UNSUPERVISED LEARNING

21.-26. März 1999

organized by

Joachim M. Buhmann, Wolfgang Maass,
Helge Ritter, Naftali Tishby

What is unsupervised learning and how does it relate to the well founded theory of supervised learning? These questions have been discussed during this seminar which brought together neural modellers, statisticians, computational learning theorists ("COLT people") and theoretical computer scientists and physicists. The field of machine learning with its broad range of pattern recognition applications in data mining and knowledge discovery, in information retrieval and in classical areas like speech and image processing, computational linguistics or robotics is confronted with various problems beyond classification and regression. The search for structure in large data sets requires automatic inference tools which can also provide quality guarantees to validate the results.
The discussions after the talks and in special discussion sessions circled around two main issues of unsupervised learning:

1. What does it mean to detect structure in a data set and how can we quantify it?

2. How can we provide guarantees that the detected structure generalizes from one sample set to a second one?

It is unrealistic to expect a general answer to the first question. A general theory of structure has not been developed yet and attempt like the inference on the basis of Kolmogorov complexity are debated. One might even argue that such a goal is completely elusive since it encompass the program of natural science, engineering and the humanities. The different talks, therefore, covered a wide spectrum of special suggestions how structure could be defined and detected ranging from trees in image analysis, informative projections like PCA or ICA representations of high dimensional data, clusters in vectorial data and histograms as well as groups in relational data or principal surfaces.
It became apparent in the discussion of simulation results that fluctuations in the data should have little influence on the learned structures. This requirement might be enforced by bounding techniques as they have been developed for the computational learning theory of supervised learning or by information theoretic compression ideas. The challenges of unsupervised learning for the COLT and

the modeling community seem to crystallize around the questions how optimally generalizing structures in data can be discovered and how they are characterized and validated in terms of robustness, compactness (description length) and efficiency for learning.
What does biology teach us about unsupervised learning? Apart from the miracle how supervised learning might be organized in the brain at the neuronal level, the biological substrate seems to support unsupervised learning and related modeling ideas (or is at least compatible with them) by a potentially large computational power in synapses. Furthermore, spike trains might not only serve as a robust communication protocol but possibly provides probabilistic inference with an appropriate data format.

Joachim M. Buhmann

Contents

1 Uniform convergence and computational complexity analysis of a simplistic density estimation task

Shai Ben-David
Technion, Haifa, Israel

We investigate a model for a sub-task of density estimation. The model, presented originally in a paper with Lindenbaum [1], applies to this task a paradigm that is borrowed from (supervised) computational learning theory. The learner fixes a collection of domain subsets (the 'hypothesis class'), and then, upon receiving the (unlabeled) data, searches for a member in this class that exhibits high sample density (w.r.t. some fixed reference distribution over the domain).
This paradigm, while being rather simplistic, allows a complete analysis of the uniform convergence rate of learning. Assuming the the data-generating distribution belongs to some family of 'nice' distributions, we provide an upper bound for the sample complexity of learning in this model in terms of the VC-dimension of the hypothesis class, this bound is independent of the data-generating distribution, and is complemented by a matching lower bound.
The next step in this work is an investigation of the computational complexity of the learning task. Here our results are of pessimistic nature; For several basic hypothesis classes, we prove the NP-hardness of the task of finding approximations to the densest hypothesis in the class.

References

[1] Ben-David, S., and Lindenbaum, M., "Learning Distributions by their Density Levels – A paradigm for Learning Without a Teacher". *Journal of Computer and Systems Sciences*, Volume 55, No. 1, 1997, pp 171-182.

