Big Data

(Data-mining)

Instructor: Thanh Binh Nguyen

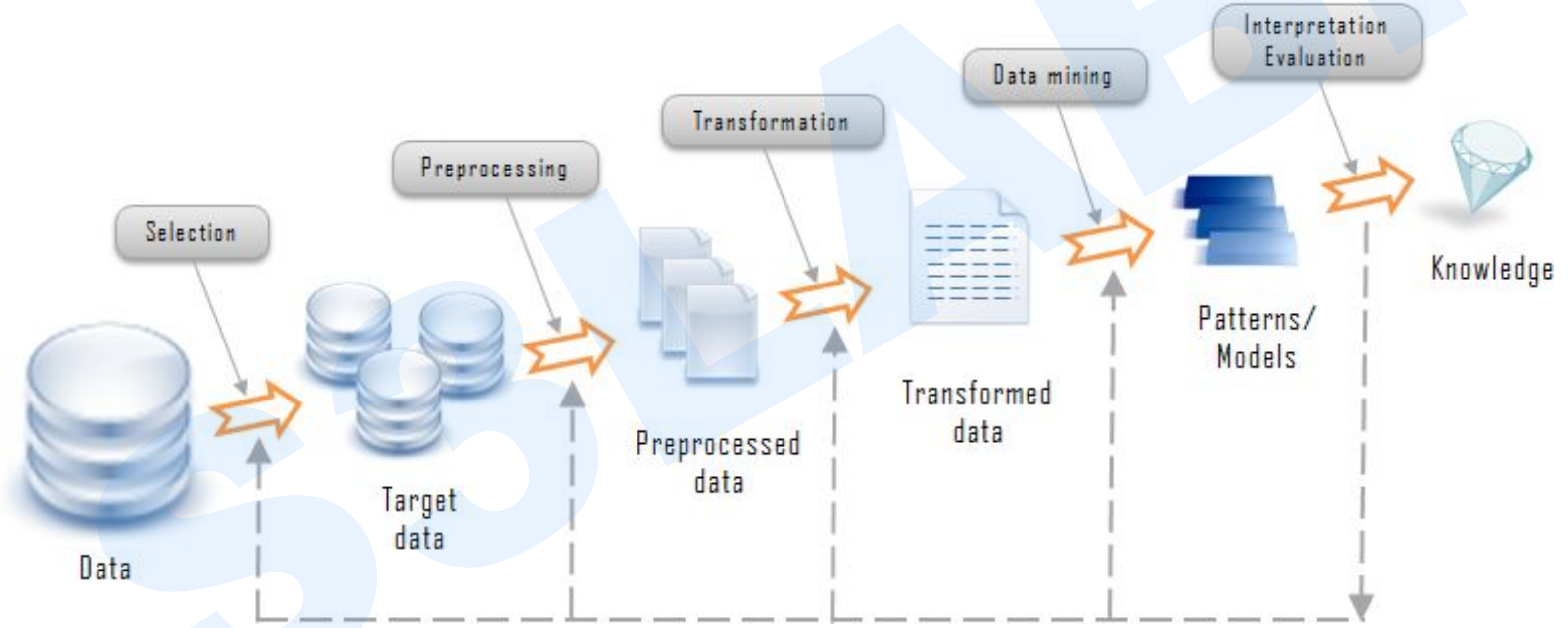
September 1st, 2019

S3Lab
Smart Software System Laboratory

“Big data is at the foundation of all the megatrends that are happening today, from social to mobile to cloud to gaming.”

– Chris Lynch, Vertica Systems

Knowledge Discovery in Database



What is Data Mining

- Data mining is basically one of the steps in the process of **knowledge discovery in database** (**KDD**)
- The computer-aid process that digs and analyzes enormous sets of data and then extracting the knowledge or information out of it. By its simplest definition, data mining automates the detections of relevant patterns in the database.

Data Mining: Different Perspectives

- Data to be mined
  - Object-oriented/relational, spatial, time-series, text, multimedia, heterogeneous, legacy, WWW
- Knowledge to be mined
  - Characterization, discrimination, association, classification, clustering, trend/deviation, outlier analysis, etc.
  - Multiple/integrated functions and mining at multiple levels
- Techniques utilized
  - Database-oriented, data warehouse, machine learning, statistics, visualization, etc.
- Applications adapted
  - Retail, telecommunication, CRM, banking, fraud analysis, forecasting, bio-data mining, stock market analysis, text mining, Web mining, etc.

Implementation Process

Implementation Process

Business Understanding

Determine Business Objectives
Background
Business Objectives
Business Success Criteria

Situation Assessment
Inventory of Resources
Requirements, Assumptions, and Constraints
Risks and Contingencies
Terminology
Costs and Benefits

Determine Data Mining Goal
Data Mining Goals
Data Mining Success Criteria

Produce Project Plan
Project Plan
Initial Asessment of Tools and Techniques

Data Understanding

Collect Initial Data
Initial Data Collection Report

Describe Data
Data Description Report

Explore Data
Data Exploration Report

Verify Data Quality
Data Quality Report

Data Preparation

Data Set
Data Set Description

Select Data
Rationale for Inclusion / Exclusion

Clean Data
Data Cleaning Report

Construct Data
Derived Attributes
Generated Records

Integrate Data
Merged Data

Format Data
Reformatted Data

Modeling

Select Modeling Technique
Modeling Technique
Modeling Assumptions

Generate Test Design
Test Design

Build Model
Parameter Settings
Models
Model Description

Assess Model
Model Assessment
Revised Parameter Settings

Evaluation

Evaluate Results
Assessment of Data Mining Results w.r.t. Business Success Criteria
Approved Models

