CS 84010 Big Data Analytics

Fall 2013 - Hanghang Tong

Course Information

We are in the age of big data - data that is both large and complex. Big data analytics is the science of analyzing the data, generating insights, and making decisions by addressing all three dimensions of the big data challenges including variety, velocity and volume. It is essential behind many high impact real applications, such as social networks analysis, finance and business intelligence, climate modeling, health care, political science and so on.

This class aims to provide a comprehensive overview of recent advance in machine learning and data mining to analyze big data. Selected topics include big data clustering and classification, anomaly and fraud detection, time-series analysis, big graph mining, and massive-scale data analytics; as well as case studies in social networks analysis, healthcare, business intelligence, etc.

- Instructor: Hanghang Tong (tong at ccny dot cuny dot edu)
  - GC 4420, x8196
  - Office hour: 12:45-1:45pm Monday, 4420
- Class meets: Mon 2-4pm, 3212
- Late policy:
  - each person has 2 slip days in total for the whole semester. After that, 20% deduction per day of delay
  - no penalty if medical emergence (need doctor’s notes)
- Big Data Seminar (10/1st, bi-weekly afterwards): Science Center, room 4102

Schedule

<table>
<thead>
<tr>
<th>Lecture</th>
<th>Date</th>
<th>Topics</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Sep 2</td>
<td>Labor day, no class</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Sep 9</td>
<td>Introduction + Basic Concepts</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Recommended reading</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1) “Challenges and Opportunities with Big Data”</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Sep 16</td>
<td>Classification 1 (Bayes Classifier, kNN) + Candidate projects</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Recommended reading</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1) Theoretic Properties of knn</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2) K-d-tree tutorial</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Sep 23</td>
<td>Classification 2 – Indexing, Logistic Regression</td>
<td></td>
</tr>
</tbody>
</table>
### Recommended reading

1. **Logistic Regression with SGD**

### Sep 30

**Recommended reading**

1. **SVM_Perf**
2. **Pegasos**

**Classification 3 - SVM**

**Proj. proposal proposal**

**seminar reading 1 due**

### Oct 7

**Clustering 1 – Kmeans and PCA**

**seminar reading 2 due**

### Oct 14

**Columbus day, no class**

### Oct 15

**Oct 14 schedule: Web Fraud Detection (Tim Pan)**

### Oct 21

**Clustering -2 – LSH, GMM and spectral clustering**

**seminar reading 3 due**

### Oct 28

**Midterm exam**

### Nov 4

**Time Series**

**Recommended reading**

1. **DnS algorithm**
2. **Spirit**

**Proj midterm**

### Nov 11

**No class, Prof. Terzi’s seminar moved to Sep. 20th**

**reading 1 due**

### Nov 18

**Big Graph Mining 1 – Co-clustering**

**Recommended reading**

1. **Co-clustering**
2. **Cross-association**

### Nov 25

**Big Graph Mining 2 – Low-rank approximation**

1. **Colibri**
2. **NMF**

**reading 2 due**

### Nov 28

**Big Graph Mining 3 – Pattern, Dissemination and Proximity**

**Mining Rare Events from Big Data (Jingrui He)**

**reading 3 due**

### Dec 11

**project presentation at noon, 4421**

### Dec 12

**Project final report due**

---

### Big Data Seminar

<table>
<thead>
<tr>
<th>Seminar</th>
<th>Date</th>
<th>Speaker</th>
<th>Topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Sep. 20th</td>
<td>Evimaria Terzi (Boston)</td>
<td>Entity Selection and Ranking in Data Mining Applications Related paper</td>
</tr>
<tr>
<td>2</td>
<td>Oct. 1</td>
<td>Fei Wang (IBM)</td>
<td>Feature Engineering for Predictive Modeling with Large Scale Electronic Medical Records: Augmentation, Densification and Selection Related Papers (a); (b)</td>
</tr>
<tr>
<td>3</td>
<td>Oct. 15</td>
<td>Tim Pan (Google)</td>
<td>Click Fraud - Challenges and Remedies Related Papers: (a) - TSum: Fast, Principled Table Summarization (b) - The Goals and Challenges of Click Fraud Penetration Testing Systems</td>
</tr>
<tr>
<td>4</td>
<td>Nov. 5</td>
<td>Han Liu (Princeton)</td>
<td>From High Dimensional Data to Big Data</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Related Paper</td>
</tr>
<tr>
<td>----</td>
<td>------</td>
<td>-------------------------</td>
<td>-------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>5</td>
<td>Nov. 12</td>
<td>Ruoming Jin (Kent)</td>
<td>Finally, Simple, Fast and Scalable Reachability Oracle!</td>
</tr>
<tr>
<td>6</td>
<td>Dec 10</td>
<td>Tao Li (FIU)</td>
<td>Learning to Understand Documents</td>
</tr>
</tbody>
</table>

(a) [Challenges of big data analysis](#)  
(b) [Huge Package](#)