MLCI 2016 – Schedule

Room 711 – DIBRIS, University of Genova, Italy

Monday, 20 June

10.00 - 13.00 **Francesco Masulli**, DIBRIS, University of Genova, Italy
Course organization – Introduction to Computational Intelligence and to Fuzzy Sets.

14.30 - 15.30 **Elena Bertozzi,** Quinnipiac University, Connecticut (USA)
Help Me Help Myself - Using Play to Empower Players and Motivate Pro-health Behaviors

15.30 - 17.30 **Stefano Rovetta**, DIBRIS, University of Genova, Italy
Artificial neural networks. Perceptual problems. Single-unit neural networks. The problem of classification: Bayes decision theory.

Tuesday, 21 June

10.00 - 13.00 **Stefano Rovetta**, DIBRIS, University of Genova, Italy
Characterization and evaluation of classifiers. Discriminative and generative classifiers. Sequential data. Classification of sequences. Learning as optimization. Representation problems.

14.00-15.00 **Nikesh Bajaj,** *PhD Student*, DITEN, University of Genova, Italy
Ensemble approaches for classification and regression.

15.00 – 17.00 **Francesco Masulli**, DIBRIS, University of Genova, Italy
Unsupervised learning: the clustering problem.

Wednesday, 22 June

10.00 – 13.00 **Francesco Masulli,** DIBRIS, University of Genova, Italy
Fuzzy clustering.

14.00 – 15.00 **Bukahally S. Harish,** S. J. College of Engineering, Mysore, India
Cluster Based Symbolic Representation and Feature Selection for Text Classification.

15.00 – 16.00 **Amr Abdullatif**, *PhD Student*, DIBRIS, University of Genova, Italy
Clustering methods to model and analyze non-stationary data streams and its applications.

16.00 -17.00 **Stefano Rovetta**, DIBRIS, University of Genova, Italy
Multi-layer neural networks. The error back-propagation algorithm. Principal Component Analysis and subspace methods. Autoencoders.

Thursday, 23 June

10.00 – 13.00 **Stefano Rovetta**, DIBRIS, University of Genova, Italy
Restricted Boltzmann Machines and Deep learning. Other feature learning methods.

14.00 – 15.00 **Bukahally S. Harish**, S. J. College of Engineering, Mysore, India
Symbolic Representation of Text Document.

15.00 - 16.00 **Francesco Masulli**, DIBRIS, University of Genova, Italy
Kernel clustering. Spectral clustering.

Talk by Prof. *Elena Bertozzi* - *20* Jun 2016

Title: *Help Me Help Myself - Using Play to Empower Players and Motivate Pro-health Behaviors*
Speaker: **Prof. Elena Bertozzi**
Institution: Game Design & Development - Visual and Performing Arts - Quinnipiac University, Connecticut (USA).

Abstract: Over the past 18 years, Dr. Elena Bertozzi and her team have worked with scientists, artists and healthcare professionals on a variety of games that incentivize players to seek and achieve positive behavioral change in healthcare. Bertozzi has written extensively on gender, sexuality and technological self-efficacy. Motivated by her experiences in using games to address previously intractable problems, she studies ways in which interactive technologies can guide players towards better decision making based on accurate knowledge. This talk will discuss how to design, develop and deploy impactful games for health. Bertozzi will provide examples of recent projects, discuss lessons learned, and common roadblocks to success. Her group at Quinnipiac University specializes in leveraging current technologies to produce low-cost 2D games that can be delivered over the most accessible device for the target audience.

Talk by Mr. Nikesh Bajaj - *21* Jun 2016

Title: *Ensemble approaches for classification and regression*
Speaker: **Mr. Nikesh Bajaj,** *PhD Student*
Institution: DITEN, University of Genova, Italy

Abstract: I wish to discuss two problems I worked on: (1) National Data Science Bowl - Predict ocean health. It is a classification problem with 121 classes of plankton population. Data: 30K images of different size for training, 100K images for testing; (2) Liberty Mutual Group's Property Inspection Prediction (Hazard score). It is a regression problem, where 32 anonymous features where given. Center idea would be to share the different approaches and models used for these problems.