Review Process
Review of Process

Determine Next Steps
List of Possible Actions
Decision

Deployment

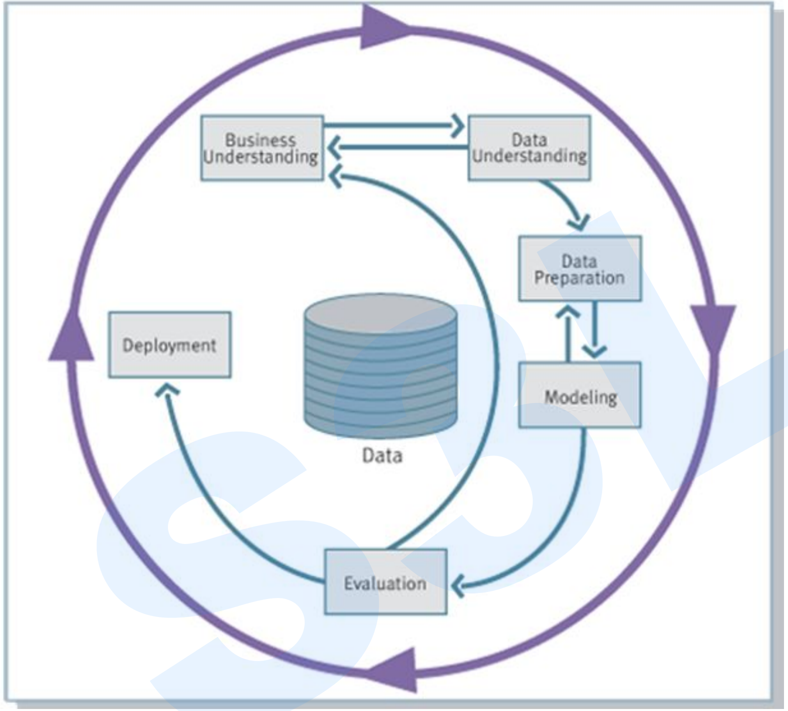
Plan Deployment
Deployment Plan

Plan Monitoring and Maintenance
Monitoring and Maintenance Plan

Produce Final Report
Final Report
Final Presentation

Review Project
Experience Documentation

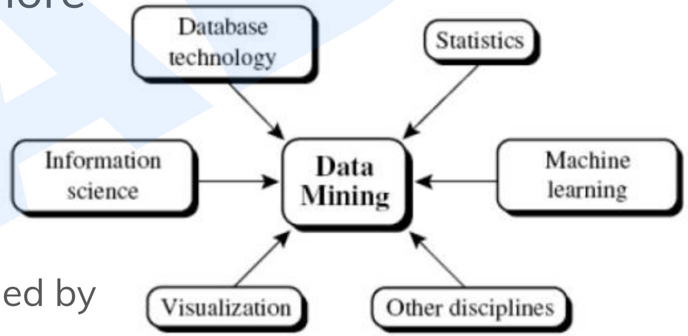
Implementation Process



- Business Understanding + Data Understanding + Data Preparation 80% of the time
- Modeling (applying mining algorithm) 20%

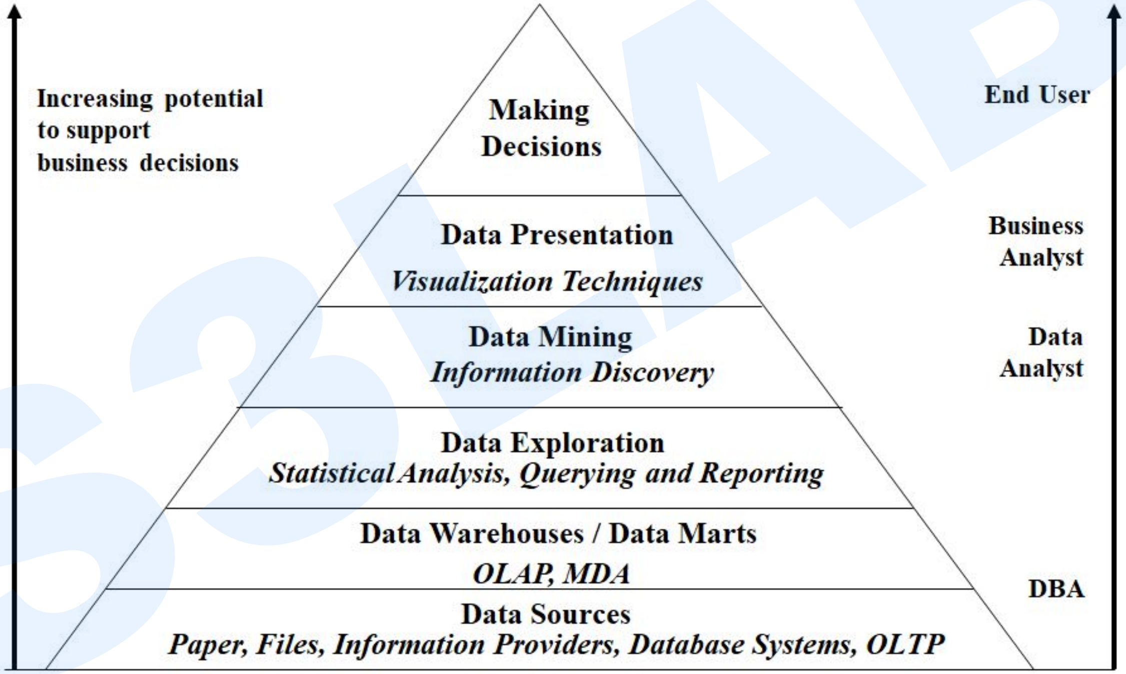
Confluence of Multiple Disciplines

- Overlaps with machine learning, statistics, artificial intelligence, databases, visualization but more stress on
  - scalability of number of features and instances
  - stress on algorithms and architectures whereas foundations of methods and formulations provided by statistics and machine learning.
  - automation for handling large, heterogeneous data
- Distinctions are fuzzy



