Basic Data Mining Techniques

Overview

- Data & Types of Data
- Fuzzy Sets
- Information Retrieval
- Machine Learning
- Statistics & Estimation Techniques
- Similarity Measures
- Decision Trees

What is Data?

- Collection of data objects and their attributes
- An attribute is a property or characteristic of an object
  - Examples: eye color of a person, temperature, etc.
  - Attribute is also known as variable, field, characteristic, or feature
- A collection of attributes describe an object
  - Object is also known as record, point, case, sample, entity, or instance

Attributes

Objects

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Attribute Values

- Attribute values are numbers or symbols assigned to an attribute
- Distinction between attributes and attribute values
  - Same attribute can be mapped to different attribute values
    - Example: height can be measured in feet or meters
  - Different attributes can be mapped to the same set of values
    - Example: Attribute values for ID and age are integers
    - But properties of attribute values can be different
      - ID has no limit but age has a maximum and minimum value

Types of Attributes

- There are different types of attributes
  - Nominal
    - Examples: ID numbers, eye color, zip codes
  - Ordinal
    - Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height in {tall, medium, short}
  - Interval
    - Examples: calendar dates, temperatures in Celsius or Fahrenheit.
  - Ratio
    - Examples: temperature in Kelvin, length, time, counts

Properties of Attribute Values

- The type of an attribute depends on which of the following properties it possesses:
  - Distinctness: $= \neq$
  - Order: $< >$
  - Addition: $+ -$
  - Multiplication: $* /$

  - Nominal attribute: distinctness
  - Ordinal attribute: distinctness & order
  - Interval attribute: distinctness, order & addition
  - Ratio attribute: all 4 properties

Attribute Type	Description	Examples	Operations
Nominal	The values of a nominal attribute are just different names, i.e., nominal attributes provide only enough information to distinguish one object from another. (=, ≠)	zip codes, employee ID numbers, eye color, sex: {*male, female*}	mode, entropy, contingency correlation, χ^2 test
Ordinal	The values of an ordinal attribute provide enough information to order objects. (<, >)	hardness of minerals, {*good, better, best*}, grades, street numbers	median, percentiles, rank correlation, run tests, sign tests
Interval	For interval attributes, the differences between values are meaningful, i.e., a unit of measurement exists. (+, -)	calendar dates, temperature in Celsius or Fahrenheit	mean, standard deviation, Pearson's correlation, *t* and *F* tests
Ratio	For ratio variables, both differences and ratios are meaningful. (*, /)	temperature in Kelvin, monetary quantities, counts, age, mass, length, electrical current	geometric mean, harmonic mean, percent variation

Attribute Level	Transformation	Comments
Nominal	Any permutation of values	If all employee ID numbers were reassigned, would it make any difference?
Ordinal	An order preserving change of values, i.e., *new_value = f(old_value)* where *f* is a monotonic function.	An attribute encompassing the notion of good, better best can be represented equally well by the values {1, 2, 3} or by {0.5, 1, 10}.
Interval	*new_value =a * old_value + b* where a and b are constants	Thus, the Fahrenheit and Celsius temperature scales differ in terms of where their zero value is and the size of a unit (degree).
Ratio	*new_value = a * old_value*	Length can be measured in meters or feet.

Discrete and Continuous Attributes

- Discrete Attribute
  - Has only a finite or countably infinite set of values
  - Examples: zip codes, counts, or the set of words in a collection of documents
  - Often represented as integer variables.
  - Note: binary attributes are a special case of discrete attributes
- Continuous Attribute
  - Has real numbers as attribute values
  - Examples: temperature, height, or weight
  - Practically, real values can only be measured and represented using a finite number of digits
  - Continuous attributes are typically represented as floating-point variables

Types of data sets

- Record
  - Data Matrix
  - Document Data
  - Transaction Data
- Graph
  - World Wide Web
  - Molecular Structures
- Ordered
  - Spatial Data
  - Temporal Data
  - Sequential Data
  - Genetic Sequence Data

Characteristics of Structured Data

- Dimensionality
  - Curse of Dimensionality
- Sparsity
  - Only presence counts
- Resolution
  - Patterns depend on the scale

Record Data

- Data that consists of a collection of records, each of which consists of a fixed set of attributes

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Data Matrix

- If data objects have the same fixed set of numeric attributes, then the data objects can be thought of as points in a multi-dimensional space, where each dimension represents a distinct attribute
- Such data set can be represented by an m by n matrix, where there are m rows, one for each object, and n columns, one for each attribute

Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

Document Data

- Each document becomes a `term' vector,
  - each term is a component (attribute) of the vector,
  - the value of each component is the number of times the corresponding term occurs in the document.

	team	coach	pla y	ball	score	game	wi n	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

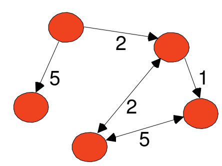
Transaction Data

- A special type of record data, where
  - each record (transaction) involves a set of items.
  - For example, consider a grocery store. The set of products purchased by a customer during one shopping trip constitute a transaction, while the individual products that were purchased are the items.

