Probability Estimation in Prognostics

Making accurate binary classification, also termed pattern recognition or class prediction has a long successful history in statistics. Beyond a simple statement that a subject belongs to one group or another, more detailed information is the probability for belonging to one of the groups. Logistic regression is often used to solve this problem. But logistic regression assumes a fully and correctly specified parametric model. If the model is incorrect, substantial biases in probability estimates can occur. Machine learning approaches which are essentially nonparametric are an alternative and can yield consistent probability estimates. Such machines include versions of k-nearest neighbors, neural nets, random forest, support vector machines, and many others.

In this 1-day workshop, we provide a comprehensive overview on machine learning approaches to probability estimation for dichotomous variables.

When: Friday September 16, 2011
9.00 – 17.00
Where: Universität Zürich, Schönberggasse 11, Building SOE, Level F, Room F-02
Map at: http://www.plaene.uzh.ch/gebaeude/SOE_list.html#ankermap
Admission fee: Free admission for members of the International Biometric Society
150 € for non-members
Registration: needs to be made through https://www.conftool.pro/cen2011
You do not need to register for the conference of the Central European Network in order to participate in the Workshop.
Maximum number of participants:
48
Contact Information:
Univ.-Prof. Dr. Andreas Ziegler
University of Lübeck, Maria-Goeppert-Str. 1, 23562 Lübeck, Germany
Email: ziegler@imbs.uni-luebeck.de
Phone: +49 451 500-2780

Outline of the Workshop

09.00 – 10.30	
Michael Kohler	**On nonparametric regression with random design**
10.30 – 10.45	Health break
10.45 – 12.15	
Yufeng Liu	**The Support Vector Machine and its recent developments**
12.15 – 13.00	Lunch break
13.00 – 14.30	
Gérard Biau	**Random forests**
14.30 – 14.45	Health break
14.45 – 16.15	
James D. Malley	**Probability Machines**
16.15 – 16.30	Health break
16.30 – 17.00	
Jochen Kruppa	**Using R and Random Jungle for probability estimation**

Adjourn

Abstracts

On nonparametric regression with random design

Michael Kohler

Department of Mathematics, Technical University Darmstadt, Germany

Let Y be a univariate random variable with existing variance, and $\boldsymbol{X}$ be a d dimensional random vector of covariates. For a sample of size n, we consider the problem of estimating the regression function $m(\boldsymbol{x}) = \mathbb{E}(Y|\boldsymbol{X} = \boldsymbol{x})$. In this talk we give an overview on known results on this topic, in particular on the principle for constructing nonparametric estimates and on results concerning consistency and rate of convergence. As an application, we use nonparametric regression to define plug-in estimates for pattern recognition and study the properties of these estimates.

The Support Vector Machine and its recent developments

Yufeng Liu

University of North Carolina at Chapel Hill, United States of America

The Support Vector Machine (SVM) has been a popular margin-based technique for classification problems in both machine learning and statistics. It has weak distributional assumptions and great flexibility in dealing with high dimensional data. In this talk, I will present various aspects of the SVM as well as its recent developments. Issues including statistical properties of the SVM, robust SVM, multicategory SVM will be covered. Furthermore, recent developments on class probability estimation of the SVM will be discussed.

Random forests

Gérard Biau

University of Pierre and Marie Curie, Paris, France

Random forests are a scheme proposed by Leo Breiman in the 00's for building a predictor ensemble with a set of decision trees that grow in randomly selected subspaces of data. Despite growing interest and practical use, there has been little exploration of the statistical properties of random forests, and little is known about the mathematical forces driving the algorithm. In this talk, we offer an in-depth analysis of a random forests model suggested by Breiman in 2004, which is very close to the original algorithm. We show in particular that the procedure is consistent and adapts to sparsity, in the sense that its rate of convergence depends only on the number of strong features and not on how many noise variables are present.

