ICMS Workshop on the Vapnik-Chervonenkis Dimension List of Abstracts

Invited talks

VC-dimension, sample-compression schemes and learning

Shai Ben-David

Technion

ABSTRACT Given a class of sets, $C \subseteq 2^X$, a sample-compression scheme for C is a mapping from (input) sequences of examples, labeled according to their membership in some set from C, to short sub-sequences. It is required that the labels of the examples in the input sequence can be reconstructed from the short sub-sequence.

The VC-dimension of a class C is a lower bound on the size of compressed sub-sequences of any compression scheme for C. The question whether the VC-dimension implies also an upper bound (on the minimal sufficient size of compressed images) is an intriguing open problem. In this talk we shall show that, for certain families of finite VC-dimension classes, compression schemes do exist.

Another interesting aspect of sample-compression schemes is their relationship to learning. We shall introduce several variants of the compression problem, and show implications between the existence of compression schemes of certain types (for a class C) and the learnability (of that class) in a corresponding learning model. For example, jointly with Bshouty and Kushilevitz, we have recently proved the exact learnability, from equivalence queries only, of classes of geometric concepts. A key component in that proof is the construction of strong compression schemes for half-spaces in $\mathbb{R}^n$.

VC Dimension, Covering Numbers and Worst-case Prediction of Individual Sequences

David Haussler

University of California at Santa Cruz

ABSTRACT Suppose you are given a set of functions, here called a "comparison set," in which each function is a different method for predicting whether the stock market will go up or down on a given day, given the values of certain economic indicators for that day. Suppose further that over the next N days, you must predict the direction of the stock market each day, and your goal is to predict as well as that function in the comparison set that does best in predicting the market for these coming N days, no matter what the market does. Of course, this goal is unattainable, so you settle for a prediction method that is guaranteed to predict "almost" as well as the function that

does best over the coming N days. Such a prediction method is called a worst-case (or "universal") predictor for individual sequences. The worst-case difference between your predictive performance and the performance of the function in the comparison set that predicts best is called your worst-case regret. An average case regret can be defined similarly. These and related quantities are of fundamental interest in statistics. How small you can make your regret depends on the size of the comparison set. We outline how this tradeoff works. We discuss how for infinite comparison sets, the size can be fruitfully measured in terms of certain kinds of covering numbers, and, in turn, these covering numbers can be upper-bounded using the VC dimension.

Geometric applications: discrepancy and derandomization

Jiří Matoušek

Charles University, Prague

ABSTRACT We survey some applications of the Vapnik-Chervonenkis dimension and related concepts (shatter functions etc.) in combinatorial and computational geometry.

Most of the talk is devoted to geometric *discrepancy theory* (or theory of *irregularities of distribution*). The problems in discrepancy theory have usually been formulated in a "continuous" setting: One asks how uniformly can an n-point set be distributed in the unit cube, where the uniformity of distribution is measured with respect to some class of geometric shapes. The current best asymptotic upper bounds for several problems of this type are obtained by purely combinatorial methods.

Derandomization means the construction of deterministic algorithms from randomized ones (randomized algorithms are usually much easier to design). For many geometric algorithms dealing with objects in a Euclidean space of a small dimension, an asymptotically efficient derandomization can be done using a deterministic algorithm for computing ε-approximations and ε-nets.

Tutorial

Tutorial on PAC learning and related frameworks

Anselm Blumer

Tufts University, MA

ABSTRACT Recent interest in the computational complexity of learning has resulted in a large number of frameworks that allow different aspects of learning to be quantified. This tutorial will focus on the Probably Approximately Correct (PAC) framework, in which the learner is given random examples according to a fixed but unknown probability distribution, where each example consists of a point and its classification. The learner tries to select a hypothesis that will have high probability of correctly classifying new points drawn from the same distribution. Variations on this framework include restrictions on allowable probability distributions, attribute or classification noise in the examples, and restrictions on hypothesis classes.

Classification noise is often studied in the Statistical Query (SQ) framework, in which the learner makes queries about the expected value of a functional and is given an answer within a specified tolerance. Results from this framework can be translated into the PAC framework.

A third related framework is exact learning via membership and equivalence queries. A membership query is just a request for the classification a given point. An equivalence query gives a hypothesis and is returned a counterexample in case the hypothesis is incorrect. Learning using equivalence queries may be affected by restrictions on the allowable hypotheses. Relationships between this framework and the PAC framework will be discussed.

