COMP4670 Lecture 3: Security and Privacy

Security: tools and techniques to protect confidentiality, integrity, and availability
Privacy: the right of an individual to control the collection, use, disclosure, and retention of their personal information

Data Mining: Brief Review

- process of analyzing data from different perspectives and discovering useful information and knowledge
- standard methods and algorithms to move forward in the path of DIKW (Data, Information, Knowledge, Wisdom)
- finding correlations or patterns among large amounts of data

Information Hierarchy

- **data** (know nothing)
  - symbols or observations reflecting differences in the world, that represent properties of objects, events and their environments
  - lowest level of abstraction
  - of no use until they are useable, relevant form
- **information** (know what)
  - meaningful and processed data or facts which conclusions can be drawn by human or computer
  - when data is processed into an answer to an inquiry, it becomes information
- **knowledge** (know how)
  - information that is justifiably considered true
  - allows to promote information to a controlling role to transform information into instructions
- **widsom** (know why)
  - critical use of knowledge to make intelligent decisions
  - ability to make sound judgments and decisions and increase effectiveness

Association Rule Mining

- algorithm for discovering interesting rules or relations between variables in large datasets
- let $I = \{i_1, i_2, ..., i_n\}$ be a set of n binary attributes called *items*
- let $T = \{t_1, t_2, ..., t_m\}$ be a set of transactions
- a *rule* is defined as an implication of the form where $X \Rightarrow Y$ and $X, Y \subseteq I$ and $X \cap Y = \emptyset$
- the itemsets X and Y are called antecedent (LH side) and consequent (RH side) of the rule $X \Rightarrow Y$ respectively
- we are usually looking for interested rules
- the *support* $SUPP(X)$ of an itemset X is defined as the proportion of transactions in the dataset which contain the itemset X
- the *confidence* of a rule $X \Rightarrow Y$ is defined as $CONF(X \Rightarrow Y) = \frac{SUPP(X \cup Y)}{SUPP(X)}$
- association rules are usually required to satisfy a *minimum support* and a *minimum confidence*
- association rule generation splits up into two separate steps:
  1. minimum support is applied to find all frequent itemsets
  2. frequent itemsets and the minimum confidence constraint are used to form rules

Example

Transaction ID	Milk	Bread	Butter	Beer
1	1	1	0	0
2	0	1	1	0
3	0	0	0	1
4	1	1	1	0
5	0	1	0	0

Suppose our association rule is $\{Milk, Bread\} \Rightarrow \{butter\}$

milk	bread	butter
1	1	0
0	1	1
1	1	1
0	1	0

$$Support_{X \Rightarrow Y} = \frac{\sum_s Count_{x,y}}{|S|} = \frac{1}{5}$$

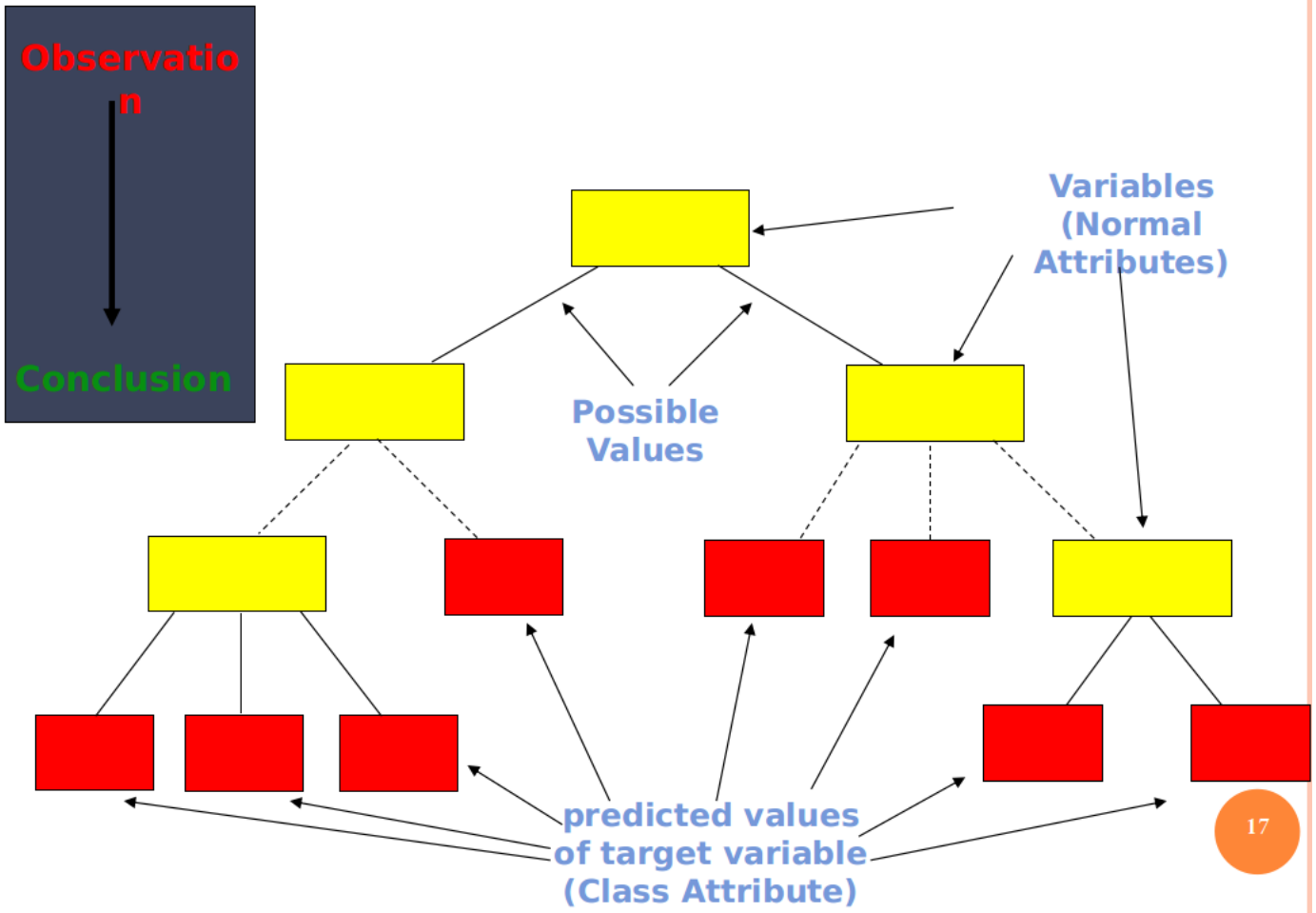
$$Support_X = \frac{\sum_s Count_x}{|S|} = \frac{2}{5}$$

