03 - Data Mining

Dr Shafaq Khan
School of Computer Science
University of Windsor

Submission Deadlines

- **Project proposal**: Sec 2: Jan 30; Sec 3: Jan 31; Sec 4: Feb 1
- **Lab 1**: Sec 2: Jan 25; Sec 3: Jan 26; Sec 4: Jan 27 (11:59 PM)

- **Assignment 1:**
  - **Certificate submission before the start of your class next week**
  - **Quiz on Assignment 1 during your next Class**

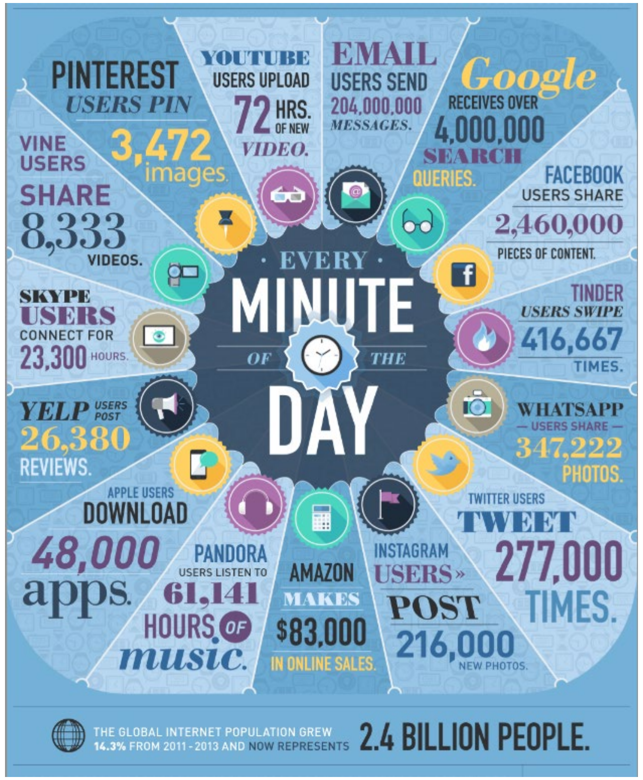
Agenda

- Data Mining
  - Commercial & Scientific Viewpoint
- Definition of Data Mining
- Major Data Mining Techniques
  - Clustering
  - Association Rules
  - Regression
  - Classification
- Lab 1
- Summary and Conclusion

Why Data Mining?

Commercial Viewpoint

- **Lots of data is being generated**
- **Web data**
  - Yahoo has Peta Bytes of web data
  - Facebook has billions of active users purchases at department/grocery stores, e-commerce
  - Amazon handles millions of visits/day
  - Bank/Credit Card transactions
- **Competitive Pressure is Strong**
  - Provide better, customized services for an edge



https://time.com/73581/data-generated-online-every-minute-domo/

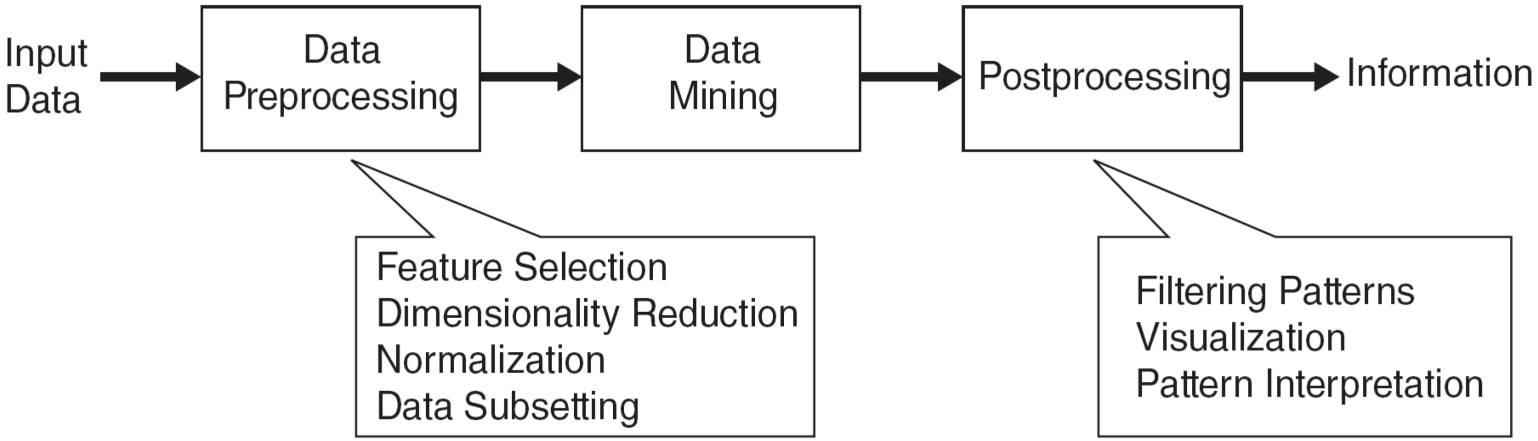
Scientific Viewpoint

- **Data collected and stored at enormous speeds**
  - Remote sensors on a satellite

    NASA EOSDIS archives over petabytes of earth science data / year
  - Telescopes scanning the skies
    - Sky survey data
    - High-throughput biological data
  - Scientific simulations
    - Terabytes of data generated in a few hours
- **Data mining helps scientists**
  - in automated analysis of massive datasets
  - in hypothesis formation

What is Data Mining?

- Multiple Definitions
  - Analyzing massive amounts of data to **uncover hidden trends, patterns, and relationships**; to form computer models to simulate and explain the findings; and then to use such models **to support business decision making**.
  - **Non-trivial extraction** of previously unknown and potentially useful information from data

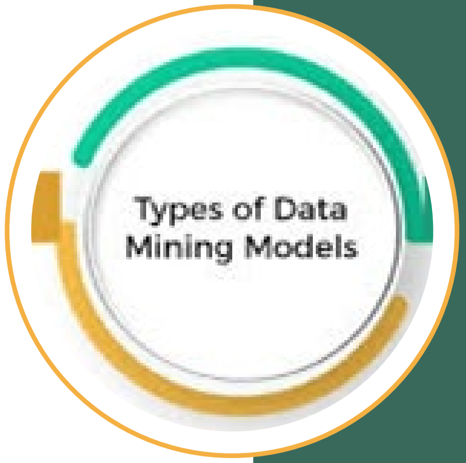


Introduction to Data Mining, 2nd Edition, Tan, Steinbach, Karpatne, Kumar

Data Mining Tasks

Predictive methods (Supervised learning)

- Use some variables to predict unknown or future values of other variables
  - Classification, which is used for discrete target variables, and
  - Regression, which is used for continuous target variables



Descriptive methods (Unsupervised learning)

- Find human-interpretable patterns that describe the data
  - Example: Clustering, Association analysis