Contributed talks

Probabilistic generalisation of functions, and uniform convergence

Martin Anthony

London School of Economics

ABSTRACT In this talk, I shall survey recent applications in learning theory of "scale-sensitive" dimensions. These dimensions are extensions, applicable to real-valued functions, of the VC-dimension.

Pattern classification and scale-sensitive versions of the VC-dimension for neural networks

Peter L. Bartlett

Australian National University

ABSTRACT The estimation error in a pattern classification problem depends on the Vapnik-Chervonenkis dimension of the function class. If the function class consists of thresholded versions of real-valued functions, and misclassification probability is estimated using the number of training examples that are labelled to the correct side of the threshold by some positive margin, recent results have shown that the estimation error instead depends on a scale-sensitive version of the VC-dimension known as the fat-shattering dimension. This talk describes a bound on this dimension for feed-forward sigmoidal neural networks that, in contrast with the VC-dimension, does not depend on the number of parameters if the size of the parameters is bounded. It also describes an extension of the estimation error inequality that gives a similar bound on misclassification probability that holds uniformly over all values of this bound on the parameters. This leads to a complexity regularization technique in which the complexity term involves the size of the parameters. Thus, arguments involving the fat-shattering dimension support the use of popular neural network heuristics that attempt to keep the parameters small during training.

Ideals over the reals defined by notions of shattering

Shai Ben-David

Technion

ABSTRACT We discover close relationships between the notions of VC-dimension and shattering, and the classical notions of measure and category of sets of real numbers. Using the common correspondence between reals and sets of natural numbers, we identify *sets* of reals with *classes* of subsets of $\mathbb{N}$ and obtain a notion of VC-dimension for subsets of $\mathbb{R}$.

We apply this duality both ways. On one hand, we employ the notion of shattering to construct a proper extension of Lebesgue measure over the reals. On the other direction, we invoke techniques of real analysis to show that the family of classes of infinite VC-dimension can be divided into three distinct levels of complexity, where each of the two lower levels is a σ-ideal (in the set-theoretic sense). This talk is based on joint work with Leonid Gurvits.

A graph-theoretic generalization of the Sauer-Shelah Lemma.

Nicolò Cesa-Bianchi

DSI, University of Milan

ABSTRACT We show a natural graph-theoretic generalization of the Sauer-Shelah lemma. This result is applied to bound the ℓ_∞ and L_1 packing numbers of classes of functions whose range is an arbitrary, totally bounded metric space.

(Joint work with David Haussler, UC Santa Cruz.)

A power law for the sample size when learning in the presence of noise

Paul Fischer

Universität Dortmund

ABSTRACT In this talk we investigate the sample size behaviour of PAC learning algorithms in the presence of malicious noise. We mainly focus on the use of randomised hypotheses for learning and high malicious noise rates. It is shown that the information theoretic upper bound $\eta_{\text{det}} = \varepsilon/(1+\varepsilon)$ on the tolerable noise rate which holds for deterministic hypotheses can be replaced by $\eta_{\text{rand}} = 2\varepsilon/(1+2\varepsilon)$, where ε is the desired accuracy. Moreover we show that a quadratic power law $1/\Delta^2$, where $\eta = \eta_{\text{rand}} - \Delta$ is the malicious noise rate, engages when the noise rate approaches this bound. Especially we present a strategy which uses randomised hypotheses for learning the powerset of d elements using an optimal sample size of order $d\varepsilon/\Delta^2$ (ignoring logarithmic factors) and tolerating a noise rate $\eta = \eta_{\text{rand}} - \Delta$. A similar result can be proved for the class of unions of k intervals with $d = 2k$ the VC-dimension of this class. We complement these results by proving that no learner can output ε-accurate randomised (or deterministic) hypotheses with a noise rate larger or equal to η_{rand}, this disregarding sample size.

Combinatorial and topological properties of VC-dimension and its generalizations

Leonid Gurvits

NEC Research Institute, Princeton

ABSTRACT In a first part of the talk I will discuss a number of inequalities which look very much like Sauer's lemma. All of them are based on interpolational dimension, i.e., for a given set of vectors, say X, $\mathrm{ID}(X)$ is a minimal degree d such that any interpolational problem on X can be solved by some polynomial of degree not greater than d. (In a case of Pollard's dimension one needs slightly different notion). Basically,the main idea to get an upper bound on a size of the set X is to get some upper bound on $\mathrm{ID}(X)$ and after that just to count the number of corresponding monomials. Moreover,it is possible to obtain almost by free lower bounds for cardinalities of projections of the set X. I will show how this method can be applied to "shorten" inclusion-exclusion formulas and to bound VC-dimension of so called cell-wise constant functions, the later bound uses also Milnor's theory based bounds. In a second part I will talk about the following problem: for a given set F of bounded measurable function when does the closure of a convex hall $\mathrm{CO}(F)$ of F in L_2 topology coincide with the closure of $\mathrm{CO}(F)$ in L_∞ topology.

