Data Science/Data Analytics and Scaling to Big Data with MathWorks

Using Data Analytics to turn large volumes of complex data into actionable information can help you improve design and decision-making processes. However, developing effective analytics and integrating them into business systems can be challenging. In this seminar you will learn approaches and techniques available in MATLAB® to tackle these challenges.

**Data Science/Data Analytics (Part 1) — Fri, Feb 6, 2-3pm | Klaus 1116**

* **Access and Explore data**: Accessing, exploring, and analyzing data stored in files, the web, and databases
* **Data Munging**: Techniques for cleaning, exploring, visualizing, and combining complex multivariate data sets
* **Developing predictive models**: Prototyping, testing, and refining predictive models using machine learning methods
* **Integrating analytics with systems**: Integrating and running analytics within enterprise business systems and interactive web applications

**Scale to Big Data (Part 2) — Fri, Feb 6, 3-4pm | Klaus 2443**

* Work with out-of-memory datasets with MATLAB
* MapReduce algorithms and Hadoop integration in MATLAB
Data Mining Concepts & Tasks

Duen Horng (Polo) Chau
Georgia Tech

Partly based on materials by Professors Guy Lebanon, Jeffrey Heer, John Stasko, Christos Faloutsos
Final words about

Data Integration
Freebase
(a graph of entities)

“…a large collaborative knowledge base consisting of metadata composed mainly by its community members…”

Wikipedia.
Crowd-sourcing Approaches: Freebase

[Image of Freebase website]

http://wiki.freebase.com/wiki/What_is_Freebase%3F
We need ways to identify the many ways that one thing may be called. How?
Entity Resolution
(A hard problem in data integration)

Polo Chau
P. Chau
Duen Horng Chau
Duen Chau
D. Chau
Why is Entity Resolution so Important?
Find Your iPhone

iPhone 5s
- 4" Retina Display
- True Tone Flash
- Slow-mo Video
- Touch ID
Shop iPhone 5s

iPhone 5c
- 4" Retina Display
- 8 MP Camera
- 1080p HD Video
Shop iPhone 5c

iPhone 5
- 4" Retina Display
- 8 MP Camera
- 1080p HD Video
- Face Time, Siri
Shop iPhone 5

Apple iPhone 4S A1387, Sprint, 16GB, White, Clean ESN
$54.00
2 bids
2m left (Today 7:53AM)

Apple iPhone 4 - 8GB - Verizon Straight Talk...
$109.95
Buy It Now
Free shipping

Apple iPhone 4 - 8GB - Black (Verizon)...  
$78.95
Buy It Now
Free shipping

Apple iPhone 4 - 16GB - Black (Verizon) Smartphone - Black or White - Good
USA SELLER *** WARRANTY *** ACCESSORIES INCLUDED
$79.88
Buy It Now
Free shipping

Apple iPhone 4 - 16GB - Black (Verizon) Smartphone 7.1.2 MC676LL/A Clean ESN
$79.00
0 bids
2m left (Today 7:54AM)

Apple iPhone 4 - 8GB - White (Verizo...
$79.95
Buy It Now
Free shipping
D-Dupe

Interactive Data Deduplication and Integration
TVCG 2008

University of Maryland
Bilgic, Licamele, Getoor, Kang, Shneiderman

http://www.cs.umd.edu/projects/linqs/ddupe/ (skip to 0:55)
Numerous similarity functions

Excellent read: http://infolab.stanford.edu/~ullman/mmds/ch3a.pdf

- Euclidean distance
  Euclidean norm / L2 norm

- Manhattan distance

- Jaccard Similarity
  e.g., overlap of nodes’ #neighbors

\[ \text{Jaccard similarity of sets } S \text{ and } T \text{ is } \frac{|S \cap T|}{|S \cup T|} \]

- String edit distance
  e.g., “Polo Chau” vs “Polo Chan”

- Canberra distance
Core components: Similarity functions

Determine how two entities are similar.

D-Dupe’s approach:

**Attribute similarity + relational similarity**

\[
sim(e_i, e_j) = (1 - \alpha) \times sim_A(e_i, e_j) + \alpha \times sim_R(e_i, e_j),
\]

\[
0 \leq \alpha \leq 1,
\]

Similarity score for a pair of entities
Attribute similarity (a weighted sum)

\[ sim_A(e_i, e_j) = \sum_{k=1}^{n} w_k \times sim_{-\text{fun}}(e_i \cdot a_k, e_j \cdot a_k), \]

\[-1 \leq w_k \leq 1 \quad \text{and} \quad \sum_{k=1}^{n} |w_k| = 1,\]
Data Mining Concepts & Tasks
A critical skill in data science is the ability to decompose a data-analytic problem into pieces such that each piece matches a known task for which tools are available. Recognizing familiar problems and their solutions avoids wasting time and resources reinventing the wheel. It also allows people to focus attention on more interesting parts of the process that require human involvement—parts that have not been automated, so human creativity and intelligence must come into play.
1. Classification or class Probability Estimation

Predict which of a (small) set of classes an entity belong to.

- Cancer testing (yes, no)
- Movie genre (action, drama, etc.)
- Sports (win, loss)
- Email spam filter (spam, or not)
- Gesture detection (pinch, swipe…)
- Planet zone habitable or not
- Gene prediction
- News types (sports, entertainment)
- Virus scanning (malware or not)
2. Regression ("value estimation")

Predict the **numerical value** of some variable for an entity.

- point value of wine (50-100)
- credit score (start with classification; default or not)
- stock prices — wall street
- relationship between price and sales
- weather
- sports and game scores
3. Similarity Matching

Find similar entities (from a large dataset) based on what we know about them.

• recommending items you may want to buy
• find similar gene sequences (that may be repeating, or does similar things)
• online dating
• building auditing (energy consumption)
• patent search
• carpool matching (find people to carpool)
• detecting fake identities
4. Clustering (unsupervised learning)

Group entities together by their similarity. (User provides # of clusters)

- cluster people into demographics groups (young, old, etc)
- cluster people by accents (y’all, you all)
- hierarchical clustering for metabolomics
- clustering images on the web (cat?)
- ~ = dimensionality reduction
5. Co-occurrence grouping

(Many names: frequent itemset mining, association rule discovery, market-basket analysis)

Find associations between entities based on transactions that involve them (e.g., bread and milk often bought together)

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

http://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/
6. Profiling / Pattern Mining / Anomaly Detection (unsupervised)

Characterize *typical* behaviors of an entity (person, computer router, etc.) so you can find *trends* and *outliers*.

Examples?
- computer instruction prediction
- removing noise from experiment (data cleaning)
- detect anomalies in network traffic
- moneyball
- weather anomalies (e.g., big storm)
- google sign-in (alert)
- smart security camera
- embezzlement
- trending articles
7. Link Prediction / Recommendation

Predict if two entities should be connected, and how strongly that link should be.

linkedin/facebook: people you may know

amazon/netflix: because you like terminator… suggest other movies you may also like
8. Data reduction ("dimensionality reduction")

Shrink a large dataset into smaller one, with as little loss of information as possible

1. if you want to visualize the data (in 2D/3D)
2. faster computation/less storage
3. reduce noise
Start Thinking About Project!

- What problems do you want to solve?
- Using what (large) datasets?
- What techniques do you need?