Major Data Mining Techniques

- Clustering
- Association Rules
- Regression
- Classification

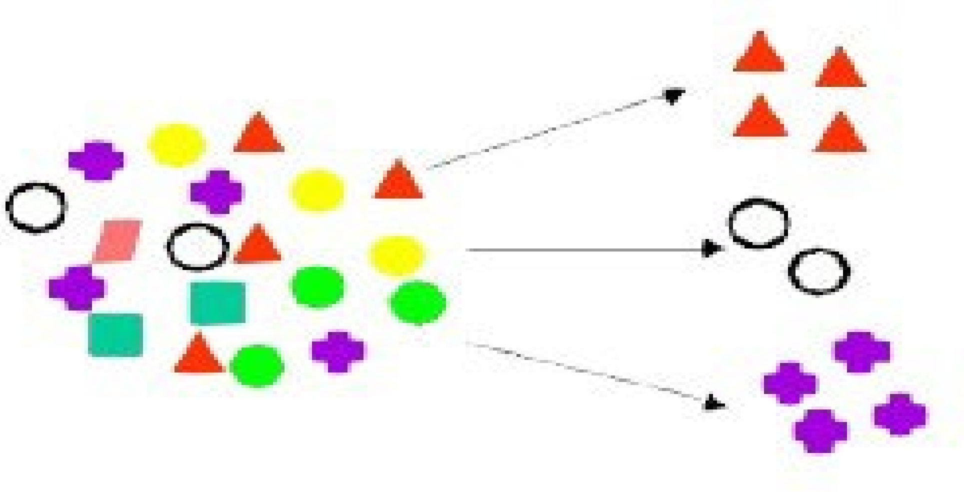
Clustering

What Kind of Problem do I Need to Solve? How do I Solve it?

The Problem to Solve	The Category of Techniques	Analytical methods
I want to group items by similarity. I want to find structure (commonalities) in the data	**Clustering**	**K-means clustering** DBSCAN(density-based)
I want to discover relationships between actions or items	Association Rules	Apriori AIS
I want to determine the relationship between the outcome and the input variables	Regression	Linear Regression Logistic Regression
I want to assign (known) labels to objects	Classification	Naïve Bayes Decision Trees, K-NN

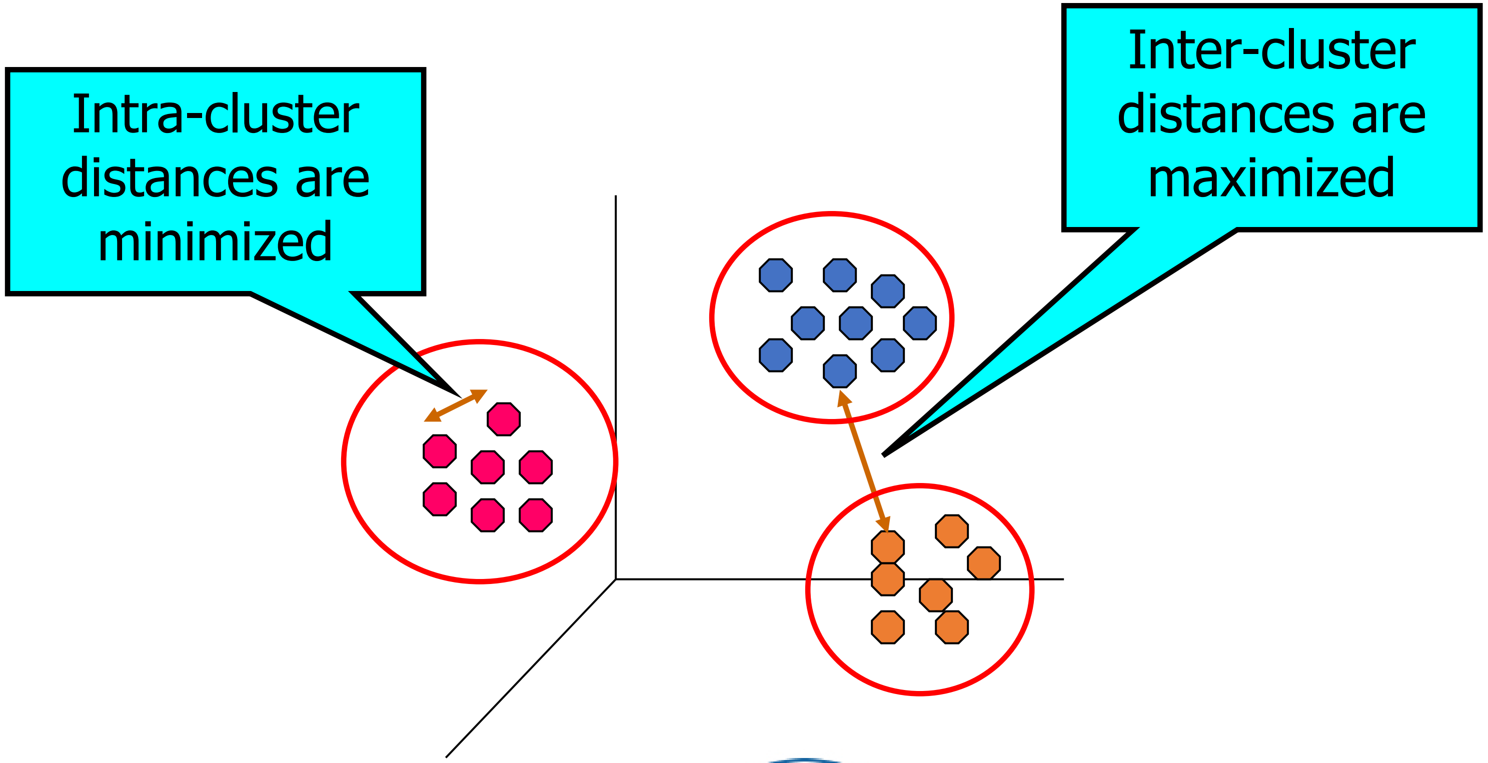
Clustering

- Objective: grouping members that have similar characteristics together
- Widely applied on market segmentation, document clustering, health, business and science



What is Cluster Analysis?

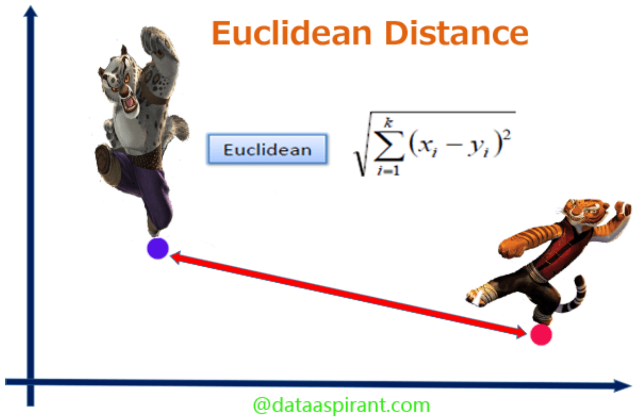
Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups



