PGDBD-Core course II

IT 602 Data analytics and modeling

Review of basic concepts in probability and mathematical statistics. Graph and network theory concepts.

Data Mining basic concepts and nature of the data mining problems.
Data summarization and visualization.
Data preprocessing: Data cleaning. Data integration, Data reduction. Data transformation and discretization.
Data classification techniques: Decision tree, Bayesian classification.
Performance measures and improvement of classification techniques. Ensemble learning.
Bagging boosting and random forests.
Advanced classification: Bayesian belief networks, Neural networks, Support vector machines
Regression techniques. Multiple linear regression, Regularization. Ridge regression and LASSO. Elastic Nets.
Mixture models.
Elementary ideas of Deep learning. Boltzman machine, CNN, LSTM and other variants.

Cluster analysis of large scale data sets: partitioning methods, hierarchical clustering. Clustering graphs and network data

Practicals:

1. Re-implementing selected best machine learning tasks for selected applications in benchmark data sets such as in UCI and Kaggle.
2. Writing practical codes to solve selected biological problems using traditional machine learning algorithms such as SVM, MLR and neural networks.
3. Writing basic codes to implement CNN and LSTM in Tensorflow framework.
4. Assessing scalability and overhead in parallelization of simple tasks and one case study with advanced problem.