2 Generalized Clustering Criteria for Kohonen Maps

Hans-Hermann Bock
Technical University of Aachen, Germany

Kohonen's self-organizing map (SOM) visualizes the structure of data points $x_1, x_2, ... \in R^p$ by (a) implicitly constructing m 'mini-classes' $C_1, ..., C_m$ of data points, (b) representing each class C_i by a center $z_i \in R^p$, (c) assigning each center z_i to one of the m vertices P_i of an s-dimensional lattice $\mathcal{G}$ such that (d) 'similar' centers (Euclidean metric in R^p) are assigned to 'neighbouring' vertices in $\mathcal{G}$ (path distance δ in $\mathcal{G}$). This is performed (e) by recursively updating the centers $z_j^{(n)}$ after observing x_{n+1}.
The paper embeds Kohonen's approach into the classical clustering framework, proposes a generalized, probabilistically motivated 'model-based SOM' (MSOM) where mini-classes are characterized by class-specific *model parameters* z_i (e.g., covariance matrices, hyperplanes, regression or interaction coefficients, not only points in R^p), and discusses the asymptotic convergence of the model estimates to an optimum configuration for an increasing sample size.
All results are based on the interplay between the *discrete* clustering criterion: $g_n(\mathcal{C}, \mathcal{Z}) := \sum_{i=1}^m \sum_{k \in C_i} \phi_i(x_k; \mathcal{Z}) \rightarrow \min_{\mathcal{C},\mathcal{Z}}$ (for a fixed sample size n) and its asymptotic *continuous* counterpart: $g(\mathcal{B}, \mathcal{Z}) := \sum_{i=1}^m \int_{B_i} \phi_i(x; \mathcal{Z}) f(x) dx \rightarrow \min_{\mathcal{B},\mathcal{Z}}$ where $\phi_i(x; \mathcal{Z}) := -\sum_{j=1}^m K(\delta(P_i, P_j)) \log f(x; z_j)$ is a generalized distance of x to the i-th class which depends smoothly on the system $\mathcal{Z} = (z_1, ..., z_m)$ of class parameters, and minimization is over all systems $\mathcal{Z}$ and all m-partitions $\mathcal{C} = (C_1, ..., C_m)$ of $\{x_1, ..., x_n\}$ and all m-partitions $\mathcal{B} = (B_1, ..., B_m)$ of R^p, respectively. The density $f(x; z_i)$ is chosen class-specific, whereas $f(x)$ is the underlying distribution of the i.i.d. data points x_k.

Optimization is performed either by a k-means algorithm (which alternates between a modified maximum-likelihood estimate of the z_i and the construction of minimum-ϕ_i-distance partitions), a sequential MacQueen-type algorithm, or by stochastic approximation. The latter approach incorporates the data points in sequential order and uses the updating formula

$$z_j^{(n+1)} = z_j^{(n)} - \alpha_n \nabla_{z_j} \phi_{i^*}(x_{n+1}; \mathcal{Z}^{(n)}) = z_j^{(n)} - \alpha_n K(\delta(P_{i^*}, P_j)) \nabla_z \log f(x_{n+1}; z)_{|z=z_j^{(n)}}$$

where i^* is the class with minimum distance $\phi_i(x_{n+1}; \mathcal{Z}^{(n)})$. Typical examples include ellipsoidal clusters, class-specific hyperplanes (principal component clustering, PCA-SOMs), regression hyperplanes (two types of regression SOMs), log-linear models (for qualitative data, entropy SOMs) etc.

3 Empirical Risk Approximation: An Induction Principle for Unsupervised Learning

Joachim M. Buhmann
Universität Bonn, Germany

Unsupervised learning algorithms are designed to extract structure from data samples. The quality of a structure is measured by a cost function which is usually minimized to infer optimal parameters characterizing the hidden structure in the data. Reliable and robust inference requires a guarantee that extracted structures are typical for the data source, i.e., similar structures have to be extracted from a second sample set of the same data source. Lack of robustness is known as overfitting from the statistics and the machine learning literature. In this talk I characterize the overfitting phenomenon for a class of histogram clustering models which play a prominent role in information retrieval, linguistic and computer vision applications. Learning algorithms with robustness to sample fluctuations are derived from large deviation results and the maximum entropy principle for the learning process. The hypothesis class for histogram clustering contains function which depend on probabilities of features given a cluster and on assignments of data to clusters. This hypothesis class is covered by an γ-net which measures the optimally achievable approximation precision ϵ for learning. Bernstein's large deviation inequality quantifies the trade-off between the size of the γ-balls and the precision ϵ. The analysis yields a control parameter for the ball size which is known as temperature in statistical physics. The theory, therefore, validates continuation methods like simulated or deterministic annealing as optimally robust approximation schemes from a statistical learning theory point of view. The large deviation bound can be used as a conservative estimate of the number of clusters, i.e., a too large number of clusters exhibits overfitting phenomena in form of increasing deviations of empirical from expected risk.

References

[1] Technical Report IAI-TR 98-3 (1998), Informatik III, Universität Bonn, `http://www-dbv.informatik.uni-bonn.de/abstracts/buhmann.TR98.html`

4 On some properties of infinite VC dimension systems

Alexey Chervonenkis
ICS RAS, Moscow, Russia

The case, when (S is a system of sets $A \subset X$, given $P(x)$)

$$H^S(l)/l = \frac{\mathbf{E} \log \Delta^S(x_1, \ldots, x_l)}{l} \longrightarrow C > 0 \quad \text{for} \quad l \to \infty$$

is considered. It is shown that in this case there exists a set $T \subset X, P(T) = 1$ and

$$\Delta^S(x_{i_1}, \ldots, x_{i_K}) = 2^K$$

almost for certain $(x_{i_1}, \ldots, x_{i_K}) \in T$.

