# Data Mining

Unit # 1



# Course Outlines

- Classification Techniques
  - Classification/Decision Trees
  - Naïve Bayes
  - Neural Networks
- Clustering
  - Partitioning Methods
  - Hierarchical Methods
- Feature Selection
- Model Evaluation
- Patterns and Association Mining
- Text Mining
- Papers/Case Studies/Applications Reading

## Software/Data Repository

- KNIME
- R
- Data on Kaggle Website
  - http://www.kaggle.com/



## Useful Information

- Course Wiki
  - http://cse505fall2014.wikispaces.com/
- Text/Reference Books
  - Introduction to Data Mining by Tan, Steinbach and Kumar (2006)
  - Data Mining Concepts and Techniques by Han and Kamber (2011)
  - Data Mining: Practical Machine Learning Tools and Techniques by Witten and Frank (2011)

# Marks Distribution

- Midterms (2) 40
- Finals 40
- Projects (2) 20



# Data Science

- Data science incorporates varying elements and builds on techniques and theories from many fields, including mathematics, statistics, data engineering, pattern recognition and learning, advanced computing, visualization, uncertainty modeling, data warehousing, and high performance computing with the goal of extracting meaning from data and creating data products.
- A practitioner of data science is called a data scientist.

# What is a Data Scientist (Source: IBM)

- "A data scientist is somebody who is inquisitive, who can stare at data and spot trends." (Anjul Bhambri, Vice President of Big Data Products at IBM).
- The data scientist will sift through all incoming data with the goal of discovering a previously hidden insight, which in turn can provide a competitive advantage or address a pressing business problem.



BEST JOBS 2012

## Best new jobs in America

CNNMoney

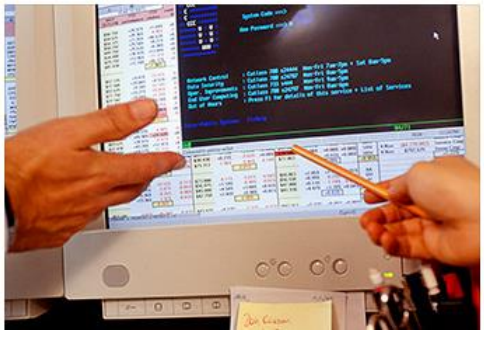
Recommend 893 Tweet 140 Share 262 +1 19 Email Print

### IT Data Scientist

3 of 6 BACK NEXT

10-year job growth: **18.7%**
Median pay: **$98,600**

**What they do all day?** Tech firms like LinkedIn, Facebook and Twitter are at the heart of the big data movement. Their users are generating loads of information by the second. Turning those heaps of data into business value falls to data scientists, who apply various tools and methods to find meaningful patterns and insights in large data sets.



## CNN Money (Cont'd)

- **How to get the job?** An affinity for numbers is key, as well as a command of computing, statistics, math and analytics. One can't underestimate the importance of soft skills either. Data scientists work closely with management and need to express themselves clearly.
- **What makes it great?** This is a cutting-edge field. The information explosion is spurring types of analysis that have never been performed before. The skill set is unique, and employers are willing to pay for qualified candidates. Six-figure paydays aren't uncommon.



## Big Data (Wikipedia)

- Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process the data within a tolerable elapsed time.
- Big data sizes are a constantly moving target, as of 2012 ranging from a few dozen terabytes to many petabytes of data in a single data set.

## Big Data Scientist

(http://www.bigdata-startups.com/job-description-big-data-scientist/)

- The big data scientist is said to be the sexiest job in the 21st century.
- Successful big data scientists will be in high demand and will be able earn very nice salaries.
- But to be successful, big data scientists need to have a wide range of skills that until now did not even fit into one department.



## Big Data Scientist (Cont'd)

- They need to have statistical, mathematical, predictive modelling as well as business strategy skills to build the algorithms necessary to ask the right questions and find the right answers.
- They also need to be able to communicate their findings, orally and visually.
- They need to understand how the products are developed and even more important, as big data touches the privacy of consumers, they need to have a set of ethical responsibilities.

## Why Would You Like to Invest in Analytics

- To optimize business operations
- To identify business risks
- To predict new business opportunities
- To comply with law or regulatory requirements



## Analytics

- The discovery and communication of meaningful patterns in data.
- Two major types are
  - Descriptive Analytics
  - Predictive Analytics

# Descriptive vs. Predictive Analytics

- Descriptive Analytics
  - what happened and why did it happen
  - Referred to as "unsupervised learning" in machine learning
- Predictive Analytics
  - what will happen
  - Referred to as "supervised learning" in machine learning



# Predictive analytics

- Predictive analytics techniques can be further divided into classification and prediction techniques.
- Classification Techniques
  - Classification Trees
  - Naïve Bayes
  - Random Forest
  - Neural Networks
  - Support Vector Machine
- Prediction
  - Regression Analysis
  - Time Series Analysis

# Descriptive Analytics

- Clustering
- Association Rule Mining
- Text Mining (to build word cloud, document clustering, etc.)



