Course Name: Mining Massive Data Sets (3-1-0-4-4)

Unit 1 - Basics
Introduction to machine learning - Bayesian Decision Theory, MLE; Feature Selection - Information Gain, Derived features: PCA, LDA & PLS, Aggregation/Coding; Handling Missing Values; Class Imbalance: Differential loss functions, differential slack, Sampling based approaches.

Unit 2 – Ensemble Methods
Boosting - Adaboost, LogitBoost, Gradient Boosting; Bagging - Simple methods, Random Forest, Subsampling; Stacking; Decision Fusion.

Unit 3 – Scaling Up
Map-Reduce; Locality Sensitive Hashing; Large Scale optimization; Page Rank on large graphs.

Unit 4 – Clustering
Hierarchical Clustering; k-means; CURE; DBScan; Non-Eucledian spaces; Spectral clustering

Unit 5 – Frequent Pattern Mining
Apriori Algorithm; Large datasets; Limited-pass algorithms; Frequent sub-sequence mining

Unit 6 – Stream Mining
The stream data model, Sampling Data in a stream, Filtering Streams, Bloom Filter, Counting in streams,

Text Books

1. Mining Massive Datasets. Anand Rajaraman, Jeffrey D. Ullman, and Jure Leskovec. Cambridge.

Practicals (6 hours per week * 12 weeks)
Based on Units 1 - 6

Outside class (6 hours per week * 12 weeks)

References:

1. Elements of Statistical Learning. Hastie, Tibshirani, and Friedman. Springer.
2. Pattern Recognition and Machine Learning. Christopher Bishop.
3. Data Mining: Tools and Techniques, 3rd Edition. Jiawei Han and Michelline Kamber.

Video Lectures: Will be announced in the class.