Laboratoire de l'Informatique du Parallélisme

École Normale Supérieure de Lyon
Unité Mixte de Recherche CNRS-INRIA-ENS LYON
n° 8512

Linear Ensemble Methods for Multiclass Discrimination

Yann Guermeur
Hélène Paugam-Moisy

December 1998

Research Report N° 1998-52

École Normale Supérieure de Lyon
46 Allée d'Italie, 69364 Lyon Cedex 07, France
Téléphone : +33(0)4.72.72.80.37
Télécopieur : +33(0)4.72.72.80.80
Adresse électronique : `lip@ens-lyon.fr`

Linear Ensemble Methods for Multiclass Discrimination

Yann Guermeur
Hélène Paugam-Moisy

December 1998

Abstract

We consider the case of a discrimination problem for which several classifiers are available. The object of this study is the combination of the class posterior probability estimates provided by these classifiers, in order to obtain better estimates and concomitantly a better discriminant function. The ensemble method is the multivariate linear regression. We specify the optimization problem to which training amounts and we characterize the sets of optimal solutions corresponding to the main convex objective functions. Means to assess and control the generalization capabilities are also investigated.

Keywords: Discrimination, probability estimation, linear regression

Résumé

Nous considérons le cas d'un problème de discrimination pour lequel plusieurs classifieurs sont disponibles. L'objet de cette étude est la combinaison des estimations des probabilités *a posteriori* des classes fournies par ces classifieurs, afin d'obtenir de meilleures estimations et, de manière concomitante, une meilleure fonction de discrimination. Le modèle de combinaison considéré est la régression linéaire multivariée. Nous spécifions le problème d'optimisation auquel se réduit l'apprentissage et caractérisons les ensembles de solutions optimales correspondant aux principales fonctions objectif convexes. Nous étudions également le moyen d'évaluer et de contrôler les performances en généralisation.

Mots-clés: Discrimination, estimation de probabilités, régression linéaire

1 Introduction

Model combination is a subject that has been intensively investigated in statistics during the last three decades [3]. However, theoretical evidence has been mainly developed for forecasting [8] and regression [7], whereas the specificities of discrimination have seldom been taken into account [16]. We study here the combination of discriminant systems which estimate the class posterior probabilities, by means of the model of Multivariate Linear Regression (MLR). We demonstrate that optimal solutions providing new posterior probability estimates can be obtained by restricting the domain of feasible solutions to a convex polyhedron and choosing adequately the objective function. This leads to the definition of a convex programming problem simple to solve. Results are given that characterize the sets of optimal solutions with respect to the nature of the objective function. Finally, the question of the generalization ability is tackled throughout a computation of an upper bound of the sample complexity.

2 Specification of the optimization problem

Let us consider a space of description X and a set of Q categories C_k $(1 \leq k \leq Q)$. For each element of X, the outputs of P classifiers are available to predict the corresponding class. Let f_j be the function computed by the j^{th} classifier: $f_j(x) = [f_{j,k}(x)] \in \mathbb{R}^Q$. Its k^{th} component $f_{j,k}(x)$ approximates $p(C_k|x)$. For each example x and each classifier f_j, $f_j(x)$ thus belongs to U with:

$$U = \left\{ u \in \mathbb{R}_+^Q / 1_Q^T u = 1 \right\}$$

The Q components of vector 1_Q are equal to 1. U is the non-negative part of the hyperplane of $\mathbb{R}^Q$ defined by $1_Q^T u = 1$. Let $F(x) = [f_j(x)]$, $(1 \leq j \leq P)$ $(F(x) \in U^P)$. The MLR model studied here, parametrized by $v = [v_k] \in \mathbb{R}^{PQ^2}$, computes the function g given by:

$$g(x) = \begin{bmatrix} g_1(x) \\ \vdots \\ g_k(x) \\ \vdots \\ g_Q(x) \end{bmatrix} = \begin{bmatrix} v_1^T \\ \vdots \\ v_k^T \\ \vdots \\ v_Q^T \end{bmatrix} F(x)$$

Let $\mathcal{J}$ be the set of objective functions the minimization of which ensures that the combiner will actually estimate the class posterior probabilities. Then, finding a solution to our combination problem amounts to solving the following optimization problem:

Problem 1 *Given a convex objective function $J \in \mathcal{J}$, find among the functions g taking their values in U one which minimizes $J(v)$.*