# Analytics Example

| | |
|---|---|
| Grouping items by similarity | Clustering |
| Discovering relationships between items | Association rules |
| Determining relationship between outcome and the input variables | Regression |
| Analyzing text data to find trending terms, sentiment analysis, document classification, etc. | Text analytics |
| Assigning label/class to records | Classification |

# BI vs. Data Analytics

- Business Intelligence (BI) focuses on using a consistent set of metrics to measure past performance and inform business planning.
- Data Analytics refers to a combination of analytical and machine learning techniques used for drawing inferences and insight out of data




# Analytics Eco System

- Data Savvy Professionals
  - Business User
  - Business Intelligence Analyst
- Technology and Data Enablers
  - Data Engineer
  - Database Administrator
- Deep Analytical Talent
  - Data Scientist

# Analytics

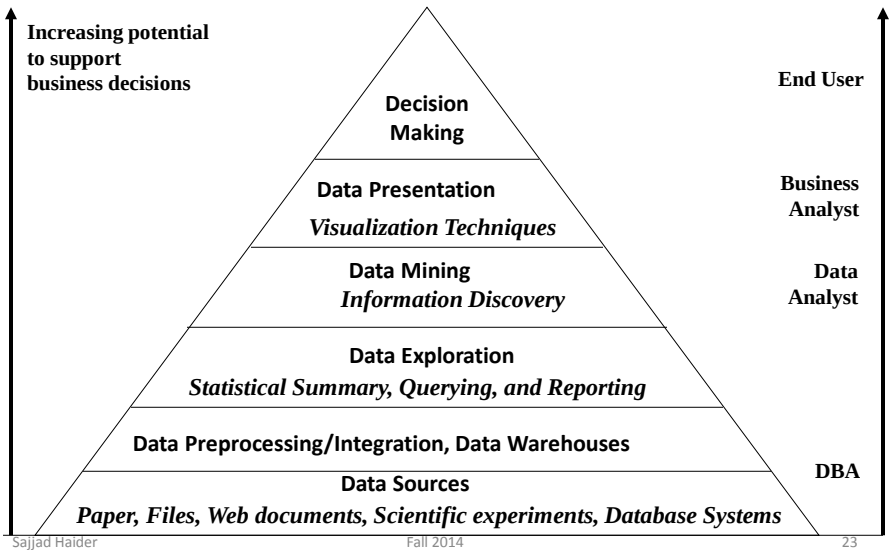
- Data Mining (Statistical Models)
- Computer Programming
- Operations Research



# What Is Data Mining?

- Data mining (knowledge discovery from data)
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
  - Data mining: a misnomer?
- Alternative names
  - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.
- Watch out: Is everything "data mining"?
  - Simple search and query processing
  - (Deductive) expert systems

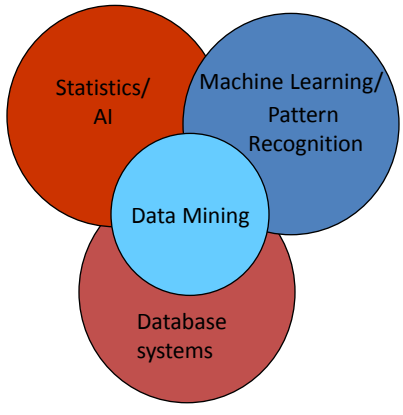
# Data Mining and Business Intelligence






# Origins of Data Mining

- **Draws ideas from machine learning/AI, pattern recognition, statistics, and database systems**
- **Traditional Techniques may be unsuitable due to**
  - **Enormity of data**
  - **High dimensionality of data**
  - **Heterogeneous, distributed nature of data**



# Are All the "Discovered" Patterns Interesting?

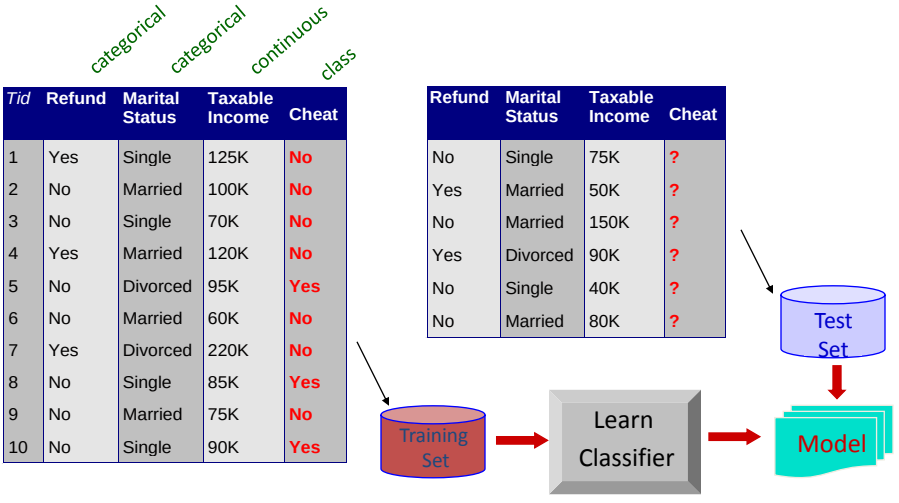
- Data mining may generate thousands of patterns: Not all of them are interesting
  - Suggested approach: Human-centered, query-based, focused mining
- **Interestingness measures**
  - A pattern is interesting if it is easily understood by humans, valid on new or test data with some degree of certainty, potentially useful, novel, or validates some hypothesis that a user seeks to confirm
- **Objective vs. subjective interestingness measures**
  - Objective: based on statistics and structures of patterns, e.g., support, confidence, etc.
  - Subjective: based on user's belief in the data, e.g., unexpectedness, novelty, actionability, etc.