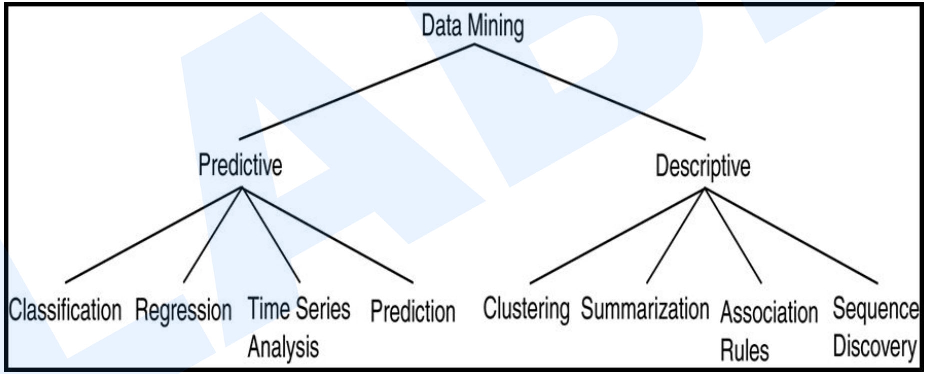
Confluence of Multiple Disciplines

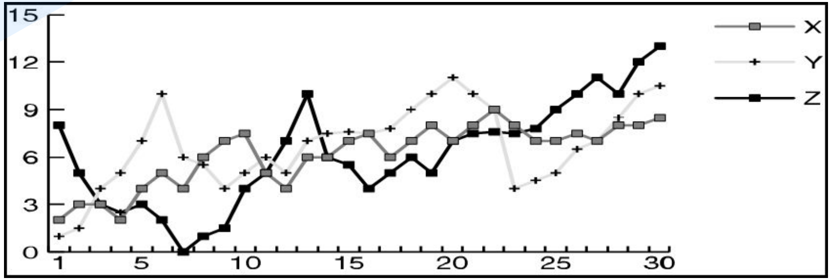
Data Mining and Business Intelligence



Basic Operations

- Predictive:
  - Regression
  - Classification
  - Collaborative Filtering
- Descriptive:
  - Clustering / similarity matching
  - Association rules and variants
  - Deviation detection

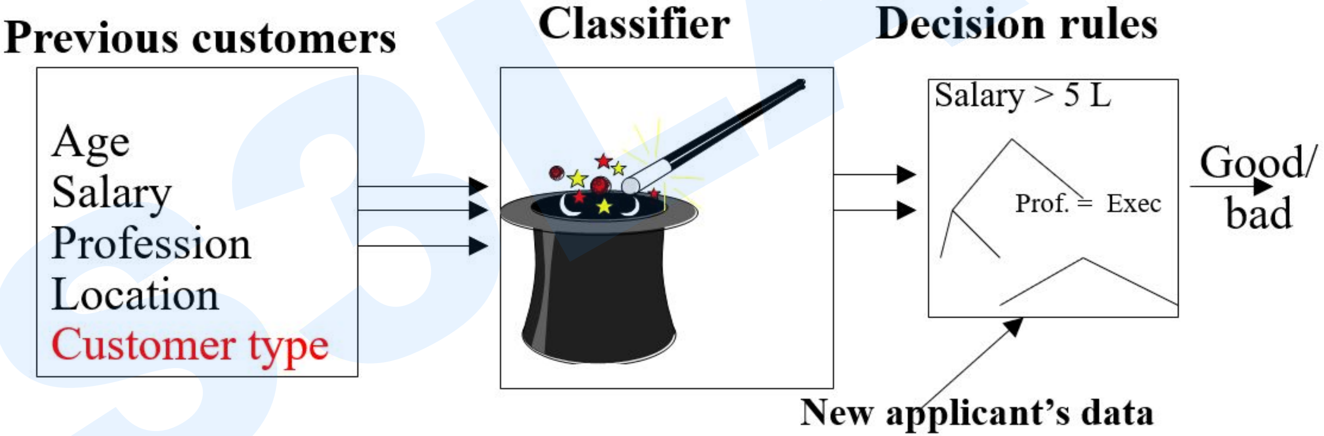




Basic Operations

Classification (Supervised learning)

- Given old data about customers and payments, predict new applicant's loan eligibility.



Basic Operations

Classification - methods

- Goal: Predict class $C_i = f(x_1, x_2, .. X_n)$
- Regression: (linear or any other polynomial)
    - $a * x_1 + b * x_2 + c = C_i$. (find values to best fit the data)
- Nearest neighbour
- Decision tree classifier: divide decision space into piecewise constant regions.
- Probabilistic / generative models
- Neural networks: partition by non-linear boundaries

Basic Operations

Classification (Supervised learning) - Nearest neighbor

- Define proximity between instances, find neighbors of new instance and assign majority class
- Case based reasoning: when attributes are more complicated than real-valued.