Intra-cluster distances are minimized

Inter-cluster distances are maximized

‘Closeness’ is measured by Euclidean distance

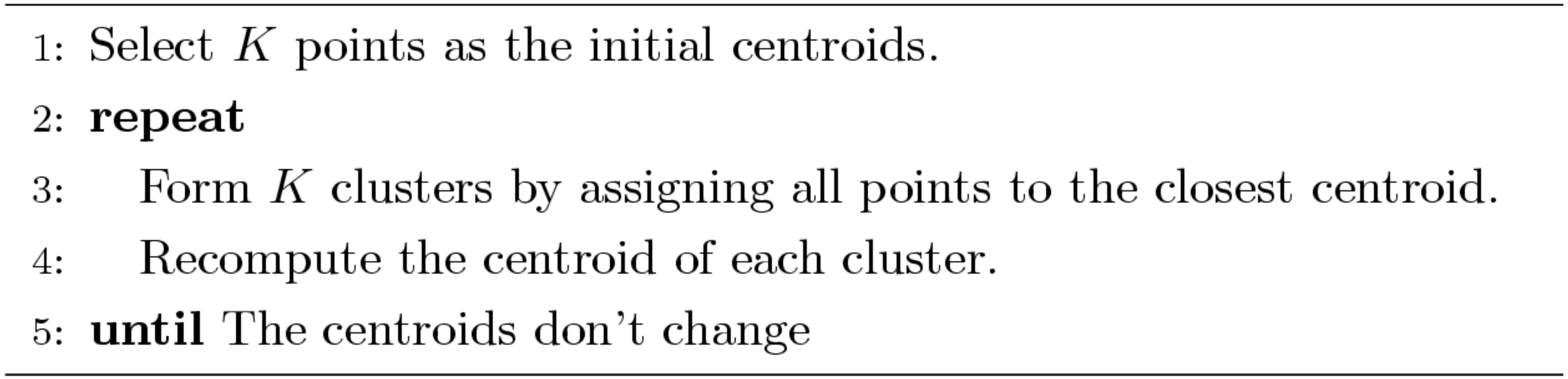


K-means Clustering

Characteristics

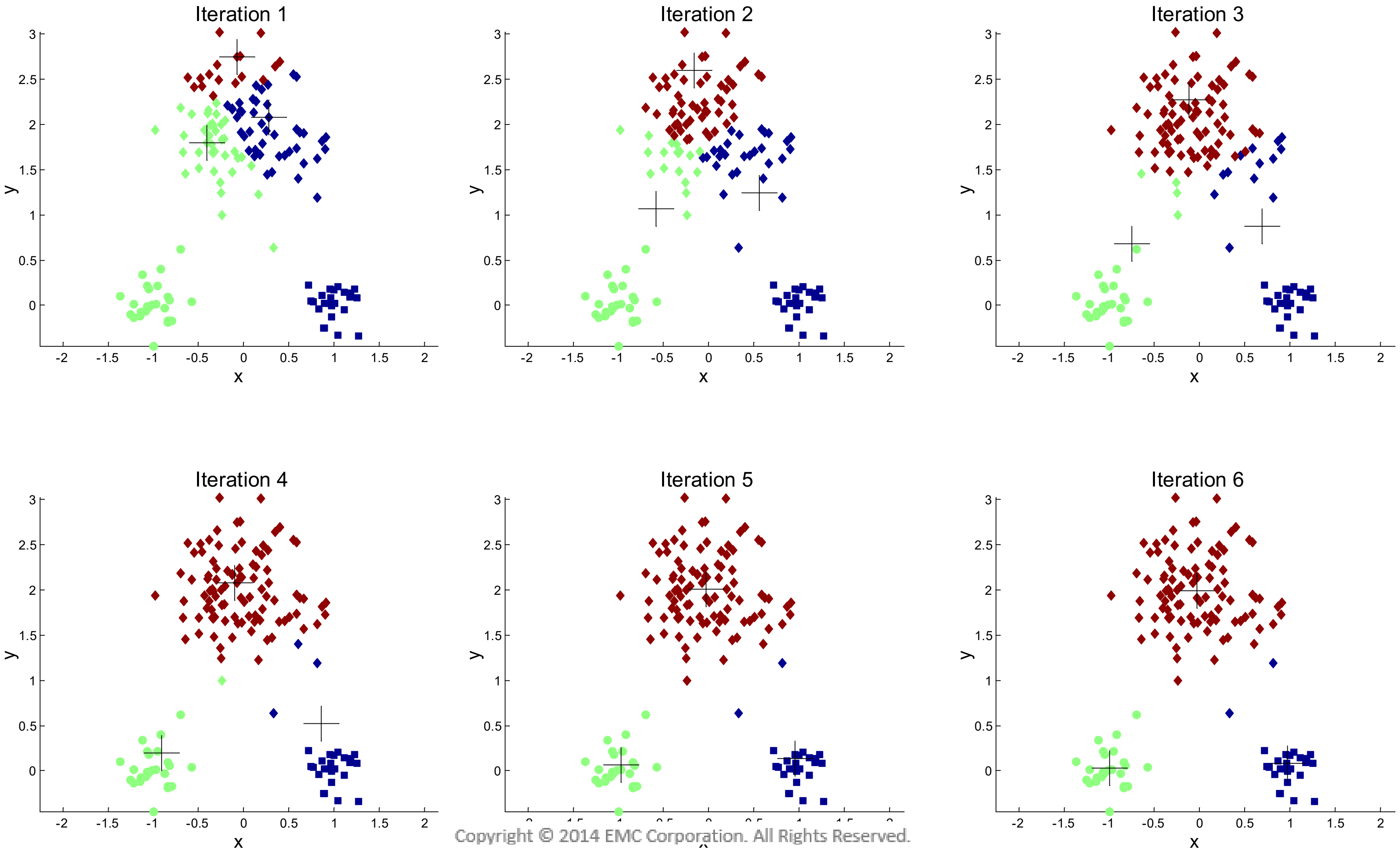
- Partitional clustering approach
- Each cluster is associated with a **centroid** (center point)
- Each point is assigned to the cluster with the closest centroid
- Number of clusters, K, must be specified
- The basic algorithm is very simple

Algorithm:

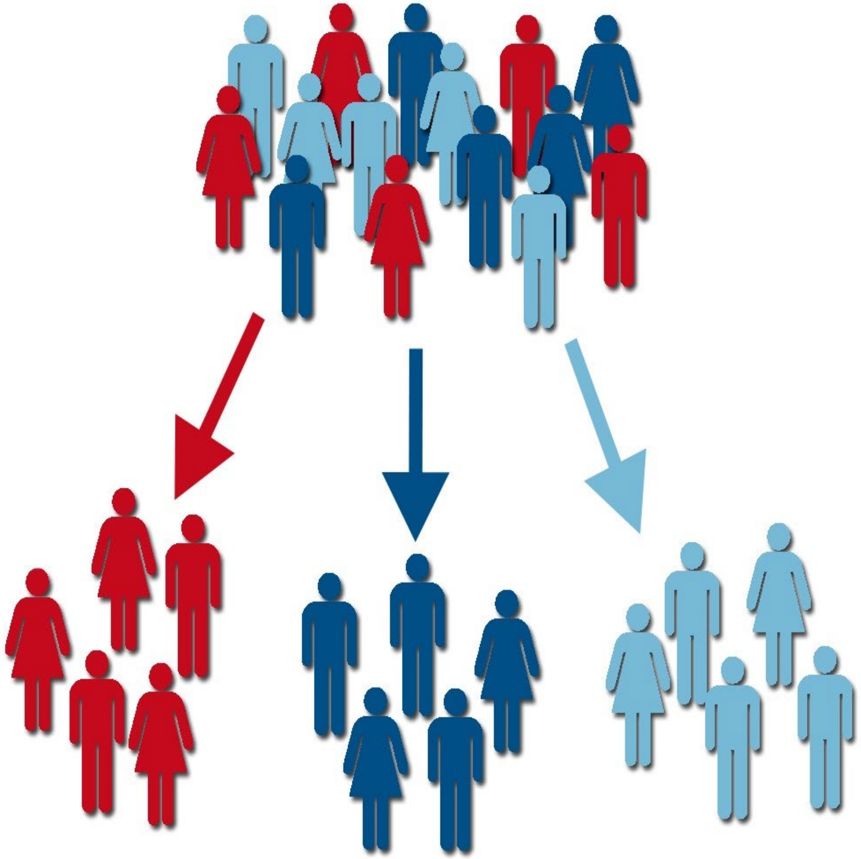


```
Select $K$ points as the initial centroids.
repeat
    Form $K$ clusters by assigning all points to the closest centroid.
    Recompute the centroid of each cluster.
until The centroids don't change
```

Importance of Choosing Initial Centroids



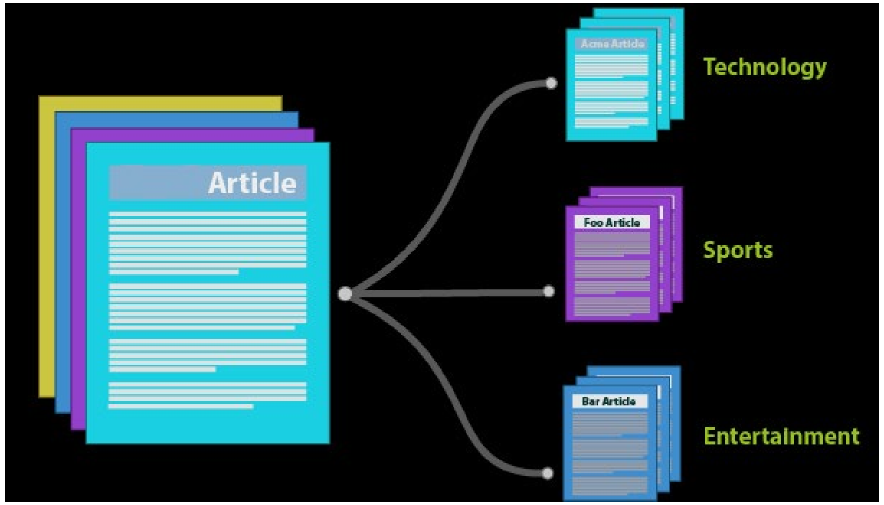
Clustering- Application 1: Market Segmentation



www.medium.com

- **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

www.towardsdatascience.com

Quick Check

Is it possible that assignment of observations to clusters does not change between successive iterations in K-Means

[✓ A. Yes]

B. No

C. Can’t say

D. None of these

Lab Exercise - Clustering

- This Lab is designed to investigate and practice Clustering.

  After completing the tasks in this lab you should be able to:

  - Use RStudio environment to code and visualize clustering models.

Association Rules

What Kind of Problem do I Need to Solve? How do I Solve it?

The Problem to Solve	The Category of Techniques	Analytical methods
I want to group items by similarity. I want to find structure (commonalities) in the data	Clustering	K-means clustering
I want to discover relationships between actions or items	Association Rules	**Apriori**
I want to determine the relationship between the outcome and the input variables	Regression	Linear Regression Logistic Regression
I want to assign (known) labels to objects	Classification	Decision Trees

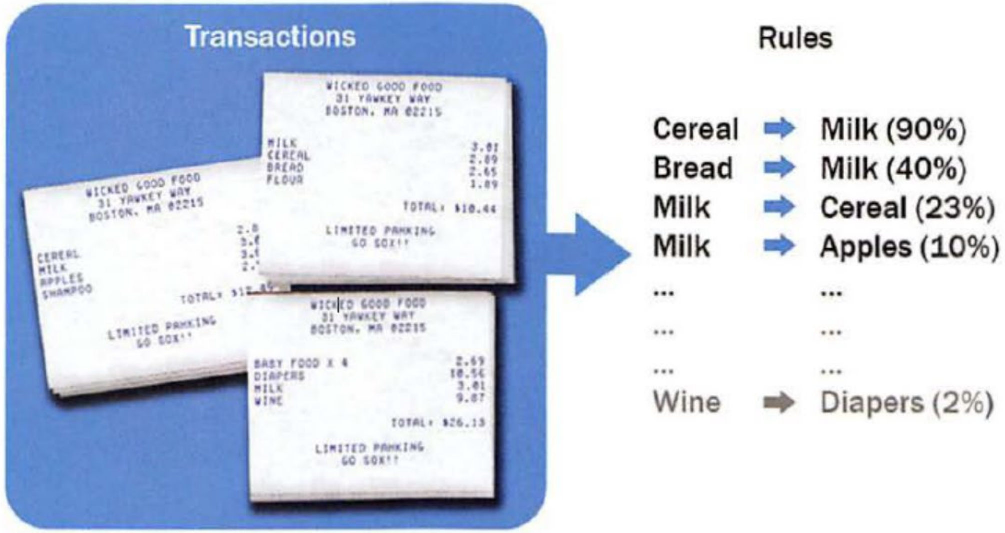
Association Rules

- Which of my products tend to be purchased together?
- Of those customers who are similar to this person, what products do they tend to buy?
- Of those customers who have purchased this product, what other similar products do they tend to view or purchase?
- Discover "interesting" relationships among variables in a large database
  - Rules of the form “If X is observed, then Y is also observed"
  - The definition of "interesting“ varies with the algorithm used for discovery

- Not a predictive method; finds similarities, relationships

Association Rules (also called market basket analysis)

The goal with association rules is to discover interesting relationships among the items.



