Statistical Learning
Apprentissage statistique
(Org: **Yoshua Bengio** (Montreal))

LOUBNA BENABBOU, Université Laval
Generalization bounds for multiclass classifiers

Classification methods based on statistical learning theory are mostly concerned with binary (two-class) classifiers. It is sometimes argued that multi-class (more than two classes) problems can be reduced to binary ones through sequential dichotomization. The drawbacks of such approaches are obvious. We seek here simultaneous solutions (classifier determinations) to a multiple classification problem with real loss function on classification error cases, reflecting the possibly variable gravity of misdiagnoses and/or a decision not to classify an object. We demonstrate a general *reduction principle* and show in particular that asymmetry in the loss function is necessary and sufficient for the multi-class problem not to be reducible to a bi-class one. We then propose two generalization bounds specifically designed for the multi-class setting. These bounds are numerical, and are tight by construction.

YOSHUA BENGIO, U. Montreal
On the Challenge of Learning Abstractions

We argue that learning to be intelligent involves the learning of highly varying functions, in a mathematical sense. We present results suggesting strongly that most currently popular statistical learning approaches to learning flexible functions have fundamental limitations that render them inappropriate for learning highly varying functions. The first issue concerns the representation of such functions with what we call shallow model architectures. We discuss limitations of shallow architectures, such as so-called kernel machines, boosting algorithms, decision trees, and one-hidden-layer artificial neural networks. Mathematical results in circuits complexity theory helps us understand the issue. The second issue is more focused and concerns kernel machines with a local (e.g. Gaussian) kernel.

We show that they have a limitation similar to those already proved for older non-parametric methods, and connected to the so-called curse of dimensionality. Though it has long been believed that efficient learning in deep architectures is difficult, recently proposed computational principles for learning in deep architectures may offer a breakthrough. An idea that emerges from the experiments performed with these algorithms is that in order to optimize a highly non-convex functions, humans and machines could be exploiting the pedagogical approach: learn simple concepts first, and when they are mastered use them to express and learn more abstract concepts. Selecting training examples and the order in which they are presented (just like teachers do with children) could be a way to guide this difficult optimization problem by solving a series of gradually more complex problems embedded in each other.

JULIE CARREAU, Université de Montréal
Hybrid Pareto models for asymmetric fat-tailed data

Density estimators that can adapt for asymmetric heavy tails are required in many applications such as finance and insurance. We put forward a non-parametric density estimator that brings together the strengths of non-parametric density estimation and of Extreme Value Theory. A hybrid Pareto distribution that can be used in a mixture model is proposed to extend the generalized Pareto (GP) to the whole real axis. Experiments on simulated data show the following. On one hand, the mixture of hybrid Paretos converges faster in terms of log-likelihood and provides good estimates of the tail of the distributions when compared with other density estimators including the GP distribution. On the other hand, the mixture of hybrid Paretos offers

an alternate way to estimate the tail index which is comparable to the one estimated with the standard GP methodology. The mixture of hybrids is also evaluated on the Danish fire insurance data set.

ALI GHODSI, University of Waterloo, 200 University Ave. W., Waterloo, ON, N2L 3G1
An SVD-based approach to nonnegative matrix factorization

Nonnegative matrix factorization (NMF) was introduced as a tool for data mining by Lee and Seung in 1999. NMF attempts to approximate a matrix with nonnegative entries by a product of two low-rank matrices, also with nonnegative entries. We propose an algorithm called R1D (rank-one downdate) for computing a NMF that is motivated by singular value decomposition. This computes the dominant singular values and vectors of adaptively determined submatrices of a matrix.

Preliminary computational tests indicate that this method is able to successfully identify features in realistic datasets.

Joint work with Stephen Vavasis and Michael Biggs.

NICOLAS LE ROUX, Université de Montréal, Québec, Canada
Representational Power of Restricted Boltzmann Machines and Deep Belief Networks

Deep Belief Networks (DBN) are generative neural network models with many layers of hidden causal variables, recently introduced by Hinton et al., along with a greedy layer-wise unsupervised learning algorithm. The building block of a DBN is a probabilistic model called a Restricted Boltzmann Machine (RBM), used to represent one layer of the model. Restricted Boltzmann Machines are interesting because inference is easy in them, and because they have been successfully used as building blocks for training deeper models.

We show that RBMs are universal approximators of discrete distributions. A first theorem shows that adding hidden units yields improved modeling power, while a second theorem shows that an RBM can model any discrete distribution.

We then study the question of whether DBNs with more layers are strictly more powerful in terms of representational power. This suggests another criterion for DBNs, obtained by considering that the top layer can perfectly fit its input.

RUSLAN SALAKHUTDINOV, University of Toronto
Non-linear Dimensionality Reduction using Neural Networks

Scientists, working with large amounts of high-dimensional data are constantly facing the problem of dimensionality reduction: how to discover low-dimensional structure from high-dimensional observations. The compact representation can be used for exploratory data analysis, preprocessing, data visualization, and information retrieval.

One way to discover low-dimensional structure is to convert high-dimensional data to low-dimensional codes by training a multi-layer neural network with a small central layer to reconstruct high-dimensional input vectors. Gradient descent can be used for fine-tuning the weights in such "autoencoder" networks, but this only works well if the initial weights are close to a good solution. In this talk we will describe an effective way of initializing the weights which allows deep autoencoder networks to learn low-dimensional codes that work much better than widely used Principal Components Analysis.

When trained on large document corpus, autoencoders are capable of extracting low-dimensional "semantic" codes that allow for much more accurate and faster retrieval than Latent Semantic Analysis, a well-known document retrieval method based on Singular Value Decomposition.

DANA WILKINSON, University of Waterloo
Learning Useful Subjective Representations

In a variety of domains it is desirable to learn a representation of an environment defined by a stream of sensori-motor experience. In many cases such a representation is necessary as the observational data is too plentiful to be stored in a computationally feasible way. In other words, the primary feature of a learned representation is that it must be compact, summarizing information in a way that alleviates storage and retrieval demands.

This admits a new way of phrasing the problem: as a variation of dimensionality reduction. There are a variety of well-studied algorithms for the dimensionality reduction problem. We argue that any of these can be useful for learning compact representations as long as additional constraints to the problem are respected, namely that the resulting representation is useful in the context of the actions which generated the observations.

Here, we formalize the problem of learning a subjective representation, clearly articulating solution features that are necessary for a learned representation to be "useful"; the actions must correspond to simple and consistent transformations in the learned representation. Further, we briefly present a possible solution to the newly defined problem and demonstrate its effectiveness for reasoning, planning and localization.