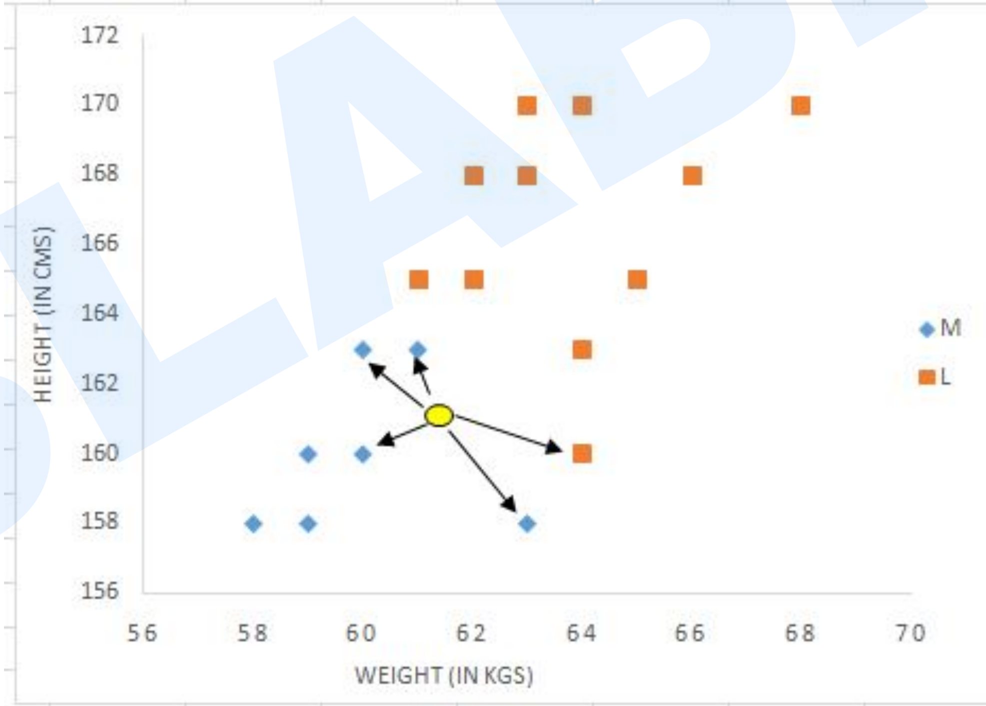
  - **Pros**
    - *Fast training*
  - **Cons**
    - *Slow during application.*
    - *No feature selection.*
    - *Notion of proximity vague*

Basic Operations

Classification (Supervised learning) - Nearest neighbor

=SQRT((A21-A6)^2+(B21-B6)^2)

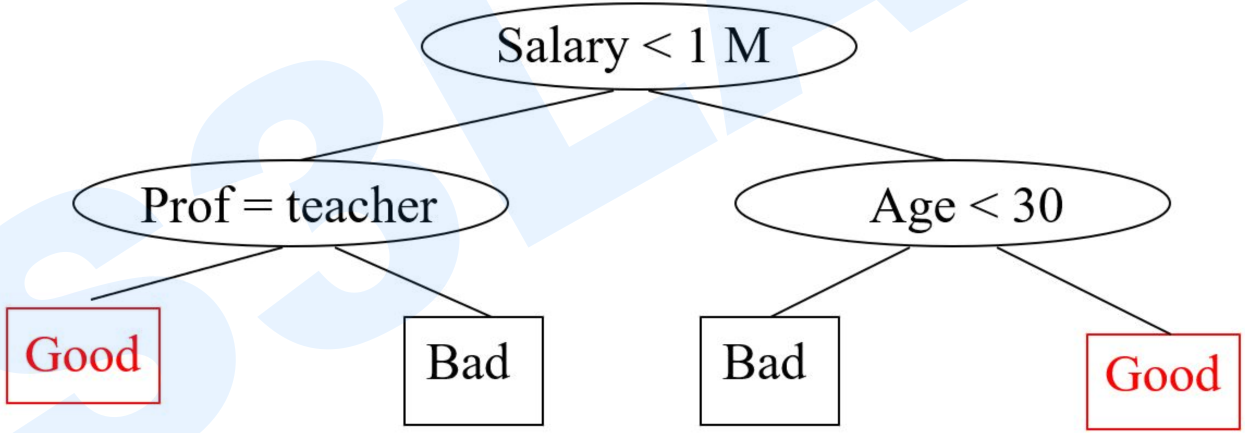
	A	B	C	D	E
1	Height (in cms)	Weight (in kgs)	T Shirt Size	Distance	
2	158	58	M	4.2	
3	158	59	M	3.6	
4	158	63	M	3.6	
5	160	59	M	2.2	3
6	160	60	M	1.4	1
7	163	60	M	2.2	3
8	163	61	M	2.0	2
9	160	64	L	3.2	5
10	163	64	L	3.6	
11	165	61	L	4.0	
12	165	62	L	4.1	
13	165	65	L	5.7	
14	168	62	L	7.1	
15	168	63	L	7.3	
16	168	66	L	8.6	
17	170	63	L	9.2	
18	170	64	L	9.5	
19	170	68	L	11.4	
20					
21	161	61			



