SMAI-Spring 2016 Major Project Ideas

Project Distribution:

Each of the teams have to give **5 preferences** for the project. We have broadly classified the topics in 5 section. Select **one project from each section**. The section are as follows:

1- 8 : Classification & Clustering
9-16: Prediction & Recognition
17-24: Detection & Recommendation
25-32: Textual & Search Processing
33-40: Image Analysis & Miscellaneous

Detailed description of each of the project is given below :

Project Number	Project Details
1	Project Name: Machine Learning Classification of Kidney and Lung Cancer Types. Project Description:In this project, we used machine learning algorithms to classify 4 different kidney and lung cancer types based on their methylation profiles, and show that our two best performing models have an accuracy exceeding 90%. Techniques used here are SVM, Naive Bayes, Knn, K-means. Project Reference: http://cs229.stanford.edu/proj2013/JainMenZhou-MachineLearningClassificationofKidneyandLungCancerTypes.pdf
2	Project Title: Question Classification Project Description: Classify the question to the anticipated type of the answer, reducing the search space to identify the correct answer. Current trend is to use semantic features and do comparative study on different classifiers to achieve benchmark results.

3	Project Title: Document Classification and Clustering Project Description: Text documents can be classified into certain categories based on content, author, year of publication etc. A document classification system takes a sample document as input and maps it to one of these categories. On the other hand, retrieval systems return a set of similar documents. Some of the most popular techniques used are Clustering, Bag of Words (BoW), Inverted Index, KNN References: 1. https://www.stat.washington.edu/people/wxs/Learning-papers/MuruaStuetzleTantrumSieberts-A4.pdf 2. http://kmi.open.ac.uk/publications/pdf/kmi-00-14.pdf
4	Project Title: Learning Social Circles Project Description : Our personal social networks are big and cluttered, and currently there is no good way to organize them. Social networking sites allow users to manually categorize their friends into social circles (e.g. 'circles' on Google+, and 'lists' on Facebook and Twitter), however they are laborious to construct and must be updated whenever a user's network grows. We define a novel machine learning task of identifying users' social circles. We pose the problem as a node clustering problem on a user's ego network, a network of connections between her friends. References (if any) : https://cs.stanford.edu/people/jure/pubs/circles-tkdd14.pdf
5	Project Title: Tag Assignment for Competitive Programming Problems Description: Multi-label classification is an important paradigm in machine learning. In this project you would have to scrape coding websites like - CodeChef, Codeforces etc., and come up with your own set of features to label programming problems. This tag assignment problem is multi-class as a program can involve multiple concepts like discrete maths, dynamic programming etc. References: http://psiexp.ss.uci.edu/research/papers/RubinEtAl_2011_MLJ_SpecialIssue_2ndResubmission_V14p1.pdf
6	Project Title: Object Recognition (Classification/ Retrieval) Project Description: Object classification refers to those systems, which, given an image of an object are able to classify it. On the other hand, a good example of an

	object retrieval system is Google Image search. The user provides a query image and the system retrieves a set of similar images. Some of the most popular techniques used are SIFT, Bag of Visual Words, Clustering, Support Vector Machines. A large number of open source libraries are available.
7	Project Title: Face Detection in images Project Description: Face detection is a computer technology being used in a variety of applications that identifies human faces in digital images. Implement various methods (eg: using different classifiers) for face detection.
8	Project Name: Link Prediction in Social Networks using Supervised Learning Project Description: Social network analysis has attracted much attention in recent years. Link prediction is a key research directions within this area. In this research, we study link prediction as a supervised learning task. Along the way, a set of features is identified that are key to the superior performance under the supervised learning setup. Then, different classes of supervised learning algorithms are compared in terms of their prediction performance using various performance metrics Project Ref: http://www.siam.org/meetings/sdm06/workproceed/Link%20Analysis/12.pdf
9	Project Title:Predicting risk of mortality from time-series data of patients Description: This is an application based project . You would be provided with time-series data of patients' vitals and other lab tests information and are required to develop a statistical model to predict mortality rate of patients from the data, given some constraints, for e.g. your model should predict at least 5 hours before the actual death. This has practical applications as it would allow hospitals to concentrate their resources on patients at risk well in advance. References: http://physionet.org/challenge/2012/papers/
10	Project Name: Hacking the Hivemind: Predicting Comment Karma on Internet Forums Project Description: Virality on online platforms has a number of social and monetary implications. To measure virality, many websites such as Hacker News and Reddit allow users to "upvote" or "downvote" others' comments, and these votes are aggregated into a single numeric "karma" score for each comment. While these karma scores are a numeric proxy for the popularity and virality of a user's views, they are affected by a multitude of factors, and as a result, virality is notoriously difficult to