Probability Machines

James D. Malley
Center for Computational Bioscience, Center for Information Technology, National Institutes of Health, Bethesda, Maryland, United States of America

Many statistical learning machines can provide an optimal classification for binary or category outcomes. However, for personalized medicine probabilities are better suited for risk estimation using individual patient characteristics. It is noted that any statistical learning machine that is consistent for the nonparametric regression problem is also consistent for the probability estimation problem. These will be called probability machines. For evaluating probabilistic forecasting it is known that strictly proper scores are preferred, one example being the Brier statistic. It is shown that any consistent machine also minimizes the expected Brier score, so that evaluation of any probability machine is transparent. Probability machines discussed include classification and regression random forests, and nearest-neighbor schemes, all of which use any collection of predictors with arbitrary statistical structure. Also considered is the classical parametric approach of logistic regression, and the learning machine logitboost. Two synthetic and two real data sets illustrate the use of these machines, and several extensions of these machines are considered for other data structures and prognostic diagnosis.

Using R and Random Jungle for probability estimation

Jochen Kruppa
Institute of Medical Biometry and Statistics, University of Lübeck, University Hospital Schleswig-Holstein, Campus Lübeck, Germany

This talk provides an overview of practical probability estimation generated by different approaches – regression random forest, classification random forest, LogitBoost, k nearest neighbors, bagged k nearest neighbors, and logistic regression. All these methods are freely accessible in R packages. We demonstrate for each approach an easy to implement and fast way to obtain individual probabilities. Furthermore, we illustrate the probability estimation in Random Jungle (Schwarz 2010 Bioinformatics), a novel random forest application in C++ with a generalized framework for tree growing. Random jungle offers different types of tree building mechanics, such as CART, LOTUS and conditional trees. Probability estimation is simple to perform, and the Random Jungle thus framework allows fast relations of machine learning and probability prediction.

Biosketches

Gérard Biau was born in France in 1973. He obtained his Ph.D. from the University Montpellier II in 2000 and joined University Pierre et Marie Curie – Paris VI in 2001, where he is currently professor in the Probability and Statistics Team. His research interests include dynamical systems and chaos, nonparametric estimation, pattern recognition, and high-dimensional statistical learning.

Yufeng Liu obtained his Ph.D in statistics from The Ohio State University in June 2004 and joined the faculty at University of North Carolina at Chapel Hill in July 2004, where he is currently associate professor in statistics. He holds a joint appointment with the Carolina Center for Genome Sciences at UNC, and he is also a member of the UNC Lineberger Comprehensive Cancer Center. His research interests include statistical machine learning, nonparametric statistics, high dimensional data analysis, and bioinformatics.

Michael Kohler received degrees in computer science and mathematics from the University of Stuttgart in 1995, and the Ph.D degree in mathematics from the University of Stuttgart in 1997. In 1998, he worked as a Visiting Scientist at the Stanford University, Stanford, USA. From 2005 until 2007 he was Professor of Applied Mathematics at the University of Saarbrücken, since 2007 he is Professor of Mathematical Statistics at the Technical University Darmstadt. He coauthored with L. Györfi, A. Krzyżak and H. Walk the book *A Distribution-Free Theory of Nonparametric Regression*, (Springer; 2002). His main research interests include nonparametric statistic, nonparametric regression, pattern recognition and data mining.

Jochen Kruppa received his MSc. in plant biotechnology from the Leibniz University of Hannover in 2009 with the focus on statistics. He joined the University of Lübeck at the Institute of Medical Biometry and Statistics (IMBS) in 2009, where he is currently working on learning machines and genome-wide association studies.

James D. Malley is as research mathematical statistician at the National Institutes of Health since 1977, working with a wide range of biomedical research collaborators. His original training was in noncommutative ring theory, resulting in two Springer monographs on statistical applications of Jordan algebras (1986, 1994). More recently he coauthored (with KG Malley and S Pajevic) a text on learning machines, *Statistical Learning for Biomedical Data* (Cambridge; 2011). He has also published on the foundations of quantum mechanics, the connections between classical probability and quantum outcomes, and is currently studying the possibility of quantum probability machines.