**CDSC 608 – Applied Machine Learning (7 ECTS)**

**I. Aims**

On completion of the course students will be expected to:

* Have a good understanding of the two numerical approaches to learning (optimization and integration) and how they relate to maximum likelihood and the Bayesian approach.
* Have an understanding of how to choose a probabilistic model to describe a particular type of data.
* Know how to evaluate a learned model in practice.
* Understand the role of machine learning in massive-scale automation.
* Have a good understanding of the problems that arise when dealing with very small and very big data sets, and how to solve them.
* Understand the mathematics necessary for constructing novel machine learning solutions.
* Be able to design and implement various machine learning algorithms in a range of real-world applications.

**II. Contents**

Prediction is at the heart of almost every scientific discipline, and the study of generalization (that is, prediction) from data is the central topic of machine learning and statistics, and more generally, data mining. Machine learning and statistical methods are used throughout the scientific world for their use in handling the "information overload" that characterizes our current digital age. Machine learning developed from the artificial intelligence community, mainly within the last 30 years. This course provides a selection of the most important topics from both of these subjects.

The course will start with machine learning algorithms, followed by statistical learning theory, which provides the mathematical foundation for these algorithms. It also gives an overview of concepts, techniques, and algorithms in machine learning, beginning with classification and linear regression and ending up with more recent topics such as boosting, support vector machines, Hidden Markov Models (HMMs), and Bayesian networks. We will then bring this theory into context, through the history of ML and statistics. This provides the transition into Bayesian analysis.

**Concepts**

* Parametric and multivariate methods
* Training and testing, cross-validation
* Dimensionality reduction
* Clustering
* Nonparametric methods
* Multilayer perceptrons
* Overfitting/ underfitting, structural risk minimization, bias/variance trade-off
* Regularized learning equation
* Conjugate priors and exponential families
* Algorithms (some covered in more depth than others)
* Apriori (for association rule mining)
* k-NN (for classification)
* k-means (for clustering)
* Naive Bayes (for classification)
* Decision trees (for classification)
* Perceptron (for classification)
* SVM (for classification)
* AdaBoost and RankBoost (classification and ranking)
* Hierarchical Bayesian modeling (for density estimation), including sampling techniques

**III. Learning Outcomes**

Subject-specific Knowledge:

* After studying this module, students will be familiar with some of the fundamentals of the methods in machine learning and which are useful for solving various physical problems.

Subject-specific Skills:

* In addition to the acquisition of subject-specific knowledge, students will be able to apply the acquired machine learning skills to solve various physical problems.
* Explain many machine learning methods and their advantages and disadvantages
* Implement the methods or know where to obtain them from
* Use existing library software
* Have a working knowledge of most of the methods
* Determine the most appropriate learning method for a specific application

**IV. Modes of teaching and learning**

* The mode of the teaching-learning process will involve lectures (50%) and practical work (50%) on a computer which involves about 90% student participation.
* The module will be offered through block teaching, one and a half hours per day, five days a week for six effective weeks.
* Two weeks are reserved for examinations and processing of the examinations.
* The lectures provide the means to give a concise, focused presentation of the subject matter of the module. When appropriate, the lectures will be supported by the distribution of written materials, or by relevant links on websites.
* Regular problem exercises will give students the chance to develop their theoretical understanding and problem-solving skills.
* Students will be able to obtain further help in their studies by approaching their lectures, either after lectures or at other mutually convenient times. The programming language to be used in this course will be either python or R.

**References**

* Kevin P. Murphy. Machine Learning: A Probabilistic Perspective, MIT Press 2012.
* Christopher M. Bishop. Pattern Recognition and Machine Learning, Springer 2007.
* T. Hastie, R. Tibshirani, and J. Friedman. The Elements of Statistical Learning. Springer 2011.
* Alpaydin, E. 2010.  Introduction to Machine Learning, 2nd Edition, MIT Press Cambridge.
* Bishop, Christopher M. Pattern Recognition and Machine Learning. Springer.
* Mitchell, Tom M. 1997. Machine Learning. McGraw-Hill
* Dietterich, T. G. 2003. “Machine Learning.” In Nature Encyclopedia of Cognitive Science. London: Macmillan.
* Duda, R. O. et al.  2001. Pattern Classification. 2nd Ed. New York: Wiley.
* Manning, C. D., and H. Schutze. 1999. Foundations of Statistical Natural Language Processing. Cambridge, MA: MIT Press.
* McLachlan, G. J. 1992. Discriminant Analysis and Statistical Pattern Recognition. New York: Wiley.
* Rencher, A. C. 1995. Methods of Multivariate Analysis. New York: Wiley.