5 A new approach for pairwise clustering

Ran El-Yaniv
Technion, Haifa, Israel

We present a novel pairwise clustering method. Given a proximity matrix of pairwise distances (i.e. pairwise similarity or dissimilarity estimates) between data points, our algorithm extracts two most prominent components in the data set. The algorithm, which is completely nonparametric, iteratively employs a two-step transformation on the proximity matrix. The first step represents each point by its relation to *all* other data points and then the second step re-estimates the pairwise distances using a statistically motivated proximity measure. Using this transformation, the algorithm progressively partition the data points, until it finally converges to two components. Although the algorithm is simple and intuitive, it generates a complex dynamics on the proximity matrices. Based on this bipartition procedure we devise an hierarchical clustering algorithm, which employs the basic bipartition algorithm in a straightforward divisive manner. The hierarchical clustering algorithm copes with the model validation problem using a simple and general cross-validation approach. This cross-validation method can be combined with various hierarchical clustering methods.

6 Simultaneous Clustering and Dimensionality Reduction

Zoubin Ghahramani
University College London, Great Britain

I presented a model that combines clustering and dimensionality reduction in a Bayesian framework. The clustering model used was the mixture of Gaussians model, which contains as a special case k-means vector-quantization. The dimensionality reduction model used was factor analysis, which contains as a special case principal components analysis. By combining these two models into a mixture of factor analysers (MFA) it is possible to cluster data using a metric obtained from a local dimensionality reduction, and conversely, to infer the local dimensionilty reduction from the clustering. This model can also be thought of as a reduced parameter method of fitting a mixture of Gaussians in high dimensions. The model can be generalised to deal with time series, resulting in a switching state-space model. However, it suffers from two important problems.

The first problem is local optima in the likelihood. This often occurs when there are too many Gaussians dedicated to modeling data in one part of space, and two few in another part; the Gaussians cannot move from one part of space to another without going through lower-likelihood regions. I presented a Split and Merge modification of the EM algorithm (SMEM) which capitalises on the fact that the expected log likelihood can written as a direct sum over the clusters. The results show that it can very reliably escape local optima that trap the standard EM algorithm. Results also show that local optima can have a severe impact on performance on real problems such as digit classification and image compression, which SMEM can overcome.

The second problem is model selection: how to select the optimal number of clusters and the optimal dimensionality of each factor analyser. We approach this problem by formulating the model in a Bayesian framework and integrating over the parameters of the model. In this way, the evidence, $P(\text{data}|\text{model})$, for models with more degrees of freedom is penalised since more complex models can model a larger class of data. We use a variational approximation to obtain a lower bound on the (intractable) log evidence. This variational Bayesian approach can be combined with the Split and Merge approach since the lower bound on the log evidence can also be written as a direct sum. Work done is collaboration with Naonori Ueda, Matthew Beal and Geoffrey Hinton.

7 Using spike-timing for high-dimensional representations

Geoffrey Hinton
University College London, U.K.

To represent the pose of an object, neurons in the brain need to represent a point in a 6-D space. To combine evidence from noisy observations, neurons need to represent a probability distribution over the 6-D space. The distribution can be sharply peaked at any point in the space, so mixture models are infeasible.

Instead of averaging the density represented by each neuron, the brain may use a semantics in which the individual densities are multiplied and neural activities are exponents on the individual densities. (This is just averaging in the log probability domain). With this semantics each neuron can represent a very broad density and a population of active neurons can represent a very sharp density.

This still leaves the problem of how spiking neurons can represent continously varying exponents on the individual densitites. The relatively slow and smooth time course of EPSP"s (excitatory post-synaptic potentials) suggests that we can view each spike as a representation of a time-varying analog value of the exponent. So by controlling the precise times of spikes it is possible to achieve analog control over the density functions they represent.

8 Unsupervised Learning from Text by Probabilistic Latent Semantic Analysis

Thomas Hofmann
University of California Berkeley, USA

My talk introduces a principled statistical framework for learning from text and language based on probabilistic latent class models. The main focus is on a novel technique for factor analysis of two-mode data, called Probabilistic Latent Semantic Analysis, and its applications in information retrieval and statistical language modeling. Compared to standard Latent Semantic Analysis, which

stems from linear algebra and is based on a Singular Value Decomposition of the term-document matrix, the proposed novel technique makes use of a mixture decomposition which results in a well-founded problem formulation in terms of maximum likelihood estimation. Probabilistic Latent Semantic Analysis is able to distinguish between different types of word usage and automatically detects topics along with their characteristic words. Experiments in automated indexing and in language modeling indicate substantial performance gains over standard Latent Semantic Analysis. The talk also addresses the relationship to document/word clustering approaches and discuss how these can be uniformly modeled in the latent class framework.

