Data Analytics 1

Vikram Pudi
vikram@iiit.ac.in
IIIT Hyderabad

Data Systems Evolution

- Traditional Database Systems
  - Indexing
  - Query languages
  - Query optimization
  - Transaction processing
  - Recovery …
- Relational / SQL

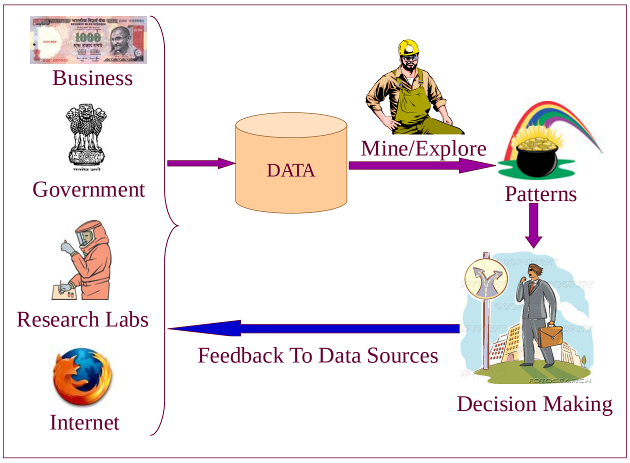
Post-Relational Revolution

- New organizations of data
  - Object oriented (OO) [Zope] and object-relational (OR) systems [SQLAlchemy]
  - Semi-structured [XML, JSON] and unstructured data (Text)
  - Vertical / Column stores [Cassandra]
  - Unnormalized relations: Document Databases [MongoDB]
  - Key-value Stores [Redis]
  - Graph Databases [Neo4j]
- New functionality
  - Distribution & Heterogeneity (multi-databases, interoperability)
  - Active databases (triggers) and deduction
  - ERP packages (application-oriented tasks common to many organizations)
  - ***Data analysis*** *(*data warehouses and ***data mining****)*
- More complex data domains (e.g., design, geography, molecular biology, social networks)
- Relaxation of ACID test for DBMS

Data Mining

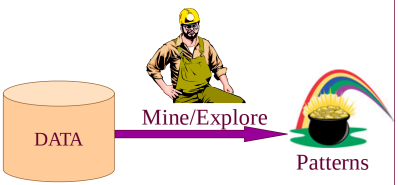
= Automated discovery of *interesting patterns* in large datasets

- Researchers identified several kinds of interesting patterns in an adhoc manner
  - classification and regression models, clusters, association rules, frequent patterns, sequential patterns, time-series patterns, summaries, cyclic patterns, hierarchical patterns, max-patterns, closed patterns, multi-dimensional patterns, etc.



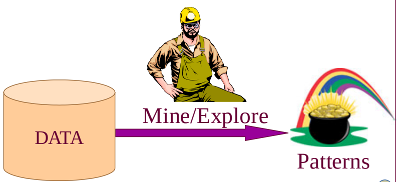
Data Science / KDD Life-cycle

- Domain understanding
- Data Preprocessing
  - Data integration
  - Cleaning
  - Selection
  - Transformation
- *Data Mining*
- Post Mining
  - Presentation / Visualization
  - Evaluation
  - Decision making



Data Science / KDD Life-cycle

- Domain understanding
- Data Preprocessing
  - Data integration
  - Cleaning
  - Selection
  - Transformation
- *Data Mining*
- Post Mining
  - Presentation / Visualization
  - Evaluation
  - Decision making



DATA

Mine/Explore

Patterns

Many knowledge discovery applications need *exact*, *interpretable* knowledge for decision making

Types of Patterns

- Associations
  - *Coffee* buyers usually also purchase *sugar*
- Clustering
  - Segments of customers requiring different promotion strategies
- Classification
  - Customers expected to be *loyal*

Association Rules

That which is infrequent is not worth worrying about.

Association Rules

D :

Transaction ID	Items
1	Tomato, Potato, Onions
2	Tomato, Potato, Brinjal, Pumpkin
3	Tomato, Potato, Onions, Chilly
4	Lemon, Tamarind

Rule: Tomato, Potato   Onion (confidence: 66%, support: 50%)

Support(*X*) = |transactions containing *X*| / |D|
Confidence(*R*) = support(R) / support(LHS(R))

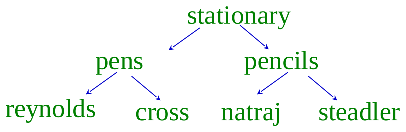
Problem proposed in [AIS 93]: Find all rules satisfying user given minimum support and minimum confidence.

