Data Mining

Jeff M. Phillips

January 8, 2014

Data Mining

What is Data Mining?

- Finding structure in data?
- Machine learning on large data?
- Unsupervised learning?
- Large scale computational statistics?

Data Mining

What is Data Mining?

- Finding structure in data?
- Machine learning on large data?
- Unsupervised learning?
- Large scale computational statistics?

- What you can recover from data and what you cannot recover.
- Algorithms for how to recover it efficiently.

Data Mining

What is Data Mining?

- Finding structure in data?
- Machine learning on large data?
- Unsupervised learning?
- Large scale computational statistics?

- What you can recover from data and what you cannot recover.
- Algorithms for how to recover it efficiently.

- How to think about data analytics.

Modeling versus Efficiency

Two Intertwined Objectives:

- Model Data Correctly
- Process Data Efficiently

Other Data Mining Course

Every university teaches data mining differently!

Other Data Mining Course

Every university teaches data mining differently!

What flavor is offered in this class:

- Focus on techniques for *very* large scale data
- Broad coverage ... with recent developments
- Formally and generally presented (proof sketches)
- ... but useful in practice (e.g. internet companies)
- Probabilistic algorithms: connections to CS and Stat

Other Data Mining Course

Every university teaches data mining differently!

What flavor is offered in this class:

- Focus on techniques for *very* large scale data
- Broad coverage ... with recent developments
- Formally and generally presented (proof sketches)
- ... but useful in practice (e.g. internet companies)
- Probabilistic algorithms: connections to CS and Stat
- *not specific software tools / programming languages*

Other Data Mining Course

Every university teaches data mining differently!

What flavor is offered in this class:

- Focus on techniques for *very* large scale data
- Broad coverage ... with recent developments
- Formally and generally presented (proof sketches)
- ... but useful in practice (e.g. internet companies)
- Probabilistic algorithms: connections to CS and Stat
- *not specific software tools / programming languages*

Maths: Linear Algebra, Probability, High-dimensional geometry

Outline

Statistical Principals:

- 1. **Understanding random effects**

Data and Distances:

- 2. **Similarity** (find duplicates and similar items)
- 3. **Clustering** (aggregate close items)

Structure in Data:

- 3. **Clustering** (aggregate close items)
- 4. **Regression** (patterns in data)
- 5. **Noisy Data** (anomalies in data)

Controlling for Noise and Uncertainty:

- 5. **Noisy Data** (anomalies in data)
- 6. **Link Analysis** (prominent structure in large graphs)
- 7. **Summaries** (concise representation)

Statistical Principals

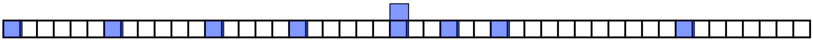
What happens as data is generated with replacement
{IP addresses, words in dictionary, edges in graph, hash table}

- When do items collide?
- When do you see all items?
- When is the distribution almost uniform?

Statistical Principals

What happens as data is generated with replacement
{IP addresses, words in dictionary, edges in graph, hash table}

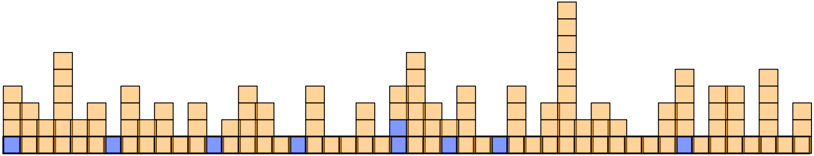
- When do items collide?
- When do you see all items?
- When is the distribution almost uniform?



Statistical Principals

What happens as data is generated with replacement
{IP addresses, words in dictionary, edges in graph, hash table}

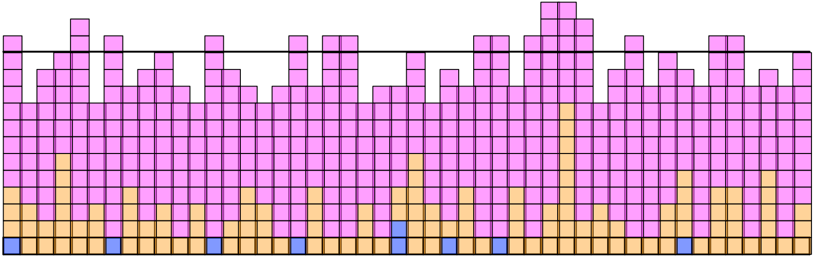
- When do items collide?
- When do you see all items?
- When is the distribution almost uniform?