9 Stochastic neural networks

H.J. Kappen
University of Nijmegen, The Netherlands

Networks of stochastic binary neurons are a (very) abstract model for information processing in the cortex. Under rather mild conditions, the stochastic dynamics converges to a time independent probability distribution. If the connectivity in the network is symmetric, this stationary distribution is the Boltzmann-Gibbs distribution.
The computation of statistics from the equilibrium distribution, such as mean firing rates or correlations, is generally intractable, requiring an exponentially large sum over all configurations (states) of the network. Mean field theory, is an efficient approximation method to compute these statistics.
For symmetric networks, the mean field theory can be derived as a Taylor series expansion of the free energy associated with the Boltzmann distribution. In the asymmetric case, however, no free energy exists and the derivation of mean field theory is less known. In this presentation, we recast the mean field approximation in an information geometric framework. This interpretation shows immediately how to derive mean field theory for non-symmetric stochastic networks, or in fact for any other class of probability distributions.
In addition, we compute the radius of convergence of the Taylor series, which provides a criterion to assess the validity of the mean field approximation. We illustrate the method for Boltzmann Machine learning.

10 Learning the Parts of Objects with Nonnegative Matrix Factorization

Daniel D. Lee
Bell-Laboratories, Murray Hills, USA

An unsupervised learning algorithm that we call nonnegative matrix factorization (NMF) is able to learn the parts of objects. This is in contrast to other algorithms like principal components analysis (PCA) and vector quantization (VQ) which learn holistic, not parts-based, representations. When all three algorithms are viewed as techniques for matrix factorization, NMF is distinguished from PCA and VQ by its use of nonnegativity constraints. These constraints lead to a parts-based representation because they allow only additive, not subtractive, combinations. We demonstrate the NMF algorithm learning the parts of human face images, as well as semantic features of text.

11 ICA Mixture Models for Unsupervised Classification of Non-Gaussian Sources and Automatic Context Switching in Blind Signal Separation

TeWon Lee
Salk Institute, San Diego, USA

An unsupervised classification algorithm is derived from an ICA mixture model assuming that the observed data can be categorized into several mutually exclusive data classes. The components are generated by linear mixtures of independent non-Gaussian sources. The algorithm finds the independent sources, the mixing matrix for each class and also computes the class membership probability for each data point. The new algorithm improves classification accuracy compared with Gaussian mixture models. When applied to blind source separation in nonstationary environments, the method can switch automatically between learned mixing matrices. The algorithm can learn efficient codes to represent images containing both natural scenes and text. This method shows promise for modeling structure in high-dimensional data and has many potential applications.

12 Open Problems Regarding Unsupervised Learning in Neural Networks with Dynamic Synapses

Wolfgang Maass
Technical University of Graz, Austria

We consider mathematical models for synapses ("dynamic synapses") that incorporate recent experimental data about the intrinsic temporal dynamics of biological synapses.
In joint work with Eduardo Sontag we have established a complete characterization of all filters that can be approximated by feedforward neural networks with standard sigmoidal neurons in combination with such dynamic synapses: These are exactly those filters that can be approximated by Volterra series.
We also report recent results of computer simulations of such networks, where we have carried out experiments with simple learning rules for the "hidden parameters" that determine the dynamic behaviour of a dynamic synapse (joint work with Thomas Natschlaeger and Tony Zador). We have shown that these networks can learn to predict quite complex sequences, and that they can also learn to act like RBF-networks in the temporal domain; i.e. they can learn to recognize frequently occurring temporal patterns.
Further details are available from `http://www.cis.tu-graz.ac.at/igi/maass/`

13 Distortion Bounds for Vector Quantizers with Finite Codebook Size

Ron Meir
Technion, Haifa, Israel

Upper and lower bounds are presented for the distortion of the optimal N-point vector-quantizer applied to k-dimensional signals. Under certain smoothness conditions on the source distribution, the bounds are shown to hold for each and every value of N, the code-book size. These results extend bounds derived in the high resolution limit, which assumes that the number of code vectors is arbitrarily large. Two approaches to the upper bound are presented. The first, constructive construction, achieves the correct asymptotic rate of convergence as well as the correct dependence on the source density, although leading to an inferior value

for the constant. The second construction, based on a random coding argument, is shown to additionally achieve a value of the constant which is much closer to the best known result derived within the asymptotic theory. Lower bound results derived in the paper are again shown to possess the correct asymptotic form and yield a constant which is almost indistinguishable from the best value achieved in the asymptotic regime. Finally, application of the results to the problem of source coding, yields upper bounds on the distortion rate function for a wide class of processes.

References

[1] Ron Meir and Vitaly Maiorov, Distortion Bounds for Vector Quantizers with Finite Codebook Size, IEEE Trans. Inf. Theory, July 1999

14 Blind Source Separation

Klaus-Robert Müller
GMD FIRST.IDA, Berlin, Germany

A blind source separation algorithm using only second order correlations in time is proposed and applied to denoise SQUID measurements of evoked responses in the peripheral nervous system. Artifacts in magnetoneurography (MNG) data due to endogenous biological noise sources, e.g. heart signal can be four orders of magnitude higher than the signal of interest. Nevertheless ICA successfully separates artifacts from signal components and projecting out the artifact signals yields a significant improvement of the neuro-magnetic source analysis. Assumptions and priors and their implications are discussed. (in collaboration with G.Nolte, G. Curio, B.-M. Mackert (FU Berlin))
Joint work with A. Ziehe and P. Philips.

