Pattern Discovery

Methods & Applications

DSA 8102: Data mining, data storage and retrieval

Strathmore UNIVERSITY

D. Owuor (Ph.D)

Objectives

- Describe the pattern mining concepts.
- Describe sequential pattern mining, constraint-based mining, graph pattern mining.
- Demonstrate basic classification and clustering algorithms.
- Demonstrate the Apriori algorithm in R using the market basket analysis.

Preliminary

- Data mining uses mathematical analysis to derive patterns and trends that exist in data (Microsoft, 2019).
- Discovering patterns in Big data is a non-trivial task without data mining.
- Some of the most fundamental data mining tasks are:
  - clustering, classification, pattern mining etc.

Data Mining Tasks

Strathmore UNIVERSITY

- Prediction methods
  - use some variables to predict unknown or future values of other variables.

Earthquakes
Weather forecast
Cancer detection

- Description methods
  - find human-interpretable patterns that describe data.

Gene description
Demographic description
Land description
Transaction description

Strathmore
UNIVERSITY

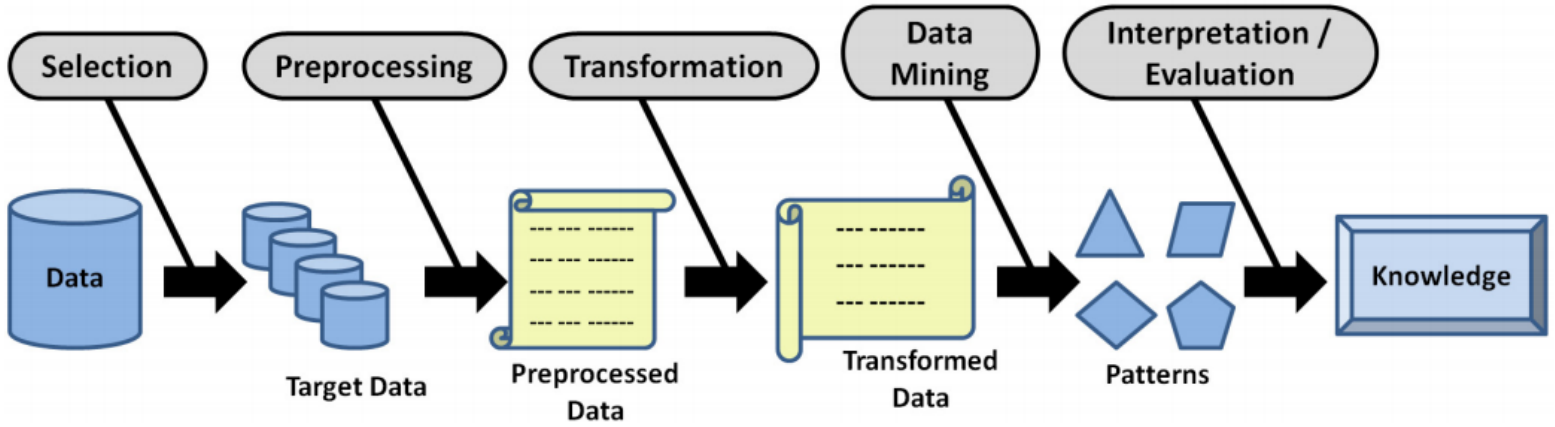
Descriptive Data Mining

Strathmore
UNIVERSITY

Pattern Mining

Patterns

- The term “pattern” refers to a subset of the data expressed in the form of rules (Gullo, 2015).
- A pattern means that your data are correlated into a relationship.
- **Pattern mining** concentrates on identifying rules (i.e., association) that describe specific patterns within the data (Britannica, n.d).



Patterns during the KDD Process (Gullo, 2015)

Key Terms

- **Item:** An item is any particular object.
  - Example: {Milk}
- **Item set:** Set of items that occur together or a collection of items.
  - Example: {Milk, Diaper, Beer}
- **Association rule:** This is a technique used to uncover how items are associated to each other. *An implication expression of the form*, X → Y, where X and Y are item sets.
  - Example: {Milk, Diaper} → {Beer}
- **Frequent patterns:** given a user-defined threshold, frequent item sets appear (in the data set's transactions) more than the threshold.

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

Transactions of items

TID	Milk	Bread	Butter	Beer
1	1	0	1	1
2	1	1	1	0
3	0	1	1	0
4	1	0	0	1
5	1	1	1	1

Association Rules

Strathmore UNIVERSITY

- Let the rule discovered be:
  {Milk, Diaper} → {Beer}
- Beer as consequent => can be used to determine what should be done to boost its sales.
- Milk in the antecedent => can be used to see which products would be affected if the store discontinues selling milk.
- Milk in antecedent and Beer in consequent => can be used to see what products should be sold with Milk to promote sale of Beer!

