Introduction to Data Science

CS 5665
Utah State University
Department of Computer Science
Instructor: Prof. Kyumin Lee

Practicum

- This Thursday
  - AWS (Anuj)
  - Tableau (Astha)
  - Pandas (Prakhar & Hans)
  - OpenRefine (Ashwani)

Project Teams

- Sahit Katragadda, Sree Vidya Susarla, Shivakesh Reddy Annepally
- Arshdeep Singh, Venkatesh Kadali, Rohit Gopalan
- Sidhant Chatterjee, Sirisha Rani Deekonda
- Jake Felzien, Mike Larsen, and Yancy Knight
- Bhagyashree, Meiling, Anuj
- Mounika, Akhil, Pravallika
- Prakhar Amlathe, Amitesh Mahajan, Vaibhav Sahu
- Aditya, Hans Gunther, Karun Joseph
- Astha Tiwari, Kamna Yadav, Ashwani Chahal
- Ruchi Chauhan, Vishal Sharma
- Nick, Jacob, Vahe

HW1

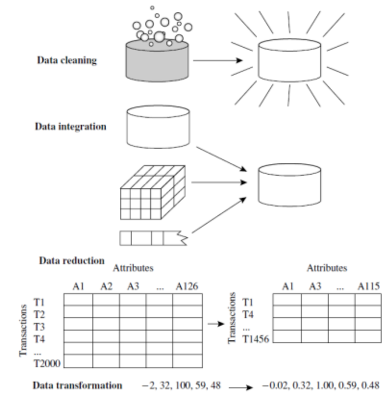
- https://usu.instructure.com/courses/431417/assignments/2117107
- Due date: September 22

Previous Class…

Data Integration
→ Correlation coefficient, Covariance

Data Reduction
→ Dimensionality Reduction, Numerosity Reduction

Previous Class…



Data Transformation and Data Discretization

Data Transformation

- A function that maps the entire set of values of a given attribute to a new set of replacement values so that each old value can be identified with one of the new values
- Why conduct data transformation?
  - The resulting mining process may be more efficient, the patterns found may be easier to understand
- Data Transformation Methods
  - Smoothing: Remove noise from data
  - Attribute/feature construction
    - New attributes constructed from the given ones
  - Aggregation: Summarization,
  - Normalization: Scaled to fall within a smaller, specified range
    - min-max normalization
    - z-score normalization
    - normalization by decimal scaling
  - Discretization: raw values of numeric attributes (e.g., age) replaced by interval labels (e.g., 0-10, 11-20, etc.) or conceptual labels (e.g., youth, adult, senior)

Normalization

- **Min-max normalization**: to [new_min$_A$, new_max$_A$]

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

  - Ex. Let income range \$12,000 to \$98,000 normalized to [0.0, 1.0]. Then \$73,600 is mapped to $\frac{73{,}600 - 12{,}000}{98{,}000 - 12{,}000}(1.0 - 0) + 0 = 0.716$
- **Z-score normalization** (μ: mean, σ: standard deviation):

$$v' = \frac{v - \mu_A}{\sigma_A}$$

  - Ex. Let μ = 54,000, σ = 16,000. Then $\frac{73{,}600 - 54{,}000}{16{,}000} = 1.225$
- **Normalization by decimal scaling**

$$v' = \frac{v}{10^j}$$ Where j is the smallest integer such that Max(|ν'|) < 1

Discretization

- Three types of attributes
  - Nominal: values from an unordered set, e.g., color, profession
  - Ordinal: values from an ordered set, e.g., military or academic rank
  - Numeric: real numbers, e.g., integer or real numbers
- Discretization: Divide the range of a continuous attribute to intervals
  - Interval labels can then be used to replace actual data values
  - Reduce data size by discretization
  - Supervised vs. unsupervised
  - Split (top-down) vs. merge (bottom-up)
  - Discretization can be performed recursively on an attribute
  - Prepare for further analysis e.g., classification