15 Topographic Clustering Methods

Klaus Obermayer
Technical University of Berlin, Germany

Topographic clustering algorithms and Kohonen maps are well known tools for statistical data analysis. Starting from the concept of a stochastic autoencoder, I derived a general topographic clustering method for pairwise data which is based on maximum entropy inference. This ansatz allows the grouping of data objects characterized by proximity matrices as well as a non-linear embedding of the data in a low-dimensional space for the purpose of visualization. If proximity is measured via an Euclidean distance measure between feature vectors our ansatz reduces to an annealed variant of Kohonen's self-organizing map (SOM). In doing this we obtain a robust learning rule for SOM which makes the conventional annealing in the width of the neighborhood function superflous. Results from a theoretical analysis are complemented by examples from image compression and from the visualization of proximity data. Finally, I showed that pairwise topographic clustering methods perform well even if the proximity matrix is sparse and more than 80 percent of the entries are missing. In using active strategies for data selection, based for example on the expected reduction in the clustering cost, this percentage can be made even higher.

16 Independent Component Analysis by Unsupervised Learning

Erkki Oja
Helsinki University of Technology, Finland

The recent work in the author's research group on using Independent Component Analysis (ICA) for various signal decomposition tasks is reviewed. ICA belongs to the group of linear digital signal transform methods. The goal is rather unique: to make a transform into a signal space in which the signals are statistically independent. Sometimes independence can be attained, especially in blind source separation in which the original signals are linear mixtures of independent source components and the goal of ICA is to invert the unknown mixing operation. Even when independence is not possible, the ICA transformation produces useful com-

ponent signals that are nongaussian, similar to the Projection Pursuit technique, or whose density allows sparse coding. The ICA transformation is also related to the structure of the found signals as measured by Kolmogorov complexity or its approximations.
After discussing the ICA criterion and its relations to other linear signal transforms, the FastICA algorithm is reviewed. It is a computationally efficient method for finding a subset or all of the component signals. Then, an application is covered: the decomposition of digital images into sparse codes by which they can be compressed and their noise can be attenuated.

17 Retarded Learning in high dimensional spaces — a toy model

Manfred Opper
Aston University, Birmingham, U.K.

I discuss a simple toy model of unsupervised learning which is formulated in the framework of density estimation. I assume that data are generated with equal weight from two separared Gaussians in a D-dimensional space. The structure which has to be estimated is the direction which connects the two centers. One might expect that learning of this direction progresses gradually as more and more examples are observed. This is however no longer true when the dimension D becomes very large and the distance between centers is properly scaled with D such that the problem remains nontrivial. Within a Bayesian scenario, I give a simple lower bound on the KL loss. I find that when the number of examples/D is below a critical value, the loss of the Bayes optimal estimator converges to that of a trivial estimator which predicts an unstructered distribution.

18 Clustering and Low-Dimensional Representation of Large Data-Sets

Günther Palm
University of Ulm, Germany

A new algorithm is presented that combines k-means clustering (albeit with a variable number of cluster-centers) with an on-line non-linear low-dimensional projection of the cluster centers. The algorithm is derived from a cost function that combines the usual cost function for k-means (quantization error) with the stress functional used for multidimensional scaling or **Sammon mapping** (preservation of distances). The algorithm is discussed in terms of some desirable properties of data exploration algorithms in general, namely simplicity, few parameters, intuitive display, incrementality, interactivity, convergence, stability and robustness.

19 Complexity control in local PCA representations

Helge Ritter
Bielefeld University, Germany

Local PCA models offer an attractive approach to represent many types of data distributions. With local PCA's, the overall model complexity is composed of two parts which can be geometrically characterized: a "horizontal" complexity, measured by the number of centers at which local PCAs are placed, and a "vertical" complexity measured by the number of PCA dimensions used at each center. We present an approach in which both types of complexity are controlled by a single, global noise level parameter. This yields two advantages over previous approaches: (*i*) we can use the noise level for a (two phase) deterministic annealing scheme to obtain good maximum likelihood parameter values for centers and subspace directions, and (*ii*) the method can flexibly assign different subspace dimensionalities to each center, if the geometric structure of the data set requires to do so. We discuss the approach in the statistical framework of EM-likelihood

maximization and demonstrate its performance on both artificial data sets as well as on a larger real-world data set from a digit classification task.
Joint work with P. Meinicke.