Rule Quality:
Evaluation Metrics

Strathmore UNIVERSITY

- **Support count (σ):** Frequency of occurrence of an item set.
  - σ({Milk, Diaper, Beer}) = 2
  - σ({Milk, Diaper}) = 3
- **Support (s):** Fraction of transactions that contain an item set.
  - s({Milk, Diaper, Beer}) = 2/5 = 0.4
  - s({Milk, Diaper}) = 3/5 = 0.6
  - s({Beer}) = 3/5 = 0.6
- **Confidence (c):** Measures how often items in Y appear in transactions that contain X.
  - c (X → Y) = σ({Milk, Diaper, Beer}) / σ({Milk, Diaper}) = 2/3 = 0.67
- **Lift (l):** This is the ratio of the confidence of the rule and the expected confidence.
  - l = 0.67 / 0.6 = 1.11

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

$$Rule\ X \Longrightarrow Y$$

$$Support = \frac{Frequency\ (X,Y)}{N}$$

$$Confidence = \frac{Frequency\ (X,Y)}{Frequency(X)}$$

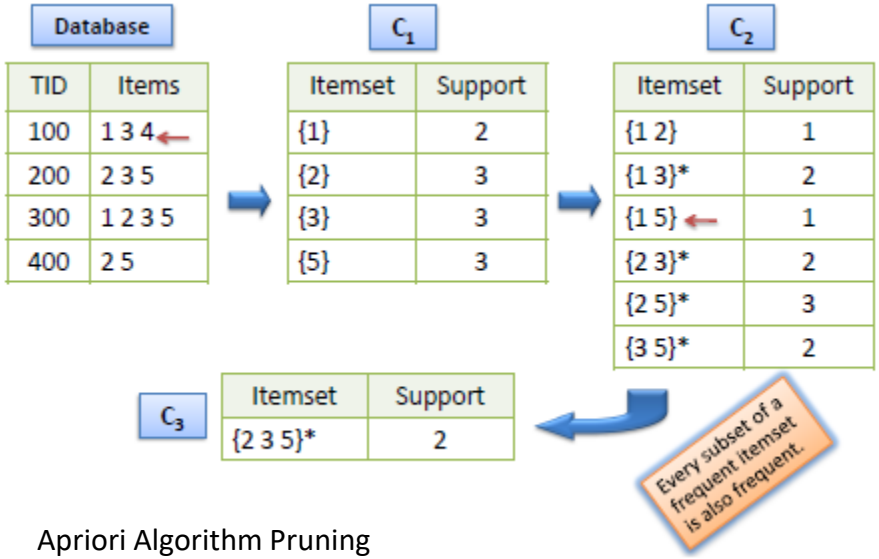
$$Lift = \frac{Confidence(X \rightarrow Y)}{Expected\ Confidence}$$

$$Expected\ Confidence = Support(Y)$$

Apriori Algorithm

Strathmore UNIVERSITY

- This is used to mine frequent item sets and association rules (Agrawal and Srikant, 1994).
- First algorithm that was proposed for frequent item set mining (Muliono et al. 2019).
- Designed to work on databases that contain transactions.
- **Advantage:**
  - Easy to understand
  - Results are intuitive and easy to understand
- **Disadvantage:**
  - It requires a higher computation if the item sets are very large and if the min. support is kept very low
  - Entire database (transactions) needs to be scanned



Apriori Algorithm Pruning

Exercise

Strathmore UNIVERSITY

1. Discuss FP-growth as an alternative to Apriori algorithm.
2. Figure 2 contains transactional items, answer the following:
   a. What is the support and support count of the items marked in red?
   b. What is the confidence of the item set marked in red?

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}

Strathmore
UNIVERSITY

Frequent Pattern Mining Domains

Sequential Patterns Mining

Strathmore UNIVERSITY

CustID	TransID	Transactions
1	100	a,b,c,d
3	111	a,f,d,c
1	122	d,e,p
3	133	b,f,s,a
1	144	b,c,d,e
3	155	a,f,d,c
1	166	a,e,p

We can represent all Customer 1 and Customer 3 transactions sequences, given some time reference:

CustID	Sequences
1	(a,b,c,d) (d,e,p) (b,c,d,e) (a,e,p)
3	(a,f,d,c) (b,f,s,a) (a,f,d,c)

- Sequential pattern mining (SPM) is the process of finding frequent sequences of item sets in a dataset to identify patterns of ordered events.
  - Example: S = {a1, a2, a3, a4, a5}
  - In this sequence, S, a1 comes before a2, then a3, …
- It generally intends to discover meaningful subsequences from a group of sequences.
  - **Occurrence Frequency and time-bound are key!**
- **Applications areas:**
  - Customer shopping sequences: buy a computer, then CDROM, and then digital camera within 3 months
  - Fraud detection
  - Telephone calling patterns
  - Natural disasters (e.g. earthquakes)
  - Stock markets

Constraint Pattern Mining

Strathmore UNIVERSITY

- Constraint-based mining involves searching for patterns or model space restricted by constraints (Nijssen, 2021).
- Goals of *constraint* pattern matching:
  - Increase the effectiveness of the search (ignore trivial patterns)
  - Reduce the number of patterns that are presented to the user
  - Make knowledge discovery more effective and useful
- Types of Constraints:
  - Knowledge type constraint: classification, association, etc.
  - Data constraint (using SQL-like queries): find product pairs sold together in stores in Pretoria in 2018.
  - Dimension / level constraint: In relevance to price, brand, customer, location, category, time, etc.
  - Rule (or pattern) constraint: Small sales (price < KES10) and fetch Big sales (Sum > KES 500)
  - Interestingness constraint: strong rules when (min_support ≥ 3%, min_confidence ≥ 60%)

