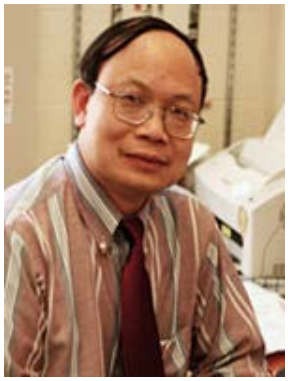
Title: Kernel Machine Learning for Big Data



S.Y. Kung,

Princeton University

The intensive computing need in big data will undoubtedly necessitate special hardware and software technologies for high performance (parallel and/or distributed) systems whose architectural platform must be closely dependent upon a novel "big-data" algorithmic paradigm. While not yet well-defined, big data is known to be characterized by 3Vs: Volume, Variety and Velocity. This talk shall explore the 3Vs from a kernel learning perspective.

Regarding the "volume" of data, there are two separate issues (1) large training data size and (2) high feature dimensionality. As to large data size, we shall review various (statistical and algebraic) approaches. Examples include: divide-and-conquer; K-means to handle the data partitioning, and selection criterion base on kernel matrix. The typical answer to high feature dimensionality is dimension reduction, which is being used as an effective antidote to counteract two feature-dimension-related problems: computation costs and data over-training. For unsupervised learning scenarios, a classical reduction method is Principal Component Analysis (PCA) . We shall show that PCA's trace-norm optimization can be extended to supervised learning applications. More exactly, by incorporating the SNR metric (of Fisher Discriminant Analysis) into the formulation, we can derive DCA (Discriminant Component Analysis) which may be viewed as the supervised learning counterpart of PCA.

The second V-issue (variety) is inevitable for big data, which by definition has many divergent types of sources, from physical (sensor/IoT) to social and cyber (web) types. Some of the data may be fuzzy, unreliable, or heterogeneously formatted. In more severe scenario, the data could be defective, messy, and partially missing. This prompts a relatively new application paradigm of incomplete data analysis (IDA). For big data, it is inevitable to encounter missing/defective entries in the columns or rows of the original data matrix. Consequently, if the traditional ``total availability" criterion were adopted too many defective columns or rows would be discarded or, equivalently, too few be retained for learning analysis. It then makes sense

In order to maximize the data utilization, the ``total availability" should be replaced by a less restrictive notion of ``pairwise availability", leading to Kernel Approach to Incomplete Data Analysis (KAIDA). KAIDA focuses on deriving correlation between data entries co-existent in both partial vectors in each pair. It is our opinion that imputation amidst highly sparse data tends to prone to uncertainty and other adverse effects. Thus, we shall advocate a non-imputed kernel approach. Furthermore, experimental results will demonstrate the strong resilience of the proposed approach against high data sparsity.

The third V-issue (velocity) is already partially addressed by the aforementioned techniques using dimension reduction, data partitioning, and selection criterion. Nevertheless, it is worth noting that the data size N tends to be enormously large for big data, and, consequently, kernel learning may be better performed in its intrinsic space rather than its empirical space. For example, the complexity of SVM learning in the empirical space will be of the order N^2, while the complexity of kernel ridge regression (KRR) in the intrinsic space grows linearly with N, which is clearly the best possible scenario.

This talk will give a balanced coverage between the theoretical foundation and practical consideration.