Rules

Cereal ➡ Milk (90%)
Bread ➡ Milk (40%)
Milk ➡ Cereal (23%)
Milk ➡ Apples (10%)
... ...
... ...
... ...
Wine ➡ Diapers (2%)

Each of the uncovered rules is in the form X -> Y, meaning that when item X is observed, item Y is also observed. X is also called ‘**Antecedent**’ and Y is called as ‘**Consequent**’

Association Rules - Apriori

- Earliest of the association rule algorithms
- Specifically designed for mining over transactions in databases
- Used over itemsets: sets of discrete variables that are linked:
    - Retail items that are purchased together
    - A set of tasks done in one day
    - A set of links clicked on by one user in a single session

- **Our Example: Apriori**

Association Rules

- Each transaction can be viewed as the shopping basket of a customer that contains one or more items. This is also known as an item set.
- The term *itemset* refers to a collection of items or individual entities that contain some kind of relationship.
- Examples: A set of retail items purchased together in one transaction, a set of hyperlinks clicked on by one user in a single session, or a set of tasks done in one day.
- An item set containing *k* items is called a *k-itemset.*

  k-itemset = { item 1, item *2, . . .* item k}

Apriori Algorithm - Support

- One major component of Apriori is ***support***.
- Given an item set *L,* the ***support*** of L is the percentage of transactions that contain *L.*
- For example, if 80% of all transactions contain item set {bread}, then the support of {bread} is 0.8. Similarly, if 60% of all transactions contain itemset {bread, butter}, then the support of {bread, butter} is 0.6.
- Support is used to filter out items that are less frequently bought.

Frequent itemset

- **Frequent itemset**: a set of items L that appears together "often enough“: Formally: meets a minimum support criterion
- If the **minimum support** is set at 0.5, any itemset can be considered a frequent item set if at least 50% of the transactions contain this itemset. In other words, the support of a frequent itemset should be greater than or equal to the minimum support.
- Example: For the previous example, both {bread} and {bread, butter} are considered frequent item sets at the minimum support 0.5. If the minimum support is 0.7, only {bread} is considered a frequent itemset.

Frequent Itemsets

- There are typically Billions or even Trillions (or more) of transactions (which is usually not the cause of concern)
- It is the no of items which can be a cause of concern
    - With n=3, there are 8 subsets **to search**
    - With n=5, there are 32 subsets
    - With n=10, there are 1024 subsets
    - With n=20, there are 1,048,576 subsets
    - With n=100, there are 1,267,650,600,228,229,401,496,703,205,376 subsets
        - **Which one of these subsets is the most frequent itemset?**
- Running time can be reduced by focusing on subsets of certain sizes or types (or other constraints)

Frequent Itemset

ITEMS= { BREAD ,MILK, EGGS}

- { }
- { Bread}
- { Milk}
- { Eggs}
- { Bread ,Milk}
- { Bread ,Eggs}
- { Milk, Eggs}
- { Bread ,Milk, Eggs}

FREQUENCIES OF ITEMSETS (Support values)

- { Bread} = 5
- { Milk} = 5
- { Eggs} = 7
- { Bread ,Milk} = 2
- { Bread ,Eggs} = 4
- { Milk, Eggs} = 5 // Most Frequent Itemset
- { Bread ,Milk, Eggs}= 2

- **TRANSACTIONS**
- T1{ Bread, Eggs}
- T2{ Bread}
- T3{ Milk, Eggs ,Bread}
- T4{ Eggs, Milk}
- T5{ Bread, eggs}
- T6{ Bread, Milk, Eggs}
- T7{ Milk, Eggs,}
- T8{ Eggs ,Milk,}

Discovery of Association Rules

- **FREQUENCIES OF ITEMSETS**
- { Bread} = 5
- { Milk} = 5
- { Eggs} = 7
- { Bread ,Milk} = 2
- { Bread ,Eggs} = 4
- { Milk, Eggs} = 5 // Most Frequent Itemset
- { Bread ,Milk, Eggs}= 2

- After computing the frequencies of each itemset, association rules can be drawn based on **threshold values**(min support value or frequency) and **the size** of the subsets.
- Ex: In the above example, if the **min_size of the itemset=2** and the **threshold is 3**, the following itemset can be considered for association rules
  - { Bread ,Eggs} = 4 and { Milk, Eggs} = 5
- The following association rules can be inferred
  - Bread-> Eggs
  - Mill-> Eggs

Apriori property

- ***Apriori Property: Any subset of a frequent itemset is also frequent***
  - It has at least the support of its superset
- For example, if 60% of the transactions contain {bread, jam}, then at least 60% of all the transactions will contain {bread} or {jam}.

- ***Prune***: Pruning away means discarding all transactions of the item sets that have a support less than minimum support threshold (or the minimum support criterion).
- Example: Discarding transactions that appear in fewer than 50% of the transactions.

Apriori Algorithm

The Apriori algorithm takes a bottom-up iterative approach to uncovering the frequent itemsets by first determining all the possible items (o r **1-itemsets**, for example {bread}, {eggs}, {milk}, .. .) and then identifying which among them are frequent. It identifies and retains those itemsets that appear in at least 50% of all transactions and discards (or "prunes away") the itemsets that have a support less than 0.5 or appear in fewer than 50% of the transactions.

In the next iteration of the Apriori algorithm, the identified frequent 1-itemsets are paired into **2-itemsets** (for example, {bread, eggs}, {bread , milk}, {eggs, milk), ...) and again evaluated to identify the frequent 2-itemsets among them.

At each iteration, the algorithm checks whether the support criterion can be met; if it can, the algorithm grows the itemset, repeating the process until it runs out of support or until the itemsets reach a predefined length.

A Sketch of the Algorithm

- If L_k is the set of frequent k-itemsets:
  - Generate the candidate set C_{k+1} by joining L_k to itself
  - Prune out the (k+1)-itemsets that don't have minimum support Now we have L_{k+1}
- We know this catches all the frequent (k+1)-itemsets by the apriori property
  - a (k+1)-itemset can't be frequent if any of its subsets aren't frequent
- Continue until we reach k_{max}, or run out of support
- From the union of all the L_k, find all the rules with minimum confidence

Example on Association Rules, example

Support

- Transaction1: {Apple, Juice, Rice, Chicken}
- Transaction2: {Apple, Juice, Rice}
- Transaction3: {Apple, Juice}
- Transaction4: {Apple, Grapes}
- Transaction5: {Milk, Juice, Rice, Chicken}
- Transaction6: {Milk, Juice, Rice}
- Transaction7: {Milk, Juice}
- Transaction8: {Milk, Grapes}

Support=freq(X,Y)/N

Support(Apple) = 4/8

Confidence

How likely item Juice is purchased when item Apple is purchased, expressed as {Apple -> Juice}. This is measured by the proportion of transactions with item Apple, in which Juice also appears. In Table 1, the confidence of {apple -> Juice} is ?

$$Confidence(X \rightarrow Y) = \frac{Support(X \wedge Y)}{Support(X)}$$

Confidence {Apple -> Juice} = ?

Support {Apple, Juice}/ Support {Apple}

(3/8)/(4/8)*=(3/8)*(8/4)=3/4=0.75= 75%

Association Analysis: Use Cases

Market-basket analysis
Rules are used for sales promotion, shelf management, and inventory management

Recommender Systems
"People who bought what you bought also purchased….“.

Medical Diagnostics : Rules are used to find combination of patient symptoms and test results associated with certain diseases More accurate diagnosis

Discovering web usage patterns
People who land on page X click on link Y 76% of the time.

Census Data

Plan efficient public services

Support public policy-making

Lab Exercise - Association Rules

This Lab is designed to investigate and practice Association Rules.