Customer A: {TV} … {DVD Player}

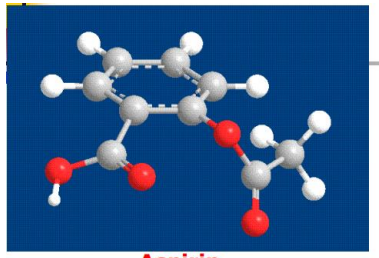
Customer B: {TV} … {DVD Player}

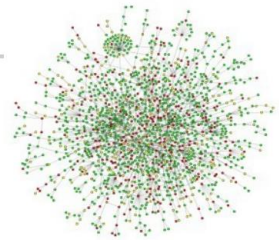
The sequential pattern of interest is {TV}, {DVD Player} which suggests that people who buy TV will also soon buy DVD player.

Timing constraints: After 10 years, those who bought TV may not purchase a DVD

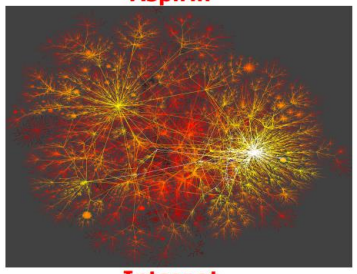
Graph Pattern Mining

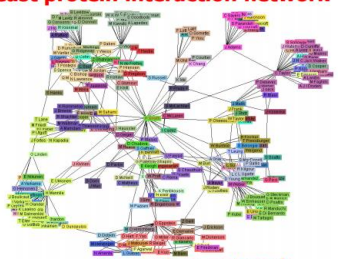
- Most of existing mining algorithms are based on Flat transaction representation, i.e., sets of items.
- **Motivation:** Datasets with structures do not fit well in flat transactions.
  - E.g., protein sequences, chemical compounds etc.
  - Graphs are suitable for capturing arbitrary relations between the various elements. E.g., how connected articles are on the Internet, etc.
- Graph Mining is the problem of discovering repetitive subgraphs occurring in the input graphs
- **Application areas:**
  - Detection of financial crimes (Jedrzejek et. Al.)
  - Drug development (Christian Borgelt et al.)
  - Customer behaviour analysis (Yada, 2004)

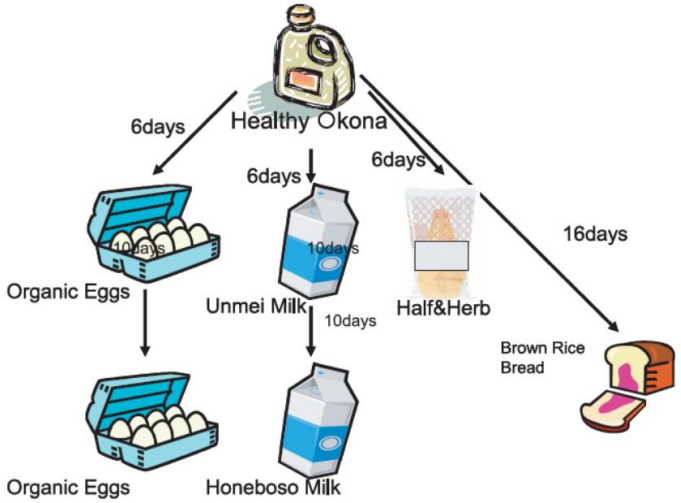

Aspirin



Yeast protein interaction network


Internet


Co-author network



Gradual Pattern Mining

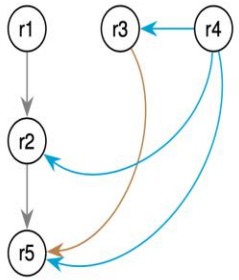
Strathmore UNIVERSITY

id	date (day/month)	exercise (hours)	stress levels
r1	01/06	1	4
r2	04/06	2	2
r3	05/06	3	3
r4	10/06	1	2
r5	12/06	4	1

gradual pattern : $\{(exercise, \uparrow), (stress, \downarrow)\}$

GRITE - **GR**adual **I**Temset **E**xtraction

- depth-first precedence graph



- support : $\frac{3}{5}$

GRAANK - **GRA**dual r**ANK**ing

- Kendall's τ gradual ranking (concordant pairs)
- 10 possible ordering pairs : [r1,r2], [r1,r3], [r1,r4], [r1,r5], [r2,r3], [r2,r4], [r2,r5], [r3,r4], [r3,r5], [r4,r5]
- 5 concordant pairs : [r1,r2], [r1,r5], [r2,r5], [r3,r5], [r4,r5]
- support : $\frac{5}{10}$

- Mines correlations between attributes of a data set through gradual rules/patterns
- Main algorithms: GRAANK and GRITE (Laurent 2009)

