Syllabus for PBL Program

1. Course Details

<table>
<tr><td>Title</td><td colspan="3">Machine Learning</td></tr>
<tr><td>Mode</td><td colspan="3">Online lectures and workshop</td></tr>
<tr><td>Targeted Students</td><td colspan="3">Undergraduate or advanced high school students who would like to have a career in data science. Students should have solid backgrounds in math, statistics, and a working level in Python coding.</td></tr>
<tr><td rowspan="4">Prerequisites</td><td rowspan="2">High School Students</td><td>Required course/Knowledge</td><td>Strong math, beginning Python programming</td></tr>
<tr><td>Recommended Materials for preparing for the course</td><td>Foundations of Predictive Analytics, Wu, Coggeshall. Covers the mathematical fundamentals of popular ML methods. Nice overall description of modeling processes.
The Elements of Statistical Learning, Hastie, Tibshirani, Friedman. Excellent treatise on machine learning methods.
Pattern Classification, Duda, Hart, Stork. Very nice descriptive book of basic statistical machine learning concepts.</td></tr>
<tr><td rowspan="2">College Students</td><td>Required course/Knowledge</td><td>Math through calculus, basic statistics, intermediate Python skills</td></tr>
<tr><td>Recommended Materials for preparing for the course</td><td>Foundations of Predictive Analytics, Wu, Coggeshall. Covers the mathematical fundamentals of popular ML methods. Nice overall description of modeling processes.
The Elements of Statistical Learning, Hastie, Tibshirani, Friedman. Excellent treatise on machine learning methods.
Pattern Classification, Duda,</td></tr>
</table>

			Hart, Stork. Very nice descriptive book of basic statistical machine learning concepts.

2. Program Introduction and Objectives

Introduction	● **Background** Machine Learning is the use of statistical modeling algorithms to solve practical quantitative problems around large data sets. The mainline practices are building either supervised or unsupervised algorithms that can be used for data analysis, predictions and forecasts. Machine Learning is used extensively throughout scientific and business disciplines for both research and practical solutions to common or unusual business problems. The main processes in machine learning are data exploration, analysis, cleaning, building expert variables, applying linear or nonlinear fitting algorithms and evaluation of results. There are many kinds of statistical and machine learning algorithms including linear and logistic regressions, decision trees, boosted trees, random forests, neural nets (shallow and deep), support vector machines, k nearest neighbors, Bayesian networks, and clustering algorithms. ● **Aims** 1. Understand the fundamental issues and steps in building applied machine learning algorithms 2. Understand the basic architecture and high-level workings of today's most common ML algorithms 3. Understand the important issues in building ML models including overfitting, measures of goodness, feature selection, data cleaning, and feature engineering 4. Complete a hand-on machine learning project on a practical real-world data set to solve an applied problem ● **Description** In this course we will explore all important aspects of applied machine learning (ML). We will cover the basics of the most popular ML algorithms including • Linear, logistic regression • Decision trees • Random forests • Boosted decision trees

<table>
<tr><td colspan="2"></td><td colspan="2">Neural networks
Deep learning
Support vector machines The course is divided into five weeks of lectures followed by one week of direct guidance on a hands-on machine learning project, and the following final week the students will present the results of their projects. Each week in the first five weeks we will have 90-minute lectures on the architecture and details of these machine learning algorithms, as well as other important topics in building supervised and unsupervised models. The students will choose a group project that they will execute during the next four weeks, with guidance from the instructor. It is important that the students have the ability to build and run Python code. It is recommended that each student have a Jupyter notebook environment on their laptops to execute the assignments and to complete the project. Students don’t need deep Python coding skills, but they should have the capabilities of executing, modifying and extending code. </td></tr>
<tr><td rowspan="3">Course Objectives</td><td>Theoretical</td><td colspan="2">We will learn the fundamental principals in behind the mainline machine learning algorithms at a high level. This includes the architecture, how data is handled, how the algorithm uses the data to build/learn the desired data relationships, the important user-selected hyperparameters in the models and what they do.</td></tr>
<tr><td rowspan="2">Practical</td><td>Software/Skills</td><td>Jupyter notebooks with Python</td></tr>
<tr><td>Details</td><td>The students will build, modify and execute Python notebooks for a wide range of modern machine learning algorithms on practical and real-world data sets.</td></tr>
<tr><td colspan="2">Teaching Method</td><td colspan="2">The class will be mostly lectures with open opportunity for question at any time. We will cover the basic structure and use of these mainline ML algorithms. There will be a number of homework problems requiring execution of Python ML libraries such as found in sklearn. The students will work in teams on a final project and make a presentation of their project results.</td></tr>
</table>

Program Materials	The book by Wu and Coggeshall is highly recommended. All other materials (slides) will be supplied during classes. Students will need to be able to execute Python notebooks on their own computers.