Statistical Principals

What happens as data is generated with replacement
{IP addresses, words in dictionary, edges in graph, hash table}

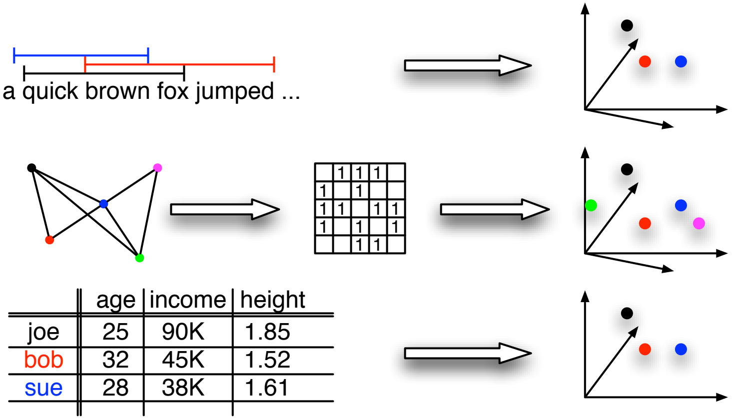
- When do items collide?
- When do you see all items?
- When is the distribution almost uniform?



Raw Data to Abstract Representations

How to measure similarity between data?

Key idea: data $\rightarrow$ point

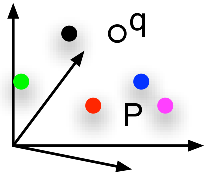


	age	income	height
joe	25	90K	1.85
bob	32	45K	1.52
sue	28	38K	1.61

Similarity

Given a large set of data P.
Given new point q, is q in P?

Given a large set of data P.
Given new point q, what is *closest* point in P to q?



Clustering

How to find groups of similar data.

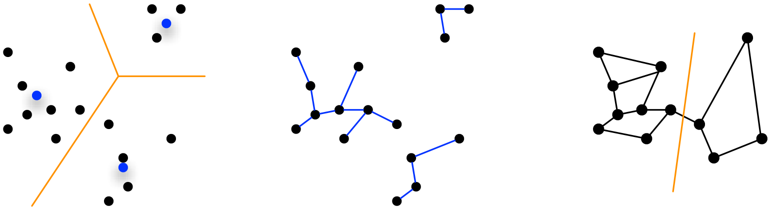
- do we need a representative?
- can groups overlap?
- what is structure of data/distance?

Clustering

How to find groups of similar data.

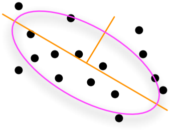
- do we need a representative?
- can groups overlap?
- what is structure of data/distance?

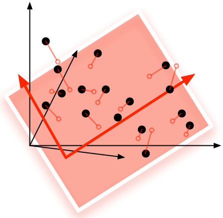
- **Hierarchical clustering** : When to combine groups?
- ***k*-means clustering** : k-median, k-center, k-means++
- **Graph clustering** : modularity, spectral



Regression

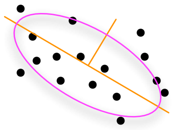
Consider a data set $P \in \mathbb{R}^d$, where d is BIG!
Want to find representation of P in some $\mathbb{R}^k$
$\mu(P) \to Q \in \mathbb{R}^k$ so $\|p_i - p_j\| \approx \|q_i - q_j\|$
$Q \in \mathbb{R}^k$ should capture most data in P

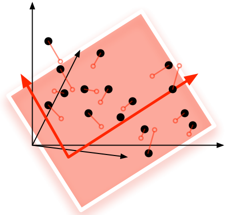




Regression

Consider a data set $P \in \mathbb{R}^d$, where d is BIG!
Want to find representation of P in some $\mathbb{R}^k$
$\mu(P) \to Q \in \mathbb{R}^k$ so $\|p_i - p_j\| \approx \|q_i - q_j\|$
$Q \in \mathbb{R}^k$ should capture most data in P