A necessary and sufficient condition for the components of $g(x)$ to sum to one can be established by considering two vectors $F(x_1)$ and $F(x_2)$ in U^P which only differ by two components, namely:

$$f_{l,m}(x_1) = f_{l,m}(x_2) + \epsilon, \; f_{l,n}(x_1) = f_{l,n}(x_2) - \epsilon.$$

Let $v_{k,l,m}$ be the general term of v_k, i.e. the coefficient associated with the m^{th} output of classifier f_l in the regression computed to estimate $p(C_k|x)$. We have then $1_Q^T\left(g(x_1)-g(x_2)\right)=\epsilon\sum_{k=1}^{k=Q}\left(v_{k,l,m}-v_{k,l,n}\right)=0$, whence we deduce the conditions $\forall(l,m,n),\ \sum_{k=1}^{k=Q}(v_{k,l,m}-v_{k,l,n})=0$. They can be expressed in a more compact form. Favouring the last component of vectors $f_l(x)$, we get for instance $\forall(l,m),\ \sum_{k=1}^{k=Q}(v_{k,l,m}-v_{k,l,Q})=0$. Under these constraints, the sum of the outputs is constant. For it to be equal to 1, it is necessary and sufficient to impose further $1_{PQ^2}^T v=Q$. The condition of non-negativity can be written for each function g_k as: $\sum_{l=1}^{l=P} min_m(v_{k,l,m})\geq 0$. It takes the form of a system of Q^{P+1} linear inequalities. This springs from the fact that since the indices $m(k,l)$ satisfying $v_{k,l,m(k,l)}=min_m(v_{k,l,m})$ are not known *a priori*, for each of the Q possible values of k, the Q^P different forms that the inequality $\sum_{l=1}^{l=P} min_m(v_{k,l,m})\geq 0$ can take must be taken into account. Each of the resulting Q^{P+1} inequality thus generates a constraint. The size of this system constitutes the limiting factor for the resolution of Problem 1. However, we have established in [9] that solutions would exist in the non-negative part of the convex domain defined by the set of aforementioned constraints. To find such solutions, it is sufficient to solve Problem 2:

Problem 2 *Minimize $J(v)$ subject to $v\in V$ with*

$$V=\left\{\ v\in\mathbb{R}_+^{PQ^2}\ /\left\{\begin{array}{l}\forall(l,m),\ \sum_{k=1}^{k=Q}(v_{k,l,m}-v_{k,l,Q})=0\\ 1_{PQ^2}^T v=Q\end{array}\right.\right\}$$

This proposition, which holds irrespective of the nature of the convex objective function J, is a consequence of the redundancy of the predictors. With this simplification, the training procedure consists in solving a simple convex programming problem, which can be done efficiently by means of standard algorithms such as the *gradient projection method* [13] (see [9] for details).

3 Choice of the training criterion

Among the objective functions that can be considered, the quadradic cost and the cross-entropy satisfy the general conditions for outputs to be interpreted as probabilities (see for instance [6]). For discrimination problems, implementing the principle of maximum likelihood amounts to minimizing the cross-entropy error function. The least-squares criterion, on the contrary, is ordinarily derived from the principle of maximum likelihood under hypotheses that do not make sense in the context of discrimination. The first criterion should thus *a priori* be preferred here. However, we observed empirically that choosing either of them makes little difference for performance (see for instance [9]). Furthermore, it is easier to characterize the optimal solutions for the quadratic loss. With this criterion, the following important result holds:

Theorem 1 *All the solutions to Problem 2 obtained by the least-squares method correspond to a single function g.*

This theorem is a direct consequence of the following lemma.

Lemma 1 *The difference between two least-squares solutions of Problem 2 is a vector of* $\left(\left(U^P\right)^\perp\right)^Q$.

