EEL709: Project Presentations

15 April – 1 May, 2013

Presentations on all days will be between 12:00–13:00 in III-LT4.

Monday, 15 April

- **Exploiting Infinite Unlabelled Data: Semi-supervised Object Recognition on CIFAR-10**
  Vinayak Agarawal, Sherjil Ozair, Aayush Ahuja

  Object recognition finds its applications in a variety of tasks ranging from image search to self-driven cars. The number of images available on the Internet is increasing rapidly, and with it the need to correctly classify it for retrieval. However, most of the images available are unlabeled, or are labeled incorrectly. Thus, there is a need to develop fast semi-supervised learning methods which can take advantage of a large set of unlabeled data along with a relatively less number of labeled data.

  At first, we focus on learning features from unlabeled data using various feature learning algorithms like the RBM, K-means, Auto-encoders etc and analyze the effect of changes in the model setup: the receptive field size, number of features, the step-size between extracted features, and the effect of whitening. We observe that the model setup plays an important role in the success.

  Based on the insights developed from our analysis of the various classes of algorithms, we present a novel algorithm which optimizes for both efficiency and accuracy, as demonstrated by the experiments, and is ideally suited to exploit large corpus of unlabelled data.

- **Community Detection and Course Recommendation in Academic Networks**
  Jobin Wilson, Anupama Ray, Shraddha Chaudhary

  This work presents a novel scheme for community detection within an academic system by modeling groups based on similar academic interests. It also provides personalized course recommendations based on individual interests and crowds behavior. Academic information of each member and course details are extracted from the publicly available course list and prospectus of an institution. LDA based topic modeling is applied on course descriptions to discover topic distributions across each course. Individual interest towards each topic is calculated from topic distributions of courses for which he has enrolled and these vectors are clustered to discover communities. We also cluster course-topic distribution vectors to discover groups of similar courses. For each student, a co-occurrence based collaborative filtering approach is used to recommend top K courses based on his/her enrollments and the crowd behavior about the courses. Precision, recall and F-score are calculated through cross-validation. This work can be extended to consider additional features such as gender and type of enrolled programme to refine community grouping. It can be integrated with existing social networks to recommend sample question papers, solutions and research ideas. The idea remains novel since currently there are no such web based system which provide such capabilities without users providing explicit information.

Tuesday, 16 April

- **StackOverflow Question Classification**
  Dilpreet Singh Chahal, Vishnu Gupta, Arjun Attam

  This project aims to examine success of questions on Stack Overflow, a question answer website on a wide range of topics in computer programming. The website serves as a platform for users to ask and answer questions, and, through membership and active participation, to vote questions and answers up or down. Using methods of Machine Learning and extraction of relevant features, we hope to predict the success of a question and further identify features to identify good questions, which implies getting up votes and answers, posted on the Stack Overflow website. The website encourages users to ask "practical, answerable questions based on actual problems"chatty, open-ended questions are discouraged, as they diminish the usefulness of the website and push good questions off the front page. Consequently, such questions are "closed", as compared to "open", which is the status for good questions. Several distinct types of questions on the website can be distinguished: factual (seeking objective data); advice, (seeking recommendations); opinion, (seeking others viewpoints), and non-questions (spam). Questions that are non-constructive, off-topic and too localised are also discouraged, and closed by the moderators. In essence, we have a multiclass classification problem and we will be using methods like Logistic Regression, Support Vector Machines and Naive Bayes to classify questions as accurately as possible.

- **Decoding Captchas**
  Keshav, Manish Singh Rawat

  A captcha is a program that can generate and grade tests that humans can pass but current computer programs cannot. The term captcha stands for Completely Automated Public Turing Test To Tell Computers and Humans Apart. Our project focusses on decoding text captchas. Typically text captcha recognition involves two steps namely, segmenting the captcha image to obtain individual characters and then recognising those characters via a character recognition program. We concentrate on the character recognition part in our project and are currently able to recognise digits (0–9) efficiently. We are using PCA to extract features from the character image and then using a 3-layer neural network to classify the features appropriately. MNIST dataset of digits is used to train the neural network.

- **Job Salary Prediction**
  Vaibhav Khandelwal, Alok Singhal

  In this project, we aim to predict salaries of different jobs offered by different UK companies. This data can further be used by the companies to make an estimate of the salaries and also by job seekers to count their worth.