	predict. In this project, we apply machine learning techniques to build systems that can accurately predict comments' success, using karma as a proxy. We further analyze the content-based and metadata features that are most weighted in determining a comments' popularity in order to provide insight into the role of such features in comment virality. Project Ref: http://cs229.stanford.edu/proj2014/Daria%20Lamberson,Leo%20Martel,%20Simon%20Zheng,Hacking%20the%20Hivemind.pdf
11	Project Name: Using Facebook Profiles to Predict Sexual Orientation. Project Description: We emphasize the value of implicit data by creating a machine learning algorithm that uses basic information, photos, and published text on Facebook profiles to predict sexual orientation in males. We constructed a model with Naïve Bayes classifiers and a Support Vector Machine, performing on different types of data. Project Ref: http://cs229.stanford.edu/proj2015/019_report.pdf
12	Project Title: Predicting Significance of publication using bibliographic text data Project Description : You will be given a huge dataset (not necessarily given, you might have to crawl Google scholar or Microsoft Academic Search) and will see how you can use the crawled information to predict the significance of that publication References (if any) : http://cse.iitkgp.ac.in/~pawang/papers/cikm15.pdf
13	Project Title: Stock Price Trend Forecasting using Supervised Learning methods Project Description: Predicting the stock price trend by interpreting the seemly chaotic market data has always been an attractive topic to both investors and researchers. Among those popular methods that have been employed, Machine Learning techniques are very popular due to the capacity of identifying stock trend from massive amounts of data that capture the underlying stock price dynamics. In this project, we applied supervised learning methods to stock price trend forecasting References: 1. http://cs229.stanford.edu/proj2013/DaiZhang-MachineLearningInStockPriceTrendForecasting.pdf 2. http://cs229.stanford.edu/proj2012/ShenJiangZhang-StockMarketForecastingusingMachineLearningAlgorithms.pdf

14	Project Title: Computation of Visual saliency model Project Description: Identifying the type of stimuli that attracts human visual attention has been an appealing topic for scientists for many years. This project aims to mark the salient regions in a given image using machine learning References (if any) : A computational visual saliency model based on statistics and machine learning. Lin RJ, Lin WS
15	Project Title: Product Search Evaluation Project Description: Companies like Home Depot use search relevancy of their product search algorithms as an implicit measure of how quickly they can get customers to the right products . Currently, human raters evaluate the impact of potential changes to their search algorithms, which is a slow and subjective process. . Help them improve their customers' shopping experience by developing a model that can accurately predict the relevance of search results. References : https://www.kaggle.com/c/home-depot-product-search-relevance
16	Project Title: Hand Gesture Recognition Project Description: Hand gesture recognition system that can be used for interfacing between computer and human using hand gesture. In natural Human Computer Interactions (HCI), visual interpretation of gestures can be very useful.
17	Project Title: Human Detection in images Project Description: In surveillance systems, it is a basic task to detect and track the presence of a human figure. Several applications like traffic monitoring, event detection etc. rely on a standard Human Detection Algorithm. The challenges include change in lighting conditions, occlusions, viewpoint variance, poor video quality etc. References (if any) : Histograms of Oriented Gradients for Human Detection "Navneet Dalal and Bill Triggs"
18	Project Title: Object Detection Project Description: The task is detecting instances of semantic objects of a certain class (such as humans, buildings, or cars) in images and videos. Object detection has applications in many areas of computer vision, including image retrieval and video surveillance. The challenges include change in lighting conditions, occlusions, viewpoint variance, poor video quality etc.

19	Project Name: FarmX: Leaf based disease identification in farms. Project Description: The input is a picture of a diseased leaf along with the healthy and diseased portions. The output is the name of the disease that is affecting the leaf. In this project we evaluate several machine learning techniques to (i) Identify the diseased area (We used K-Means and Gaussian Mixture) and (ii) Identify the disease (We used Linear SVM, Quadratic SVM, K-Means and LDA) by classifying among four classes of diseases. Project Ref: http://cs229.stanford.edu/proj2015/161_report.pdf
20	Project Name:Deep Learning Networks for Off-Line Handwritten Signature Recognition Project Description: Reliable identification and verification of off-line handwritten signatures from images is a difficult problem with many practical applications. This task is a difficult vision problem within the field of biometrics because a signature may change depending on psychological factors of the individual. Motivated by advances in brain science which describe how objects are represented in the visual cortex, advanced research on deep neural networks has been shown to work reliably on large image data-sets. In this paper, we present a deep learning model for offline handwritten signature recognition which is able to extract high-level representations. We also propose a two-step hybrid model for signature identification and verification improving the misclassification rate in the well-known GPDS database. Project Ref: https://www.cisuc.uc.pt/publication/showfile?fn=1366024295_Deep_Learning_Networks_for_Off-Line_Handwritten_Signature_Recognition.pdf
21	Project Title: Handwritten Digit Recognition Project Description: We will be tweaking the feature representation as well as the classifier to see their impact on performance. We will use MNIST and USPS datasets for this purpose. Implementation and testing of various classifier algorithm for digit recognition. References: 1. http://www.eecs.berkeley.edu/Pubs/TechRpts/2009/EECS-2009-159.pdf 2. http://cs229.stanford.edu/proj2014/Shaohan%20Xu,%20Qi%20Wu,%20Siyuan%20Zhang,%20Application%20of%20Neural%20Network%20In%20Handwriting%20Recognition.pdf