The proof of lemma 1 is straightforward. Let $s = \{(x_i, y_i)\}$, $(1 \leq i \leq N)$ be the training sample, where y_i is the canonical coding of the category of x_i, i.e. $y_i = [\delta_{k,l}]$, $(1 \leq k \leq Q) \iff x_i \in C_l$, where δ denotes the Kronecker symbol. Let $F(X) \in M_{N,PQ}(\mathbb{R})$ denote the matrix of explanatory variables (its lines are the vectors $F(x_i)^T$). Define $A = \frac{1}{N} I_Q \otimes F(X)^T F(X)$, where $\otimes$ denotes the Kronecker product. Let $Y_k = [y_{ik}] \in \mathbb{R}^N$, $(1 \leq k \leq Q)$ be the vectors of desired outputs and $b = \frac{1}{N}[Y_1^T F(X), \ldots, Y_k^T F(X), \ldots, Y_Q^T F(X)]^T$. $J(v)$ (or more precisely its sample-based estimate) is then:

$$J(v) = \frac{1}{2} v^T A v - b^T v + \frac{1}{2}$$

Let v and $v + w$ be two least-squares solutions of Problem 2. We can then write $J(v+w) = J(v)$, from which it comes $\frac{1}{2} w^T A w + (Av - b)^T w = 0$. Notice that $Av - b = \nabla J(v)$. By hypothesis, w is neither an ascent nor a descent direction at v, thus $(Av - b)^T w = 0$ and consequently $w^T A w = 0$. Since $w^T A w = 0 \iff F(X) w_k = 0_N$, $(1 \leq k \leq Q)$, $w \in Ker(A)$ and more generally $w \in \left(\left(U^P\right)^\perp\right)^Q \square$.

It springs from Theorem 1 that the choice of one particular least-squares solution has no incidence on the generalization performance.

Although the loss function induced by the L_1 distance does not satisfy the general conditions for outputs to be interpreted as probabilities, it can be used here, in view of the following result:

Theorem 2 *A solution to Problem 2 of minimal modules is obtained by selecting the best expert with respect to this criterion.*

The use of the Minkowski-1 error measure thus specifies a model selection method rather than a model combination method. The proof is based upon a study of the set of *extreme points* or *vertices* of the *polyhedron* V, set which contains at least one optimal solution to the corresponding linear programming problem. Indeed, these extreme points satisfy:

$$\forall \left((k_1, l_1, m_1), (k_2, l_2, m_2)\right), (1 \leq k_1, k_2, m_1, m_2 \leq Q), (1 \leq l_1 < l_2 \leq P)$$

$$v_{k_1,l_1,m_1} v_{k_2,l_2,m_2} = 0$$

This can be proved by contradiction. Let us consider a vector $v \in V$ which does not exhibit this property. Let f_{l_1} and f_{l_2} denote two classifiers which are taken into account in the combination associated with vector v (i.e. the corresponding coefficients are not all equal to zero). Thus, since v satisfies $\forall (l, m)$, $\sum_{k=1}^{k=Q} (v_{k,l,m} - v_{k,l,Q}) = 0$, we have:

$$\forall i \in \{1, 2\}, \forall m \in \{1, \ldots, Q\}, \exists k \in \{1, \ldots, Q\} / v_{k,l_i,m} > 0$$

By definition of V, we have further $v_{k,l_i,m} \in]0, 1[$. For each pair (l_i, m), let $k(l_i, m)$ be any of the indices k such that $v_{k,l_i,m} \in]0, 1[$. Let ϵ be any value small enough to satisfy: $\forall(l_i, m),\ 0 \leq v_{k(l_i,m),l_i,m} - \epsilon < v_{k(l_i,m),l_i,m} + \epsilon \leq 1$. Let $v^{(1)}$ and $v^{(2)}$ be two vectors of V deduced from v in the following way:

$$\forall(i,j) \in \{1,2\}^2,\ \forall m \in \{1,\dots,Q\}\,,\ v^{(j)}_{k(l_i,m),l_i,m} = v_{k(l_i,m),l_i,m} + (-1)^{(i+j)}\epsilon$$

The values of the components of $v^{(1)}$ and $v^{(2)}$ unspecified by this expression are the values of the components of v of same index. By construction, both $v^{(1)}$ and $v^{(2)}$ belong to V, and $v = \frac{1}{2}\left(v^{(1)} + v^{(2)}\right)$, whence we conclude that v is not an extreme point of V $\square$.

4 Complexity control

Although the learning process described in the previous sections only amounts to estimating probability densities on a finite sample, the criterion of interest, the generalization ability in terms of recognition rate, can be both estimated and controlled. To perform this task, we use results derived in the framework of the computational learning theory and the statistical learning theory introduced by Vapnik [17]. Roughly speaking, they represent corollaries of uniform laws of large numbers. The measure of confidence in the generalization capabilities we use is called the *sample complexity* of the model. In order to define it, we must first introduce additional notations.