TID	*Items*
1	**Bread, Coke, Milk**
2	**Beer, Bread**
3	**Beer, Coke, Diaper, Milk**
4	**Beer, Bread, Diaper, Milk**
5	**Coke, Diaper, Milk**

Graph Data

- Examples: Generic graph and HTML Links



Data Mining

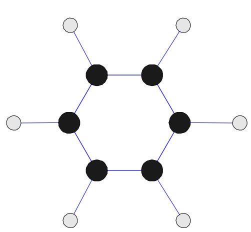
Graph Partitioning

Parallel Solution of Sparse Linear System of Equations

<a href="papers/papers.html#ffff">
N-Body Computation and Dense Linear System Solvers

Chemical Data

Benzene Molecule: C_6H_6



Ordered Data

Sequences of transactions

Items/Events

(A B) (D) (C E)
(B D) (C) (E)
(C D) (B) (A E)

An element of the sequence

Ordered Data

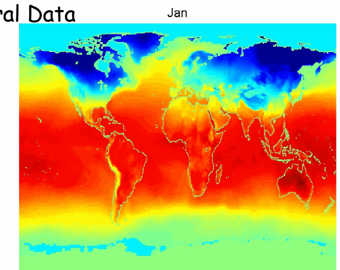
Genomic sequence data

```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```


Ordered Data

Spatio-Temporal Data

Average Monthly Temperature of land and ocean



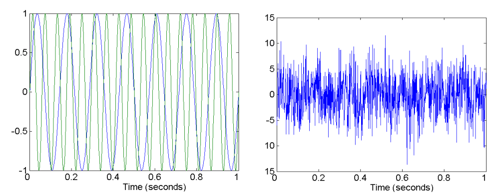
Data Quality

- What kinds of data quality problems?
- How can we detect problems with the data?
- What can we do about these problems?

- Examples of data quality problems:
  - noise and outliers
  - missing values
  - duplicate data

Noise

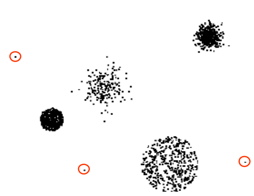
- Noise refers to modification of original values
  - Examples: distortion of a person's voice when talking on a poor phone and "snow" on television



Two Sine Waves   Two Sine Waves + Noise

Outliers

- Outliers are data objects with characteristics that are considerably different than most of the other data objects in the data set



Missing Values

- Reasons for missing values
  - Information is not collected (e.g., people decline to give their age and weight)
  - Attributes may not be applicable to all cases (e.g., annual income is not applicable to children)

- Handling missing values
  - Eliminate Data Objects
  - Estimate Missing Values
  - Ignore the Missing Value During Analysis
  - Replace with all possible values (weighted by their probabilities)

Duplicate Data

- Data set may include data objects that are duplicates, or almost duplicates of one another
  - Major issue when merging data from heterogeneous sources
- Examples:
  - Same person with multiple email addresses
- Data cleaning
  - Process of dealing with duplicate data issues

Data Preprocessing

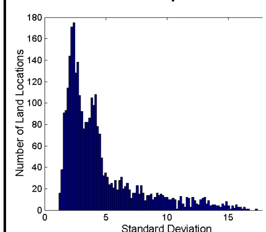
- Aggregation
- Sampling
- Dimensionality Reduction
- Feature subset selection
- Feature creation
- Discretization and Binarization
- Attribute Transformation

Aggregation

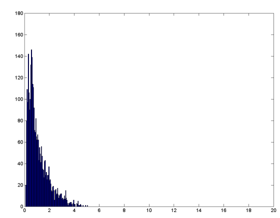
- Combining two or more attributes (or objects) into a single attribute (or object)
- Purpose
  - Data reduction
    - Reduce the number of attributes or objects
  - Change of scale
    - Cities aggregated into regions, states, countries, etc
  - More "stable" data
    - Aggregated data tends to have less variability

Aggregation

Variation of Precipitation in Australia



Standard Deviation of Average Monthly Precipitation



Standard Deviation of Average Yearly Precipitation

Sampling

- Sampling is the main technique employed for data selection.
  - It is often used for both the preliminary investigation of the data and the final data analysis.
- Statisticians sample because obtaining the entire set of data of interest is too expensive or time consuming.
- Sampling is used in data mining because processing the entire set of data of interest is too expensive or time consuming.