# Classification: Definition

- Given a collection of records (*training set* )
  - Each record contains a set of *attributes*, one of the attributes is the *class*.
- Find a *model* for class attribute as a function of the values of other attributes.
- Goal: previously unseen records should be assigned a class as accurately as possible.
  - A *test set* is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

# Classification Example






# Classification: Application 1

- Direct Marketing
  - Goal: Reduce cost of mailing by *targeting* a set of consumers likely to buy a new cell-phone product.
  - Approach:
    - Use the data for a similar product introduced before.
    - We know which customers decided to buy and which decided otherwise. This *{buy, don't buy}* decision forms the *class attribute*.
    - Collect various demographic, lifestyle, and company-interaction related information about all such customers.
      - Type of business, where they stay, how much they earn, etc.
    - Use this information as input attributes to learn a classifier model.

From [Berry & Linoff] Data Mining Techniques, 1997

# Classification: Application 2

- Fraud Detection
  - Goal: Predict fraudulent cases in credit card transactions.
  - Approach:
    - Use credit card transactions and the information on its account-holder as attributes.
      - When does a customer buy, what does he buy, how often he pays on time, etc
    - Label past transactions as fraud or fair transactions. This forms the class attribute.
    - Learn a model for the class of the transactions.
    - Use this model to detect fraud by observing credit card transactions on an account.



# Classification: Application 3

- Customer Attrition/Churn:
  - Goal: To predict whether a customer is likely to be lost to a competitor.
  - Approach:
    - Use detailed record of transactions with each of the past and present customers, to find attributes.
      - How often the customer calls, where he calls, what time-of-the day he calls most, his financial status, marital status, etc.
    - Label the customers as loyal or disloyal.
    - Find a model for loyalty.

From [Berry & Linoff] Data Mining Techniques, 1997

# Clustering Definition

- Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that
  - Data points in one cluster are more similar to one another.
  - Data points in separate clusters are less similar to one another.
- Similarity Measures:
  - Euclidean Distance if attributes are continuous.
  - Other Problem-specific Measures.

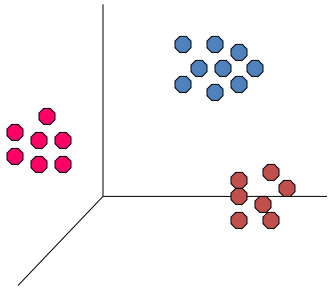


# Illustrating Clustering

☒Euclidean Distance Based Clustering in 3-D space.

Intracluster distances are minimized

Intercluster distances are maximized



# Clustering: Application 1

- Market Segmentation:
  - Goal: subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix.
  - Approach:
    - Collect different attributes of customers based on their geographical and lifestyle related information.
    - Find clusters of similar customers.
    - Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters.



# Clustering: Application 2

- Document Clustering:
  - Goal: To find groups of documents that are similar to each other based on the important terms appearing in them.
  - Approach: To identify frequently occurring terms in each document. Form a similarity measure based on the frequencies of different terms. Use it to cluster.
  - Gain: Information Retrieval can utilize the clusters to relate a new document or search term to clustered documents.

# Association Rule Discovery: Definition

- Given a set of records each of which contain some number of items from a given collection;
  - Produce dependency rules which will predict occurrence of an item based on occurrences of other items.

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

Rules Discovered:
{Milk} --> {Coke}
{Diaper, Milk} --> {Beer}



# Association Rule Discovery: Application 1

- Marketing and Sales Promotion:
  - Let the rule discovered be
    *{Bagels, … } --> {Potato Chips}*
  - Potato Chips as consequent => Can be used to determine what should be done to boost its sales.
  - Bagels in the antecedent => can be used to see which products would be affected if the store discontinues selling bagels.
  - Bagels in antecedent *and* Potato chips in consequent => Can be used to see what products should be sold with Bagels to promote sale of Potato chips!