Talk by Mr. Amr Abdullatif - *22* Jun 2016

Title: *Clustering methods to model and analyze non-stationary data streams and its applications.*
Speaker: **Mr. Amr Abdullatif,** *PhD Student*
Institution: DIBRIS, University of Genova, Italy

Abstract: Data streams are quickly becoming a major paradigm in the realm of data science. They arise naturally from continuously observed phenomena in an increasing number of fields, from the web, to wearable sensors, to intelligent transportation systems, to smart homes and cities. But, in addition to this, the size of any collection of “big data” makes single-pass methods a necessity, turning these data effectively into a special case of streaming data. Data streams are always related to time, although to different degrees. They may represent actual time series or quasi-stationary phenomena that feature longer-term variability, e.g., changes in statistical distribution or a cyclical behavior. In these non-stationary conditions, any model is expected to be appropriate only in a neighborhood of the point in time where it has been learned. Its validity may decrease smoothly with time (concept drift), or there may be sudden changes, for instance, when switching from one operating condition to a new one (concept shift). The idea of our work is to study a clustering process that should adapt to streaming data, learning continuously from the input patterns as they arrive. Based on this idea our work addresses the problem of fuzzy clustering in non-stationary data streams, by using a new outlier density estimation method which is applied to detecting changes in the distribution of a clustering algorithm (Graded Possibilistic clustering approach). Based on this idea, two methods are proposed, one for batch learning using a sliding window, and one for online learning by stochastic maximum possibilistic entropy. We propose an "outlierness index" by measuring how much the total mass of membership (sum of membership) to clusters is less than one. Based on this, we propose to measure the "outlier density" for the frequency and intensity at which outliers occur. These measures are then used to differentiate between three learning regimes (Concept drift, Concept shift, and Outliers) .

Talk by Prof. Bukahally S. Harish - *22* Jun 2016

Title: *Cluster Based Symbolic Representation and Feature Selection for Text Classification*
Speaker: **Prof. Bukahally S. Harish**
Institution: S. J. College of Engineering, Mysore, India

Abstract: In automatic text classification, it has been proved that the term is the best unit for text representation and classification. Though a text document expresses vast range of information, unfortunately, it lacks the imposed structure of a traditional database. Therefore, unstructured data, particularly free running text data has to be transformed into a structured data. After converting an unstructured data into a structured data, we need to have an effective representation model to build an efficient classification system. Although many text document representation models are available in literature, frequency based Bag of Word (BOW) model gives effective results in text classification task. Unfortunately, BOW representation scheme has its own limitations. Some of them are: high dimensionality, loss of correlation and loss of semantic relationship that exists among the terms in a document. Also, in conventional supervised classification an inductive learner is first trained on a training set, and then it is used to classify a testing set, about which it has no prior knowledge. However, for the classifier it would be ideal to have the information about the distribution of the testing samples before it classifies them. To deal with this problem of learning from training sets of different sizes, we exploited the information derived from clusters of the term frequency vectors of documents. Clustering has been used in the literature of text classification as an alternative representation scheme for text documents. Given a classification problem, the training and testing documents are both clustered before the classification step. Further, these clusters are used to exploit the association between index terms and documents. Conventionally, the feature vectors of term document matrix (very sparse and very high dimensional feature vector describing a document) are used to represent the class. Later, this matrix is used to train the system using different classifiers for classification. Generally, the term document matrix contains the frequency of occurrences of terms and the values of the term frequency vary from document to document in the same class. Hence to preserve these variations, a new interval representation for each document is given. Thus, the variations of term frequencies of document within the class are assimilated in the form of interval representation. Moreover conventional data analysis may not be able to preserve intraclass variations but unconventional data analysis such as symbolic data analysis will provide methods for effective representations preserving intraclass variations. In one of our work, we propose a new method of representing documents based on clustering of term frequency vectors. For each class of documents we proposed to create multiple clusters to preserve the intraclass variations. Term frequency vectors of each cluster are used to form a symbolic representation by the use of interval valued features. Subsequently we proposed a novel symbolic method for feature selection. The corresponding symbolic text classification is also presented.

Talk by Prof. Bukahally S. Harish - *23* Jun 2016

Title: *Symbolic Representation of Text Documents*
Speaker: **Prof. Bukahally S. Harish**
Institution: S. J. College of Engineering, Mysore, India

Abstract: Nowadays many real time text mining applications have received a lot of research attention. Some of the most widely applications are: Spam filtering, Categorization of emails, Directory maintenance, Ontology mapping, Document retrieval, Routing, Filtering etc. World Wide Web contains huge volume of documents which are in electronic form. The documents which are available in electronic form (WWW), sometimes lack organization of information. The lack of organization of information motivates people to automatically manage the huge amount of information, and thus it requires the implementation of sophisticated learning agents which are capable of classifying relevant information and thereby increasing the organization of text over WWW. The possible solution to manage these unorganized data is to deploy, effective representation methods. Further, the newly employed representation methods are used for various data mining tasks.