Application Areas

Strathmore UNIVERSITY

- **Recommender systems**
  - present users with selected and personalized subset of items from a huge set of distinct candidate items (Deldjoo et al. 2020; Beheshti et al. 2020).
- **Intrusion detection**
  - classifies the characteristics of signatures used in misuse intrusion detection (Obeidat and AlZubi, 2019; Aldwairi et al., 2020).
- **Transaction data systems**
  - understand sequences in customers' transaction history -- what they have previously bought (important for promotions) (Li et al. 2018; Sarma and Roy, 2010)
- **Business process logs**
  - discover process models from event logs in both software processes and business(Bogarin & Cerezo, 2018)
- **Spatial data**
  - arrangement of individual entities in space and the geographic relationships among them (Shekhar, Evans, Kang, & Mohan, 2011)
- **Biological sequences**
  - help biologists understand the functions of and relationships among different genes (Wu et al. 2013; Chen and Wu, 2013)

Exercise

1. Explain the difference between frequent pattern mining and gradual pattern mining?
2. Constraint patterns are richer in knowledge than sequential patterns. Explain.

Strathmore
UNIVERSITY

Predictive Data Mining

Pre-requisite

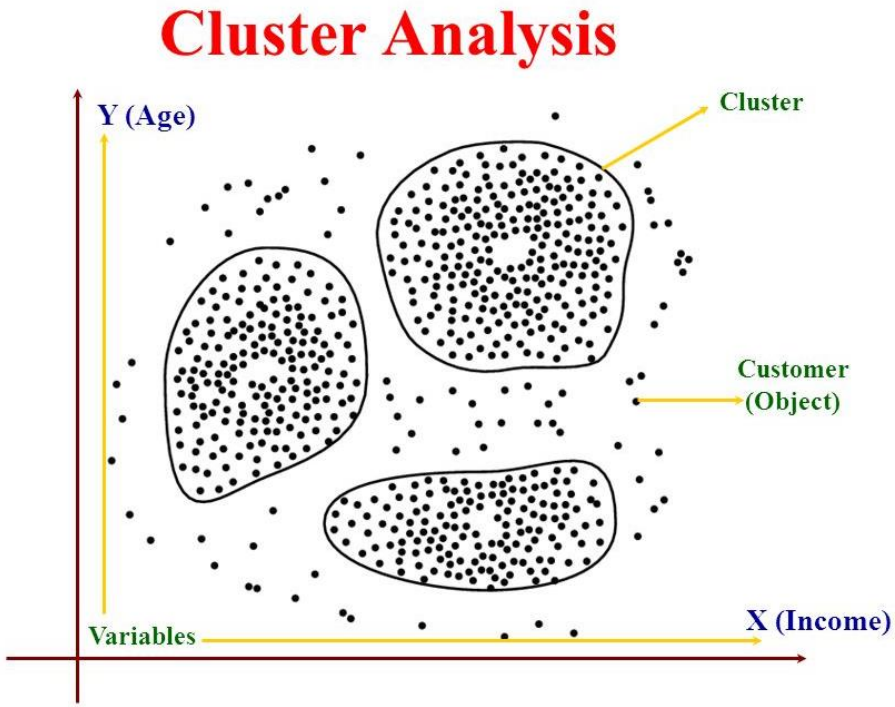
- Apart for descriptive analysis, another data mining task involves measuring ***how alike (similar)*** or ***how unalike (dissimilar)*** two objects are.
- **Proximity measures** is one way that can be used to check how close objects are to one another using distance.
- **Correlation measures** checks relationship/connection between 2 variables (i.e., a change in one variable causes a change in another variable)
- **Similarity measure:**
  - Crisp (0 – dissimilar or 1 – similar)
  - Fuzzy (scale of 0 – 1, how similar/dissimilar)

Strathmore
UNIVERSITY

Clustering

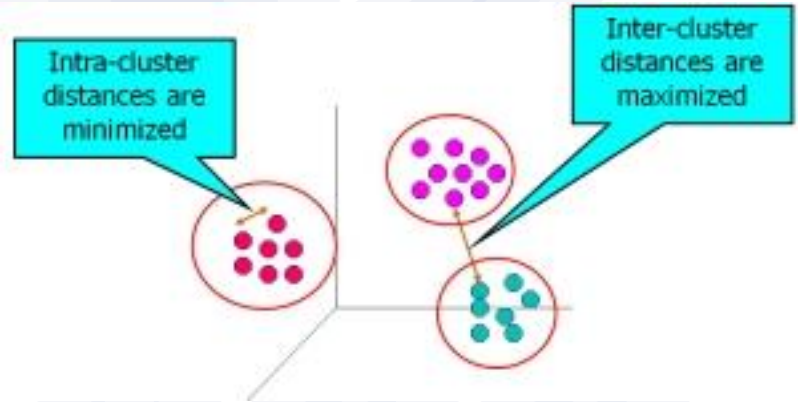
Cluster Analysis

- Clustering is attempting to group objects with similar traits together; such that:
  - objects in the same group are more similar to themselves and are dissimilar to objects in other groups
- **Application areas:**
  - Marketing and personalized advertisements: grouping customers with similar tastes
  - Content analysis: classify documents, search results etc.
  - Online fraud detection: cluster common habits in authentic users
  - List goes on …



Src: https://analyticsbuddhu.wordpress.com/2016/11/01/types-of-cluster-analysis-and-techniques-using-r/

Cluster Analysis (Cont.)



