Doctoral dissertations

Mutual Dependency-Based Modeling of Relevance in Co-Occurrence Data

Eerika Savia

Dissertation for the degree of Doctor of Science in Technology on 18 June 2010.

External examiners:
Tapio Salakoski (University of Turku, Finland)
David R. Hardoon (University College London, UK)
Opponent:
Michal Rosen-Zvi (IBM Reasearch Lab, Haifa, Israel)

Abstract:
In the analysis of large data sets it is increasingly important to distinguish the relevant information from the irrelevant. This thesis outlines how to find what is relevant in so-called co-occurrence data, where there are two or more representations for each data sample.

The modeling task sets the limits to what we are interested in, and in its part defines the relevance. In this work, the problem of finding what is relevant in data is formalized via dependence, that is, the variation that is found in both (or all) co-occurring data sets was deemed to be more relevant than variation that is present in only one (or some) of the data sets. In other words, relevance is defined through dependencies between the data sets.

The method development contributions of this thesis are related to latent topic models and methods of dependency exploration. The dependency-seeking models were extended to nonparametric models, and computational algorithms were developed for the models. The methods are applicable to mutual dependency modeling and co-occurrence data in general, without restriction to the applications presented in the publications of this work. The application areas of the publications included modeling of user interest, relevance prediction of text based on eye movements, analysis of brain imaging with fMRI and modeling of gene regulation in bioinformatics. Additionally, frameworks for different application areas were suggested.

Until recently it has been a prevalent convention to assume the data to be normally distributed when modeling dependencies between different data sets. Here, a distribution-free nonparametric extension of Canonical Correlation Analysis (CCA) was suggested, together with a computationally more efficient semi-parametric variant. Furthermore, an alternative view to CCA was derived which allows a new kind of interpretation of the results and using CCA in feature selection that regards dependency as the criterion of relevance.

Traditionally, latent topic models are one-way clustering models, that is, one of the variables is clustered by the latent variable. We proposed a latent topic model that generalizes in two ways and showed that when only a small amount of data has been gathered, two-way generalization becomes necessary.

In the field of brain imaging, natural stimuli in fMRI studies imitate real-life situations and challenge the analysis methods used. A novel two-step framework was proposed for analyzing brain imaging measurements from fMRI. This framework seems promising for the analysis of brain signal data measured under natural stimulation, once such measurements are more widely available.

Advances in the Theory of Nearest Neighbor Distributions

Elia Liitiäinen

Dissertation for the degree of Doctor of Science in Technology on 22 October 2010.

External examiners:
Mathew D. Penrose (University of Bath, UK)
Dafydd Evans (Cardiff University, UK)
Opponents:
Luc Devroye (McGill University, Montreal, Canada)

Abstract:
A large part of non-parametric statistical techniques are in one way or another related to the geometric properties of random point sets. This connection is present both in the design of estimators and theoretical convergence studies. One such relation between geometry and probability occurs in the application of non-parametric techniques for computing information theoretic entropies: it has been shown that the moments of the nearest neighbor distance distributions for a set of independent identically distributed random variables are asymptotically characterized by the R´enyi entropies of the underlying probability density. As entropy estimation is a problem of major importance, this connection motivates an extensive study of nearest neighbor distances and distributions.

In this thesis, new results in the theory of nearest neighbor distributions are derived using both geometric and probabilistic proof techniques. The emphasis is on results that are useful for finite samples and not only in the asymptotic limit of an infinite sample.

Previously, in the literature it has been shown that after imposing sufficient regularity assumptions, the moments of the nearest neighbor distances can be approximated by invoking a Taylor series argument providing the connection to the R´enyi entropies. However, the theoretical results provide limited understanding to the nature of the error in the approximation. As a central result of the thesis, it is shown that if the random points take values in a compact set (e.g. according to the uniform distribution), then under sufficient regularity, a higher order moment expansion is possible. Asymptotically, the result completely characterizes the error for the original low order approximation.

Instead of striving for exact computation of the moments through a Taylor series expansion, in some cases inequalities are more useful. In the thesis, it is shown that concrete upper and lower bounds can be established under general assumptions. In fact, the upper bounds rely only on a geometric analysis.

The thesis also contains applications to two problems in nonparametric statistics, residual variance and R´enyi entropy estimation. A well-established nearest neighbor entropy estimator is analyzed and it is shown that by taking the boundary effect into account, estimation bias can be significantly reduced. Secondly, the convergence properties of a recent residual variance estimator are analyzed.

Probabilistic Analysis of the Human Transcriptome with Side Information

Leo Lahti

Dissertation for the degree of Doctor of Science in Technology on 17 December 2010.

External examiners:
Juho Rousu (University of Helsinki, Finland)
Simon Rogers (University of Glasgow, UK)
Opponent:
Volker Roth (Universität Basel, Switzerland)

Abstract:
Recent advances in high-throughput measurement technologies and efficient sharing of biomedical data through community databases have made it possible to investigate the complete collection of genetic material, the genome, which encodes the heritable genetic program of an organism. This has opened up new views to the study of living organisms with a profound impact on biological research.