$$Confidence_{X \Rightarrow Y} = \frac{Support_{X \Rightarrow Y}}{Support_X} = \frac{\frac{1}{5}}{\frac{2}{5}} = \frac{1}{2}$$

Decision Tree

- a tree-like structure in which
  - an *internal node* represents *test* on an attribute
  - each *branch* represents *outcome* of test
  - each *leaf node* represents *class label*
  - a *path from root to leaf* represents *classification rules*
- a tree structure wherein
  - *leaves* represent *classifications*
  - *branches* represent *conjunctions* of features that lead to those classifications
- *ID3*, Iterative Dichotomizer 3, is a decision tree induction algorithm developed by Quinlan

Predictive Model

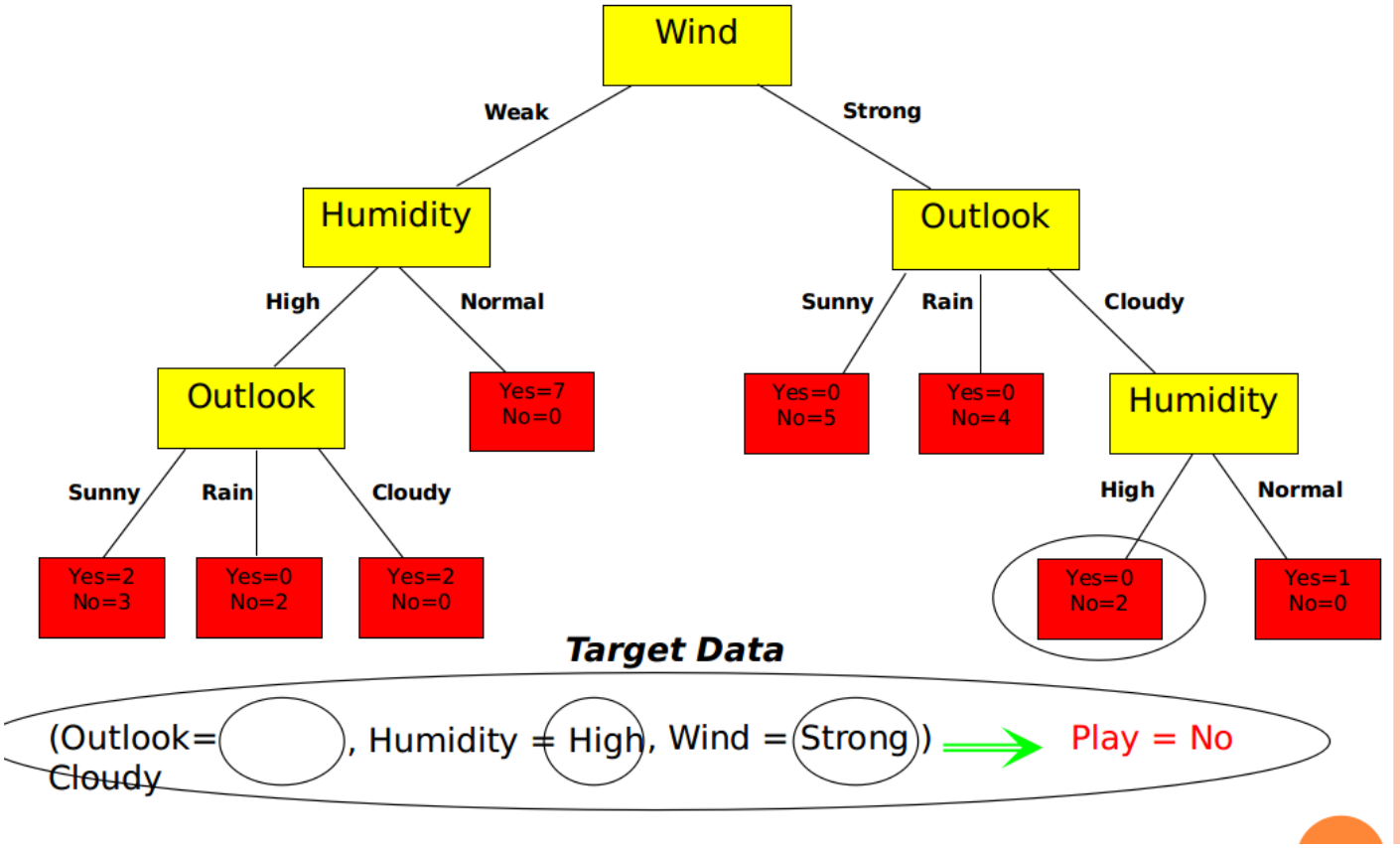


Day	Outlook	Humidity	Wind	Play
1	Sunny	High	Weak	No
2	Sunny	High	Weak	No
3	Cloudy	High	Strong	No
4	Rain	Normal	Strong	No
5	Rain	Normal		
6	Rain	High	Weak	
7			Weak	Yes
8	Sunny		Strong	No
9	Sunny	Normal	Strong	No
10	Sunny	High	Strong	No
11	Rain	Normal	Weak	Yes
12	Cloudy	Normal	Weak	Yes
13	Cloudy	High	Weak	Yes
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.

Class (or Dependent Attribute

Normal (or Independent) Attributes

Example:



K-Means Clustering

ALGORITHM: k-means clustering algorithm

1. Determine k entities as the initial means
2. *repeat*
3. assign each data entity to the closest mean
4. reconstruct the mean of each cluster
5. *until* means do not change

Machine Learning

- prediction, based on known properties learned from the training data
- two types of data involved:
  - training data
  - testing data
- standard techniques and algorithms
  - artificial neural networks
  - back-propagation
  - bayesian networks
  - extreme learning machine

Applications

- financial data analysis: credit fraud detection, trend analysis, analyzing profitability, etc
- marketing activities
- targeted adveretising
- healthcare and biomedical: disease progress analysis, adverse drug reactions, evaluation of effectiveness of medical treatments

When & Why Privacy is Needed?

- privacy acts
- financial competition
- top-secret data

Privacy-Preserving Data Mining

Main approaches: (change of data leak higher in approach #1)

1. Randomization and Anonymization

- challenge: accuracy vs privacy (privacy up, accuracy down)
- uses various techniques:
  - suppression
  - aggregation
  - anonymization