  At first, we focused on building the entire pipeline, starting from basic features and then regressing the data. This step was essential to see the whole system working, since we were then the first-hand users of Python Scikit Learn toolkit. The next step in this lane was to specialize each and every block of this pipeline.

  Our presentation will focus mainly on the problems encountered in the feature extraction block and how by the use of different techniques, weve managed to get the best of the features. It incorporates the use of nltk (Natural Language Toolkit), bi-grams bag of words, categorical feature extraction v/s bag of words representation and handling of missing data. Each step has been supported with the corresponding score obtained by application of regressors on the data to highlight the importance of each feature. Further, classifier is also applied to the data by organizing the normalized salary into a number of groups. The performance is also tabulated with the classification results.

- **Object Classification on Caltech 101**
  Ankita, Pritesh Kumar, Ankit Nayan

The target of classifying various categories of objects is one of the pioneers in the field of image processing and machine learning. Various algorithms have been developed to capture the appearance and shape of the object images. We have implemented BOW (Bag Of Words) model for object classification. It uses SURF descriptors of keypoints to get best features of most salient points in the image. A codebook is generated and a dictionary is made based on all collected keypoints. K-means clustering algorithm is then run to assign each keypoint to the nearest codeword. A histogram for each image is made that counts the frequency of keypoints in the dictionary. This feature vector is then tested in various classifiers and the results thus obtained are analysed.

Wednesday, 17 April

- **Social Network User Interest Prediction**
  Parijat Mazumdar, Abhishek Gupta

  In this project we revisit the problem of creating recommender system in a social network setting. Put forward simply, the goal of our project is to recommend items to users to maximize the Click-Through Rate with the additional constraint of maximum 3 recommendations per user. We have chosen the popular collaborative filtering (CF) model to build our recommender system. In collaborative filtering, past choices of various users and popularity of items (user-item interactions) are used to train a model. While CF has quite a few different variants, we focus on a matrix factorization based method. In this method each entity (e.g.: user/item) is modeled as a vector of latent factors with the constraint that all the entities are of same dimension. Apart from this we also try to integrate the user social network graph information into the CF model, analyze different observations and draw out useful insights about users/items.

  Our project has been adapted from the KDD Cup 2012 Track 1 competition. The background, problem statement and dataset can be found at `http://www.kddcup2012.org/c/kddcup2012-track1`.

- **Comparative Approach and Ensemble Learning on Multiple Datasets**
  Govind Prasad Bidua, Soniya Singhal, Ronak Purushottam Gupta

  In classification problem, it has been observed that the accuracy of particular method is data dependent and if the type of data is imbalanced (amount of training samples of class are hugely skewed), single classifier may not give the desired performance. In this scenario, combining different classifiers (Ensemble Methods) can help to improve the performance. Adaboost and Bagging are the two most important ensemble methods.

  In this project, we have applied the different type of classification methods such as logistic regression, neural network and support vector machines on balanced and imbalanced datasets. Having observed that the single classifier is not giving the desired performance, we have applied two ensemble methods (Adaboost and Bagging) and achieved the significant improvement in the performance.

- **Song Preference Prediction**
  Gaurav Mishra, Shweta Mahendra Karwa

  We aim to build a song recommender system from user listening-histories available on `www.kaggle.com/c/msdchallenge`. One popular approach, called collaborative filtering, involves finding users from the training data who have preferences most similar to that of the target user and use a weighted combination of their ratings to predict song preferences for the target user. Another approach, called content-based filtering, involves finding similarities amongst songs based on user ratings and other available meta-data, and suggesting songs most similar to the songs that the target user prefers most. We have implemented, tested and analyzed the merits and demerits of these methods and some of their variants, and their specific applicability on the sparse implicit ratings that is available to us.

Monday, 22 April

- **Transfer Learning for Video Classification**
  Arihant Jain, Siddharth Srivastava, Sumit Soman

  Transfer learning is a new learning paradigm which enables us to transfer knowledge gained in one domain to other related domains.These approaches are useful in scenarios where one domain has large amount of labelled data and another domain has either zero or very few labelled examples.This paradigm is motivated by the nature of human learning, as people generally transfer the skills learnt in one domain to other related domains. In this project, we have attempted to compare traditional machine learning approaches on images, namely, Clustering and Principal Component Analysis with the transfer learning approaches. The transfer learning techniques have been used for scenarios where we have ample amount of labelled data in source domain and nearly zero labelled data in the target domain . Specifically, we have looked into the transductive transfer learning setting, where we use domain adaptation and covariance shift techniques in order to do classification in the target domain using trained classifiers in the source domain.