Basic Operations

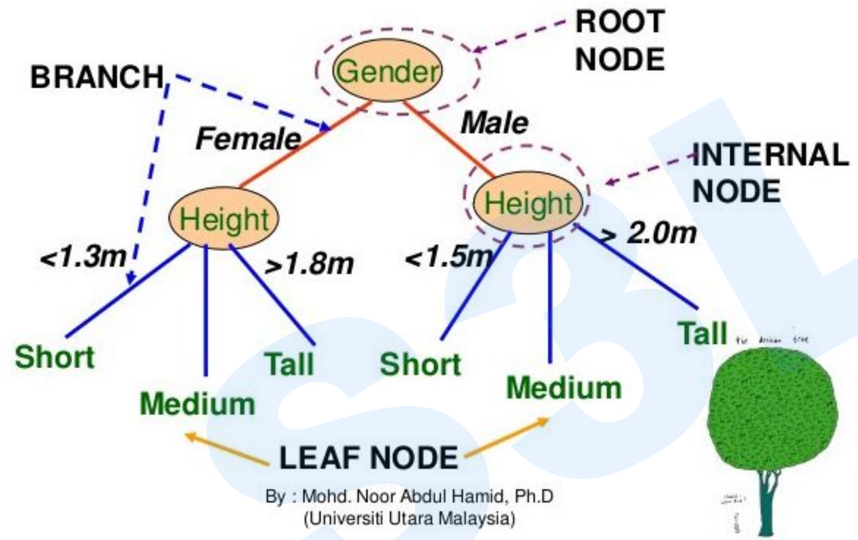
Classification (Supervised learning) - Decision trees

- Tree where internal nodes are simple decision rules on one or more attributes and leaf nodes are predicted class labels.



Basic Operations

Classification (Supervised learning) - Decision trees



Basic Operations

Classification (Supervised learning) - Decision tree classifiers

- Widely used learning method
- Easy to interpret: can be re-represented as if-then-else rules
- Approximates function by piecewise constant regions
- Does not require any prior knowledge of data distribution, works well on noisy data.
- Has been applied to:
  - Classify medical patients based on the disease,
  - Equipment malfunction by cause,
  - Loan applicant by likelihood of payment.

Basic Operations

Classification (Supervised learning) - Decision tree classifiers

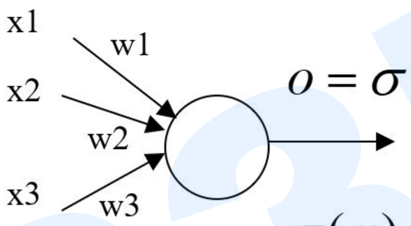
- **Pros**
  - *Reasonable training time*
  - *Fast application*
  - *Easy to interpret*
  - *Easy to implement*
  - *Can handle large number of features*
- **Cons**
  - *Cannot handle complicated relationship between features*
  - *simple decision boundaries*
  - *problems with lots of missing data*

Basic Operations

Classification (Supervised learning) - Neural Network

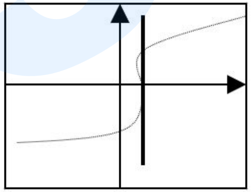
- Set of nodes connected by directed weighted edges

Basic NN unit

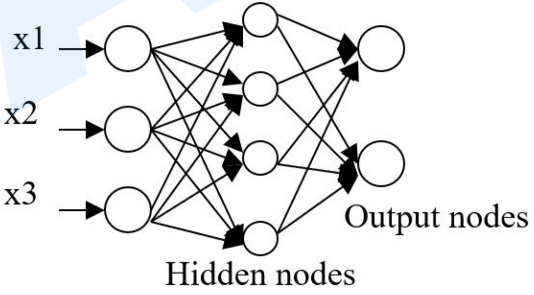


$$o = \sigma(\sum_{i=1}^{n} w_i x_i)$$

$$\sigma(y) = \frac{1}{1+e^{-y}}$$



A more typical NN

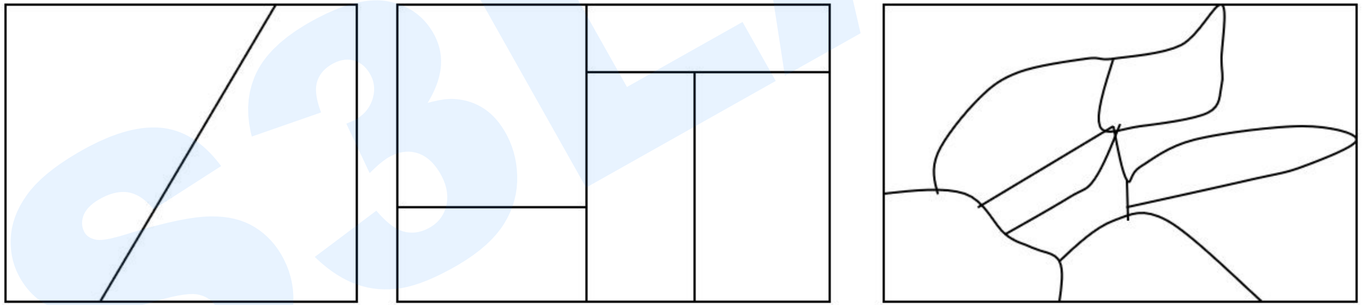


Basic Operations

Classification (Supervised learning) - Neural network

- Useful for learning complex data like handwriting, speech and image recognition

Decision boundaries:



Linear regression Classification tree Neural network

Basic Operations

Classification (Supervised learning) - Neural Network

- **Pros**
  - Can learn more complicated class boundaries
  - Fast application
  - Can handle large number of features
- **Cons**
  - Slow training time
  - Hard to interpret
  - Hard to implement: trial and error for choosing number of nodes