- **Category:**
  - unsupervised learning classifier
  - predictive/descriptive
- **Main Objectives:**
  - intra-cluster distance is minimized.
  - inter-cluster distance is maximized.
- **Challenges:**
  - determining optimum number of clusters.
  - determining the minimum inter/intra distance

Clustering Algorithms

Strathmore UNIVERSITY

1. **Connectivity Models**
   - Group object (data points) based on the distance between them
   - Objects with small distances between them are classified as similar and vice versa.
   - Can be further categorized based on proximity measure: Euclidean distance, Manhattan distance etc.
   - Examples: threshold-based clustering
2. **Centroid Models**
   - Similarity is derived from the distance from the centroid of the cluster.
   - Number of centroids/clusters is user-defined
   - Examples: K-Means clustering
3. **Distribution Models**
   - Considers the probability of an object belonging to a certain cluster (NOT proximity distance)
   - May require a probability function to determine similarity
4. **Density Models**
   - Scans the distribution of data points in the space and clusters according to density of the points

Exercise

1. Descriptive data analysis techniques tell us more about correlation between variables. Discuss using real examples.
2. Cluster analysis tell us more about proximity of objects. Discuss using real examples.

Strathmore
UNIVERSITY

Supervised Techniques

Key Terms

Strathmore UNIVERSITY

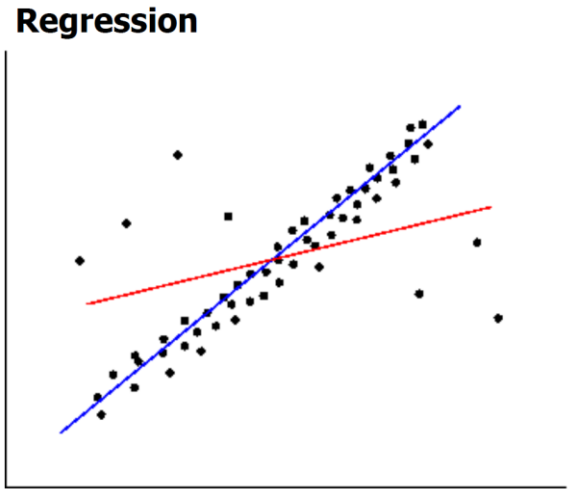
- **Target label/class:** variable to be predicted/identified.
- **Predictors:** variables used to learn the boundary. conditions that can be used identify each target label
- **Model (Classification/Prediction):** has the boundary conditions.
- **Classifier:** algorithm used to build the model.
- **Training data set:** data set with both target variable and predictor variables, and it is used to build the model.
- **Testing data set:** data set with both target variable and predictor variables, and it is used to test model's performance. Labels are hidden from model to see how accurately it can identify them.
- **Prediction:** data set without target variable. The ultimate goal of the model.

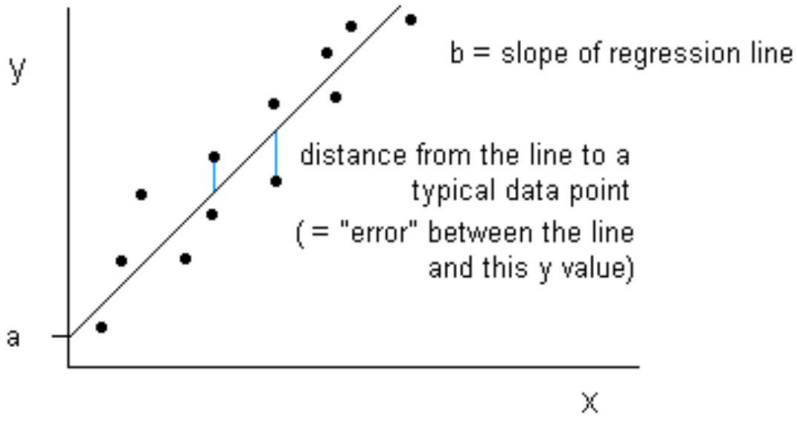
Strathmore
UNIVERSITY

Regression

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.
- **Aim:** learn "*line-of-best-fit*".
- Greatly studied in statistics, neural network fields.
- Applications:
  - 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.
  - List goes on …





