Scalable Data Science

SWAYAM Prabha Course Code: R17

PROFESSOR'S NAME	Dr. Anirban Dasgupta and Dr. Sourangshu Bhattacharya
DEPARTMENT	Computer Science and Engineering
INSTITUTE	IIT Kharagpur
COURSE OUTLINE	Consider the following example problems: One is interested in computing summary statistics (word count distributions) for a set of words which occur in the same document in entire Wikipedia collection (5 million documents). Naive techniques, will run out of main memory on most computers. One needs to train an SVM classier for text categorization, with unigram features (typically ~10 million) for hundreds of classes. One would run out of main memory, if they store uncompressed model parameters in main memory. One is interested in learning either a supervised model or nd unsupervised patterns, but the data is distributed over multiple machines. Communication being the bottleneck, naïve methods to adapt existing algorithms to such a distributed setting might perform extremely poorly. In all the above situations, a simple data mining / machine learning task has been made more complicated due to large scale of input data, output results or both. In this course, we discuss algorithmic techniques as well as software paradigms which allow one to develop scalable algorithms and systems for the common data science tasks. **Course Outline** 1. Background: Introduction | Probability: Concentration inequalities | Linear algebra: PCA, SVD |Optimization: Basics, Convex, GD | Machine Learning: Supervised, generalization, feature learning, clustering. 2. Memory efficient data structures: Hash functions, universal / perfect hash families | Bloom filters | Sketches for distinct count | Misra-Gries sketch | Statistical Mechanics an overview. 3. Memory efficient data structures (contd.): Count Sketch, Count-Min Sketch | Approximate near neighbors search: Introduction, kd-trees etc | LSH families, MinHash for Jaccard, SimHash for L2. 4. Approximate near neighbors search: Extensions e.g. multi-probe, b-bit hashing, Data dependent variants | Randomized Numerical Linear Algebra

	Random projection. 5. Randomized Numerical Linear Algebra CUR Decomposition | Sparse RP, Subspace RP, Kitchen Sink. 6. Map-reduce and related paradigms Map reduce - Programming examples - (page rank, k-means, matrix multiplication) | Big data: computation goes to data. + Hadoop ecosystem. 7. Map-reduce and related paradigms (Contd.) Scala + Spark 8. Distributed Machine Learning and Optimization: ADMM + applications | Clustering | Conclusion