Basic Operations

Classification (Supervised learning) - Bayesian Learning

- Assume a probability model on generation of data.

$$\text{predicted class}: c = \max_{c_j} p(c_j \mid d) = \max_{c_j} \frac{p(d \mid c_j)\, p(c_j)}{p(d)}$$

- Apply bayes theorem to find most likely class as:

$$c = \max_{c_j} \frac{p(c_j)}{p(d)} \prod_{i=1}^{n} p(a_i \mid c_j)$$

- Naïve bayes: Assume attributes conditionally independent given class value
- Easy to learn probabilities by counting,
- Useful in some domains e.g. text

Basic Operations

Clustering (unsupervised learning)

- Unsupervised learning when old data with class labels not available e.g. when introducing a new product.
- Group/cluster existing customers based on time series of payment history such that similar customers in same cluster.
- Key requirement: Need a good measure of similarity between instances.
- Identify micro-markets and develop policies for each

Basic Operations

Clustering (unsupervised learning) - Applications

- Customer segmentation e.g. for targeted marketing
  - Group/cluster existing customers based on time series of payment history such that similar customers in same cluster.
  - Identify micro-markets and develop policies for each
- Collaborative filtering:
  - group based on common items purchased
- Text clustering
- Compression

Basic Operations

Clustering (unsupervised learning) - Similarity

- Determine similarity between two objects.
- Similarity characteristics:
  - $\forall t_i \in D, sim(t_i, t_i) = 1$
  - $\forall t_i, t_j \in D, sim(t_i, t_j) = 0$ if t_i and t_j are not alike at all.
  - $\forall t_i, t_j, t_k \in D, sim(t_i, t_j) < sim(t_i, t_k)$ if t_i is more like t_k than it is like t_j.
- Alternatively, distance measure measure how unlike or dissimilar objects are.

Basic Operations

Clustering (unsupervised learning) - Similarity

Dice: $$sim(t_i, t_j) = \frac{2\sum_{h=1}^{k} t_{ih}t_{jh}}{\sum_{h=1}^{k} t_{ih}^2 + \sum_{h=1}^{k} t_{jh}^2}$$

Jaccard: $$sim(t_i, t_j) = \frac{\sum_{h=1}^{k} t_{ih}t_{jh}}{\sum_{h=1}^{k} t_{ih}^2 + \sum_{h=1}^{k} t_{jh}^2 - \sum_{h=1}^{k} t_{ih}t_{jh}}$$

Cosine: $$sim(t_i, t_j) = \frac{\sum_{h=1}^{k} t_{ih}t_{jh}}{\sqrt{\sum_{h=1}^{k} t_{ih}^2 \sum_{h=1}^{k} t_{jh}^2}}$$

Overlap: $$sim(t_i, t_j) = \frac{\sum_{h=1}^{k} t_{ih}t_{jh}}{min(\sum_{h=1}^{k} t_{ih}^2, \sum_{h=1}^{k} t_{jh}^2)}$$

Basic Operations

Clustering (unsupervised learning) - Distances

- Numeric data: euclidean, manhattan distances
- Categorical data: 0/1 to indicate presence/absence followed by
  - Hamming distance (# dissimilarity)
  - Jaccard coefficients: #similarity in 1s/(# of 1s)
  - data dependent measures: similarity of A and B depends on co-occurrence with C.
- Combined numeric and categorical data:
  - weighted normalized distance

Basic Operations

Clustering (unsupervised learning) - Methods

- Hierarchical clustering
  - agglomerative Vs divisive
  - single link Vs complete link
- Partitional clustering
  - distance-based: K-means
  - model-based: EM
  - density-based:

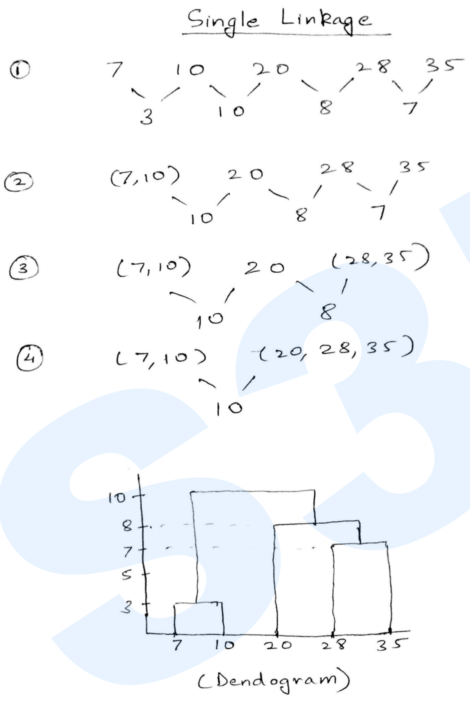
Basic Operations

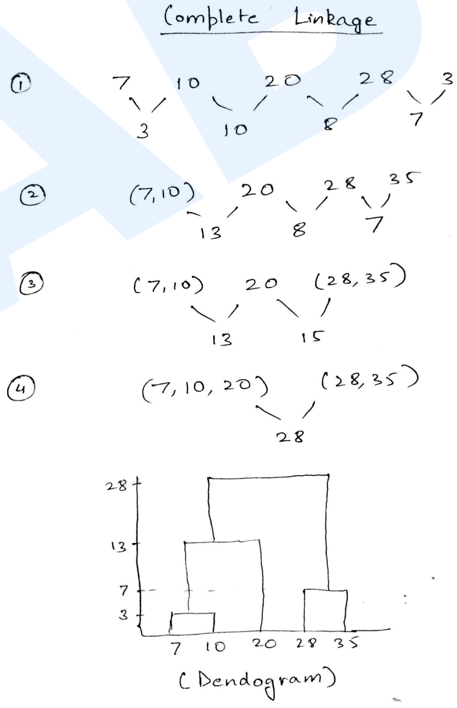
Clustering (unsupervised learning) - Agglomerative Hierarchical clustering

- Given: matrix of similarity between every point pair
- Start with each point in a separate cluster and merge clusters based on some criteria:
  - Single link: merge two clusters such that the minimum distance between two points from the two different cluster is the least
  - Complete link: merge two clusters such that all points in one cluster are “close” to all points in the other.