After completing the tasks in this lab you should be able to:

- Use R functions for Association Rule based models

Join menti
https://www.menti.com/xdzes46ta9

Regression

What Kind of Problem do I Need to Solve? How do I Solve it?

The Problem to Solve	The Category of Techniques	Analytical methods
I want to group items by similarity. I want to find structure (commonalities) in the data	Clustering	K-means clustering
I want to discover relationships between actions or items	Association Rules	Apriori
I want to determine the relationship between the outcome and the input variables	**Regression**	**Linear Regression**
I want to assign (known) labels to objects	Classification	Decision Trees

Regression

Regression analysis attempts to explain the influence that a set of variables has on the outcome of another variable of interest.

Regression analysis is useful for answering the following kinds of questions:

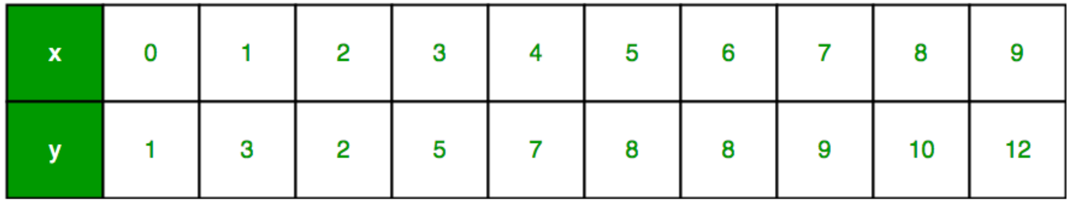
- What is a person's expected income?
- What is the probability that an applicant will default on a loan?

The outcome variable is called a ***dependent variable*** because the outcome depends on the other variables. These additional variables are sometimes called the ***input variables*** or the ***independent variables***.

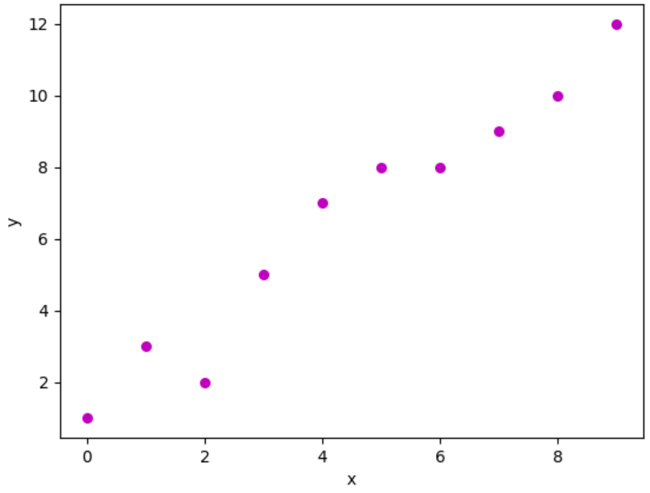
Linear Regression

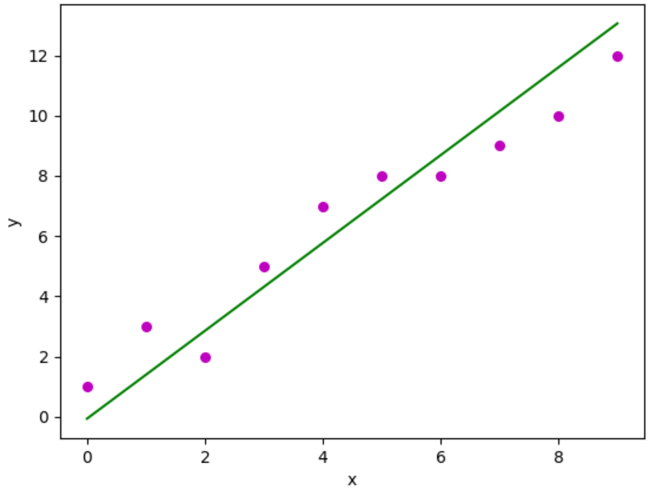
- Linear regression is an analytical technique used to model the relationship between several input variables and a continuous outcome variable.
- **Outcome** variable is continuous.
- **Input** variables can be continuous or discrete.
- Based on known input values, a linear regression model provides the expected value of the outcome variable based on the values of the input variables, but some uncertainty may remain in predicting any particular outcome.
- Model Output:
  - A set of estimated coefficients that indicate the relative impact of each input variable on the outcome
  - A linear expression for estimating the outcome as a function of input variables

Regression-Example



x	0	1	2	3	4	5	6	7	8	9
y	1	3	2	5	7	8	8	9	10	12





Regression Line
Least Squares

https://www.geeksforgeeks.org/linear-regression-python-implementation/

Regression Line

- Positive Regression
- Negative regression

A **linear regression** line has an equation of the form

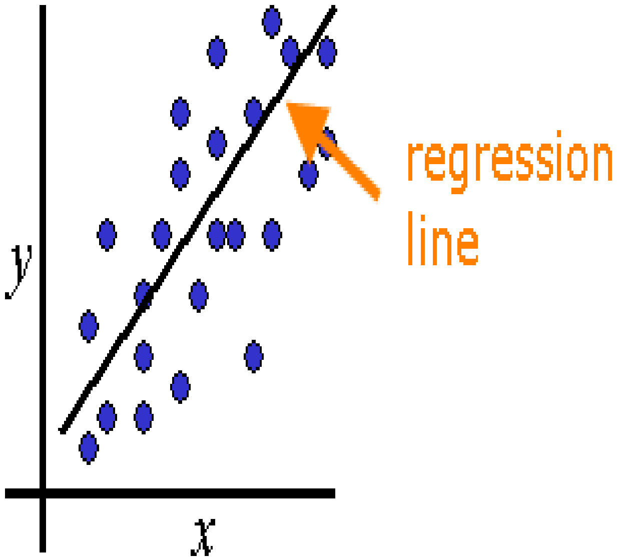
$Y = a + bX + \varepsilon$

where X is the explanatory variable and Y is the dependent variable.

b - is the slope of the line or the coefficient;

a - is the intercept (the value of y when x = 0)

ε – error term



a and b are derived mathematically

Regression- Example..

x	y	XY	X^2
1	1	1	1
2	3	6	4
3	2	6	9
4	5	20	16
5	7	35	25
6	8	48	36
7	8	56	49
8	9	72	64
9	10	90	81
10	12	120	100
55	65	454	385

x	y
1	1.9055
2	2.9265
3	3.9475
4	4.9685
5	5.9895
6	7.0105
7	8.0315
8	9.0525
9	10.0735
10	11.0945
11	12.1155
12	13.1365
13	14.1575
14	15.1785
15	16.1995
16	17.2205
17	18.2415
18	19.2625
19	20.2835
20	21.3045

$Y=a+ bX$

- a: intercept
- b: slope
- $b= (n \sum xy - \sum x \sum y) / (n \sum x^2 - \sum x \ )$
- $b= (10 \; x \; 454 – 55 \; x \; 65)/ (10 \; x \; 100 – 55)=$**1.021**
- $a= (\sum y – m \sum x \ ) / n = (65- 1.021 \; x \; 55)/10 =$**0.8845**
- **Hence Y=1.021 + (0.8845 * X)**

Example (Predict score)

- Suppose that we got java1 and java2 scores from the previous term for five students:
- Let's see if we can build a model to predict, what score that a new student can score based on her mark in java1?