Let H be a family of functions from X to a finite set Y. Let S denote the cartesian product of X and Y and p be a distribution of probability on S. Let s be a N-sample of observations ($s \in S^N$). We designate by $E_s(h)$ and $E(h)$ respectively the error on s (*observed error*) and the generalization error of a function h belonging to H.

Definition 1 *The sample complexity of a family of functions H is the smallest size N_0 of the sample such that, for a given triplet $(\gamma, \delta, \epsilon) \in [0,1]^3$, the following bound holds:*

$$p^N \left\{ s \in S^N : \exists h \in H / \left(E(h) > \epsilon \wedge E_s(h) \leq (1-\gamma)E_g(h)\right) \right\} \leq \delta$$

To bound the sample complexity of the MLR model, we used the notions of *growth function* and *Vapnik-Chervonenkis (VC) dimension* introduced by Vapnik.

Definition 2 *Let H be a set of indicator functions (values in {0,1}). Let s be a sample of X and $\Pi_H(s)$ the number of different classifications of s by the functions of H. We define the growth function Π_H by:*

$$\Pi_H(N) = \max\left\{\Pi_H(x) : x \in X^N\right\}.$$

Definition 3 *The VC dimension of a set H of indicator functions is the maximum number d of vectors that can be shattered, i.e. separated into two classes in all 2^d possible ways, using functions of H. If this maximum does not exist, the VC dimension is equal to infinity.*

For multiple-output functions, we have the following theorem (see for instance [2]):

Theorem 3 *Suppose that* $\epsilon \in]0,1[$ *and* $\gamma \in]0,1]$. *Then the probability*

$$p^N \left\{ s \in S^N : \exists h \in H / \left(E(h) > \epsilon \wedge E_s(h) \leq (1-\gamma) E_g(h) \right) \right\}$$

is at most

$$4 \Pi_{\mathcal{G}H}(2N) \exp\left(-\frac{\gamma^2 \epsilon N}{4} \right)$$

$\Pi_{\mathcal{G}H}$ is the growth function of the set of graphs (*graph space*) of the functions of H [12]. The discriminant functions corresponding to our regression model are also computed by the single hidden layer perceptron with threshold activation functions depicted in Figure 1.

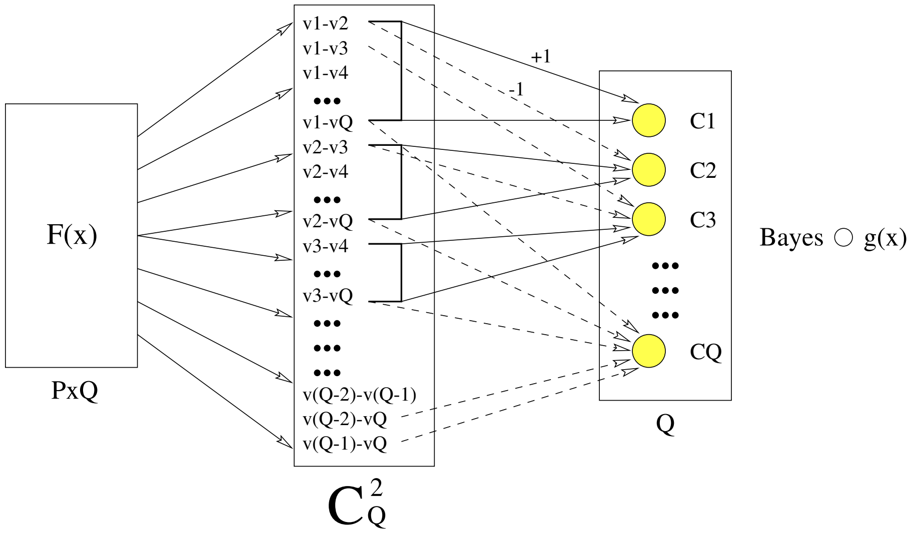

Figure 1: Architecture of a multi-layer perceptron with threshold units computing the same discriminant function as g. Each hidden unit computes a function $h_{i,j}(F(x)) = t\left(F(x)^T(v_i - v_j)\right)$, $(1 \leq i < j \leq Q)$, where $t(z) = 1$ if $z > 0$ and $t(z) = -1$ otherwise. The weights of the output layer, either $+1$ (solid lines) or -1 (dashed lines), and the biases are chosen so that the output units compute a logical AND. The number below each layer corresponds to the number of units.