3. Program Schedule

Week		Lecture Topic	Workshop and Case Study	Assignment	Reading Materials
1	**Topic**	Basics of ML modeling, Supervised, unsupervised models. What does data look like. Overfitting. Training/testing/validation data sets. Linear and logistic regressions		Get a basic Python notebook running containing various ML algorithms	Other than the background texts, all class material will be provided during classes.
	Detail				
2	**Topic**	Nonlinear ML algorithms: Decision tree, boosted trees, random forests, neural nets, SVM		Tune hyper-parameters in several nonlinear ML algorithms. Explore training/testing performance.	
	Detail				
3	**Topic**	Clustering, curse of dimensionality, feature selection, regularization, PCA, Model measures of goodness.		Run feature selection algorithms (filter, wrapper)	
	Detail				
4	**Topic**	Data preparation, filling in missing values, outliers, feature engineering, encoding of categorical fields, fuzzy matching		Describe various ML algorithms, target encoding	
	Detail				
5	Topic	Work through several applies examples such as marketing segmentation, fraud score		Modify an existing notebook to explore a particular practical example	
	Detail				
6	**Final Project Review Week**				
7	**Final Written Reporting and Oral Presentation**				

4. Assignment Schedule

Total Number of Assignments	**5: 4 individual homeworks assigned first 4 class weeks, then 5th is group presentation during final class**	
Deadline	Homework deadlines 7 days after assignment	
Mentor is needed to review and grade assignment.	Yes (x)	No ( )
A standard answer will be provided.	Yes (x) An answer guide will be provided	No ( )

5. Requirements and Evaluations of Final Written Report and Oral Presentation

Final project is a presentation of a practical machine learning algorithm applied to a particular business problem. The project will be a team/group project with a target of 3 to 5 students per team.

5.1 Final Project:

- Final Project Theme: Apply a state-of-the-art machine learning algorithm to a practical data set to solve a particular business problem.

- Final Project Format: Powerpoint presentation

- Final Project Requirements: delivered by the team, each team member delivers part of the presentation.

5.2 Oral Presentation

- Oral Presentation Requirements: Typical presentation is powerpoint slides, with professional formatting, good English.

6. Evaluation Criteria

6.1 Attendance and Participation

6.1.1 Attendance and class participation account for 20% of the final marks.
6.1.2 Students who attend fewer than 2 times (including 2 times) will be disqualified for letters of evaluation.
6.1.3 Other requirements:
Students are expected to attend and participate in every session of the program.
They are expected to answer questions posed by the instructor during the class and ask questions if something doesn't make sense.

6.2 Assignment

6.2.1 Homework assignments account for 40% of the final marks.
6.2.2 Students who don't submit assignments will be disqualified for letters of evaluation.
6.2.3 Other requirements: the final presentation, described below.

6.3 Final Presentation

6.3.1 This final project in total accounts for 40% of the final marks, in which, written presentation accounts for 20% and oral presentation accounts for 20% of the final marks.
6.3.2 Students who don't submit final written presentation before the deadline or who don't make oral presentation will be disqualified for letters of evaluation.
6.3.3 Other requirements (none)

7. Suggested Future Research Fields/Direction/Topics

Possible areas of future work for interested students:

- Feature selection algorithms
- Dealing with imbalanced data
- Methods for encoding text (NLP algorithms
- Methods for encoding images
- Generative adversarial networks
- Bayes networks using 2-d interactive distributions
- Reinforcement learning
- Semi-supervised learning
- Methods for model interpretability
- Stacking vs single pass modeling
- Nonlinear methods for forecasting: embedding methods

8. Instructor Introduction

8.1 Instructor Title: Professor

8.2 Instructor Bio

Dr. Stephen Coggeshall is the retired Chief Analytics and Science Officer at ID Analytics, an identity fraud protection company owned by LifeLock and Symantec. He was the founding CTO of ID Analytics where he built the analytics team and helped design the technical solution approach. Prior to ID Analytics Dr. Coggeshall worked for 11 years as a researcher in nuclear fusion at the Los Alamos National Laboratory. In addition to ID Analytics, Dr. Coggeshall also cofounded the analytics consulting companies CASA (acquired by HNC Software/FICO) and Los Alamos Computational Group (acquired by Morgan Stanley). His expertise is in forming and managing teams of data scientists to attack complex business problems using advanced algorithms on big data. He has deep expertise in consumer behavior modeling, optimization, forecasting and financial engineering, and spent the past 15 years focusing on identity fraud dynamics.

Dr. Coggeshall holds undergraduate degrees in math, music and physics. He has a master’s in music and a master’s and Ph.D. in nuclear engineering from the University of Illinois. He currently is a Professor at USC and UCSD teaching classes on Fraud Analytics and Business Analytics.

8.3 Instructor Profile Photo