Data Mining Tutorial

Mark A. Austin

University of Maryland

austin@umd.edu

ENCE 688P, Fall Semester 2021

October 16, 2021

Overview

Part 01

Quick Review

Artificial Intelligence (AI) and Machine Learning (ML)

Technical Implementation (2020, Google, Siemens, IBM)

- AI and ML will be deeply embedded in new software and algorithms.

Artificial Intelligence:

- Knowledge representation and reasoning with ontologies and rules. Semantic graphs. Executable event-based processing.

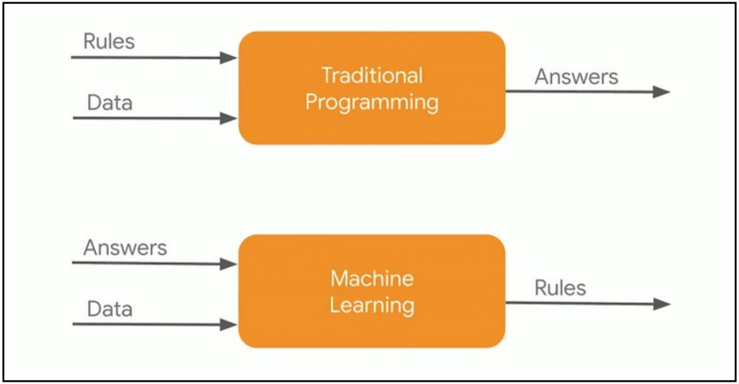
Machine Learning:

- Modern neural networks. Input-to-output prediction.
- Data mining.
- Identify objects, events, and anomalies.
- Learn structure and sequence. Remember stuff.

Man and Machine (AI-ML View)

Man	AI-ML Machine
• Good at formulating solutions to problems. • Can work with incomplete data and information. • Creative. • Reasons logically, but very slow. Forgetful. • Performance is static. • Humans make the rules, then they break them.	• Manipulates Os and 1s. • Can work with incomplete data and information. • Creative. • Fast logical reasoning. • Performance doubles every 18-24 months. • Data mining can discover the rules.

Traditional Programming vs AI-ML Workflow



Introduction to Data Mining

Numerous Definitions

Data Mining

The field of data mining addresses the question of how to best use historical data to discover general regularities and improve future decisions (Mitchell, 1999).

Data Mining

Data mining is the extraction of implicit, previously unknown, and potentially useful information – structural patterns – from data (Witten et al., 2017).

The process of discovering useful patterns from data must be automatic (or at least semi-automatic). Useful patterns allow us to make nontrivial predictions on new data.

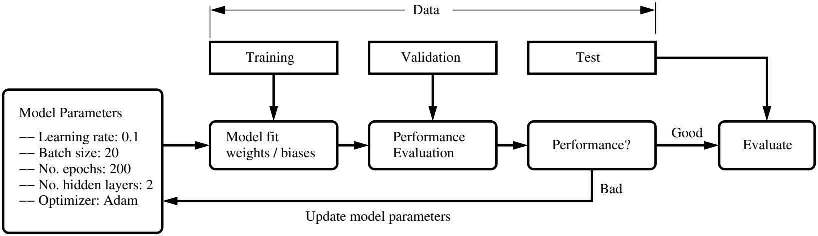
Data Mining Techniques

Working with Initial Dataset

- Data cleaning and curation
- Remove redundant features
- Identify input variables and output variable.

Preprocessed Dataset:

- Data split: 80% for training, 20% for validation and testing.



Data Mining Techniques

Training Dataset

- The sample of data used to fit the model.

Validation Dataset

- The sample of data used to provide an unbiased evaluation of the model fit on the training dataset while training the model parameters.

Testing Dataset

- The sample of data used to provide an unbiased evaluation of a final model fit on the training dataset.

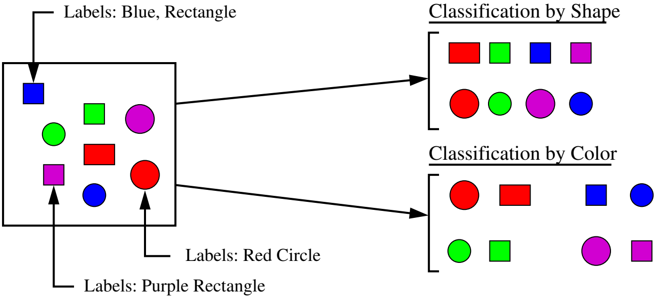
Data Mining Techniques

Data Mining Techniques

Classification Analysis

Classification analysis learns a method for predicting the instance class from pre-labeled (classified) instances.