Several articles have been devoted to bounding the growth function and VC dimension of multi-layer perceptrons [4, 12, 15, 14, 11]. Proceeding as in [4, 15], we use as upper bound on the growth function the product of the bounds on the growth functions of the individual hidden units derived from Sauer's lemma. Since the VC dimension of each hidden unit is at most equal to the dimension of the smallest subspace of $\mathbb{R}^{PQ}$ containing U^P, and this dimension is $d_E = P(Q-1)+1$, we get:

$$\Pi_{\mathcal{G}PMC}(2N) < \left(\frac{e2N}{d_E} \right)^{C_Q^2 d_E} \tag{1}$$

By substitution in the bound of theorem 3, we establish that N_0 is bounded above by:

$$\frac{8}{\gamma^2 \epsilon}\left[\ln\left(\frac{4}{\delta}\right) + C_Q^2 d_E \ln\left(\frac{48 C_Q^2}{\gamma^2 \epsilon}\right)\right] \tag{2}$$

Indeed, in order to get

$$4\Pi_{\mathcal{G}PMC}(2N)\exp\left(-\frac{\gamma^2 \epsilon N}{4}\right) \leq \delta,$$

it is sufficient to ensure that

$$\frac{\gamma^2 \epsilon N}{4} - C_Q^2 d_E \ln N + \ln\left(\frac{\delta}{4}\right) - C_Q^2 d_E \ln 6 \geq -C_Q^2 d_E \ln d_E.$$

Since

$$-C_Q^2 d_E \ln d_E \leq C_Q^2 \left(\frac{\gamma^2 \epsilon N}{8 C_Q^2} - d_E \ln\left(\frac{\gamma^2 \epsilon N}{8 C_Q^2}\right)\right),$$

a sufficient condition is given by:

$$\frac{\gamma^2 \epsilon N}{8} + \ln\left(\frac{\delta}{4}\right) - C_Q^2 d_E \ln 6 \geq C_Q^2 d_E \ln\left(\frac{8 C_Q^2}{\gamma^2 \epsilon}\right),$$

from which the bound results. This bound provides us with a measure of the confidence that one can have in the observed error. The *graph dimension* $d_{\mathcal{G}H}$ of a set of functions H is the largest integer d such that $\Pi_{\mathcal{G}H}(d) = 2^d$. It is thus the VC dimension of the graph space of H. From (1), the following bound on the graph dimension of the PMC follows:

$$d_{\mathcal{G}PMC} < 2 C_Q^2 d_E \log_2\left(e C_Q^2\right) \tag{3}$$

The bounds on the growth function and the sample complexity can be significantly improved if additional "sensible" constraints on vector v are introduced. To highlight this phenomenon, we use the following theorem due to Vapnik [17]:

Theorem 4 *Suppose the input space X is contained in a ball of radius R about some point. Consider the set of hyperplanes in canonical form with respect to X that satisfy $\|w\| \leq A$. Its VC dimension is bounded above by $\lfloor R^2 A^2 \rfloor + 1$.*

One can easily verify that the radius of the smallest hypersphere in which U^P is included is $\frac{1}{Q}\sqrt{PQ\,(Q-1)}$. Furthermore, the norm of every hyperplane associated with a hidden unit is bounded above by $\sqrt{Q}$. Consequently, Vapnik's theorem provides us with a bound on the VC dimension of each hidden unit which is exactly equal to d_E. This means that any constraint resulting in a decrease in the bound on the norm of the vectors $(v_i - v_j)$, $(1 \leq i < j \leq Q)$, will have a direct incidence on the bounds (1) and (2). In practice, we found it particularly appropriate to set upper values on coefficients $v_{k,l,m}$ [10]. However, many other possibilities can be considered.