If Patrick scored 80 in Java1, what score would he probably get in Java2 ?

	Ankit	Ali	Sam	Sandra	Navneet	Patrick
Java 1	95	85	80	70	60	80
Java 2	85	95	70	65	70	?

Example (Predict score)

	Ankit	Ali	Sam	Sandra	Navneet	Patrick
Java 1	95	85	80	70	60	80
Java 2	85	95	70	65	70	?

1

Lets use R to Plot these values:

```
Java1 <- c(95, 85, 80, 70, 60)
Java2 <- c(85, 95, 70, 65, 70)
plot(Java2, Java1)
```

2

```
# we use lm() to build a linear regression model in R
fit <- lm(Java2 ~ Java1)
Fit
Call: lm(formula = Java2 ~ Java1) Coefficients:
Intercept: 26.7808 (This is c0)
Java1: 0.6438 (This is c1)
```

According to the equation:

$$Y = c0 + c1*x1$$

Java2 = 26.7808 + 0.6438*Java1

If Patrick scored 80 in Java1, what score probably he can get in Java2 ?

Linear Regression Use Cases

Demand forecasting: Businesses and governments can use linear regression models to predict demand for goods and services.

Real estate: House sales price as function of area, number of bedrooms/bathrooms, and lot size

Medical: A linear regression model can be used to analyze the effect of a proposed radiation treatment on reducing tumor sizes. Input variables might include duration of a single radiation treatment, frequency of radiation treatment, and patient attributes such as age or weight.

Banks assess the risk of home-loan Applicants based on their age, income, expenses, occupation, number of dependents, personal status, total credit limit, etc

Data Classification

What Kind of Problem do I Need to Solve? How do I Solve it?

The Problem to Solve	The Category of Techniques	Analytical method
I want to group items by similarity. I want to find structure (commonalities) in the data	Clustering	K-means clustering
I want to discover relationships between actions or items	Association Rules	Apriori
I want to determine the relationship between the outcome and the input variables	Regression	Linear Regression
I want to assign (known) labels to objects	**Classification**	**Decision Trees**

Classifiers

Where in the shelf should I place this item?
Is this email spam?
Is this politician Democrat/Republican/Green?

- Classification: Assign labels to objects.
- Usually supervised: training set of pre-classified examples.
- Our examples:
  - Naïve Bayesian
  - **Decision Trees**
  - Time Series Analysis
  - Text Analysis

Decision Tree Classifier - What is it?

- Used for classification:
  - Returns probability scores of class membership
    - Assigns label based on highest scoring class
- **Input** variables can be continuous or discrete
- **Output**:
  - A tree that describes the decision flow.
  - Leaf nodes return either a probability score, or simply a classification.
  - Trees can be converted to a set of "decision rules“
    - "IF income < $50,000 AND mortgage_amt > $100K THEN default=T with 75% probability“

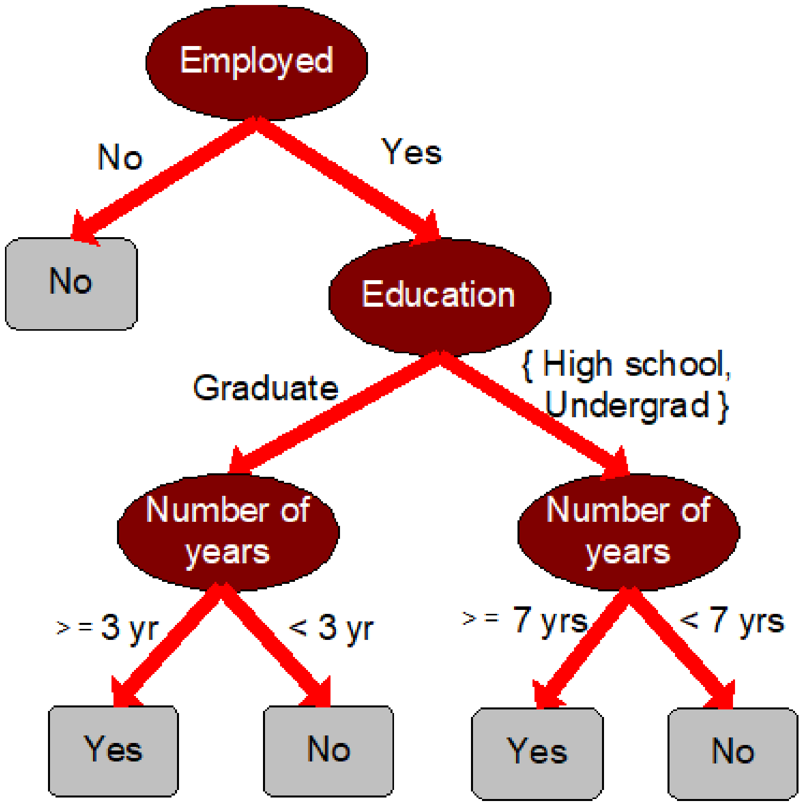
Predictive Modeling: Classification

- Find a model for class attribute as a function of the values of other attributes

Class

TID	Employed	Education	# of Years at present address	Credit Worthy
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undersgradaute	1	No
4	Yes	High School	10	Yes
5	Yes	Graduate	2	No
6	Yes	Undergraduate	8	Yes