# Association Rule Discovery: Application 2

- Supermarket shelf management.
  - Goal: To identify items that are bought together by sufficiently many customers.
  - Approach: Process the point-of-sale data collected with barcode scanners to find dependencies among items.
  - A classic rule --
    - If a customer buys diaper and milk, then he is very likely to buy beer.
    - So, don't be surprised if you find six-packs stacked next to diapers!



# Association Rule Discovery: Application 3

- Inventory Management:
  - Goal: A consumer appliance repair company wants to anticipate the nature of repairs on its consumer products and keep the service vehicles equipped with right parts to reduce on number of visits to consumer households.
  - Approach: Process the data on tools and parts required in previous repairs at different consumer locations and discover the co-occurrence patterns.

# Data Analytics Life Cycle (CRISP-DM)




# CRISP-DM (Cont'd)

- Business Understanding
  - Understand the project objectives and requirements
  - Can it be converted into a data mining problem definition
  - Were any effort made in the past? If yes, what were the findings? Why are we doing it again? What has changed?
  - Assess availability of time, technology and human resources. Do we have enough time and resources to execute the analytics project?
  - Identify the success criteria, key risks and major stake holders

# CRISP-DM (Cont'd)

- Data Understanding
  - Get familiar with the data. Is it enough to solve the stated business problem? If not, do we need to redesign the data collection process?
  - What's needed vs. what's available
  - Identify data quality problems
  - Determine the structures and tools needed
  - Discover first insights into the data



# CRISP-DM (Cont'd)

- Data Preparation
  - Construct the final dataset
  - Process likely to be repeated multiple times, and not in any prescribed order
  - Tasks include attribute selection as well as transformation and cleaning of data
  - Understand what to keep and what to discard
  - Extensive use of exploratory data analysis and visualization

# CRISP-DM (Cont'd)

- Modeling
  - Application of various modeling techniques and calibration of their parameters to optimal values
  - Documenting assumptions behind each modeling technique to get feedback from stake holders and domain experts
  - Typically require stepping back to the data preparation phase



# CRISP-DM (Cont'd)

- Evaluation
  - Test robustness of the models under consideration by gauging their performances against hold-out data
  - Analyze if the models achieve the business objectives.
  - Finalize a data mining model
  - Quantify business value and identify key findings

# CRISP-DM (Cont'd)

- Deployment
  - Typically a customer-driven stage instead of data analyst driven
  - important for the customer to understand up front the actions needed to actually make use of the created models.
  - Define process to update and retrain the model, as needed



# KDD Process: Several Key Steps

- Learning the application domain
  - relevant prior knowledge and goals of application
- Creating a target data set: data selection
- Data cleaning and preprocessing: (may take 60% of effort!)
- Data reduction and transformation
  - Find useful features, dimensionality/variable reduction, invariant representation
- Choosing functions of data mining
  - summarization, classification, regression, association, clustering
- Choosing the mining algorithm(s)
- Data mining: search for patterns of interest
- Pattern evaluation and knowledge presentation
  - visualization, transformation, removing redundant patterns, etc.
- Use of discovered knowledge

# Summary

- Data mining: Discovering interesting patterns from large amounts of data
- A natural evolution of database technology, in great demand, with wide applications
- A KDD process includes data cleaning, data integration, data selection, transformation, data mining, pattern evaluation, and knowledge presentation
- Data mining functionalities: characterization, discrimination, association, classification, clustering, outlier and trend analysis, etc.



# Fun Discussion: Best Batsman of Pakistan

- There are always discussions on the best batsman Pakistan has ever produced.
- Most of the times, the opinion is divided between Javed Miandad and Inzimam-ul-Haq.
- Who is your choice and why?
- Subjectivity vs. Objectivity in your analysis

# Best Cricketer of Pakistan (Cont'd)

| Test Record | Miandad | Inzamam |
|---|---|---|
| Average | 52.6 | 49.6 |
| Home | 61.4 | 53.8 |
| Away | 45.8 | 45.9 |
| 1st Inning | 69.1 | 51.6 |
| 2nd Inning | 45.5 | 60.9 |
| 3rd Inning | 39.8 | 51.7 |
| 4th Inning | 54.4 | 31.7 |



# Best Batsman of Pakistan (Cont'd)

| ODI | Miandad | Inzamam |
|---|---|---|
| Average | 41.7 | 39.5 |
| Home | 48.1 | 57.7 |
| Away | 41.3 | 38.2 |
| Neutral | 41.6 | 36.8 |
| 1st Inning | 42.5 | 40.0 |
| 2nd Inning | 43.9 | 41.7 |
| | | |

## Structured vs. Non-Structured Data

- Most business databases contain structured data consisting of well-defined fields with numeric or alpha-numeric values.
- Examples of semi-structured data are electronic images of business documents, medical reports, executive summaries, etc. The majority of web documents also fall in this category.
- An example of unstructured data is a video recorded by a surveillance camera in a departmental store. This form of data generally requires extensive processing to extract and structure the information contained in it.