Regression Algorithms

1. Simple linear regression
2. Lasso regression
3. Ridge regression
4. List goes on …

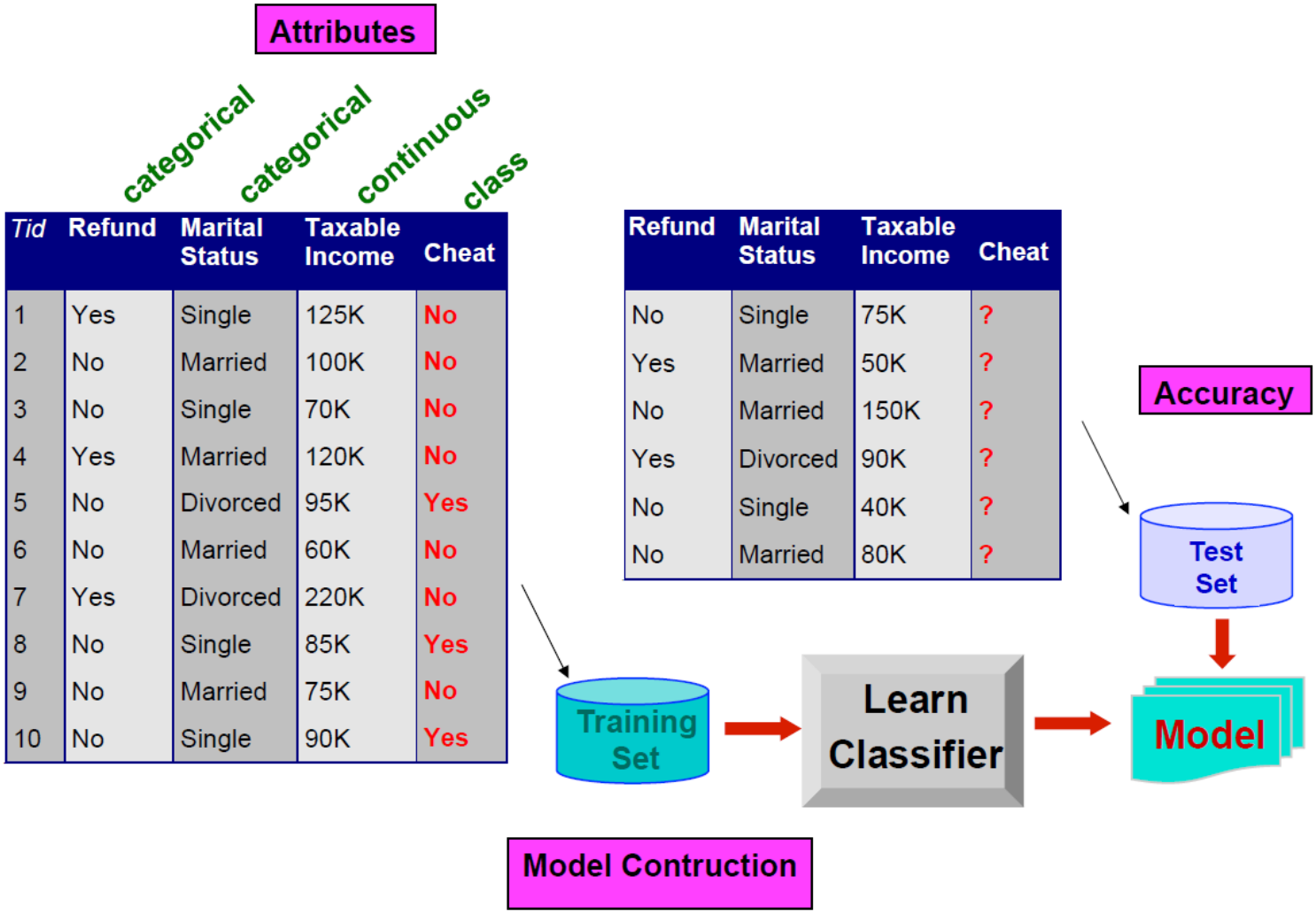
Strathmore
UNIVERSITY

Classification

Classification

- Assumptions:
  - given a collection of records (training set),
  - each record contains a set of attributes/features: one of the attributes is the class.
- Training Task:
  - find a model for the class (attribute) as a function of the values of the other attributes.
- Goal (after training):
  - apply the model on previously unseen records and it should assign them a class as accurately as possible.

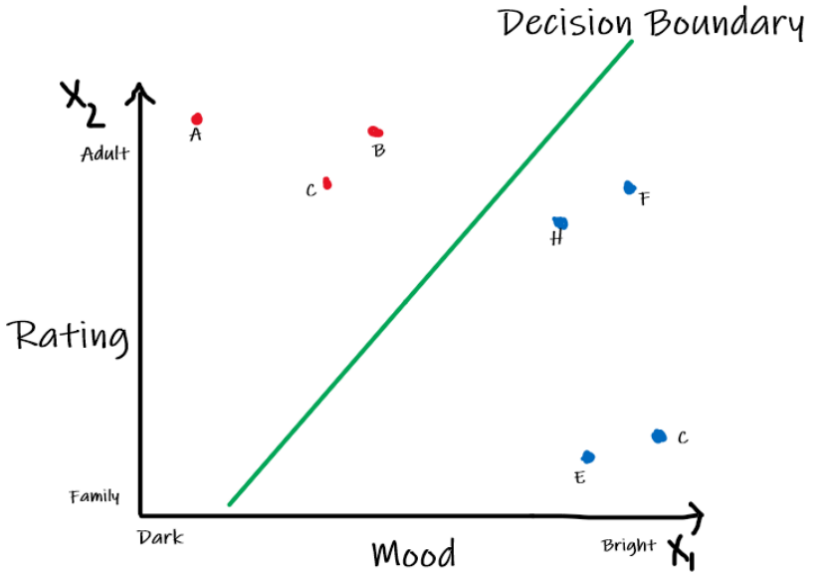
Classification Example



Classification (Cont.)