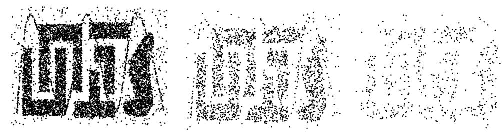
Sampling …

- The key principle for effective sampling is the following:
  - using a sample will work almost as well as using the entire data sets, if the sample is representative
  - A sample is representative if it has approximately the same property (of interest) as the original set of data

Types of Sampling

- Simple Random Sampling
  - There is an equal probability of selecting any particular item
- Sampling without replacement
  - As each item is selected, it is removed from the population
- Sampling with replacement
  - Objects are not removed from the population as they are selected for the sample.
    - In sampling with replacement, the same object can be picked up more than once
- Stratified sampling
  - Split the data into several partitions; then draw random samples from each partition

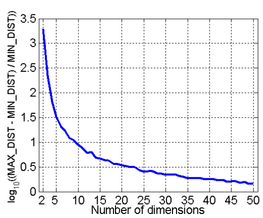
Sample Size



8000 points 2000 Points 500 Points

Curse of Dimensionality

- When dimensionality increases, data becomes increasingly sparse in the space that it occupies
- Definitions of density and distance between points, which is critical for clustering and outlier detection, become less meaningful



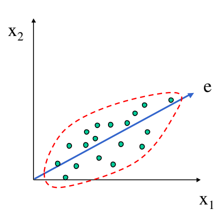
- Randomly generate 500 points
- Compute difference between max and min distance between any pair of points

Dimensionality Reduction

- Purpose:
  - Avoid curse of dimensionality
  - Reduce amount of time and memory required by data mining algorithms
  - Allow data to be more easily visualized
  - May help to eliminate irrelevant features or reduce noise
- Techniques
  - Principle Component Analysis
  - Singular Value Decomposition
  - Others: supervised and non-linear techniques

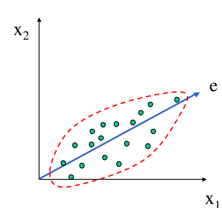
Dimensionality Reduction: PCA

- Goal is to find a projection that captures the largest amount of variation in data



Dimensionality Reduction: PCA

- Find the eigenvectors of the covariance matrix
- The eigenvectors define the new space



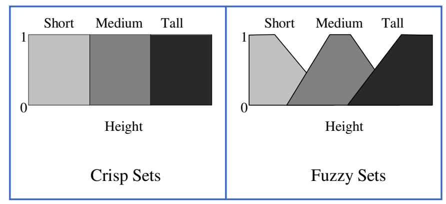
Fuzzy Sets and Logic

Fuzzy Set: Set where the set membership function is a real valued function with output in the range [0,1].
- f(x): Probability x is in F.
- 1-f(x): Probability x is not in F.

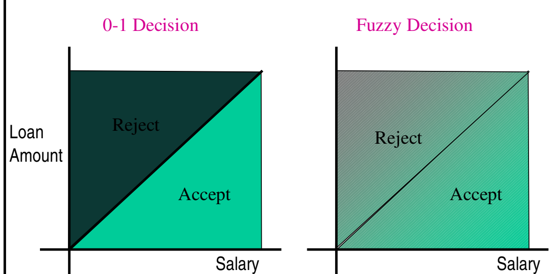
Example
- T = {x | x is a person and x is tall}
- Let f(x) be the probability that x is tall
- Here f is the membership function

DM: Prediction and classification are often fuzzy.

Fuzzy Sets



Classification/Prediction is Fuzzy



Information Retrieval

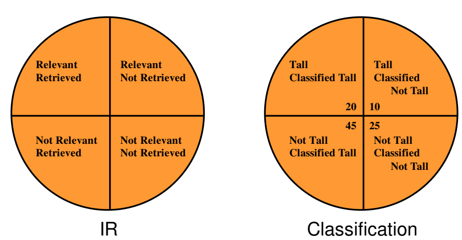
Information Retrieval (IR): retrieving desired information from textual data
- Library Science
- Digital Libraries
- Web Search Engines
- Traditionally has been keyword based
- Sample query:
  - Find all documents about "data mining".

DM: Similarity measures; Mine text or Web data

Information Retrieval (cont'd)

Similarity: measure of how close a query is to a document.
- Documents which are "close enough" are retrieved.
- Metrics:
  - *Precision* = $\frac{|\text{Relevant and Retrieved}|}{|\text{Retrieved}|}$
  - *Recall* = $\frac{|\text{Relevant and Retrieved}|}{|\text{Relevant}|}$

IR Query Result Measures and Classification



Machine Learning

- *Machine Learning (ML):* area of AI that examines how to devise algorithms that can learn.
- Techniques from ML are often used in classification and prediction.