## Structured vs. Non-Structured Data (Cont'd)

- Structured data is often referred to as traditional data, while the semi-structured and unstructured data are lumped together as non-traditional data.
- Most of the current data mining methods and commercial tools are applied to traditional data.

# SQL vs. Data Mining

- SQL (Structured Query Language) is a standard relational database language that is good for queries that impose some kind of constraints on data in the database in order to extract an answer.
- In contrast, data mining methods are good for queries that are exploratory in nature, trying to extract hidden, not so obvious information.
- SQL is useful when we know exactly what we are looking for and we can describe it formally.
- We use data mining methods when we know only vaguely what we are looking for.



# OLAP vs. Data Mining

- OLAP tools make it very easy to look at dimensional data from any angle or to slice-and-dice it.
- The derivation of answers from data in OLAP is analogous to calculations in a spreadsheet; because they use simple and given-in-advance calculations.
- OLAP tools do not learn from data, not do they create new knowledge.
- They are usually special-purpose visualization tools that can help end-users draw their own conclusions and decisions, based on graphically condensed data.

## Statistics vs. Machine Learning

- Data mining has its origins in various disciplines, of which the two most important are *statistics* and *machine learning*.
- Statistics has its roots in mathematics, and therefore, there has been an emphasis on mathematical rigor, a desire to establish that something is sensible on theoretical grounds before testing it in practice.
- In contrast, the machine learning community has its origin very much in computer practice. This has led to a practical orientation, a willingness to test something out to see how well it performs, without waiting for a formal proof of effectiveness.



## Statistics vs. Machine Learning (Cont'd)

- Modern statistics is entirely driven by the notion of a model. This is a postulated structure, or an approximation to a structure, which could have led to the data.
- In place of the statistical emphasis on models, machine learning tends to emphasize algorithms.

# Classification: Definition

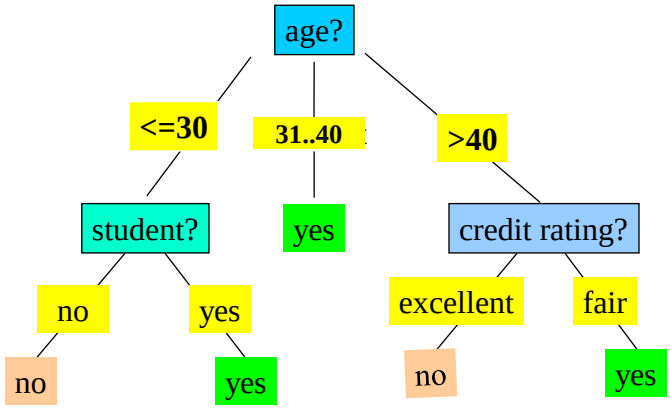
- Given a collection of records (*training set* )
  - Each record contains a set of *attributes*, one of the attributes is the *class*.
- Find a *model* for class attribute as a function of the values of other attributes.
- Goal: previously unseen records should be assigned a class as accurately as possible.
  - A *test set* is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.



# Classification: Motivation

| age | income | student | credit_rating | buys_computer |
|---|---|---|---|---|
| <=30 | high | no | fair | no |
| <=30 | high | no | excellent | no |
| 31…40 | high | no | fair | yes |
| >40 | medium | no | fair | yes |
| >40 | low | yes | fair | yes |
| >40 | low | yes | excellent | no |
| 31…40 | low | yes | excellent | yes |
| <=30 | medium | no | fair | no |
| <=30 | low | yes | fair | yes |
| >40 | medium | yes | fair | yes |
| <=30 | medium | yes | excellent | yes |
| 31…40 | medium | no | excellent | yes |
| 31…40 | high | yes | fair | yes |
| >40 | medium | no | excellent | no |

# Decision/Classification Tree





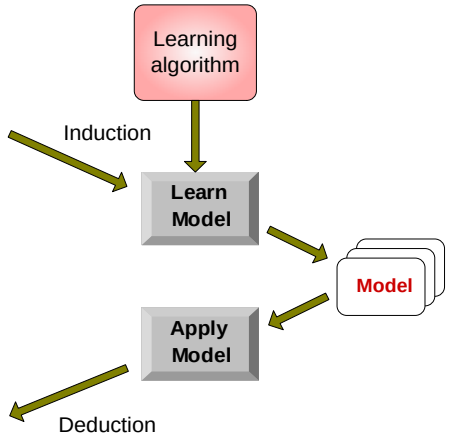

# Illustrating Classification Task

| Tid | Attrib1 | Attrib2 | Attrib3 | Class |
|---|---|---|---|---|
| 1 | Yes | Large | 125K | No |
| 2 | No | Medium | 100K | No |
| 3 | No | Small | 70K | No |
| 4 | Yes | Medium | 120K | No |
| 5 | No | Large | 95K | Yes |
| 6 | No | Medium | 60K | No |
| 7 | Yes | Large | 220K | No |
| 8 | No | Small | 85K | Yes |
| 9 | No | Medium | 75K | No |
| 10 | No | Small | 90K | Yes |

