

Teachers

- **Guest teacher:** Professor **Geoff McLachlan**, Department of Mathematics, University of Queensland, Australia.
- **Guest teacher:** Ass. Professor, **Mats Gustafsson**, Department of Signal and Systems, Uppsala University, Sweden
- **Guest teacher:** Professor **Jan Komorowski**, The Linnaeus Centre for Bioinformatics, Uppsala University, Sweden..
- Professor **Hans Liljenström**, *Department of Biometry and Informatics*, Swedish University of Agricultural Sciences.
- Professor **Dietrich von Rosen**, *Department of Biometry and Informatics*, Swedish University of Agricultural Sciences.

Language

The language of the course will be English

National contact persons

Denmark: Anders Ringgaard Kristensen, *Dina Research School*, KVL, Grønnegårdsvej 8, 1870 Frederiksberg C. E-mail: ark@dina.kvl.dk

Finland: Hannu Rita, *Department of Forest Resource Management*, P.O. Box 27 (Latokartanonkaari 7), FIN-00014 Helsingfors, E-mail: hannu.rita@helsinki.fi

Iceland: Torfi Jóhannesson, *Landbúnaðarháskólin á Hvanneyri*, IS-311 Borgarnes. E-mail: torfi@hvanneyri.is.

Norway: Knut Kvaal, *Department of Mathematical Sciences and Technology*, Agricultural University of Norway, P.O. Box 5003, N-1432 Ås. E-mail: Knut.Kvaal@imt.nlh.no.

Sweden: Ulf Olsson, *Department of Biometry and Informatics*, Swedish University of Agricultural Sciences, P.O. Box 7013, S-750 07 Uppsala. E-mail: Ulf.Olsson@bi.slu.se.

Contact information

Professor Dietrich von Rosen, *Department of Biometry and Informatics*, Swedish University of Agricultural Sciences, P.O. Box 7013, SE-750 07 Uppsala, Sweden.
E-mail: Dietrich.von.Rosen@bi.slu.se.

Course homepage:

<http://www.dina.dk/phd/s/s7/>

SLU, Faculty of Forestry, Umeå Campus, Summer school Venue

How to get there

The easiest way to get to the Faculty of Forestry is by taxi. The taxi cost (from the airport, the bus station, or the railway station) will be between 100-150 SEK, depending on how you travel to Umeå (by plane, bus, submarine or train). The site map of the faculty can be found on www.sfak.slu.se/. On the 6/6 we will provide some kind of Limousine service.



Centre of Biostochastics



The centre will host you during your stay. For more information about the centre please visit <http://biostochastics.slu.se/>. The course will use the facilities of the university

Umeå – The City of Birches

Umeå is situated in Norrland about 700km north of Stockholm. There are more than 100.000 inhabitants and 3000 birches. Umeå has two universities, Umeå University and SLU, and 25.000 students. In summer you will see the sun day and night. Nature is wonderful. You can find a huge amount of interesting animals such as reindeers, capercaillie, seals and mosquitoes. A mid night seal-safari is planned for course participants. To learn more about Umeå you could start by visiting www.umea.se.

Accommodation

Housing will mainly take place at Umeå camping <http://www.umedalen.ac/klo112/>. Umeå camping site and chalet village is located in beautiful natural surroundings by lake Nydalsjön on walking distance from SLU. It is possible to rent bikes. Participants will mostly share rooms (a number of single rooms are available on request) in cottages with cooking facilities, WC, shower and TV.

Pattern Recognition in High Dimensional Data and Complex Structures

Nordic PhD course

financed by
NOVA and NorFA



Swedish University of Agricultural Sciences

Umeå

June 6-18, 2004

Organized by
**Centre of Biostochastics,
Swedish University of Agricultural Sciences**

On behalf of
**Nordic Informatics Network in the
Agricultural Sciences**

Summer School, June 6-18, 2004

Pattern Recognition in High Dimensional Data and Complex Structures

Background

Nowadays equipment for measuring various processes within the agricultural field deliver a huge amount of output data. In a second step these data have to be analysed. One main object is to lower dimensions and identify patterns. The course will focus on the following four issues where the first two are of more theoretical character and the other consider some applications.