- **Learning How the Mind Learns**
  Sanjit Singh Batra

  What happens inside the mind of a person when he is suffering from Alzheimer's? Is it possible to "look" at the mind of a child diagnosed ADHD and comment on the validity of the diagnosis? Such questions and their answers will help not only in earlier as well as better diagnostic procedures, but also in understanding the neural changes accompanying them and in turn might aid in constructing counter-measures.

  Today, techniques like fMRI (functional Magnetic Resonance Imaging) and such have enabled us to peek into the human mind and see what's going on. The development of implicit feature extraction methods like Deep Learning through Restricted Boltzmann Machines (RBMs) make this highly tractable. Through this endeavor I hope to demonstrate as well as innovate methods for understanding the changes in the brain occurring due to medical complications.

  Another interesting application of such tools is the science of "mind reading". The changes in the mind of an individual while she looks at different objects are recorded. Then, when she thinks of these objects, they can be actively reconstructed, as though we were "reading" her mind.

Tuesday, 23 April

- **Job Salary Prediction**
  Prabhav Agrawal, Akshay Kumar, Akshay Sharma

  In this project, we work on the problem of predicting job salaries of any UK job ad based on its contents. A dataset including various fields like Title, Full Description, Company, Job Category, Location etc has been given and using various regression and classification algorithms, the salary of the ads has to be predicted.

  We have first divided salaries into 3 categories High, Medium and Low and learned features through classification algorithms like Naive Bayes, SVM. We have used Bag of Words approach for text-based features like Title and Full Description. Bag of words are either the words with large variance or the words with high frequency in the three classes. For some features like Category and Location, we have learned them as categorical features.

  In the second part, we have tried to obtain a normalized value for the predicted salary using regression algorithms like RandomForestRegressor. The metric used for measuring the correctness of predicted salaries is MeanAbsoluteError.

Our project has been adapted from the Job Salary Prediction competition on `www.kaggle.com`. The background, problem statement and dataset can be found at `www.kaggle.com/c/job-salary-prediction`.

- **Mining Twitter for Financial Event Prediction**
  Ankit Rao, Ayush Agnihotri

  In this project, we apply sentiment analysis and machine learning principles to find the correlation between "public sentiment" and "market sentiment". Market sentiments have in the past proven to be viable indicators as they contain information which is not present in the previous market data, and hence they can be used as a complimentary set of features in the stock market prediction problem. We plan to implement a simple classification problem based on some keywords relating to a particular market sentiment on the accumulated twitter data, if times permits we will also try to build up a learning system which learns the words/phrases associated with a particular market sentiment and to classify the incoming feeds accordingly. After getting the daily sentiment associated with a particular company in the market we aim to use machine learning techniques (like regression and neural networks) to correlate the two events and to develop a robust system which is able to provide us vital insight in the matter.

- **Gender Prediction from Handwriting**
  Aakansh Gupta, Utsav Bansal

  Classifying handwritten documents into a writer demographic category- gender, age or handedness is a very interesting and active research field. The application of such a classification lies mainly in forensic analysis to determine the various demographic categories and hence narrow down the class of suspects.

  In this project we aim to develop methods to predict gender of the author of a handwritten document. This is a simple classification problem of where each document is classified into male or female classes. The data set used is available through ICDAR 2013 contest on Kaggle. This consists of documents written by 282 writers and 7066 features extracted from each of them. Using these features, we model the problem into logistic regression, SVM and Neural Network implementations and analyse results. We then try the same implementations after restricting down the number of features by using techniques like mRMR, Kruskal-Wallis and Fisher score. Testing is done by cross validation and log loss metric is used as a measure of the score.

Monday, 29 April

- **Comparative Approach on Data Sets**
  Rahul Saluja, Ben Mathew John

  The comparative approach would be done on the ADULT data set. Using this data set we would see the classification approaches which suits the data set.