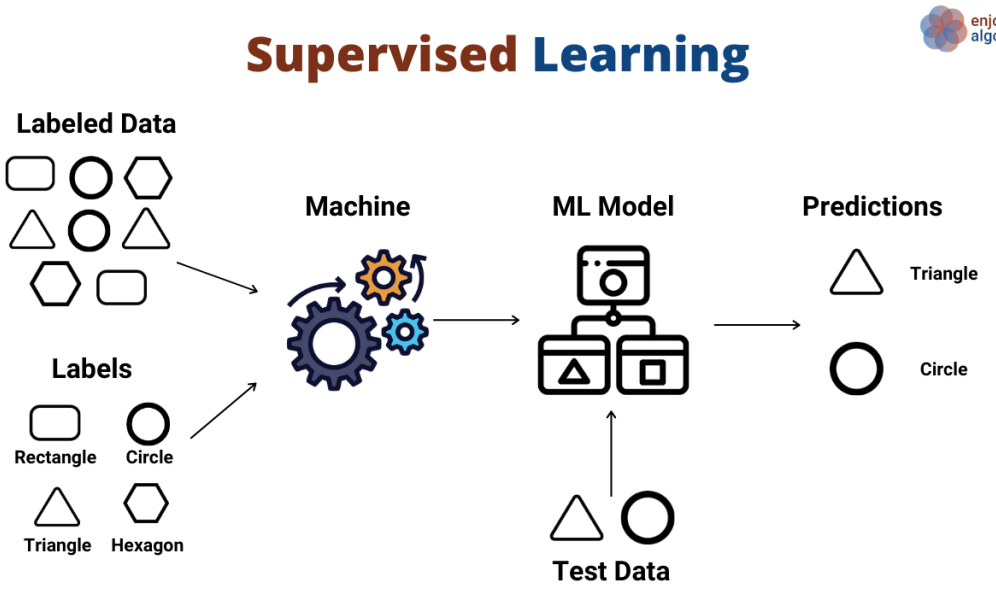
- Classification is the process of predicting a categorical label of a data set based on its features.
- **Aim:** learn "*line-of-best-split*" that separates data into classes.
- Classification can be taken to be a type of regression where data can be separated into discrete classes

Name		Mood	Rating	Class
Money Heist	A	Dark	Adult	*Dislike*
Prison Break	B	Semi-Dark	Adult	*Dislike*
Overcomer	C	Bright	Family	*Like*
House of Cards	D	Semi-Dark	Adult	*Dislike*
Selina	E	Bright	Family	*Like*
24	F	Bright	Adult	*Like*
Extraction	H	Bright	Adult	*Like*
Boss Baby	G	Bright	Children	*??*



Classification (Cont.)

- **Category:**
  - Supervised learning
  - Predictive
- **Challenges:**
  - Feature selection
  - Computation power
- **Application areas:**
  - Weather forecast
  - Pattern recognition
  - Investment suggestions
  - Disease predictions
  - Object recognition (i.e., voice, handwriting, movement)
  - List goes on …