Classification by Shape/Color (Supervised Learning)



Data Mining Techniques

Classification Problem

- **Given** a set of n attributes (ordinal or categorical), a set of k classes, and a set of labeled training instances,

$$[(i_i, l_i), \cdots, (i_j, l_j)], \quad (1)$$

where $i = (v_1, v_2, \cdots, v_n)$,
and $l \in (c_1, c_2, \cdots, c_k)$.

- **Goal** is to determine a classification rule – sequence of tests on the attributes – that predicts the class of any instance from the values of its attributes.

Note

- This is a generalization of the concept learning problem since typically there are more than two (outcome) classes.
- Data will contain scatter; may have missing values.

Data Mining Techniques

Decision Trees.

A structure that includes a root node, branches, and leaf nodes. Each internal node represents a test on an attribute; each branch represents the outcome of a test; and each leaf represents a class label.

Arbitrary Boolean Functions

- Each attribute is binary valued (true or false).
- Example trees: XOR, AND and OR, etc ...

Continuous Domains

- Each attribute is real valued (true or false).
- Tests check if $a_i >$ value.

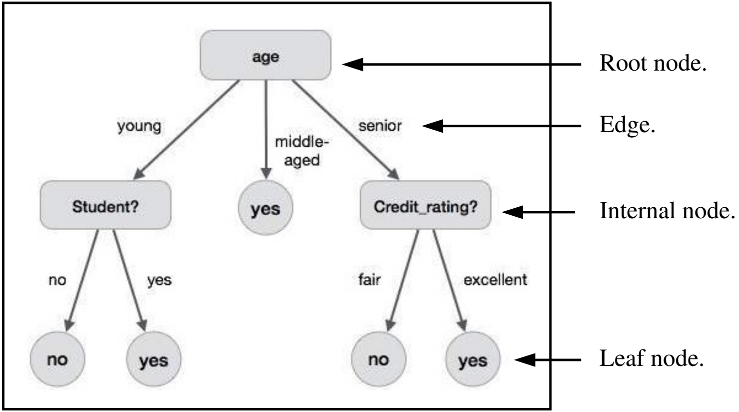
Data Mining Techniques

Sample Dataset. Will customer buy a computer?

ID	Age Group	Income	Student	Credit Rating	Buys Computer
1	young	high	no	fair	no
2	young	high	no	excellent	no
3	middle	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle	low	yes	excellent	yes
8	young	medium	no	fair	no
9	young	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	young	medium	yes	excellent	yes
12	middle	medium	no	excellent	yes
13	middle	high	yes	fair	yes
14	senior	medium	no	excellent	no

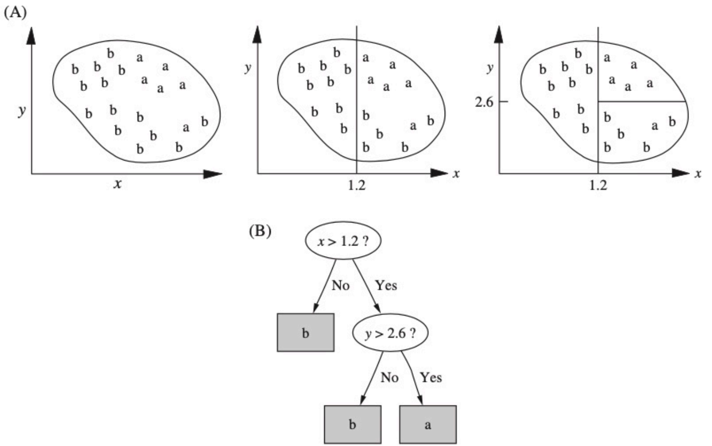
Data Mining Techniques

Sample Decision Tree (Split on Discrete Domain)



Data Mining Techniques

Covering Algorithm and Rule Construction (Split on Continuous Domain)

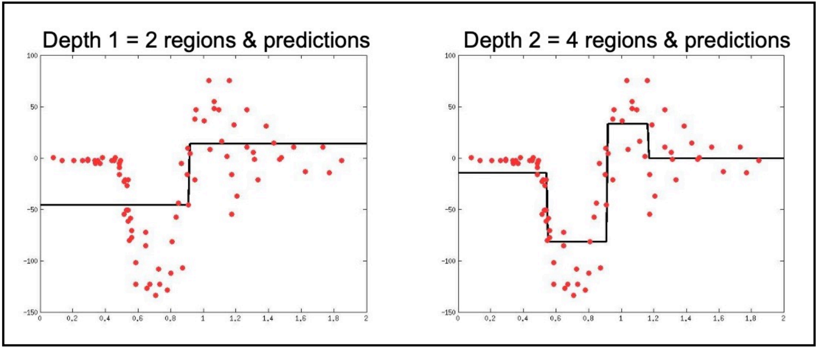


Data Mining Techniques

Decision Trees for Regression (One-Dimensional Regression)

- Goal is to predict real-valued numbers at the leaf nodes.

