CAP 6780 Big Data Analytics with Hadoop

Credits: 3 credits

Textbook, title, author, and year: (1) Data Mining: Practical Machine Learning Tools and Techniques, by I.H. Witten and E. Frank (2) Selected articles and papers are posted on the course web site.

Reference materials: None

Specific course information:

Catalog description:
The study of topics in data mining and machine learning relating to Big Data. Big Data challenges such as high dimensionality, class imbalance, quality of data, etc. will be examined and addressed. Hands-on experience with Big Data analysis in Hadoop using a high performance computing cluster.

Prerequisites: Graduate standing or permission of instructor

Specific goals for the course:

Students will learn data mining and machine learning techniques for Big Data with Hadoop. Hands-on Big Data analysis using a high performance computing cluster. Case studies with an emphasis on real world applications will be presented

Brief list of topics to be covered:
1. Introduction to Hadoop and the Hadoop ecosystem
2. Hadoop tools for data mining and machine learning
3. High performance cluster computing with Hadoop
4. Challenges with learning from Big Data
5. Quality of Data
6. High Dimensionality
7. Feature Selection
8. Class Imbalance
9. Data Sampling
10. Ensemble Learning
11. Applications and Trends in Data Analytics with Hadoop
12. Case Studies