning.

A more basic statistical challenge associated with these techniques has to do with consistency—

[1] Markov network is the AI terminology for a random field or undirected graphical model.

the convergence of the excess risk to zero as the sample size tends to infinity. While consistency for large margin binary classifiers is now well understood (Bartlett et al., 2006), the consistency problem for multi-class problems is not fully resolved. It follows from the work of Lee et al. (2004) and more recently Tewari and Bartlett (2005) that the max-margin Markov network generalization of the SVM hinge loss is inconsistent. The only known consistent methods for structured prediction problems in this class are based on conditional likelihood, where the consistency follows from standard theory. The traditional statistical thinking, which often demands consistency before anything else, perhaps deserves to be reconsidered in this context.

References

M. Alekhnovich, M. Braverman, V. Feldman, A. Klivans, and T. Pitassi. Learnability and automatizability. In *Proceedings of the 45th Foundations of Computer Science (FOCS)*, 2004.

Y. Altun, T. Hofmann, and A. J. Smola. Gaussian process classification for segmenting and annotating sequences. In *ICML-04, 21sth International Conference on Machine Learning*, 2004.

J.-Y. Audibert and A. B. Tysbakov. Fast learning rates for plug-in classifiers under the margin condition. Technical report, PMA-998, Laboratoire de Probabilités Paris 6 and 7, December 2005. arXiv:math.ST/0507180.

P. Bartlett, M. Jordan, and J. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, March 2006.

M. Belkin and P. Niyogi. Semi-supervised learning on manifolds. Technical Report TR-2002-12, University of Chicago, 2002.

Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 23(11), November 2001.

P. Bühlmann. Boosting for high-dimensional linear models. *The Annals of Statistics*, 34(2), 2006.

O. Chapelle, J. Weston, and B. Schölkopf. Cluster kernels for semi-supervised learning. In *Advances in Neural Information Processing Systems, 15*, volume 15, 2002.

S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific and Statistical Computing*, 20:33–61, 1998.

M. Collins. Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, 2002.

D. Donoho. For most large underdetermined systems of equations, the minimal ℓ^1-norm near-solution approximates the sparest near-solution. *Technical report*, *Stanford*, 2004.

P. Doyle and J. Snell. *Random Walks and Electric Networks*. Mathematical Assoc. of America, 1984.

M. Elchanan, R. O'Donnell, and R. Servedio. Learning functions of k relevant variables. *J. Comput. System Sci.*, 69(3):421–434, 2004.

J. Fan and H. Peng. Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, 32:928–961, 2004.

Y. Freund and R. E. Schapire. Experiments with a new boosting algorithm. In *Machine Learning: Proceedings of the Thirteenth International Conference*, pages 148–156, 1996.

W. Fu and K. Knight. Asymptotics for lasso type estimators. *The Annals of Statistics*, 28:1356–1378, 2000.

M. Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM*, 45(6): 983–1006, 1998.

G. Kerkyacharian, O. Lepski, and D. Picard. Nonlinear estimation in anisitropic multi-index denoising. *Probability Theory and Related Fields*, 121:137–170, 2001.

J. Kivinen and M. Warmuth. Additive versus exponentiated gradient updates for linear prediction. *Journal of Information and Computation*, 132(1):1–64, 1997.

J. Kivinen, M. Warmuth, and P. Auer. The perceptron algorithm vs. winnow: Linear vs. logarithmic mistake bounds when few input variables are relevant. *Artificial Intelligence on Relevance*, 97 (1–2):325–343, 1997.

S. Kumar and M. Hebert. Discriminative fields for modeling spatial dependencies in natural images. In *Advances in Neural Information Processing Systems 16*, 2003.

E. Kushilevitz and Y. Mansour. Learning decision trees using the Fourier spectrum. *SIAM Journal on Computing*, 22, 1993.

J. Lafferty and L. Wasserman. Rodeo: Sparse nonparametric regression in high dimensions. *http://xxx.arxiv.org/pdf/math.ST/0506342*, 2005.

J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*, pages 282–289, 2001.

Y. Lee, Y. Li, and G. Wahba. Multicategory support vector machines: Theory and application to the classification of microarray data and satellite radiance data. *Journal of the American Statistical Association*, 99(465), 2004.

O. V. Lepski and V. G. Spokoiny. Optimal pointwise adaptive methods in nonparametric estimation. *Ann. Statist.*, 25(6):2512–2546, 1997.

O. V. Lepski, E. Mammen, and V. G. Spokoiny. Optimal spatial adaptation to inhomogeneous smoothness: An approach based on kernel estimates with variable bandwidth selectors. *The Annals of Statistics*, 25:929–947, 1997.

N. Linial, Y. Mansour, and N. Nisan. Constant depth circuits, Fourier transform, and learnability. *Journal of the ACM*, 40:607–620, 1993.

A. McCallum. Efficiently inducing features of conditional random fields. In *Nineteenth Conference on Uncertainty in Artificial Intelligence (UAI03)*, 2003.

D. Pinto, A. McCallum, X. Wei, and W. B. Croft. Table extraction using conditional random fields. In *Proceedings of the 26th Annual International ACM SIGIR conference on Research and Development in Informaion Retrieval*, pages 235–242. ACM Press, 2003.

L. Pitt and L. G. Valiant. Computational limitations on learning from examples. *J. ACM*, 35(4): 965–984, 1988.

D. Ruppert and M. P. Wand. Multivariate locally weighted least squares regression. *The Annals of Statistics*, 22:1346–1370, 1994.

C. Scott and R. Nowak. Minimax-optimal classification with dyadic decision trees. *IEEE Transactions on Information Theory*, 2006.

F. Sha and F. Pereira. Shallow parsing with conditional random fields. In *Proceedings of HLT-NAACL*. Association for Computational Linguistics, 2003.

X. Shen and L. Wang. Discussion of 2004 IMS Medallion lecture: Local Rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 2006. To appear.

A. Smola and R. Kondor. Kernels and regularization on graphs. In *Conference on Learning Theory, COLT/KW*, 2003.

B. Taskar, C. Guestrin, and D. Koller. Max-margin Markov networks. In *Advances in Neural Information Processing Systems 16*, 2003.

A. Tewari and P. L. Bartlett. On the consistency of multiclass classification methods. In *Proceedings of the 18th Annual Conference on Learning Theory*, volume 3559, pages 143–157. Springer, 2005.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B, Methodological*, 58:267–288, 1996.

I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, 6:1453–1484, September 2005.

L. G. Valiant. A theory of the learnable. *Commun. ACM*, 27(11):1134–1142, 1984.

M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. Technical report, Number 649, Department of Statistics, University of California, Berkeley, 2003.

H. Zhang, G. Wahba, Y. Lin, M. Voelker, R. K. Ferris, and B. Klein. Variable selection and model building via likelihood basis pursuit. *Journal of the American Statistical Association*, 99(467): 659–672, 2005.

X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using Gaussian fields and harmonic functions. In *ICML-03, 20th International Conference on Machine Learning*, 2003.