20 Feature Extraction with Neural Pyramids

Raul Rojas, Sven Behnke
University of Berlin, Germany

In this talk, we motivate the construction of an unsupervised feature extraction system by showing first how two supervised classifiers were built. The task to be solved is the recognition of ZIP codes contained in video images of large format letters. The first classifier reconstructs the digit's strokes and recognizes it from an attributed structural graph representation. The second classifier is based on the Time Delay Neural Network architecture. The two classifiers were combined and incorporated in the postal sorting machines built by Siemens. Up to now, 150 machines have been installed in Germany.
The Neural Abstraction Pyramid is an architecture for iterative image interpretation that has been inspired by the information processing principles of the visual cortex. We present an unsupervised learning algorithm for the design of its feed-forward connectivity that is based on Hebbian weight updates and competition. It yields a hierarchy of feature detectors that produce a sequence of representations of the image content that become increasingly abstract. These representations are distributed and sparse and facilitate the interpretation of the image.
We apply the algorithm to a dataset of handwritten digits, starting from local contrast detectors. The emerging representations correspond to step edges, lines, strokes, curves, and typical digit shapes.
Joint work with Marcus Pfister, Siemens AG.

21 Optimal on–line principal component analysis

David Saad
Aston University, Birmingham, U.K.

Various techniques, used to optimise on-line principal component analysis, are investigated by methods of statistical mechanics. These include local and global optimisation of node-dependent learning-rates which are shown to be very efficient in speeding up the learning process. They are investigated further for gaining insight into the learning rates' time-dependence, which is then employed for devising simple practical methods to improve training performance. Simulations demonstrate the benefit gained from using the new methods.

References

[1] E. Schlösser, D. Saad, and M. Biehl, Optimal on–line principal component analysis, AND SOME INFOs

22 Single-Class Support Vector Machines

Bernhard Schölkopf
GMD FIRST, Berlin, Germany

Suppose you are given some dataset drawn from an underlying probability distribution P and you want to estimate a subset S of input space such that the probability that a test point drawn from P lies outside of S is bounded by some a priori specified $0 < \nu \leq 1$.
We propose an algorithm to deal with this problem by trying to estimate a function f which is positive on S and negative on the complement of S. The functional form of f is given by a kernel expansion in terms of a potentially small subset of the training data; it is regularized by controlling the length of the weight vector in an associated feature space.
We can prove that ν upper bounds the fraction of outliers (training points outside of S) and lower bounds the fraction of support vectors. Asymptotically, under some mild condition on P, both become equalities.
The algorithm is a natural extension of the support vector algorithm to the case

of unlabelled data.
Joint work with R. Williamson, A. Smola and J. Shawe-Taylor.

23 A program for learning transformation from a corpus of pairs of words.

Eli Shamir
Hebrew University, Jerusalem, Israel

Given a corpus of pairs of words $A = (x, y)$ the task is to find a cost-minimal system T of context-sensitive rewrite rules:

$$s \to t/u\_v \text{ rewrite } t \text{ for } s \text{ in the context } usv$$

so that the whole of A [or a subset covering all the x's] can be obtained by a decomposition into T-legal segments:

$$(x, y) = (s(1), t(1)) * (s(2), t(2)) * ... * (s(k), t(k))$$

Learning a transformation T is actually learning a structure in the corpus, a task which presents itself in many linguistic applications. Other areas such as DNA sequence alignments present similar problems. In linguistic, this learning task realizes Koskeniemi's two -level transformation, which is quite popular in various morphophonemic representations. Finding a cost–minimal transformation seems to be $\mathcal{NP}$–hard. Gil Broza, in a recent MSc thesis, developed an approximate solution. CANDID is an unsupervised - learning iterative program, based on the minimal description length principle, which generates an approximately minimal transformation. Similar to EM algorithms, it alternates between segmentation and maximization stages in order to approximate a cost-optimal solution. Several improvements in expanding the segmentation search space and reducing processing time are crucial for the successful performance of CANDID.
Some applications and experimental runs are described, e.g., for corpus of 'cognate' words in Spanish-French, and for learning the phonological value of 'th' in English.

24 Learning About the Distribution

John Shawe-Taylor
Royal Holloway, University of London, U.K.