22	Project Title: Trend Detection Project Description: Detecting trends/breaking news on social media stream, Twitter for example. Among the flooded stream of posts on social media, the task is to identify the trending and significant posts using hashtags and popular keywords. This can lay out as identifying ones based on user input or breaking on the charts using corpus of tweets. References: TwitterMonitor: Trend Detection over the Twitter Stream, Michael Mathioudakis
23	Project Name:Restaurant Recommendation System Project Description: There are many recommendation systems available for problems like shopping, online video entertainment, games etc. Restaurants & Dining is one area where there is a big opportunity to recommend dining options to users based on their preferences as well as historical data. Yelp is a very good source of such data with not only restaurant reviews, but also user-level information on their preferred restaurants. This report describes the work to learn to predict whether a given yelp user visiting a restaurant will like it or not. I explore the use of different machine learning techniques and also engineer features that perform well on this classification. Project Ref: http://cs229.stanford.edu/proj2014/Ashish%20Gandhe,Restaurant%20Recommendation%20System.pdf
24	Project Title: Job Recommendation Engine (Recsys challenge 2016) Project Description: The RecSys Challenge 2016 is organized by XING and CrowdRec. XING is a social network for business. People use XING, for example, to find a job and recruiters use XING to find the right candidate for a job. Given a user, the goal of the job recommendation system is to predict those job postings that are likely to be relevant to the user. References: http://2016.recsyschallenge.com/
25	Project Title:Language Identification Description: Social media is a very powerful platform for understanding opinions of people. At the same time the processing of such kind of data involves many challenges. One of the challenges is its code mixed nature. Code mixing is a common phenomenon in India where people speak more than one language. Proper functioning of any Natural language processing (NLP) tool on this kind of data is not possible without pre-processing of the text which is essentially to convert this text to a standard

	format.The first step in handling such data is to do language identification. If given data has english-Hindi code mixed data,identify which word belongs to which language,using statistical methods.
26	Project Title: Efficient Web Search Ranking Project Description: Machine learning algorithms have successfully entered large-scale real-world industrial applications (like web search ranking) . Here, the CPU cost during test-time must be budgeted and accounted for . Approaches that efficiently carry out computation of features and evaluation during test time have been proposed and this project deals with understanding and implementing such algorithms . References : Baseline Dataset - http://webscope.sandbox.yahoo.com/catalog.php?datatype=c Paper-http://webscope.sandbox.yahoo.com/files/ICML2013_CSTC.pdf
27	Project Title: Named Entity Recognition Project Description: Identification of proper names in texts, and classification into a set of predefined categories of interest. Linguistic techniques as well as Statistical methods are techniques being used. References: https://aritter.github.io/rt080-ritter.pdf
28	Project Title: Web-Filtering Algorithms Project Description: Differentiating between spam and important mails, detecting a potentially malicious program, detecting pornographic or abusive contents in web pages/documents. Machine Learning and Pattern Recognition techniques are being extensively used for all these activities. References: 1. http://cs229.stanford.edu/proj2013/GopalanMathewRaghavan-Validating%20user%20spam%20reports%20in%20Chat%20Networks.pdf 2. http://cs229.stanford.edu/proj2013/ShiraniMehr-SMSSpamDetectionUsingMachineLearningApproach.pdf