  - randomization
  - data perturbation

2. Secure Computation

- challenge: efficiency vs privacy
- uses various cryptography and security tools (building blocks)

Secure Multi-Party Computation (SMC)

Data is distributed:

- each party has a part of the whole data
- data could be partitioned: horizontally, vertically, or both
- involved parties want to operate a joint function on their private inputs
- functions could be: data mining algorithm, statistical analysis methods, mathematical functions
- concerns:
  - privacy: intermediate and/or final outputs reveal no info of private inputs
  - correctness (accuracy of final results)
  - efficiency

Parameters:

- parties behaviour: honest, semi-honest, malicious
- number of parties involved: two-party, multi-party
- parties network type: client-server, peer-to-peer, third-party
- type of final result release:
  - parties will receive the complete final output
  - parties will receive a portion of the final output

Examples:

- privacy-preserving Decision Tree
  - Information Gain
  - Gini Index
- privacy-preserving k-means Clustering
  - Secure Dot Product
  - Secure Comparison
- privacy-preserving Association Rule Mining
  - Secure Binary Dot Product
  - Cardinality of Set Intersection
  - Commutative Encryption
- privacy-preserving Neural Networks
  - Secure Dot Product
- privacy-preserving Bayesian Networks
  - Secure Exponentiation
  - Secure Factorial

Secure Mean

- n participants, each with a private number
- mean value of numbers computer securely and released to each person $M = \frac{\sum_{i=1}^{n} N_i}{n}$ where N is an array of private numbers
- none of the participants or any third party will know the private numbers of each other

Possible Issues

1. second and forth person compromise: they can reveal the third persons private number
2. presence of malicious person:
   - incorrect value can be shared by this person
   - none of the persons, except the malicious one, will receive the correct mean value

Solution for Issue 1

- data segmentation: each participant breaks her data into k segments
- multi-round protocol:
  - protocol will be performed in k rounds
  - in each round the order of the participants will be rotated

Homomorphic Encryption

An operation on the plaintexts will be mapped to another operation on the ciphertexts