Association Rule Applications

- E-commerce
  - People who have bought *Sundara Kandam* have also bought *Srimad Bhagavatham*
- Census analysis
  - *Immigrants* are usually *male*
- Sports
  - A chess end-game configuration with “*white pawn on A7*” and “*white knight dominating black rook*” typically results in a “*win for white*”.
- Medical diagnosis
  - Allergy to *latex rubber* usually co-occurs with allergies to *banana* and *tomato*

Types of Association Rules

- Boolean association rules
- Hierarchical rules



stationary

pens   pencils

reynolds   cross   natraj   steadler

reynolds   pencils

Quantitative & Categorical rules

(Age: 30…39), (Married: Yes)   (NumCars: 2)

More Types of Association Rules

- Cyclic / Periodic rules
  - Sunday   vegetables
  - Christmas   gift items
  - Summer, rich, jobless   ticket to Hawaii
- Constrained rules
  - Show itemsets whose average price > Rs.10,000
  - Show itemsets that have television on RHS
- Sequential rules
  - Star wars, Empire Strikes Back   Return of the Jedi

Classification

To be or not to be: That is the question.

- William Shakespeare

The Classification Problem

Outlook	Temp (F)	Humidity (%)	Windy?	Class
sunny	75	70	true	*play*
sunny	80	90	true	*don't play*
sunny	85	85	false	*don't play*
sunny	72	95	false	*don't play*
sunny	69	70	false	*play*
overcast	72	90	true	*play*
overcast	83	78	false	*play*
overcast	64	65	true	*play*
overcast	81	75	false	*play*
rain	71	80	true	*don't play*
rain	65	70	true	*don't play*
rain	75	80	false	*play*
rain	68	80	false	*play*
rain	70	96	false	*play*

sunny	77	69	true	?
rain	73	76	false	?

Play Outside?

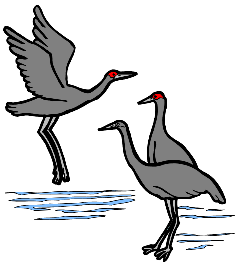
Model relationship between class labels and attributes

e.g. outlook = overcast   class = *play*

Assign class labels to new data with *unknown* labels

Applications

- Text classification
  - Classify emails into spam / non-spam
  - Classify web-pages into yahoo-type hierarchy
  - NLP Problems
    - Tagging: Classify words into verbs, nouns, etc.
- Risk management, Fraud detection, Computer intrusion detection
  - Given the properties of a transaction (items purchased, amount, location, customer profile, etc.)
  - Determine if it is a fraud
- Machine learning / pattern recognition applications
  - Vision
  - Speech recognition
  - etc.
- All of science & knowledge is about predicting future in terms of past
  - So classification is a very fundamental problem with ultra-wide scope of applications



Clustering

Birds of a feather flock together.

The Clustering Problem

Outlook	Temp (F)	Humidity (%)	Windy?
sunny	75	70	true
sunny	80	90	true
sunny	85	85	false
sunny	72	95	false
sunny	69	70	false
overcast	72	90	true
overcast	73	88	true
overcast	64	65	true
overcast	81	75	false
rain	71	80	true
rain	65	70	true
rain	75	80	false
rain	68	80	false
rain	70	96	false

Find groups of similar records.

Need a function to compute similarity, given 2 input records

Unsupervised learning

Applications

- Targetting similar people or objects
  - Student tutorial groups
  - Hobby groups
  - Health support groups
  - Customer groups for marketing
  - Organizing e-mail
- Spatial clustering
  - Exam centres
  - Locations for a business chain
  - Planning a political strategy

Take Home

- Data mining is a mature field
- Don’t waste time developing new algorithms for core tasks
- Focus on applications to challenging kinds of data
  - Streams, Distributed data, Multimedia, Web, …
- Most effort is in how to map domain problems to data mining problems
- And how to make sense of the output.

Grading Plan (Tentative)

Normal Semester

- 10% Assignments
- 30% Mid
- 25% Endsem
- 30% Projects

Online Semester

- 20% Assignments
- 10% Quiz
- 25% 30 Min Quiz
- 45% Projects