(Joint work with P. Koiran).

VC dimension and the sample complexity of error estimation techniques

Sean B. Holden

University College, London

ABSTRACT Much of the research performed to date that attempts to apply the VC dimension and related theory to machine learning problems has studied the probability that the resubstitution estimate—often referred to as the "error on the training set"—provides, in some sense, a bad estimate of the true error probability. It is well-known that in practice the resubstitution estimate tends to be highly optimistic, in the sense that it underestimates the true error probability, and it is therefore interesting to ask whether this kind of analysis can be extended to provide results regarding more sophisticated estimates, such as the cross-validation and holdout estimates. This talk studies both of the latter estimates within the standard PAC framework, presenting new bounds which apply for any consistent learning algorithm.

General upper bounds for VC dimension in terms of descriptive and computational complexity

Mark Jerrum

University of Edinburgh

ABSTRACT Using a classical result from real algebraic geometry it is possible to provide simple and very general upper bounds on the VC dimension of continuously parameterised concept classes (set systems). These bounds are polynomial in the number of real

parameters used to specify a concept (set) in the following situations: (a) the criterion for membership of an instance in a concept (point in a set) can be expressed as a formula in the first-order theory of the reals, with fixed quantification depth and exponentially-bounded length, whose atomic predicates are polynomial inequalities of exponentially-bounded degree, and (b) membership of an instance in a concept is decidable in polynomial time in a model of computation akin to that of Blum, Shub and Smale that allows exact arithmetic on reals as atomic operations.

These upper bounds cover most of the concept classes that are likely to be of practical interest. Thus, in the continuous setting, as in the discrete, the real barrier to efficient learning in the Occam sense is complexity-theoretic and not information-theoretic.

(Joint work with Paul Goldberg.)

VC dimension of feedforward and recurrent neural networks

Pascal Koiran

ENS, Lyon

ABSTRACT I will present several bounds (mostly lower bounds) on the VC dimension of various types of "neural" networks.

It will be shown that for most activation functions the VC dimension of a feedforward network can be at least quadratic in the number of parameters.

A distinctive feature of recurrent networks is that a fixed-size net can receive inputs of arbitrary length. So far, they have not been studied as much as feedforward nets from the point of view of VC dimension. I will present a few results in this direction.

(Joint work with Eduardo Sontag.)

The connection between model theory and VC dimension

Angus Macintyre

Oxford University

ABSTRACT The model-theoretic notion of o-minimality arose in the attempt to solve Tarski's Problem on the decidability of the real exponential field. The relevance of the notion to VC-dimension is that in an o-minimal structure all families of definable sets, in arbitrary dimension, have finite VC-dimension. The method provides by far the most extensive list of situations where finite VC-dimension prevails. I discuss the various examples, and connect them to the issue of definable sets having finitely many connected components. This is relevant for constructive aspects of the theory, where, for example, estimates obtained from Morse theory are close to optimal. I will discuss Wilkie's recent work which shows, inter alia, that the Pfaffian structure on the reals has the finite VC property.

On the VC-dimension of networks of spiking neurons

Michael Schmitt
Technische Universität Graz

ABSTRACT Networks of spiking neurons have gained interest as models for information processing in neural systems because they appear to be biologically more relevant than networks of threshold or sigmoidal gates. There is experimental evidence that some tasks are performed by biological systems using the timing of single firing events rather than average firing rates. In building neural net hardware one has also started to employ methods such as pulse stream VLSI where analog values are encoded by time differences between pulses.

Furthermore, recent theoretical results show that some networks of spiking neurons are computationally more powerful than networks that use traditional neuron models as computational units. For instance, one was able to exhibit functions that can be computed by a single spiking neuron but require a number of units that grows almost linearly in the number of inputs when computed by circuits of threshold, resp. sigmoidal, gates.

The ability of biological neural nets to learn and the need of artificial neural nets to be programmed motivate the investigation of the learning capabilities of networks of spiking neurons. In this talk we survey some results and report about on-going work on the VC-dimension of these networks.

(Joint work with Wolfgang Maass.)