- *Supervised Learning:* learns by example.
- *Unsupervised Learning:* learns without knowledge of correct answers.
- Machine learning often deals with small or static datasets.

DM: Uses many machine learning techniques.

Statistics

- Usually creates simple descriptive models.
- *Statistical inference:* generalizing a model created from a sample of the data to the entire dataset.
- *Exploratory Data Analysis:*
  - Data can actually drive the creation of the model.
  - Opposite of traditional statistical view.
- Data mining targeted to business users.

DM: Many data mining methods are based on statistical techniques.

Point Estimation

Point Estimate: estimate a population parameter.

- May be made by calculating the parameter for a sample.
- May be used to predict values for missing data.

Ex:
- R contains 100 employees
- 99 have salary information
- Mean salary of these is \$50,000
- Use \$50,000 as value of remaining employee's salary.

Is this a good idea?

Estimation Error

Bias: Difference between expected value and actual value.

$$\boxed{Bias = E(\hat{\Theta}) - \Theta}$$

Mean Squared Error (MSE): expected value of the squared difference between the estimate and the actual value:

$$\boxed{MSE(\hat{\Theta}) = E(\hat{\Theta} - \Theta)^2}$$

- Why square?
- Root Mean Square Error (RMSE).

Jackknife Estimate

- *Jackknife Estimate:* estimate of parameter is obtained by omitting one value from the set of observed values.
- Ex: estimate of mean for $X=\{x_1, \ldots, x_n\}$

$$\boxed{\hat{\theta}(i) = \frac{\sum_{j=1}^{i-1} x_j + \sum_{j=i+1}^{n} x_j}{n-1}}$$

$$\boxed{\hat{\theta}_{(.)} = \frac{\sum_{j=1}^{n} \hat{\theta}_{(j)}}{n}}$$

Maximum Likelihood Estimate (MLE)

- Obtain parameter estimates that maximize the probability that the sample data occurs for the specific model.
- Joint probability for observing the sample data by multiplying the individual probabilities. Likelihood function:

$$\boxed{L(\Theta \mid x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i \mid \Theta)}$$

- Maximize L.

MLE Example

- Coin toss five times: {H,H,H,H,T}
- Assuming a perfect coin with H and T equally likely, the likelihood of this sequence is:

$$\boxed{L(p \mid 1,1,1,1,0) = \prod_{i=1}^{5} 0.5 = 0.03.}$$

- However if the probability of a H is 0.8 then:

$$\boxed{L(p \mid 1,1,1,1,0) = 0.8 \times 0.8 \times 0.8 \times 0.8 \times 0.2 = 0.08.}$$

MLE Example (cont'd)

General likelihood formula:

$$L(p \mid x_1, ..., x_5) = \prod_{i=1}^{5} p^{x_i} (1-p)^{1-x_i} = p^{\sum_{i=1}^{5} x_i} (1-p)^{5-\sum_{i=1}^{5} x_i}.$$

$$l(p) = log L(p) = \sum_{i=1}^{5} x_i log(p) + (5 - \sum_{i=1}^{5} x_i)\, log(1-p)$$

$$\frac{\partial l(p)}{\partial p} = \sum_{i=1}^{5} \frac{x_i}{p} - \frac{5 - \sum_{i=1}^{5} x_i}{1-p}.$$

$$p = \frac{\sum_{i=1}^{5} x_i}{5}$$

Estimate for p is then 4/5 = 0.8

Expectation-Maximization (EM)

Solves estimation with incomplete data.

Algorithm
- Obtain initial estimates for parameters.
- Iteratively use estimates for missing data and continue refinement (maximization) of the estimate until convergence.

Expectation Maximization Algorithm

Input:
- $\Theta = \{\theta_1, ..., \theta_p\}$ //Parameters to be Estimated
- $X_{obs} = \{x_1, ..., x_k\}$ //Input Database Values Observed
- $X_{miss} = \{x_{k+1}, ..., x_n\}$ //Input Database Values Missing

Output:
- $\hat{\Theta}$ //Estimates for Θ

EM Algorithm:

```
i := 0;
Obtain initial parameter MLE estimate, Θ̂^i;
repeat
    Estimate missing data, X̂^i_miss;
    i++;
    Obtain next parameter estimate, θ̂^i to maximize data;
until estimate converges;
```


Expectation Maximization Example

$\{1,5,10,4\}$; $n = 6$ $k = 4$; Guess $\hat{\mu}^0 = 3$.

$$\hat{\mu}^1 = \frac{\sum_{i=1}^{k} x_i}{n} + \frac{\sum_{i=k+1}^{n} x_i}{n} = 3.33 + \frac{3+3}{6} = 4.33$$

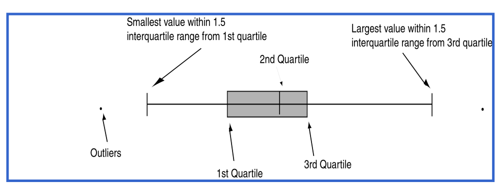
$$\hat{\mu}^2 = \frac{\sum_{i=1}^{k} x_i}{n} + \frac{\sum_{i=k+1}^{n} x_i}{n} = 3.33 + \frac{4.33+4.33}{6} = 4.77$$

$$\hat{\mu}^3 = \frac{\sum_{i=1}^{k} x_i}{n} + \frac{\sum_{i=k+1}^{n} x_i}{n} = 3.33 + \frac{4.77+4.77}{6} = 4.92$$

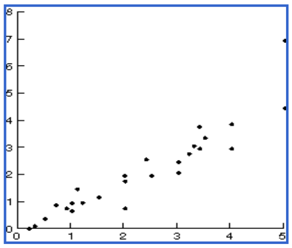
$$\hat{\mu}^4 = \frac{\sum_{i=1}^{k} x_i}{n} + \frac{\sum_{i=k+1}^{n} x_i}{n} = 3.33 + \frac{4.92+4.92}{6} = 4.97$$

Models Based on Summarization

- *Visualization:* Frequency distribution, mean, variance, median, mode, etc.
- *Box Plot:*



Scatter Diagram



Bayes Theorem

- *Posterior Probability:* $P(h_1|x_i)$
- *Prior Probability:* $P(h_1)$
- *Bayes Theorem:*

$$P(x_i) = \sum_{j=1}^{m} P(x_i \mid h_j)\, P(h_j).$$

$$P(h_1 \mid x_i) = \frac{P(x_i \mid h_1)\, P(h_1)}{P(x_i)}.$$

- Assign probabilities of hypotheses given a data value.

Bayes Theorem Example

- Credit authorizations (hypotheses):
  - h1 = authorize purchase,
  - h2 = authorize after further identification,
  - h3 = do not authorize,
  - h4 = do not authorize but contact police
- Assign twelve data values for all combinations of credit and income:

	1	2	3	4
Excellent	x_1	x_2	x_3	x_4
Good	x_5	x_6	x_7	x_8
Bad	x_9	x_{10}	x_{11}	x_{12}

- From training data: P(h1) = 60%; P(h2)=20%; P(h3)=10%; P(h4)=10%.

Bayes Example (cont'd)

Training Data:

ID	Income	Credit	Class	x_i
1	4	Excellent	h_1	x_4
2	3	Good	h_1	x_7
3	2	Excellent	h_1	x_2
4	3	Good	h_1	x_7
5	4	Good	h_1	x_8
6	2	Excellent	h_1	x_2
7	3	Bad	h_2	x_{11}
8	2	Bad	h_2	x_{10}
9	3	Bad	h_3	x_{11}
10	1	Bad	h_4	x_9

Bayes Example (cont'd)

- Calculate $P(x_i|h_j)$ and $P(x_i)$
- Ex: $P(x_7|h_1)=2/6$; $P(x_4|h_1)=1/6$; $P(x_2|h_1)=2/6$; $P(x_8|h_1)=1/6$; and $P(x_i|h_1)=0$ for all other x_i.
- Predict the class for x_4:
  - Calculate $P(h_j|x_4)$ for all h_j.
  - Place x_4 in class with largest value.
  - Ex:
    - $P(h_1|x_4) = (P(x_4|h_1)(P(h_1))/P(x_4)$
      $= (1/6)(0.6)/0.1 = 1.$
    - x_4 in class h_1.

Hypothesis Testing

- Find model to explain behavior by creating and then testing a hypothesis about the data.
- Exact opposite of usual DM approach.
- H_0 – Null hypothesis; Hypothesis to be tested.
- H_1 – Alternative hypothesis.

Chi Squared Statistic

- O – observed value
- E – Expected value based on hypothesis.

$$\chi^2 = \sum \frac{(O-E)^2}{E}.$$

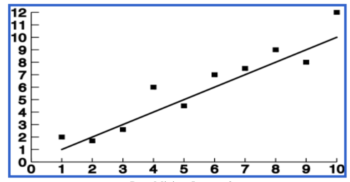
Ex:
- O = {50,93,67,78,87}
- E = 75