Training Set

| Tid | Attrib1 | Attrib2 | Attrib3 | Class |
|---|---|---|---|---|
| 11 | No | Small | 55K | ? |
| 12 | Yes | Medium | 80K | ? |
| 13 | Yes | Large | 110K | ? |
| 14 | No | Small | 95K | ? |
| 15 | No | Large | 67K | ? |

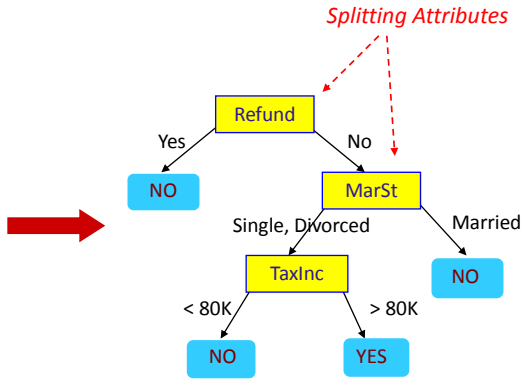
Test Set



# Example of a Decision Tree

| Tid | Refund (categorical) | Marital Status (categorical) | Taxable Income (continuous) | Cheat (class) |
|---|---|---|---|---|
| 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 |

Training Data

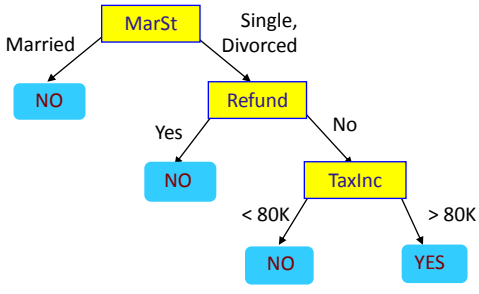



Model: Decision Tree



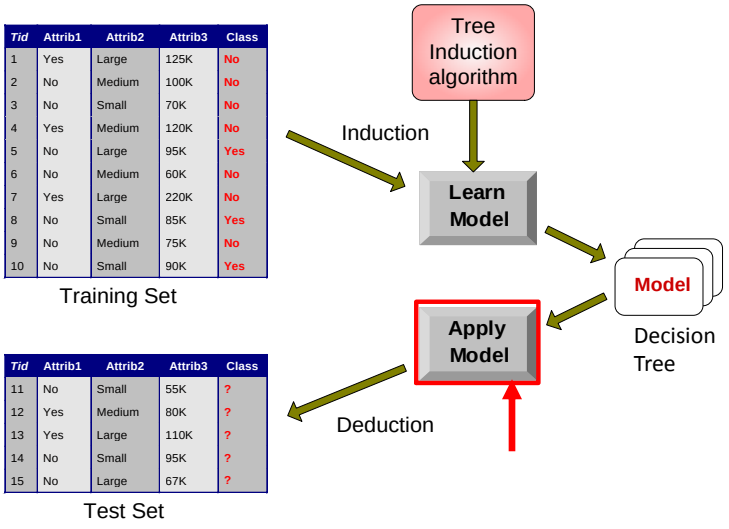
# Another Example of Decision Tree

| Tid | Refund (categorical) | Marital Status (categorical) | Taxable Income (continuous) | Cheat (class) |
|---|---|---|---|---|
| 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 |




There could be more than one tree that fits the same data!

# Decision Tree Classification Task




| Tid | Attrib1 | Attrib2 | Attrib3 | Class |
|---|---|---|---|---|
| 1 | Yes | Large | 125K | No |
| 2 | No | Medium | 100K | No |
| 3 | No | Small | 70K | No |
| 4 | Yes | Medium | 120K | No |
| 5 | No | Large | 95K | Yes |
| 6 | No | Medium | 60K | No |
| 7 | Yes | Large | 220K | No |
| 8 | No | Small | 85K | Yes |
| 9 | No | Medium | 75K | No |
| 10 | No | Small | 90K | Yes |

Training Set

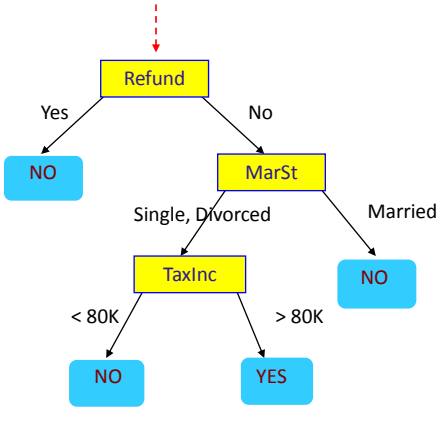
| Tid | Attrib1 | Attrib2 | Attrib3 | Class |
|---|---|---|---|---|
| 11 | No | Small | 55K | ? |
| 12 | Yes | Medium | 80K | ? |
| 13 | Yes | Large | 110K | ? |
| 14 | No | Small | 95K | ? |
| 15 | No | Large | 67K | ? |

Test Set



# Apply Model to Test Data

Start from the root of tree.