The talk addressed the question of what can be ascertained about a distribution when frequentist guarantees of reliability are required. Three questions were discussed. 1. *Large Margin Supervised Learning*: Despite the fact that we are ostensibly doing supervised learning the fact that the margin is being estimated on the training data is shown to imply that we are in fact estimating some information about the distribution. This follows from the fact that lower bounds in terms of the VC dimension are contravened, something that can only happen if we know that 'difficult' distributions have been discounted. 2. *Semi-supervised Learning and Transduction*: The use of optimal margin classifiers over the full training and test sets was analysed and the generalization bound obtained compared with the supervised large margin result. The fact that these two bounds differ only very slightly was given as circumstantial evidence that this approach to transduction is failing to capture the full benefit of the unlabelled examples. 3. *An Unsupervised Learning Result*: A result to complement the algorithm given by Bernhard Schölkopf was presented, showing that frequentist bounds can be derived for estimating the support of a distribution which show that the probability of lying outside the estimated region can be bounded by a quotient of the fat shattering dimension of the class of estimators used and the number of examples seen. This improves on the bounds for estimating level sets given by Ben-David and Lindenbaum if their results are considered only for very low levels.

25 Regularized Principal Manifolds

Alex Smola
GMD FIRST, Berlin, Germany

Many settings of unsupervised learning can be viewed as quantization problems — the minimization of the expected quantization error subject to some restrictions. This allows the use of tools such as regularization from the theory of (supervised) risk minimization for unsupervised settings. This setting turns out to be closely related to principal curves, the generative topographic map, and robust coding. We explore this connection in two ways: 1) we propose an algorithm for finding

principal manifolds that can be regularized in a variety of ways. 2) We derive uniform convergence bounds and hence bounds on the learning rates of the algorithm. In particular, we give bounds on the covering numbers which allows us to obtain nearly optimal learning rates for certain types of regularization operators. Experimental results demonstrate the feasibility of the approach.

26 Monte Carlo Algorithms for State Estimation, Learning, and Decision Making in Dynamic Systems

Sebastian Thrun
Carnegie Mellon University, Pittsburg, USA

Embedded systems — systems that interact with the physical world — face the continuous challenge of having to make decisions in real-time under uncertainty. Sensors are often inadequate to capture the complexity of the physical world; thus, embedded systems may never know the state of the world. In addition, many problem domains are characterized by tight deadlines for decision making. The world never waits for the termination of a computer program.

This raises the issue as to how to represent and reason with uncertain information efficiently. Within the probabilistic framework, uncertainty is typically represented by probability distributions. Instead of representing distributions explicitly, however, this talk proposed to approximate them using samples drawn from these distributions. Sample sets have a range of properties that make them attractive for approximate representation of uncertain knowledge: They converge to a large class of non-parametric distributions; they can cope with continuous-valued spaces; they lend themselves nicely to any-time implementation that adapt to the computational resources; they allow for efficient belief revision and decision making; and they are relatively easy to implement.

This talk discussed the application of sampling-based representations to three fundamental problems in AI: state estimation, model learning, and decision making in dynamic systems. In the context of state estimation, our approach is a version of "particle filters" or the "condensation algorithm," which previously has been applied with great success by Isard and Blake to visual tracking problems. The talk showed results obtained in a second domain: mobile robot localization and position tracking, illustrating once more that sample-based methods are well suited for state estimation in dynamic systems. In the context of model learning,

the talk presented a Monte Carlo version of Hidden Markov Models, trained using a sample-based extension of the Baum-Welsh algorithm. To avoid overfitting, the approach uses "shrinkage trees," a method previously proposed by McCallum, together with annealing and early stopping based on cross-validation. The use of sample-based representations led to an extension of HMMs, capable of learning non-parametric models in continuous spaces. Convergence proofs were presented, along with empirical results obtained in a gesture recognition domain. Finally, the talk presented an application of sample-based density representations to POMDPs (partially observable Markov decision processes), where initial results suggested improved scaling to complex physical worlds when compared to previous, exact methods.

27 The Two Sample Probem: A Unified Information Theoretic Framework for Learning

Naftali Tishby
Hebrew University, Jerusalem, Israel

Supervised and unsupervised learning can both be cast into the question of inference structures that decouple two independent samples taken with the same distribution and/or labeling hypothesis. This question is directly related to classical statistical problem known as "the two sample problem", in which one needs to estimate the likelihood that two samples are taken from the same distribution. It is also related to the problem of sample compression: find a short summary of one sample that enables good prediction of the other sample.
We introduce a novel information theoretic formulation for this problem which stems from a natural variational principle: compress one variable while preserving the mutual information to another variable. Perhaps surprisingly, this problem has an exact general implicit solution which can be expressed in terms of three self-consistent equation for the conditional distributions of the observed variables on the hidden structure. Unlike most other statistical modeling techniques, there is no explicit assumption about the data and despite the similarity to statistical estimation, this method is not covered by standard hidden variable models. The resulting self-consistent equations provide a fast converging algorithm to a solution of the equations. These equations can be solved with increasing resolution by a procedure similar to deterministic annealing, yielding hierarchical structures that approximate the "sufficient statistics" of the underlying distributions. It provides a unified framework for different important problems, such as

prediction, clustering, and learning.
This work was jointly done with W. Bialek and F. Pereira.