Dimension of query spaces and noise-robust learning.

Eli Shamir
Hebrew University, Jerusalem

ABSTRACT Kearns showed how statistical queries (SQs) lead to algorithms which are robust under classification noise. Aslam and Decatur added insight and several technical results, in particular about boosting weak learning. Still,in the application of these ideas to some concrete algorithms, e.g., Jackson's harmonic sieve, some difficulties have to be resolved:

(1) Apparently, an extended notion of SQ is needed, allowing expectations of more general real-valued functionals.

(2) For membership-queries-based algorithms, the SQs should be expressed (and empirically evaluated) as expectations with respect to (even varying) distributions which essentially differ from the underlying learning distribution.

Ultimately, the uniform convergence of the empirical evaluation of the "noise-compensated" SQs under the suitable sampling depend on:

(A) the small combinatorial dimensions of the full query space and the target family.

(B) Chernoff tail estimates.

(Joint work with Clara Shwarzmann.)

The role of the fat shattering dimension in concept learning

John Shawe-Taylor
Royal Holloway College, London

ABSTRACT The talk considers the problem of using a real valued function class to do concept learning via a fixed thresholding. If the performance of the classifier on the training set has a large margin, it is shown that the critical dimension for gauging the generalization performance is not the pseudo-dimension, but the fat shattering dimension at about the scale determined by the margin. This result is not only in agreement with the impressive performance of Vapnik's maximal margin hyperplane algorithm, but also indicates how to improve the performance of other classifiers which rely on an underlying class with suitably bounded fat shattering dimension.

Learning in the presence of malicious noise beyond the information-theoretic barrier

Hans Ulrich Simon
Universität Dortmund

ABSTRACT It is known that pac-learning in the presence of malicious noise is not possible for any noise rate equal to or greater than $\varepsilon/(1+\varepsilon)$. This results holds for any algorithm which outputs a function as hypothesis, even if it has an infinite number of labelled examples and infinite computational power. In this sense, quantity $\varepsilon/(1+\varepsilon)$ forms an information-theoretic barrier for learning in the presence of malicious noise. We show that this barrier can be broken, if we allow the algorithm to represent its hypothesis by a coin-rule. A master algorithm is presented that works for any concept class. For noise-rate $\varepsilon/(1+\varepsilon)$, the algorithm needs only $\widetilde{O}(d/\varepsilon)$ labelled examples, i.e., not significantly more examples than in the noise-free case. The result presented is part of joint work with Nicolò Cesa-Bianchi, Eli Shamir, and Paul Fischer.

Vapnik-Cervonenkis techniques for estimating the maxima of certain stochastic processes

David Steinsaltz
Technische Universität Berlin

ABSTRACT Consider the following self-service queueing problem: During the course of a day n (a large number) people visit the public swimming pool. Each person may come and go many times, but the arrivals and departures of different people are assumed to be independent (though not necessarily identically distributed). Let $X_n(t)$ be the number of people in the pool at time t. We wish to estimate $\sup_t X_n(t)$, the maximum occupancy over the course of the day.

By simple convexity arguments this problem may be compared to the situation of an empirical process with Vapnik-Cervonenkis dimension 1. By an adaptation of the traditional methods, as developed by Pollard, Devroye, et al., we then derive explicit bounds on the fluctuations of the process from its expectation, and in particular on the distribution of the maximum. We also derive limit laws for the process and for its maximum. Under fairly general conditions, including that the expectation of $X_n(t)$ have a unique quadratic maximum at a point t_0, we find that as $n \to \infty$,

$$n^{-1/2} \left(X_n(t) - E[X_n(t)] \right) \tag{1}$$

converges to a Gaussian process as $n \to \infty$, that

$$n^{-1/2} \left(\sup_t X_n(t) - \sup_t E[X_n(t)] \right) \tag{2}$$

converges to a normal variable, and that

$$n^{-1/3} \left(\sup_t X_n(t) - X_n(t_0) \right) \tag{3}$$

converges in distribution to $\sup_t B_t - at^2$, where B_t is a Brownian motion and a is a positive real number.

The same methods apply to a range of other problems, such as the following: Given n boxes and an unlimited number of balls, we sequentially distribute balls randomly to the boxes. Let $X_n(t)$ be the number of boxes containing exactly two balls, after $t \cdot n$ balls have been allocated. We consider then $\sup_t X_n(t)$. The same statements (1), (2), and (3) also apply in this case. Other examples involve sorting, random walks on graphs, and storage problems.