  The idea behind going for adult data set is to determine a person's income, which can be used by marketing firms to strategize their offers better for individuals. Almost all firms e.g.: makemyTrip have their online forms asking individuals for their age, education, work, nationality etc. Now using these features if the income of a person can be determined then they can strategize offers better. The data set is already available at the UCI repository. The Adult data set contains 48842 instances (which is quite a good number of instances), having 14 attributes: Income (nom): > 50K, <= 50K, Age (num), workclass (nom), education (nom), education-num (num), marital-status (nom), occupation (nom), relationship (nom), race (nom), sex (nom), capital-gain (num), capital-loss (num), hours-per-week (num), native-country (nom), with fields having nominal as well numeric values.

  Classification algorithms would be run on the above data set and the number of correctly classified instances in each case is determined. The algorithms would be run on WEKA

which is a collection of machine learning algorithms for data-mining tasks and has a good Graphical User Interface. Also we would look for algorithms available in the MLComp and gets the results on running data in it.

- **Learning Biological Regulatory Networks**
  Tanmay Batra, Umang Gupta, Vivek Mangal

  Reconstructing gene regulatory networks from high throughput data is a long-standing challenge. The problem of modelling gene regulatory networks is basically to find out how the concentration of the protein secreted by amino acids associated with a particular gene is affected by other genes (in our case, we assume a particular subset of genes called Transcription Factors) and external environmental factors. Genome-scale inference of transcriptional gene regulation has become possible with the advent of high-throughput technologies such as micro-arrays and RNA sequencing, as they provide the level of proteins under many tested experimental conditions. From these data, the challenge is to computationally predict direct regulatory interactions between a transcription factor and its target genes; the aggregate of all predicted interactions comprises the gene regulatory network. Many network inference methods have been developed to address this challenge like Inferelator, TIGRESS, ARACNE, etc. Our aim is to implement these inference methods, interpret the results and try to obtain a network which best explains the interactions provided in the data.

- **Gender Classification in Speech Processing**
  Baiju M Nair, Geetanjali Srivastava

  Our main objective is gender classification in speech processing. Gender classification using model like SVM and NN has been implemented based on certain acoustic feature. A benchmark dataset has been used for the evaluation to make sure the reliability of the performance. The data (120 samples with equal gender distribution) has been extracted which is readable by the MATLAB. A preprocessing of the data has been carried out to extract the data in the desired form. The data which is a long sequence is extracted frame by frame for feature extraction. Both time domain and frequency domain features are utilized for the training. The critical feature vectors for classification of speaker used at the training are energy entropy, zero crossing rate, mel-frequency and short term energy. The performance of the two models SVM and NN on the speech processing will also be demonstrated.

  The work can be extended to individual speaker classification with various noising background. Based on the results, it would be interesting to understand the algorithms that can be used to distinguish original music track from some local recordings. Sound track recognizing of the original voice of any famous singer can also be developed. Apart from this there is a wide range of medical application like automatic ECG pattern classifier to warn potential heart patient when the cross the critical level. Blockage in blood vessel, pulse monitoring can be used to understand the body state.

- **Attack Detection in Collaborative Filtering Recommender Systems**
  Anurag Tripathi, Anvaya Rai, Deepti Goel

  In this project we aim at generating fake profiles and examine the impact of these profiles on recommender system. Publicly accessible collaborative recommender system is susceptible to various types of attacks. The users enter in these systems by making a profile and then provide ratings to the items. These ratings are used to provide recommendation to the user which affect the proper functioning of the system. Here, we generate fake profile by using Bandwagon attack model and examine the tradeoff by de-correlating the fake profile. To perform experiments we use publicly available MovieLens 100K data set which consists of 100,000 ratings on 1,682 movies by 943 users and scikit-learn package implemented in python as development environment . Also we will use a hybrid method e.g. PCA and Co-occurrence matrix to identify the attack profile. The objective is to have correlation less than > 0.8 and to achieve the accuracy of Fake profile detection with $> 90\%$.