Data Discretization Methods

- **Binning**
  - Top-down split, unsupervised
- **Histogram analysis**
  - Top-down split, unsupervised
- **Clustering analysis** (unsupervised, top-down split or bottom-up merge)
- **Decision-tree analysis** (supervised, top-down split)
- **Correlation analysis** (unsupervised, bottom-up merge)

Binning Methods for Data Smoothing

❑ Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

* Partition into equal-frequency (**equi-depth**) bins:
- Bin 1: 4, 8, 9, 15
- Bin 2: 21, 21, 24, 25
- Bin 3: 26, 28, 29, 34

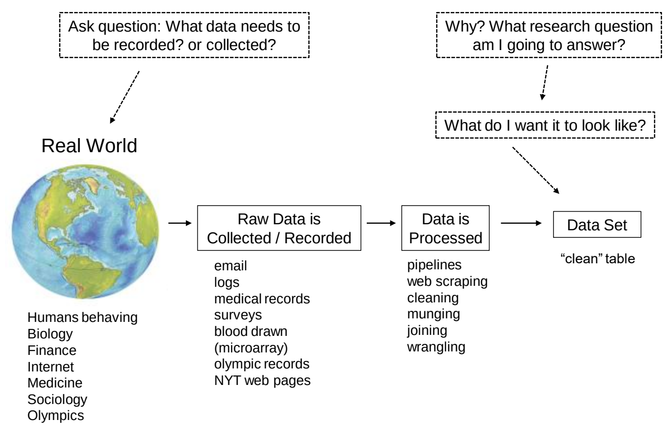
* Smoothing by **bin means**:
- Bin 1: 9, 9, 9, 9
- Bin 2: 23, 23, 23, 23
- Bin 3: 29, 29, 29, 29

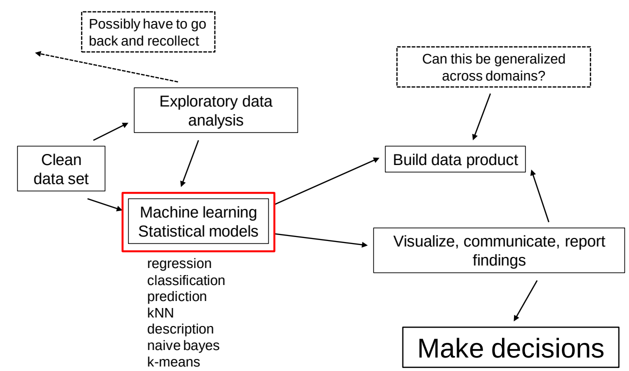
* Smoothing by **bin boundaries**:
- Bin 1: 4, 4, 4, 15
- Bin 2: 21, 21, 25, 25
- Bin 3: 26, 26, 26, 34

Summary

- **Data quality**: accuracy, completeness, consistency, timeliness, believability, interpretability
- **Data cleaning**: e.g. missing/noisy values, outliers
- **Data integration** from multiple sources:
  - Entity identification problem
  - Remove redundancies
  - Detect inconsistencies
- **Data reduction**
  - Dimensionality reduction
  - Numerosity reduction
  - Data compression
- **Data transformation and data discretization**
  - Normalization
- Read section 3 in Data Mining Concepts and Techniques

Overview of Mining and Analytics





Some classic approaches …

- Classification (predictive)
- Clustering (descriptive)
- Associate rule discovery (descriptive)
- Regression (predictive)
- Anomaly detection (predictive)

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: 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.

Classification: 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.

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.

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}

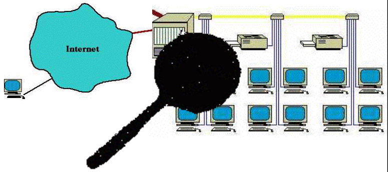
Regression

- Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency.
- Greatly studied in statistics, neural network fields.
- Examples:
  - Predicting sales amounts of new product based on advertising expenditure.
  - Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
  - Time series prediction of stock market indices.