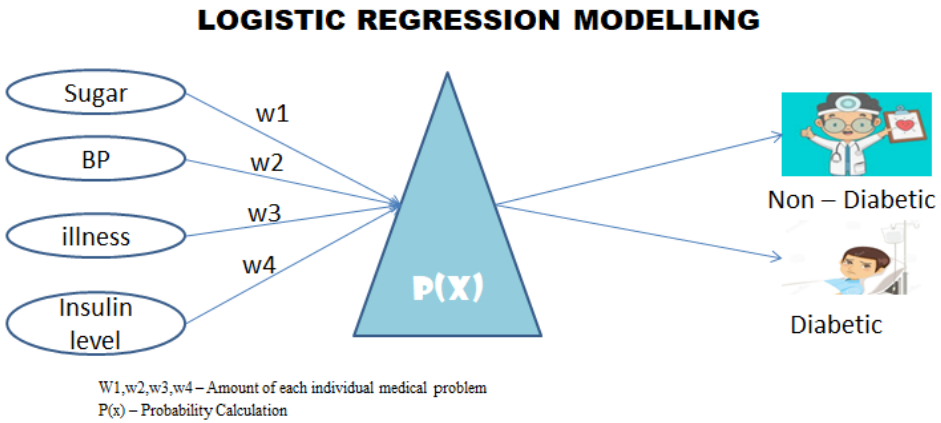
Classification Algorithms

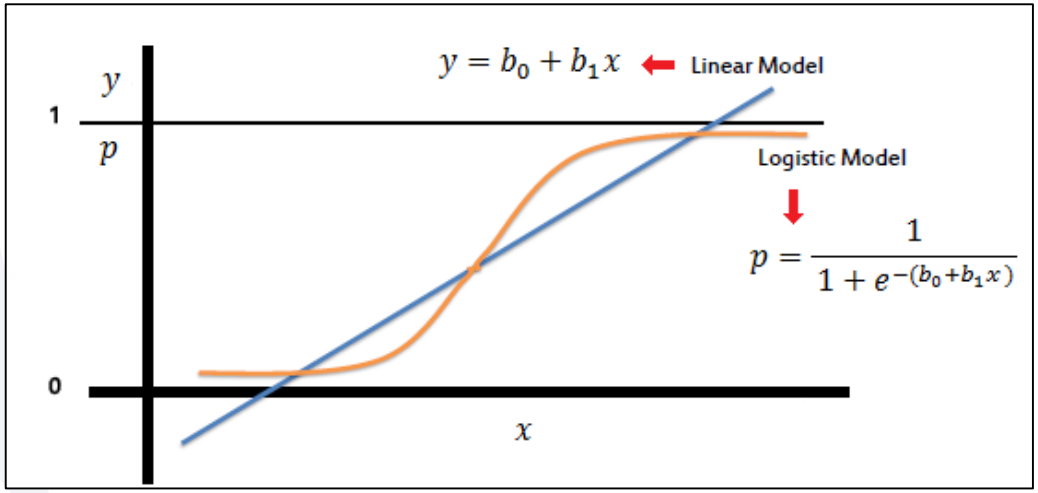
Strathmore UNIVERSITY

1. Logistic Regression
2. Decision Trees
3. Support Vector Machines
4. Naïve Bayes
5. Artificial Neural Networks
6. K-Nearest Neighbor

1. Logistic Regression

- Performs classification, not regression
- Predicts value of a categorical variable (preferably binary)
- Target variable is binary: yes/no, like/dislike, 0/1 etc.



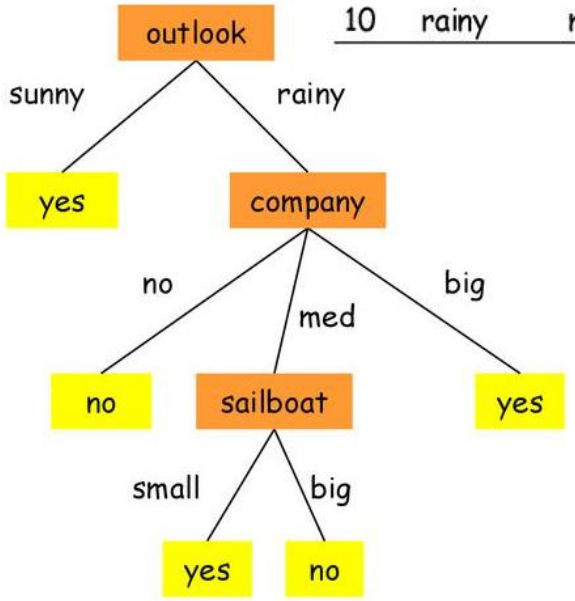


Src: https://medium.com/@slsarath2/logistic-regression-361edb3551fd

2. Decision Trees

- Decision trees represent a series of choices in the form of a tree.
- Uses predictor variables to decide which class the target variable lies in.
- Uses *divide-and-conquer* approach to divide objects repeatedly until a final decision/choice is made.

#	Attribute			Class
	Outlook	Company	Sailboat	Sail?
1	sunny	big	small	yes
2	sunny	med	small	yes
3	sunny	med	big	yes
4	sunny	no	small	yes
5	sunny	big	big	yes
6	rainy	no	small	no
7	rainy	med	small	yes
8	rainy	big	big	yes
9	rainy	no	big	no
10	rainy	med	big	no



Src: https://slideplayer.com/slide/15631470/

Exercise

1. Explain the difference between *“line of best fit”* and *“line of best split”*.
2. Using an example, describe the difference between linear regression and logistic regression.
3. Is it correct to think of decision trees as a group of nested if-else conditions? Explain your answer.

Research Perspectives

References

Gullo, F., 2015. From patterns in data to knowledge discovery: What data mining can do. *Physics Procedia*, *62*, pp.18-22.

Fournier-Viger, P., He, G., Cheng, C., Li, J., Zhou, M., Lin, J.C.W. and Yun, U., 2020. A survey of pattern mining in dynamic graphs. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10(6), p.e1372.

Essalmi, H., 2021. An Efficient Method for Mining Distributed Frequent Itemsets: MDFI. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(5), pp.895-902.

Laurent, A., Lesot, M.J. and Rifqi, M., 2009, October. Graank: Exploiting rank correlations for extracting gradual itemsets. In *International Conference on Flexible Query Answering Systems* (pp. 382-393). Springer, Berlin, Heidelberg.

https://techvidvan.com/tutorials/cluster-analysis-in-r/

Strathmore
UNIVERSITY

Thank you!

Any Questions?