1. Pattern recognition and classical statistical tools

Here it will mainly be focussed on discriminatory/ identification analysis and cluster analysis. The main ideas of principal components analyses will also be considered. The object with discriminant analysis is to decide if an object either belongs to a specific class from a family of classes or if it does not belong at all to the family. It is either based on a distance (discriminant score) or on the likelihood. In cluster analysis there are no predefined classes as in discriminant analysis. Instead one tries to split the set of observations into subsets. We will focus on hierarchical clustering methods. Each observation constitutes a cluster by itself. The two in some sense closest clusters are merged to form a new cluster that replaces the two old clusters. Merging of closest clusters is repeated until only one cluster is left. Different clustering methods varies in how the distance between two clusters is computed. Principal components reduces the dimension in the data by constructing linear functions of the data. These functions can instead of the original be used as input in other analysis, for example discriminant analysis.

2. Pattern recognition and data mining

Choosing a model of appropriate complexity is important for drawing accurate conclusions. Simple models are used for learning simple functions of the data. Complex models are required for learning complex functions. For data mining models, one way to increase the complexity of a model is to add variables. Other ways to increase complexity depend on the type of model: In regression models, one can add interactions and polynomial terms. In neural networks, one can add hidden units. In tree-based models, one can grow a larger tree.

3. Pattern recognition and images

Feature extraction from signals will be discussed. In particular remote sensing data are of interest as well as image transformations. Furthermore, signal representation relevant to machine learning is highly relevant to consider.

4. Pattern recognition and microarrays

Classification and analysis of a huge number of cellular gene expression profiles measured by means of DNA microarrays will be considered. Useful criteria for performance evaluation and methods for estimating reliability are presented. Subset selection algorithms will be put into relation to the curse of dimensionality problem.

A common thread in the above is that we have high dimensional data. It is a challenge to among others statistics to handle these kind of data. Usually high dimensional data is characterized by many variables in relation to the number of independent experimental units. On top of this one may have complex or dynamic relations. Classical statistical asymptotic results do not apply. Multiple testing is often used but it is hard work to determine significance levels. At the same time new technologies arise where the purpose is to extract information from high dimensional data and which may be classified as pattern recognition methods. In these methods often some statistical ingredient exists, which is fairly natural because data is random, but conclusions are in general not based on a probabilistic reasoning. Instead simulations convinces applicants of the correctness of the analysis.

Aim of the course

The main objective is to introduce the participants into two worlds. One is the classical statistical frame and the other is computer science. In particular participants should become knowledge in the terminology used by the different disciplines. For example training and trained models often used in scientific computations with synonyms estimation and fitted model in statistics. Emphasis will be put on what kind of conclusions can be drawn from data. For example in the analysis of micro arrays often a lot of statistical tests are performed and it soon becomes clear that for this procedure the significance level can only be directing. During the course various techniques and approaches will extensively be illustrated on computers. Participants have to work a lot with computers in order to solve problems and to understand the basic principles.

After the participation in the course, the PhD students will be able to use various methods in order to deal with huge and

complex problems. Besides getting a glimpse into statistics they should be able to apply pattern recognition methods, to understand the output of these as well as to know about their limitations.

Required knowledge

Familiarity with computers at user level, and with basic probability calculus and related concepts.

Topics and Key Words

This summer school will focus on the following main topics:

- pattern recognition
- data mining
- discriminant analysis
- cluster analysis
- signal processing
- analysis of microarrays

The material will be illustrated with various examples and supplemented with guest lectures on related issues. Throughout the course, the theory will be supplemented with exercises and computer assignments. At the end of the course, the students will work on a two-day project that involves both modelling and computing aspects.

Teaching methods and examination

Lectures alternating with intensive use of computer exercises. The availability of network connected computers is therefore essential for the benefit of the students. A small project is carried out by the students at the end of the course (individually or preferably in small groups). Examination will be based on the written project report in combination with an oral presentation. The number of credits proposed is 6 ECTS.

Financial support

The course in general is financed by NOVA and NorFA. Thus travel costs and course fee as well as accommodation and meals for Ph.D. students are covered by these general grants.

Further information and registration

A preliminary programme and an online registration form are available at

<http://www.dina.dk/phd/s/s7/>