Basic Operations

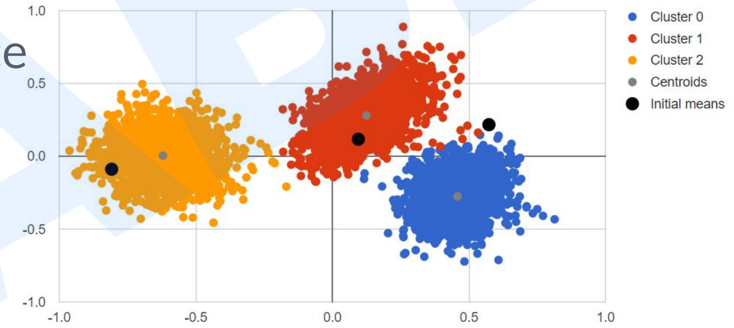
Clustering (unsupervised learning) - Agglomerative Hierarchical clustering





Basic Operations

Clustering (unsupervised learning) -> K-Means Partitional clustering



- Criteria: minimize sum of square of distance
  - Between each point and centroid of the cluster.
  - Between each pair of points in the cluster
- Algorithm:
  - Select initial partition with K clusters: random, first K, K separated points
  - Repeat until stabilization:
    - Assign each point to closest cluster center
    - Generate new cluster centers
    - Adjust clusters by merging/splitting

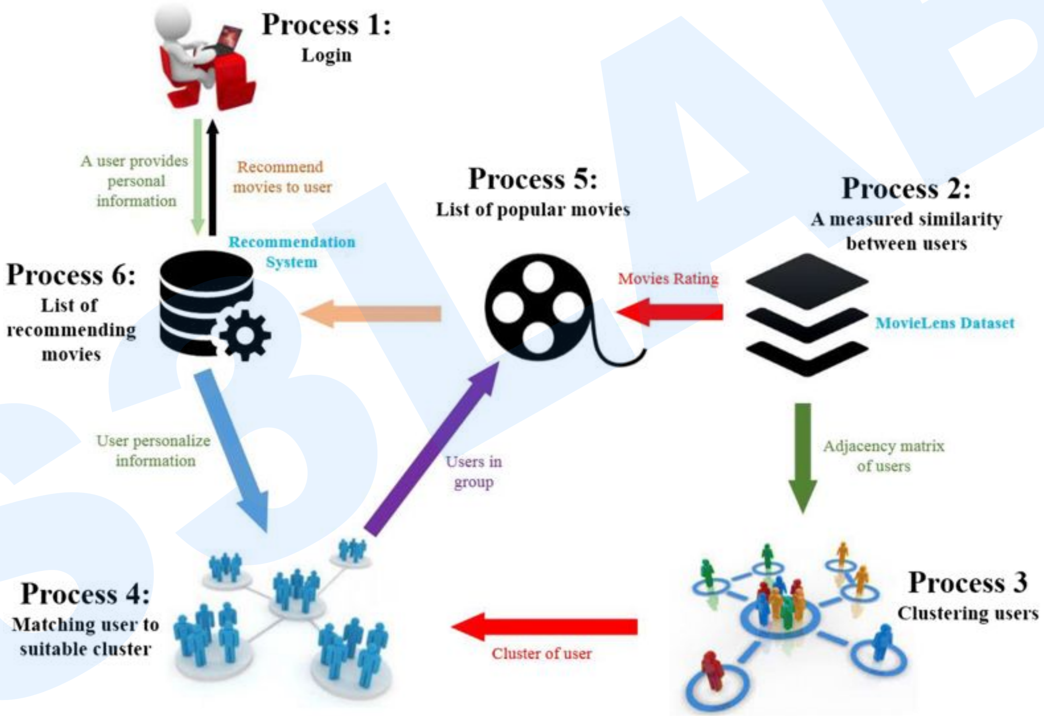
Basic Operations

Clustering (unsupervised learning) -> Collaborative Filtering

- Given database of user preferences, predict preference of new user
- Example: predict what new movies you will like based on
  - your past preferences
  - others with similar past preferences
  - their preferences for the new movies
- Example: predict what books/CDs a person may want to buy
  - (and suggest it, or give discounts to tempt customer)

Basic Operations

Clustering (unsupervised learning) -> Collaborative Recommendation



Basic Operations

Clustering (unsupervised learning) -> CR -> Cluster-based approaches

- External attributes of people and movies to cluster
  - age, gender of people
  - actors and directors of movies.
  - [May not be available]
- Cluster people based on movie preferences
  - misses information about similarity of movies
- Repeated clustering:
  - cluster movies based on people, then people based on movies, and repeat
  - ad hoc, might smear out groups