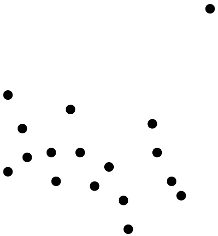


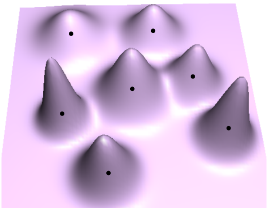
- L_2 **Regression + PCA** : Common easy approach
- **Multidimensional Scaling** : Fits in $\mathbb{R}^k$ with k small
- **Random Projections** : Faster and easier (different bounds)
- L_1 **Regression** : "Better", Orthogonal Matching Pursuit
- **Special Topic** : Compressed Sensing

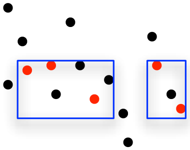
Noisy Data

What to do when data is noisy?

- **Identify it** : Find and remove outliers
- **Model it** : It may be real, affect answer
- **Exploit it** : Differential privacy *(ethics in data)*



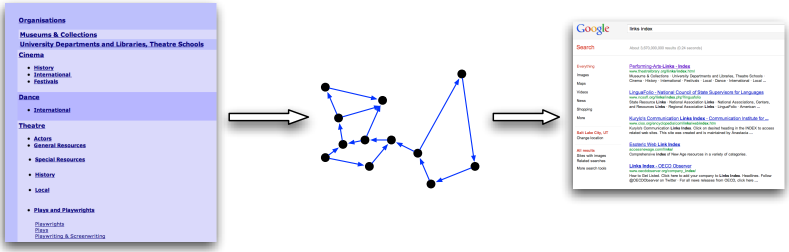




Link Analysis

How does Google Search work?
Converts webpage links into directed graph.

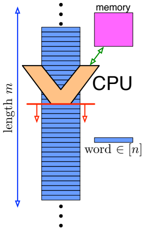
- **Markov Chains** : Models movement in a graph
- **PageRank** : How to convert graph into important nodes
- **MapReduce** : How to scale up PageRank
- **Communities** : Other important nodes in graphs

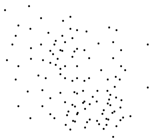
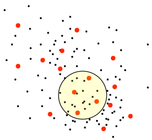
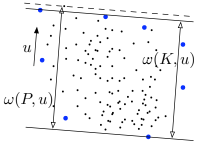


Summaries

Reducing *massive* data to small space.
Want to retain as much as possible (not specific structure)
error guarantees

- **OnePass Sampling** : Reservoir Sampling
- **Density Approximation** : Quantiles
- **MinCount Hash** : Sketching data, $\rightarrow$ abstract features
- **Spanners** : graph approximations
- **[...]** : ... on request ...



Themes

What are course goals?

- Intuition for data analytics
- How to model data (convert to abstract data types)
- How to process data efficiently (balance models with algorithms)

Themes

What are course goals?

- Intuition for data analytics
- How to model data (convert to abstract data types)
- How to process data efficiently (balance models with algorithms)

Work Plan:

- 2-3 weeks each topic.
  - Overview classic techniques
  - Focus on modeling / efficiency tradeoff
  - Special topics
  - Short homework for each (analysis + with data)
- Course Project (1/2 grade).
  - Focus on specific data set
  - Deep exploration with technique
  - Ongoing refinement of presentation + approach

On Homeworks

Managed through Canvas (will be up by end of week)

- No restriction on programming language.
- Some designed for matlab, others better in python or C++.
- Programming assignments with not too many specifications.
- Bonus Questions!

Data Group

Data Group Meeting
Thursdays @ 12:15-1:30 in MEB 3147 (LCR)

CS 7941 *Data Reading Group*
requires one presentation if taken for credit

http://datagroup.cs.utah.edu/dbgroup.php

Data Group

Data Group Meeting

Thursdays @ 12:15-1:30 in MEB 3147 (LCR)

CS 7941 *Data Reading Group*

requires one presentation if taken for credit

`http://datagroup.cs.utah.edu/dbgroup.php`

Tomorrow: Qin Zhang (Indiana University)

Subspace Embeddings and L_p-Regression Using Exponential Random Variables