Tuesday, 30 April

- **Face-Nonface Classification**
  Nitesh Gupta, Vaibhav Gaur, Vishwanath Gaur

  Face detection from cluttered images is challenging due to the wide variability of face appearances and the complexity of image backgrounds. This project proposes a classification-based method for locating frontal faces in cluttered images. To improve the detection performance, we extract texture based features i.e. Local Binary Patterns (LBP) from images as the input of the underlying two-class classifier. The texture features provide better discrimination ability than the image intensity. The underlying classifiers are Logistic Regression model and Support Vector Machine. To reduce the dimensions of LBP features (256 dimensions), we make the use of uniform LBP features which reduce the dimensionality to 59. The classifier is trained on samples of face and non-face images to discriminate between the two classes. The superior detection performance of the proposed method is justified in experiments on a large number of images. At last, we compared the results got from machine learning techniques *viz.* logistic regression and SVM.

- **Learning on Financial Time Series**
  Kanika Jain, Monalisa, Tanya Raghuvanshi

  In this project, we predict the future stock prices using nearest neighbor predictors. Applying this prediction strategy to the New York Stock Exchange (Microsoft, Google), Bombay Stock Exchange (Goldman Sachs) our results suggest that the non-linear forecasting implemented using neural networks is superior to extrapolation of past values into the immediate future based on correlation among lagged observations and error terms. In contrast, nearest neighbor methods select relevant prior observations based on their levels and geometric trajectories, not their location in time. The moving average method fails to capture peaks. On the other hand, this algorithm works fine with peaks.

- **Stock Price and Credit Risk Prediction Using Mahout/Hadoop**
  Gaurav Agarwal, Abhishek Kumar

  The credit card industry has been growing rapidly recently, and thus huge numbers of consumers credit data are collected by the credit department of the bank. The credit scoring manager often evaluates the consumers credit with intuitive experience. However, with the support of the credit classification model, the manager can accurately evaluate the applicants credit score. Support Vector Machine (SVM) classification is currently an active research area and successfully solves classification problems in many domains. Our study will use three strategies to construct the hybrid SVM-based credit scoring models to evaluate the applicants credit score from the applicants input features. Two credit datasets in UCI database are selected as the experimental data to demonstrate the accuracy of the SVM classifier. We will compare SVM classifier with neural networks, and decision tree classifiers. Additionally, combining genetic algorithms with SVM classifier, we will propose new strategy that can simultaneously perform feature selection task and model parameters optimization.

Wednesday, 1 May

- **Financial Time Series Analysis Using Pattern Recognition**
  Ravi Kumar, Naveen Kaushik, Kanwarjeet Singh Dhaliwal

  The aim of the project is to predict the interest rate and bond yields variation and stock market prices in the Indian economy using neural networks and make a comparative study on the effect of various pre-processing techniques on the performance of the neural network. We know that the financial data is the noisiest data so pre-processing techniques are focused on reduction of the effect of noise and non-stationary nature of the data.

- **Modelling Financial Time Series**
  Abhishek Sharma, Nimesh K Verma, Kumar Sourav

  A financial time series is a sequence of data points including value of a financial asset along with time, measured typically at successive points in time spaced at uniform time intervals. Time series forecasting is the use of a model to predict future values based on previously observed values. Our project aims to predict the values of a particular Stock using the historical data of that particular stock. A financial forecast is the estimated future value of a company stock or other financial instrument traded on a financial exchange. The reasonable prediction of a stock's future price could yield significant profit to the investor by investing according to the prediction.

  We have used Artificial Neural Networks and Statistical Regression Models on MATLAB Software to achieve this aim. Our presentation will mainly focus on explaining the implementation of these approaches for prediction along with their error analysis.

- **Machine Learning As an Alternative to Black-Scholes for Option Pricing**
  Avnish Kumar

- **OCR on Vehicular Images**
  Avinash Singh Bagri

  The project is basically a two-step procedure. In first part the images of vehicles having their registration numbers are taken as input followed by implementation of OCR to have the input in text form. The basic concepts of image processing i.e. thresholding, edge detection, feature extraction and classification are going to be used throughout. The implementation of OCR will be divided in two parts: Training consisting of pre-processing, feature extraction and model estimation. Second part will be Testing part which would include preprocessing, feature extraction and classification. Where preprocessing is processing the data to make it in suitable form, feature extraction is reducing the amount of data by extracting the relevant information, model estimation and classification reckon to estimating appropriating models and comparing features for closest match.

  References:

  `http://www.sersc.org/journals/IJUNESST/vol6_no1/2.pdf`

  `http://www.ele.uri.edu/~hansenj/projects/ele585/OCR/OCR.pdf`