29	Project Title: Mining top-k influential nodes in a dynamic network Project Description : This problem has vast application as it can be used to find team lead, how information cascaded in a network. Such a study can also be used by journalists to promote news. This is very important topic in network analysis.
30	Project Name: Emotion Classification on face images. Project Description: Humans can recognize intuitively emotions on people's faces, but computers or robots?. We also tackle the problem of emotion recognition on face images. In our case, emotion recognition is treated as a supervised classification problem.We have focused on the Bag-of-Words (BoW)and Fisher Vector (FV) representations. Project Ref: http://cs229.stanford.edu/proj2015/158_report.pdf
31	Project Title: Twitter Sentiment Analysis Project Description: Classification of tweets based on sentiment. The project should aim to use existing lexical resources as well as features that capture information about the informal and creative language used in microblogging. Expectation: A hybrid approach using both corpus based and dictionary based methods to determine the semantic orientation. References: A. Agarwal et al. *"Sentiment analysis of twitter data."* Proceedings of the Workshop on Languages in Social Media 23 Jun. 2011
32	Project Title: Predicting Tags for StackOverflow Questions Project Description: The question-answering site StackOverflow allows users to assign tags to questions in order to make them easier for other people to find. In this project the aim is developing a predictor that is able to assign tags based on the content of a question References: 1. http://cs229.stanford.edu/proj2013/SchusterZhuCheng-PredictingTagsforStackOverflowQuestions.pdf 2. http://cs229.stanford.edu/proj2014/Mihail%20Eric,%20Ana%20Klimovic,%20Victor%20Zhong,MLNLP-Autonomous%20Tagging%20Of%20Stack%20Overflow%20Posts.pdf

33	Project Title: Image Restoration using Markov Random Fields Description: Markov Random Fields are very powerful probabilistic models. As a part of this project you can explore this field and implement any one use case from scratch. Image restoration is just one particular example where MRFs have been very successful. References: http://research.ijcaonline.org/volume48/number8/pxc3880137.pdf.
34	Project Title: Human Detection (extension to videos) Project Description:In surveillance systems, it is a basic task to detect and track the presence of a human figure. Several applications like traffic monitoring, event detection etc. rely on a standard Human Detection Algorithm. The challenges include change in lighting conditions, occlusions, viewpoint variance, poor video quality etc. References (if any) : Histograms of Oriented Gradients for Human Detection "Navneet Dalal and Bill Triggs"
35	Project Title: Clothing Image Retrieval for Smarter Shopping Project Description: The aim is to develop a recommendation system that will take as input an image of clothing and output images of other items of the same clothing type that the user may also like. We will use color, texture, SIFT features, and object outline to determine similarity scores between pairs of images.
36	Project Name: Image Object Classification Project Description: Image Object Classification deals with detecting objects of a certain class (such as humans, buildings, or cars) in digital images (and videos) and also labelling the images accordingly.Image classification has applications in many areas of computer vision/robotic vision, including image retrieval and video surveillance. Ensemble classifiers produce better results than using a single classifier and the results can be compared. Project Ref: http://cs229.stanford.edu/proj2013/KrzesinskiWilder-ImageObjectClassification.pdf
37	Project Title: Image Recognition and Classification using Advanced Algorithms for Classification and Clustering. Project Description: The aim of this project is to experiment with classification and clustering techniques we learned in the class on real world problem. Original Yale face dataset grayscale face

	images that are captured under controlled conditions could be used for the purpose of testing out experiments and other results. References: http://cs229.stanford.edu/proj2013/SchoendorfElder-ImageClassification.pdf http://vision.ucsd.edu/content/yale-face-database
38	Project Name: Identifying Gender From Images of Faces. Project Description: The objective of this project is toidentify the gender of a person by looking at his/her photograph. This is a case of supervised learning where the algorithm is first trained on a set of female and male faces, and then used to classify new data. Techniques used here are SVM, NN.
39	Project Title: Gradient Boosting Factorization Machines Project Description : When asked about influential papers in machine learning, Peter Norvig mentioned Rendle's *Factorization Machines* as the first . Recommendation techniques have been well developed in the past decades. Most of them build models only based on user item rating matrix. However, in real world, there is plenty of auxiliary information available in recommendation systems. We can utilize these information as additional features to improve recommendation performance. We refer to recommendation with auxiliary information as context-aware recommendation. Context-aware Factorization Machines (FM) is one of the most successful context-aware recommendation models . In practice, there are tens of context features and not all the pairwise feature interactions are not useful . Gradient Boosting based factorization machines incorporate feature selection and solve the above problem . The task of this project is to implement GBFM from scratch and test its capabilities on various datasets . References :http://www.csie.ntu.edu.tw/~b97053/paper/Rendle2010FM.pdf http://dl.acm.org/citation.cfm?id=2645730
40	Project Title: Topic Modelling Project Description: To identify small number of topics that best characterize a document. LDA, basic topic model used for the field is further subjected to relaxation in assumptions to obtain better results. The task is to choose appropriate modelling assumptions and implement topic model suited with a new corpus.