Prediction of a Single Scalar Feature

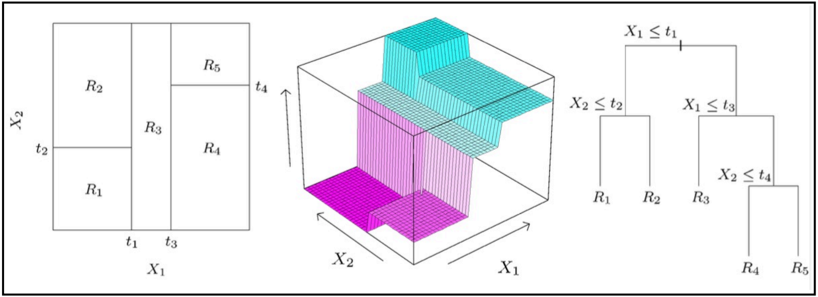


Data Mining Techniques

Decision Trees for Regression (Two-Dimensional Regression)

- Each node splits tree according to a single feature.
- Mean values of training data are predicted at leaf nodes.

Example



Data Mining Techniques

Basic Questions:

- How to choose the attribute (or value) to split on at each level of the tree?
- When should a node be declared a leaf?
- If a leaf is impure, how should it be labeled?
- If the tree is too large, how can it be pruned?

Notes on Strategy:

- When all of the data in a single node comes from the same class, can declare the node to be a leaf and stop splitting.
- When a group of data points have exactly the same attribute values, we cannot split any further. Declare the node to be a leaf, and output the class that is the majority.

Data Mining Techniques

Algorithms

- Perceptron.
- Logistic Regression.
- Decision tree algorithms (C4.5, J48)
- Support Vector Machines (SVM).
- Random Forest.

Applications

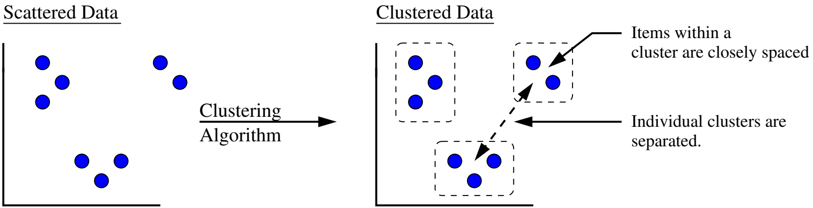
- Anomaly (Fraud) detection.
- Medical diagnosis.
- Industrial applications.

Data Mining Techniques

Clustering Problems

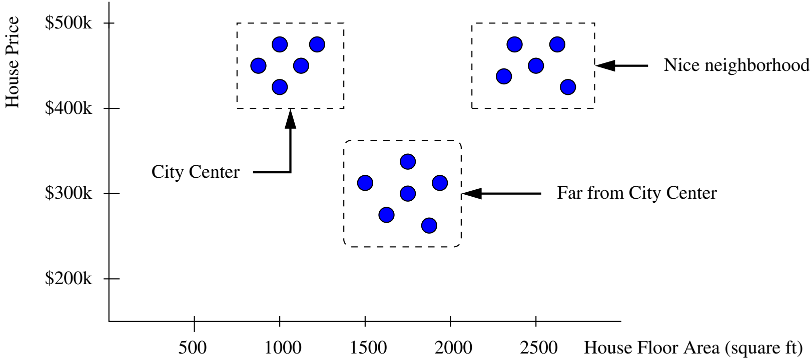
Clustering techniques apply when there is no class to be predicted, but when un-labeled instances need to be divided into common natural groups.

Clustering Process (Unsupervised Learning)



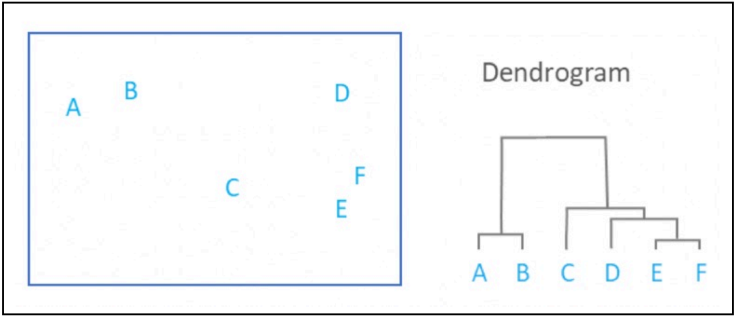
Data Mining Techniques

Example 1. Clustering of House Prices and Floor Areas



Data Mining Techniques

Example 2. Hierarchical Clustering and Dendrograms



Dendrogram

A dendrogram is a branching (tree) diagram that represents relationships of similarity among groups of entities.