Functional genomics is a subdiscipline of molecular biology that investigates the functional organization of genetic information. This thesis develops computational strategies to investigate a key functional layer of the genome, the transcriptome. The time- and context-specific transcriptional activity of the genes regulates the function of living cells through protein synthesis. Efficient computational techniques are needed in order to extract useful information from high-dimensional genomic observations that are associated with high levels of complex variation. Statistical learning and probabilistic models provide the theoretical framework for combining statistical evidence across multiple observations and the wealth of background information in genomic data repositories.

This thesis addresses three key challenges in transcriptome analysis. First, new preprocessing techniques that utilize side information in genomic sequence databases and microarray collections are developed to improve the accuracy of high-throughput microarray measurements. Second, a novel exploratory approach is proposed in order to construct a global view of cell-biological network activation patterns and functional relatedness between tissues across normal human body. Information in genomic interaction databases is used to derive constraints that help to focus the modeling in those parts of the data that are supported by known or potential interactions between the genes, and to scale up the analysis. The third contribution is to develop novel approaches to model dependency between co-occurring measurement sources. The methods are used to study cancer mechanisms and transcriptome evolution; integrative analysis of the human transcriptome and other layers of genomic information allows the identification of functional mechanisms and interactions that could not be detected based on the individual measurement sources. Open source implementations of the key methodological contributions have been released to facilitate their further adoption by the research community.

Algorithms for approximate Bayesian inference with applications to astronomical data analysis

Yoan Miche

Dissertation for the degree of Doctor of Science in Technology on 2 November 2010.

External examiners:
Thomas Villmann (Hochschule Mittweida, University of Applied Sciences, Germany)
Andrew Ker (University of Oxford, UK)
Opponent:
Tapio Seppänen (University of Oulu, Finland)

Abstract:
In the history of human communication, the concept and need for secrecy between the parties has always been present. One way of achieving it is to modify the message so that it is readable only by the receiver, as in cryptography for example. Hiding the message in an innocuous medium is another, called steganography. And the counterpart to steganography, that is, discovering whether a message is hidden in a specific medium, is called steganalysis. Other concerns also fall within the broad scope of the term steganalysis, such as estimating the message length for example (which is quantitative steganalysis).

In this dissertation, the emphasis is put on classical steganalysis of images first – the mere detection of a modified image –, for which a practical benchmark is proposed: the evaluation of a sufficient amount of samples to perform the steganalysis in a statistically significant manner, followed by feature selection for dimensionality reduction and interpretability. The fact that most of the features used in the classical steganalysis task have a physical meaning, regarding the image, lends itself to an introspection and analysis of the selected features for understanding the functioning and weaknesses of steganographic schemes.

This approach is computationally demanding, both because of the feature selection and the size of the data in steganalysis problems. To address this issue, a fast and efficient machine learning model is proposed, the Optimally-Pruned Extreme Learning Machine (OP-ELM). It uses random projections in the framework of an Artificial Neural Network (precisely, a Single Layer Feedforward Network) along with a neuron selection strategy, to obtain robustness regarding irrelevant features, and achieves state of the art performances.

The OP-ELM is also used in a novel approach at quantitative steganalysis (message length estimation). The re-embedding concept is proposed, which embeds a new known message in a suspicious image. By repeating this operation multiple times for varying sizes of the newly embedded message, it is possible to estimate the original message size used by the sender, along with a confidence interval on this value. An intrinsic property of the image, the inner difficulty, is also revealed thanks to the confidence interval width; this gives an important information about the reliability of the estimation on the original message size.

Methodologies for Time Series Prediction and Missing Value Imputation

Antti Sorjamaa

Dissertation for the degree of Doctor of Science in Technology on 19 November 2010.

External examiners:
Madalina Olteanu (Universtiy of Paris 1, France)
Vincent Wertz(Catholic University of Louvain, Belgium)
Opponent:
Guilherme Barreto (Federal University of Ceará, Brazil)

Abstract:
The amount of collected data is increasing all the time in the world. More sophisticated measuring instruments and increase in the computer processing power produce more and more data, which requires more capacity from the collection, transmission and storage. Even though computers are faster, large databases need also good and accurate methodologies for them to be useful in practice. Some techniques are not feasible to be applied to very large databases or are not able to provide the necessary accuracy.

As the title proclaims, this thesis focuses on two aspects encountered with databases, time series prediction and missing value imputation. The first one is a function approximation and regression problem, but can, in some cases, be formulated also as a classification task. Accurate prediction of future values is heavily dependent not only on a good model, which is well trained and validated, but also preprocessing, input variable selection or projection and output approximation strategy selection. The importance of all these choices made in the approximation process increases when the prediction horizon is extended further into the future.

The second focus area deals with missing values in a database. The missing values can be a nuisance, but can be also be a prohibiting factor in the use of certain methodologies and degrade the performance of others. Hence, missing value imputation is a very necessary part of the preprocessing of a database. This imputation has to be done carefully in order to retain the integrity of the database and not to insert any unwanted artifacts to aggravate the job of the final data analysis methodology. Furthermore, even though the accuracy is always the main requisite for a good methodology, computational time has to be considered alongside the precision.

In this thesis, a large variety of different strategies for output approximation and variable processing for time series prediction are presented. There is also a detailed presentation of new methodologies and tools for solving the problem of missing values. The strategies and methodologies are compared against the state-of-the-art ones and shown to be accurate and useful in practice.