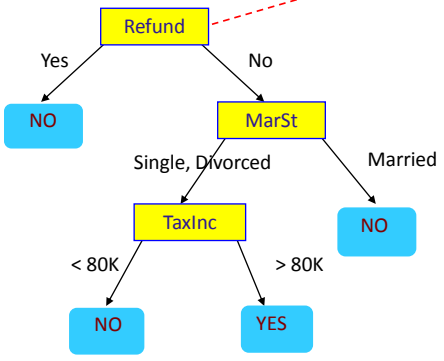

Test Data

| Refund | Marital Status | Taxable Income | Cheat |
|---|---|---|---|
| No | Married | 80K | ? |

# Apply Model to Test Data

Test Data

| Refund | Marital Status | Taxable Income | Cheat |
|---|---|---|---|
| No | Married | 80K | ? |

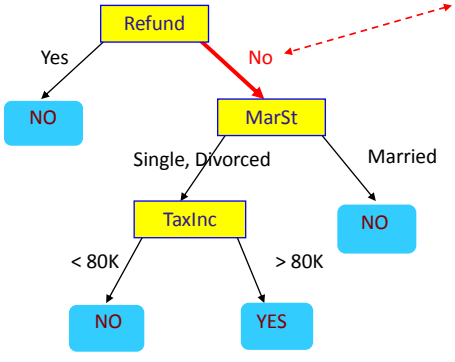





# Apply Model to Test Data

Test Data

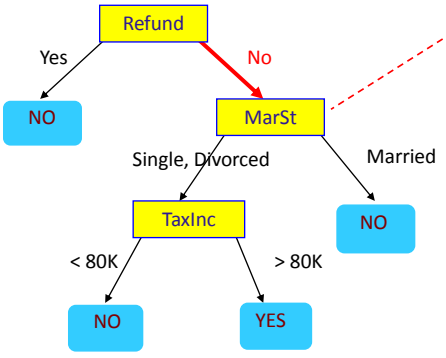
| Refund | Marital Status | Taxable Income | Cheat |
|---|---|---|---|
| No | Married | 80K | ? |



# Apply Model to Test Data

Test Data

| Refund | Marital Status | Taxable Income | Cheat |
|---|---|---|---|
| No | Married | 80K | ? |

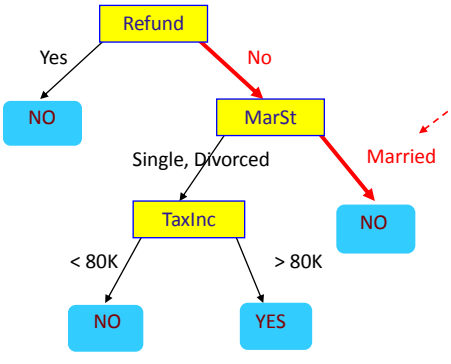





# Apply Model to Test Data

Test Data

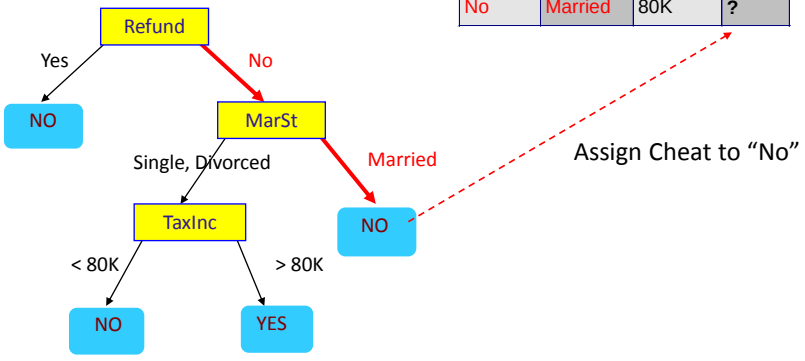
| Refund | Marital Status | Taxable Income | Cheat |
|---|---|---|---|
| No | Married | 80K | ? |



# Apply Model to Test Data

Test Data

| Refund | Marital Status | Taxable Income | Cheat |
|---|---|---|---|
| No | Married | 80K | ? |





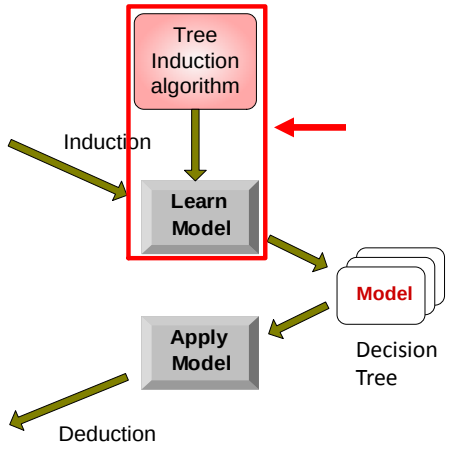

# Decision Tree Classification Task

| Tid | Attrib1 | Attrib2 | Attrib3 | Class |
|---|---|---|---|---|
| 1 | Yes | Large | 125K | No |
| 2 | No | Medium | 100K | No |
| 3 | No | Small | 70K | No |
| 4 | Yes | Medium | 120K | No |
| 5 | No | Large | 95K | Yes |
| 6 | No | Medium | 60K | No |
| 7 | Yes | Large | 220K | No |
| 8 | No | Small | 85K | Yes |
| 9 | No | Medium | 75K | No |
| 10 | No | Small | 90K | Yes |