28 Clustering in Graphical Models

Volker Tresp
Siemens AG, Munich, Germany

Structure and conditional probabilities of a Bayesian network uniquely specify the probability distribution of the domain. The locality of both structure and probabilistic information are the great benefits of Bayesian networks and require the modeler to only specify local information. On the other hand this locality might prevent the modeler —and even more any other person— to obtain an overview of the important relationships within the domain. The goal of the work presented in this paper is to provide an "alternative" view on the domain knowledge encoded in a Bayesian network which might sometimes be very helpful for providing insights into the domain. The basic idea is to calculate a mixture approximation to the probability distribution represented by the Bayesian network. The mixture component densities can be thought of as representing typical scenarios implied by the Bayesian model, providing intuition about the basic relationships. As an additional benefit, performing inference in the approximate model is very simple and intuitive and can provide additional insights. The computational complexity for the calculation of the mixture approximations critically depends on the measure which defines the distance between the probability distribution represented by the Bayesian network and the approximate distribution. Both the KL-divergence and the backward KL-divergence lead to intractable algorithms. Incidentally, the latter is used in recent work on mixtures of mean field solutions to which the work presented here is closely related. We show, however, that using a mean squared error cost function leads to update equations which can be solved using the junction tree algorithm. We conclude that the mean squared error cost function can be used for Bayesian networks in which inference based on the junction tree is tractable. For large networks, however, one may have to rely on the mean field approximations.
This work was jointly done with M. Haft and R. Hofmann.

29 Minimum Description Length Induction, Bayesianism, and Kolmogorov Complexity

Paul Vitányi
CWI and Univ. Amsterdam, The Netherlands

The relationship between the Bayesian approach and the minimum description length approach is established. We sharpen and clarify the general modeling principles MDL and MML, abstracted as the ideal MDL principle and defined from Bayes's rule by means of Kolmogorov complexity. The basic condition under which the ideal principle should be applied is encapsulated as the Fundamental Inequality, which in broad terms states that the principle is valid when the data are random, relative to every contemplated hypothesis and also these hypotheses are random relative to the (universal) prior. Basically, the ideal principle states that the prior probability associated with the hypothesis should be given by the algorithmic universal probability, and the sum of the log universal probability of the model plus the log of the probability of the data given the model should be minimized. If we restrict the model class to the finite sets then application of the ideal principle turns into Kolmogorov's minimal sufficient statistic. In general we show that data compression is almost always the best strategy, both in hypothesis identification and prediction.

30 A Stochastic Algorithm for Clustering Based on Pairwise Similarity

Michael Werman
Hebrew University, Jerusalem, Israel

We present a stochastic clustering algorithm which uses pairwise similarity of elements, based on a new graph theoretical algorithm for the sampling of cuts in graphs. The stochastic nature of our method makes it robust against noise, including accidental edges and small spurious clusters. We demonstrate the robustness and superiority of our method for image segmentation on a few synthetic examples where other recently proposed methods (such as normalized-cut) fail. In addition, the complexity of our method is lower.

31 Belief network models of images

Chris Williams
University of Edinburgh, U.K.

I will discuss some tree-structured models of images. First a balanced tree-structured belief-network model (as described in Bouman and Shapiro, 1994) is introduced. (Balanced refers to a regular architecture, such as a quad-tree in 2-d.) Compared to Markov Random Field models this has the attractive properties that it is hierarchically structured and allows efficient inference (by Pearl belief propagation). However, its rigid architecture gives rise to "blocky" artifacts. To try to overcome this problem a class of models called "dynamic trees" is introduced, where a prior is placed over tree structures. The idea is that the tree-structure will adapt to the particular input pattern. Inference is achieved on the basis of the posterior probability $P(Z|D) \propto P(Z)P(D|Z)$, where Z indexes the tree configuration and D the image data. I will discuss some experiments where we search over Z for tree interpretations of the data, and show that these give rise to interpretations which have higher posterior probability than balanced trees.

32 Covering Numbers

Robert Williamson

Covering numbers play a key role in statistical learning theory since many generalization error bounds are in terms of them. Until recently, for many classes of interest, it has only been possible to calculate upper bounds on covering numbers rather indirectly. In this talk I showed how by taking a different viewpoint, namely considering the classes of functions of interest to be the image of some unit ball under a compact operator, one can calculate covering numbers quite directly. The methods make use of the so-called entropy numbers of operators. Specifically, the way that covering numbers of the class of functions realised by support vector machines depends on the choice of kernel was explained. Some more recent developments, which bound the generalization performance of a learning machine in terms of the observed covering numbers (of the class restricted to the sample that has been seen) were also presented, and it was shown that such bounds seem to be good for model order selection.

The reported work was done (variously) jointly with Bernhard Schölkopf, Alex Smola, John Shawe-Taylor, Ying Guo and Peter Bartlett.