Basic Operations

Clustering (unsupervised learning) -> CR -> Model-based approaches

- People and movies belong to unknown classes
- P_k = probability a random person is in class *k*
- P_l = probability a random movie is in class *l*
- P_{kl} = probability of a class-*k* person liking a class-*l* movie
- Gibbs sampling: iterate
  - Pick a person or movie at random and assign to a class with probability proportional to P_k or P_l
  - Estimate new parameters
    - Need statistics background to understand details

Basic Operations

Association Rules

- Given set T of groups of items
- Example: set of item sets purchased
- Goal: find all rules on itemsets of the form a-->b such that
  - *support* of a and b > user threshold s
  - conditional probability (*confidence*) of b given a > user threshold c
- Example: Milk --> bread
- Purchase of product A --> service B

T
Milk, cereal
Tea, milk
Tea, rice, bread
cereal

Basic Operations

Association Rules

$$Rule: X \Rightarrow Y$$

$$Supprt = \frac{Frequency(X,Y)}{N}$$

$$Confidence = \frac{Frequency(X,Y)}{Frequency(X)}$$

$$Lift = \frac{Support}{Support(X) \times Support(Y)}$$

- Set of Items: I = {I1, I2,I3,..., In}
- Set of transactions: T = {T1, T2, ..., Tn}
- Each transaction has a unique id and contains subset of items
- A rule is defined as an implication of the form: X ==> Y
  - Where X, Y are itemsets, for example, X = { I1, I2 } and Y={ I5 }
  - X is called antecedent or left-hand-side (LHS)
  - and Y consequent or right-hand-side (RHS)

Basic Operations

Association Rules

- I = {milk, bread, butter, beer, diapers }
- And a rule could be {butter, bread} ==> {milk} meaning if butter and bread are bought then milk is also bought

transaction ID	milk	bread	butter	beer	diapers
1	1	1	0	0	0
2	0	0	1	0	0
3	0	0	0	1	1
4	1	1	1	0	0
5	0	1	0	0	0

Integration of Data Mining and Data Warehouse

- Data Warehousing provides the Enterprise with a memory

- Data Mining provides the Enterprise with intelligence

Integration of Data Mining and Data Warehouse

- Data mining systems, DBMS, Data warehouse systems coupling
  - No coupling, loose-coupling, semi-tight-coupling, tight-coupling
- Online analytical mining data
  - integration of mining and Online Analytical Processing (OLAP) technologies
  - Ideal platform for vertical integration but needs to be interactive instead of batch
- Interactive mining multi-level knowledge
  - Necessity of mining knowledge and patterns at different levels of abstraction by drilling/rolling, pivoting, slicing/dicing, etc.
- Integration of multiple mining functions
  - Characterized classification, first clustering and then association

Integration of Data Mining and Data Warehouse

Coupling Data Mining with DB/DW systems

- No coupling—flat file processing, not recommended
- Loose coupling
  - Fetching data from DB/DW
- Semi-tight coupling—enhanced DM performance
  - Provide efficient implement a few data mining primitives in a DB/DW system, e.g., sorting, indexing, aggregation, histogram analysis, multiway join, precomputation of some stat functions
- Tight coupling—A uniform information processing environment
  - DM is smoothly integrated into a DB/DW system, mining query is optimized based on mining query, indexing, etc.

Integration of Data Mining and Data Warehouse

Vertical Integration - Mining on the web

- Web log analysis for site design:
  - what are popular pages,
  - what links are hard to find.
- Electronic stores sales enhancements:
  - recommendations, advertisement:
  - Collaborative filtering: Net perception, Wisewire
  - Inventory control: what was a shopper looking for and could not find..

Integration of Data Mining and Data Warehouse

State of art in Mining OLAP Integration

- Decision trees [Information discovery, Cognos]
  - find factors influencing high profits
- Clustering [**Pilot software**]
  - segment customers to define hierarchy on that dimension
- Time series analysis: [Seagate's Holos]
  - Query for various shapes along time: eg. spikes, outliers
- Multi-level Associations [Han et al.]
  - find association between members of dimensions
- Sarawagi [VLDB2000]

Major Issues in Data Mining

- Mining methodology
  - Mining different kinds of knowledge from diverse data types, e.g., bio, stream, Web
  - Performance: efficiency, effectiveness, and scalability
  - Pattern evaluation: the interestingness problem
  - Incorporation of background knowledge
  - Handling noise and incomplete data
  - Parallel, distributed and incremental mining methods
  - Integration of the discovered knowledge with existing one: knowledge fusion

Major Issues in Data Mining

- User interaction
  - Data mining query languages and ad-hoc mining
  - Expression and visualization of data mining results
  - Interactive mining of knowledge at multiple levels of abstraction
- Applications and social impacts
  - Domain-specific data mining & invisible data mining
  - Protection of data security, integrity, and privacy

Some Success Stories

- Network intrusion detection using a combination of sequential rule discovery and classification tree on 4 GB DARPA data
  - Won over (manual) knowledge engineering approach
  - http://www.cs.columbia.edu/~sal/JAM/PROJECT/ provides good detailed description of the entire process
- Major US bank: customer attrition prediction
  - First segment customers based on financial behavior: found 3 segments
  - Build attrition models for each of the 3 segments
  - 40-50% of attritions were predicted == factor of 18 increase
- Targeted credit marketing: major US banks
  - find customer segments based on 13 months credit balances
  - build another response model based on surveys
  - increased response 4 times -- 2%

Q & A

Cảm ơn đã theo dõi

Chúng tôi hy vọng cùng nhau đi đến thành công.