Training Set

| Tid | Attrib1 | Attrib2 | Attrib3 | Class |
|---|---|---|---|---|
| 11 | No | Small | 55K | ? |
| 12 | Yes | Medium | 80K | ? |
| 13 | Yes | Large | 110K | ? |
| 14 | No | Small | 95K | ? |
| 15 | No | Large | 67K | ? |

Test Set



# Tree Induction

- Greedy strategy.
  - Split the records based on an attribute test that optimizes certain criterion.
- Issues
  - Determine how to split the records
    - How to specify the attribute test condition?
    - How to determine the best split?
  - Determine when to stop splitting



# How to Specify Test Condition?

- Depends on attribute types
  - Nominal
  - Ordinal
  - Continuous
- Depends on number of ways to split
  - 2-way split
  - Multi-way split

# How to determine the Best Split

- Greedy approach:
  - Nodes with homogeneous class distribution are preferred
- Need a measure of node impurity:

| C0: 5<br>C1: 5 | C0: 9<br>C1: 1 |
|---|---|
| Non-homogeneous,<br>High degree of impurity | Homogeneous,<br>Low degree of impurity |



# Measures of Node Impurity

- Gini Index
- Entropy
- Misclassification error

# Measure of Impurity: GINI

- Gini Index for a given node t :

$$GINI(t) = 1 - \sum_j [p(j \mid t)]^2$$

(NOTE: $p(j \mid t)$ is the relative frequency of class j at node t).

  - Maximum (1 - $1/n_c$) when records are equally distributed among all classes, implying least interesting information
  - Minimum (0.0) when all records belong to one class, implying most interesting information

| C1 | 0 |
|---|---|
| C2 | 6 |
| Gini=0.000 | |

| C1 | 1 |
|---|---|
| C2 | 5 |
| Gini=0.278 | |

| C1 | 2 |
|---|---|
| C2 | 4 |
| Gini=0.444 | |

| C1 | 3 |
|---|---|
| C2 | 3 |
| Gini=0.500 | |



# Examples for computing GINI

$$GINI(t) = 1 - \sum_j [p(j \mid t)]^2$$

| C1 | 0 |
|---|---|
| C2 | 6 |

P(C1) = 0/6 = 0 P(C2) = 6/6 = 1

Gini = $1 - P(C1)^2 - P(C2)^2 = 1 - 0 - 1 = 0$

| C1 | 1 |
|---|---|
| C2 | 5 |

P(C1) = 1/6 P(C2) = 5/6

Gini = $1 - (1/6)^2 - (5/6)^2 = 0.278$

| C1 | 2 |
|---|---|
| C2 | 4 |

P(C1) = 2/6 P(C2) = 4/6

Gini = $1 - (2/6)^2 - (4/6)^2 = 0.444$

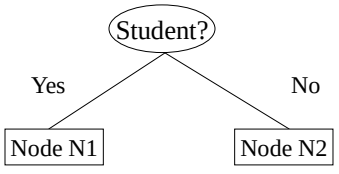
# Classification: Motivation

| age | income | student | credit_rating | buys_computer |
|---|---|---|---|---|
| <=30 | high | no | fair | no |
| <=30 | high | no | excellent | no |
| 31…40 | high | no | fair | yes |
| >40 | medium | no | fair | yes |
| >40 | low | yes | fair | yes |
| >40 | low | yes | excellent | no |
| 31…40 | low | yes | excellent | yes |
| <=30 | medium | no | fair | no |
| <=30 | low | yes | fair | yes |
| >40 | medium | yes | fair | yes |
| <=30 | medium | yes | excellent | yes |
| 31…40 | medium | no | excellent | yes |
| 31…40 | high | yes | fair | yes |
| >40 | medium | no | excellent | no |



# Binary Attributes: Computing GINI Index

- Splits into two partitions
- Effect of Weighing partitions:
  - Larger and Purer Partitions are sought for.




**Gini(N1)**
**= 1 – (6/7)² – (1/7)²**
**= 0.24**

**Gini(N2)**
**= 1 – (3/7)² – (4/7)²**
**= 0.49**

**Gini(Student)**
**= 7/14 * 0.24 +**
**7/14 * 0.49**
**= ??**

## GINI Index for Buy Computer Example

- Gini (Income):
- Gini (Credit_Rating):
- Gini (Age):