Data Mining Techniques

Algorithms

- K-means clustering.
- Hierarchical clustering.

Applications

- Preprocessing step for many scientific applications.
- Natural language processing.
- Market segmentation.
- Netflix/movie recommendations.

Data Mining Techniques

Association

Association is a data mining function that discovers the probability the co-occurrence of items (or patterns) in a collection of data.

Association Rules

- Identify relationships between co-occurring items can be expressed as association rules (e.g., if X, then Y).

Key Challenges

- How to identify useful correlations among all correlations?
- Correlation relationships are not the same as dependency relationships – *if X, then Y* does not *imply if Y, then X* !
- Historical data does not necessarily predict the future.

Data Mining Techniques

Goals of Predictive Analysis

- For a customer who purchases product A, what other products will they purchase?
- Will coupons increase same-store sales?
- Will a reduced price mean higher sales?

Retail Strategies

- Put most frequently purchased item (e.g., milk) at the back of the store.
- Co-locate items that are bought together – can lead to increase in sales for both.

Data Mining Techniques

Example 1. iPhone Color and Personality Traits.

Phone Color	Personality Traits
Green	Fresh, harmonious, healthy, hopeful.
Blue	Confident, dependable, trustworthy.
Yellow	Happy, honorable, intelligent.
Pink	Compassionate, energetic, playful.
White	Balanced, calm, clean.

Customers want to select an iPhone Color that correlates with their personality traits.

Data Mining Techniques

Example 2. Urban Legend from early 1990s: Diapers and Beer

ID	Items
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{Milk, Diapers, Beer, Cola}
4	{Bread, Milk, Diapers, Beer}
5	{Bread, Milk, Diapers, Cola}
...	...

market basket transactions

Examples of Association Rules

- $\{Diapers\} \longrightarrow \{Beer\}$,
- $\{Milk, Bread\} \longrightarrow \{Eggs, Coke\}$,
- $\{Beer, Bread\} \longrightarrow \{Milk\}$.

Data Mining Techniques

Itemset and **k-Itemset**

- A collection of one or more items (e.g., $\{Milk, Bread\}$.
- k-Itemset is an itemset containing k items.

Support Count σ

- Frequency of ocurrence of an itemset.
- Example: $\sigma(\{Milk, Bread, Diaper\}) = 2$.

Support

- Indicates how frequently the if/then relationship appears in the data.

Association Rule

- Expression of the form X $\longrightarrow$ Y, where X and Y are itemsets.

Data Mining Techniques (Rule Evaluation Metrics)

Support (s)

- Fraction of transactions that contain both X and Y.
- Support(s) $= \frac{\sigma\{Milk, Diaper, Beer\}}{T} = 2/5 = 0.4$.

Confidence (c)

- Measures how often items in Y appear in transactions that contain X.
- Confidence(c) $= \frac{\{Milk, Diaper, Beer\}}{\{Milk, Diaper\}} = 2/3 = 0.67$.

Data Mining for Association Rules

Given a set of transactions T, find all rules having:

- Support(s) $\geq$ min support threshhold.
- Confidence(c) $\geq$ min confidence threshold.

Data Mining Techniques (Brute-Force Enumeration)

Brute-Force Enumeration

- Compute support and confidence for all possible association rules.
- Prune rules that do not meet min support/confidence thresholds.

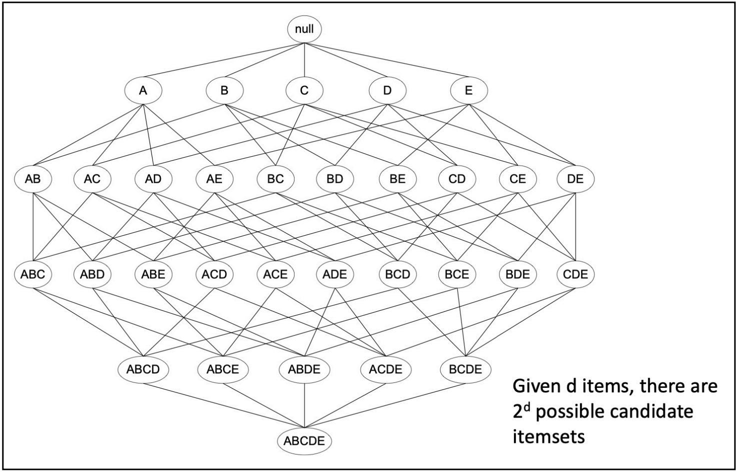
TID	*Items*
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Example of Rules:

{Milk,Diaper} → {Beer} (s=0.4, c=0.67)
{Milk,Beer} → {Diaper} (s=0.4, c=1.0)
{Diaper,Beer} → {Milk} (s=0.4, c=0.67)
{Beer} → {Milk,Diaper} (s=0.4, c=0.67)
{Diaper} → {Milk,Beer} (s=0.4, c=0.5)
{Milk} → {Diaper,Beer} (s=0.4, c=0.5)

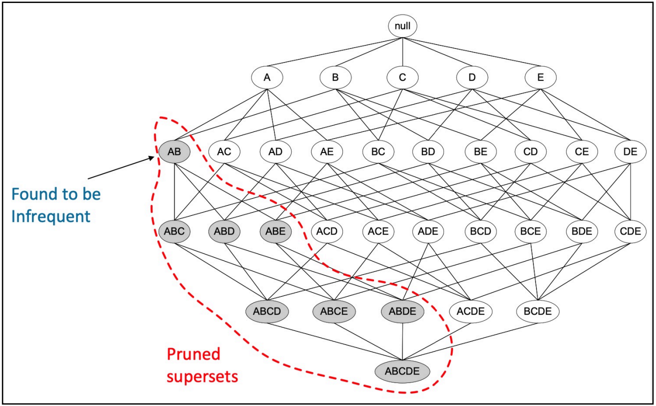
Data Mining Techniques (Brute-Force Enumeration)

Computational Complexity: Given d items, there are 2^d possible candidate itemsets.



Data Mining Techniques (Brute-Force Enumeration)

Need strategies to reduce computational effort by systematically pruning the low scoring items from candidate space.



Data Mining Techniques

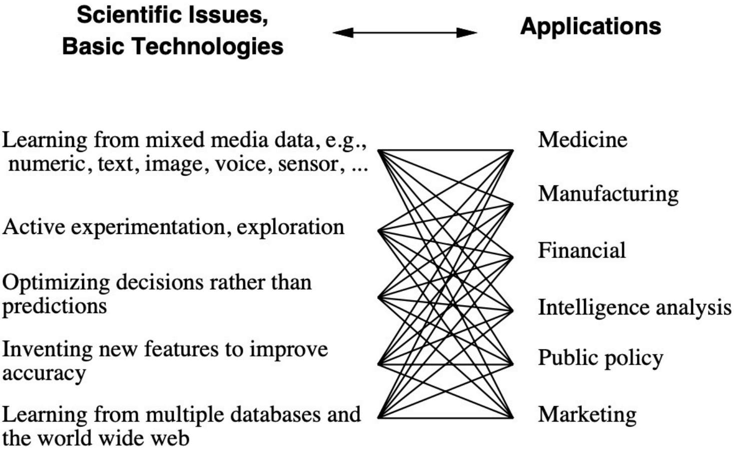
Algorithms (see Chapter 6 of Witten et al.)

- **Apriori**: Follows a generate-and-test methodology for finding frequent item sets, generating successively longer candidate item sets, and then scanning the item sets to see if they meet threshold limits.
- **Frequent Pattern Trees**: Begins by counting the number of times individual items – attribute-value pairs – occur in the dataset. This is a single pass. Then, a (sorted) tree structure is constructed with the goal of identifying large (frequent) item sets.

Applications

- Weather prediction,
- Medical diagnosis,
- Purchasing habits of retail customers.

Scientific Research Enabling Applications



Source: Mitchell, 1999.

References

- Jaynes E.T., Information Theory and Statistical Mechanics. II, Phys. Rev. 108, 171, October 1957.
- Kapur J.N., Maximum-Entropy Models in Science and Engineering, John Wiley and Sons, 1989.
- Mitchell T.M., Machine Learning and Data Mining, Communications of the ACM, Vol. 42., No. 11, November 1999.
- Russell S., and Norvig P., Artificial Intelligence: A Modern Approach (Third Edition), Prentice-Hall, 2010.
- Shanon C.E., and Weaver W., The Mathematical Theory of Communication, University of Illinois, Urbana, Chicago, 1949.
- Witten I.H., Frank E., Hall M.A., and Pal C.J., Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, 2017.