Anomaly Detection

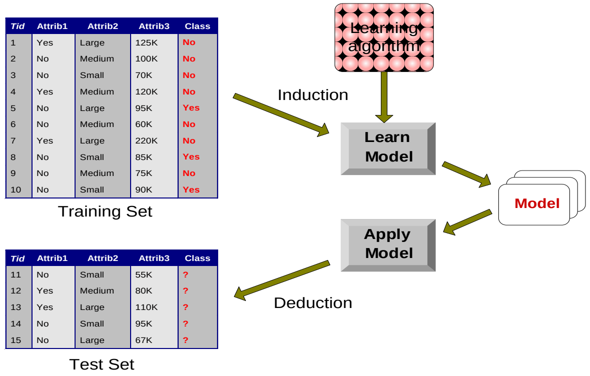
- Detect significant deviations from normal behavior

- Applications:
  - Credit Card Fraud Detection
  - Network Intrusion Detection



Mining and Analytics: Classification + Decision Trees

Illustrating Classification Task



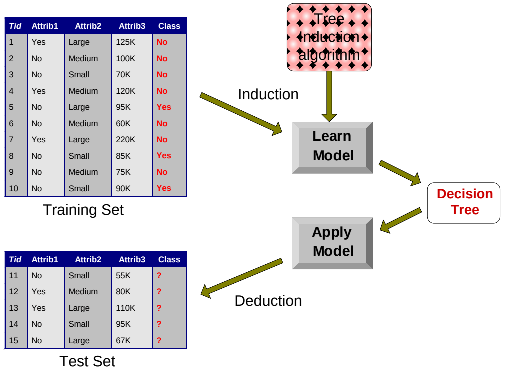
Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

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	?

Classification Techniques

- **Decision trees ← today**
- Naive Bayes
- Nearest Neighbors (KNN)
- Support Vector Machines
- …

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

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	?

What is a Decision Tree?

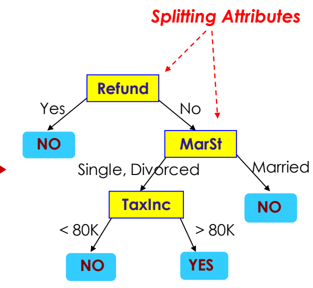
- Hierarchical structure of **nodes** and **directed edges**
  - – **Root node**: no incoming edges; zero or more outgoing edges
  - – **Internal node**: one incoming edge; two or more outgoing edges
  - – **Leaf node**: one incoming edge; no outgoing edges; **Labeled with a class**

Example of a Decision Tree

categorical categorical continuous class

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

Training Data

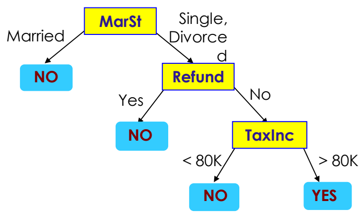


Model: Decision Tree

Another Example of Decision Tree

categorical categorical continuous class

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



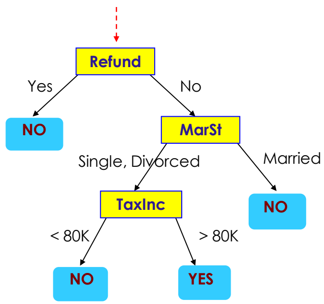
There could be more than one tree that fits the same data!

The Hope

- The decision tree (or whatever classifier we use) **generalizes** to new data!!
  - – So we can have confidence in it

Apply Model to Test Data

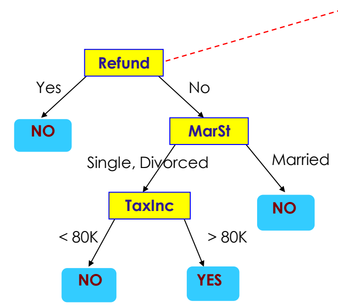
Start from the root of tree.



Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

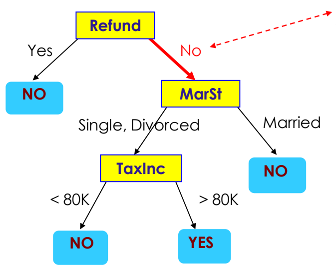
Apply Model to Test Data



Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

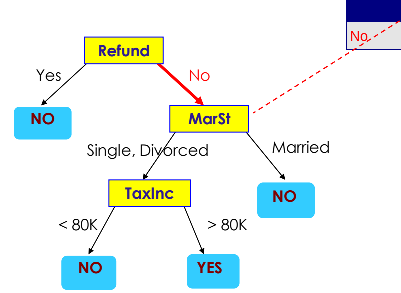
Apply Model to Test Data



Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

Apply Model to Test Data



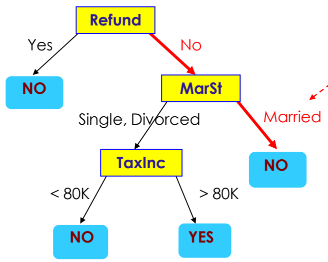
Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

Apply Model to Test Data

Test Data

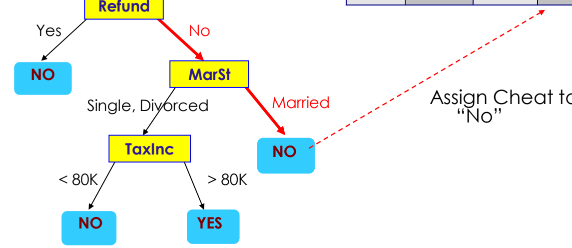
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Apply Model to Test Data

Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Why Decision Trees?

- Popular!
- Relatively inexpensive to build
- Fast to classify new data
- **Easy to interpret**

But first, we must "Learn the model"

– (i.e., build the right decision tree)

Lots of approaches

- Hunt's Algorithm
- CART
- ID3, C4.5
- SLIQ,SPRINT

- **Main ideas:**
  – Tree induction + tree pruning

General Structure of Hunt's Algorithm

- ***[Recursively apply]*** Let D_t be the set of training records that are associated with node t and y = {y_1, y_2, … y_c} be the set of class labels
  – If D_t contains records that belong the same class y_t, then its decision tree consists of a leaf node labeled as y_t
  – If D_t is an empty set, then its decision tree is a leaf node whose class label is determined from other information such as the majority class of the records
  – If D_t contains records that belong to several classes, then a **test condition** based on one of the attributes of D_t is applied to split the data into more homogenous subsets

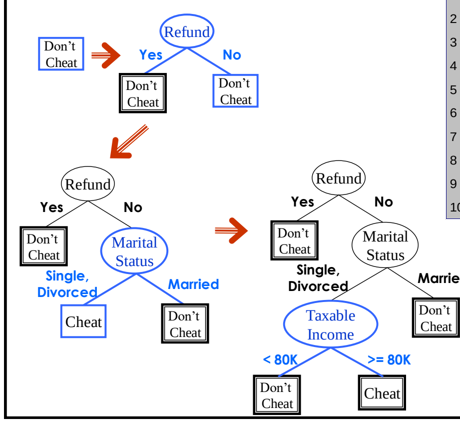
Example

- Attributes:
  – Refund (Yes, No)
  – Martial Status (Single, Divorced, Married)
  – Taxable Income (quantitative)

- Class:
  – Cheat, Don't Cheat

Hunt's Algorithm

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



Tree Induction

- Determine how to split the records
  - Use greedy heuristics to make a series of locally optimum decision about which attribute to use for partitioning the data
  - At each step of the greedy algorithm, a test condition is applied to split the data in to subsets with a more homogenous class distribution
    - How to specify test condition for each attribute
    - How to determine the best split
- Determine when to stop splitting
  - A stopping condition is needed to terminate tree growing process. Stop expanding a node
    - if all the instances belong to the same class
    - if all the instances have similar attribute values