- χ^2 = 15.55 and therefore significant

Regression

- Predict future values based on past values
- *Linear Regression* assumes that a linear relationship exists.

$$y = c_0 + c_1 x_1 + \ldots + c_n x_n$$

- Find c_i values to best fit the data



Correlation

- Examine the degree to which the values for two variables behave similarly.
- Correlation coefficient r:
  - 1 = perfect correlation
  - -1 = perfect but opposite correlation
  - 0 = no correlation

$$r = \frac{\sum (x_i - \bar{X})(y_i - \bar{Y})}{\sqrt{\sum (x_i - \bar{X})^2 \sum (y_i - \bar{Y})^2}}$$

Similarity Measures

- Determine similarity between two objects.
- Characteristics of a good similarity measure:
  - $\forall t_i \in D, sim(t_i, t_i) = 1$
  - $\forall t_i, t_j \in D, sim(t_i, t_j) = 0$ 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 measures indicate how unlike or dissimilar objects are.

Commonly Used Similarity Measures

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

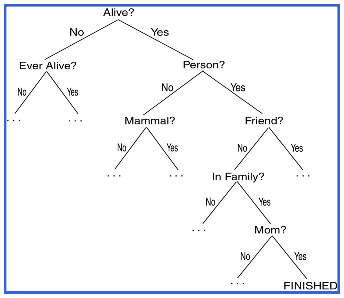
Distance Measures

Measure dissimilarity between objects

Euclidean: $dis(t_i, t_j) = \sqrt{\sum_{h=1}^{k} (t_{ih} - t_{jh})^2}$

Manhattan: $dis(t_i, t_j) = \sum_{h=1}^{k} \mid (t_{ih} - t_{jh}) \mid$

Twenty Questions Game



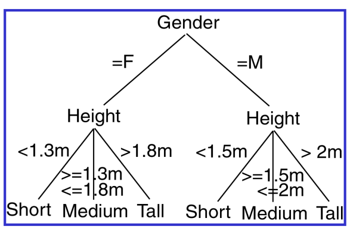
Decision Trees

Decision Tree (DT):

- Tree where the root and each internal node is labeled with a question.
- The arcs represent each possible answer to the associated question.
- Each leaf node represents a prediction of a solution to the problem.

Popular technique for classification; Leaf nodes indicate classes to which the corresponding tuples belong.

Decision Tree Example



Decision Trees

- A ***Decision Tree Model*** is a computational model consisting of three parts:
  - Decision Tree
  - Algorithm to create the tree
  - Algorithm that applies the tree to data
- Creation of the tree is the most difficult part.
- Processing is basically performing a search similar to that in a binary search tree (although DT may not always be binary).

Decision Tree Algorithm

```
Input:
  T      //Decision Tree
  D      //Input Database
Output:
  M      //Model Prediction
DTProc Algorithm:
         //Illustrates Prediction Technique using DT
  for each t ∈ D do
    n = root node of T;
    while n not leaf node do
      Obtain answer to question on n applied t;
      Identify arc from t which contains correct answer;
      n = node at end of this arc;
    Make prediction for t based on labeling of n;
```


Decision Trees: Advantages & Disadvantages

- Advantages:
  - Easy to understand.
  - Easy to generate rules from.
- Disadvantages:
  - May suffer from overfitting.
  - Classify by rectangular partitioning.
  - Do not easily handle nonnumeric data.
  - Can be quite large – pruning is often necessary.