5 Discussion

The linear combination of discriminant experts has been considered successively from the point of view of the optimization of the training performance and the evaluation of the generalization performance. Results have thus been established dealing with thoses two criteria, which pave the way for an implementation of the *Structural Risk Minimization* (SRM) inductive principle [17].
In practice, the main quality of our model is to offer a good compromise between capacity and robustness. It springs from experimental results that for many real-life problems, benefits can be expected from the use of ensemble methods more sophisticated than the common weighted averages. More complex models, such as neural networks, appear over sensitive to the fluctuations of the behaviour of the classifiers resulting for instance of the implementation of simplified procedures of *stacked generalization*. This phenomenon has been highlighted by experiments performed on an open problem in predictive structural biology, the prediction of the secondary structure of globular proteins [9, 10]. For that problem, the MLR combiner outperformed all the ensemble methods we implemented, from the majority voting to complex neural systems.
This work can be seen as a first step towards the implementation of the SRM inductive principle for multiclass discriminant models. We currently study the means to implement this principle so that the outputs could still be considered as estimates of the class posterior probabilities. Establishing the possibility to do so would be of central importance. Up to now, solving multiclass discrimination task by means of machines implementing the SRM principle implies to rely on heuristics. For instance, in the case of Support Vector Machines (SVM) [17], the number of classifiers trained is equal to the number of classes. Each classifier separates one class from the rest. The global classification is done by choosing the class with the largest classifier output value. With our approach, the heuristics can be replaced with the theorically well grounded Bayes' estimated decision rule. Furthermore, it should make it possible to use as complexity measure of the family of functions measures which have only been used so far in the framework of regression, such as the *covering number* [1, 18]. This should lead to far superior bounds on the sample complexity.

References

[1] N. Alon, S. Ben-David, N. Cesa-Bianchi and D. Haussler (1997): Scale-sensitive Dimensions, Uniform Convergence, and Learnability. *Journal of the ACM*, **Vol. 44**, N^o4, 615-631.

[2] M. Anthony (1997): Probabilistic Analysis of Learning in Artificial Neural Networks: The PAC Model and its Variants. *Neural Computing Surveys*, **Vol. 1**, 1-47.

[3] J.M. Bates and C.W.J. Granger (1969): The Combination of Forecasts. *Operational Research Quaterly*, **Vol. 20**, 451-468.

[4] E.B. Baum and D. Haussler (1989): What Size Net Gives Valid Generalization. *Neural Computation*, **Vol. 1**, 151-160.

[5] S. Ben-David, N. Cesa-Bianchi, D. Haussler and P.M. Long (1995): Characterizations of Learnability for Classes of $\{0, \ldots, n\}$-Valued Functions. *Journal of Computer and System Sciences*, 50, 74-86.

[6] C.M. Bishop (1995): *Neural Networks for Pattern Recognition.* Clarendon Press, Oxford.

[7] L. Breiman (1996): Stacked Regressions. *Machine Learning*, **Vol, 24**, 49-64.

[8] J.P. Dickinson (1975): Some Comments on the Combination of Forecasts. *Operational Research Quaterly*, **Vol. 26**, 205-210.

[9] Y. Guermeur (1997): Combinaison de Classifieurs Statistiques, Application à la Prédiction de la Structure Secondaire des Protéines. *Thèse de doctorat de l'Université Paris 6.*

[10] Y. Guermeur, C. Geourjon, P. Gallinari and G. Deléage (1998): Improved Performance in Protein Secondary Structure Prediction by Inhomogeneous Score Combination. Submitted to *Bioinformatics.*

[11] P. Koiran and E.D. Sontag (1996): Neural Networks with Quadratic VC Dimension. *Advances in Neural Information Processing Systems 8*, 197-203.

[12] B.K. Natarajan (1989): On Learning Sets and Functions. *Machine Learning*, **Vol. 4**, 67-97.

[13] J.B. Rosen (1960): The Gradient Projection Method for Nonlinear Programming. Part I. Linear Constraints. *J. SIAM*, **Vol. 8**, n^o 1, 181-217.

[14] A. Sakurai (1993): Tighter Bounds of the VC-dimension of Three-layer Networks. Proc. *WCNN'93*, 540-543.

[15] J. Shawe-Taylor and M. Anthony (1991): Sample sizes for multiple-output threshold networks. *Network: Computation in Neural Systems*, **Vol. 2**, 107-117.

[16] K. Tumer and J. Ghosh (1995): Theoretical Foundations of Linear and Order Statistics Combiners for Neural Pattern Classifiers. *TR-95-02-98*, The Computer and Vision Research Center, The University of Texas at Austin.

[17] V. Vapnik (1995): *The Nature of Statistical Learning Theory.* Springer, N.Y.

[18] R.C. Williamson, A.J. Smola and B. Schölkopf (1998): Generalization Performance of Regularization Networks and Support Vector Machines *via* Entropy Numbers of Compact Operators. *NeuroCOLT2 Technical Report Series, NC2-TR-1998-019.*