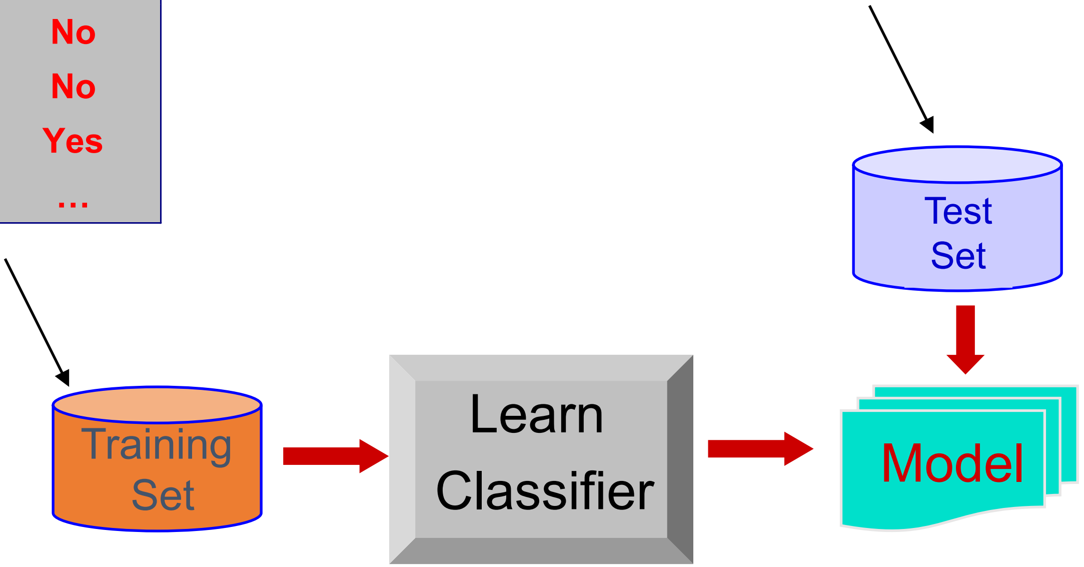
Model for predicting credit worthiness



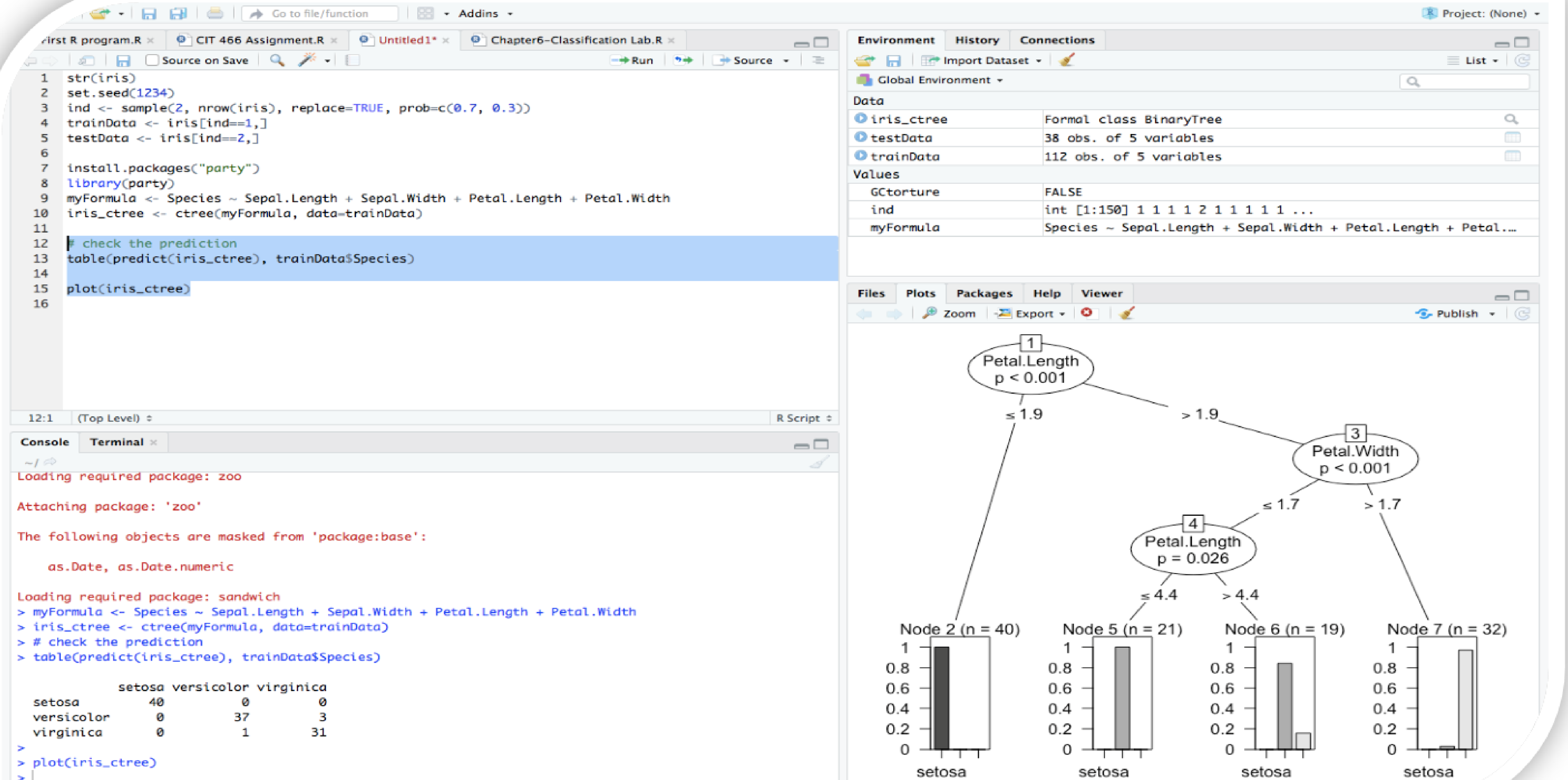
Classification Example

	categorical	categorical	quantitative	class
Tid	**Employed**	**Level of Education**	**# years at present address**	**Credit Worthy**
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undergrad	1	No
4	Yes	High School	10	Yes
…	…	…	…	…

Tid	**Employed**	**Level of Education**	**# years at present address**	**Credit Worthy**
1	Yes	Undergrad	7	?
2	No	Graduate	3	?
3	Yes	High School	2	?
…	…	…	…	…



Checking the prediction and plotting it in the decision tree



```
str(iris)
set.seed(1234)
ind <- sample(2, nrow(iris), replace=TRUE, prob=c(0.7, 0.3))
trainData <- iris[ind==1,]
testData <- iris[ind==2,]

install.packages("party")
library(party)
myFormula <- Species ~ Sepal.Length + Sepal.Width + Petal.Length + Petal.Width
iris_ctree <- ctree(myFormula, data=trainData)

# check the prediction
table(predict(iris_ctree), trainData$Species)

plot(iris_ctree)
```

Data

iris_ctree	Formal class BinaryTree
testData	38 obs. of 5 variables
trainData	112 obs. of 5 variables

Values

GCtorture	FALSE
ind	int [1:150] 1 1 1 1 2 1 1 1 1 1 ...
myFormula	Species ~ Sepal.Length + Sepal.Width + Petal.Length + Petal....

```
Loading required package: zoo

Attaching package: 'zoo'

The following objects are masked from 'package:base':

    as.Date, as.Date.numeric

Loading required package: sandwich
> myFormula <- Species ~ Sepal.Length + Sepal.Width + Petal.Length + Petal.Width
> iris_ctree <- ctree(myFormula, data=trainData)
> # check the prediction
> table(predict(iris_ctree), trainData$Species)

             setosa versicolor virginica
  setosa         40          0         0
  versicolor      0         37         3
  virginica       0          1        31
>
> plot(iris_ctree)
>
```

Classification: Application 1

- **Fraud Detection**
  - **Goal:** Predict fraudulent cases in credit card transactions.
  - **Approach:**
    - **Label past transactions as fraud or fair** transactions.
    - Learn a model to classify the transactions.
    - Use this model to detect fraud by observing credit card transactions on an account.

Classification: Application 2

https://blog.ekbana.com/customer-churn-prediction-4094fc24656b

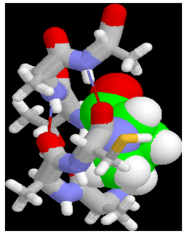
- Churn prediction for telephone customers
  - **Goal:** To predict whether a customer is likely to be lost to a competitor.
  - **Approach:**
    - Use detailed record of transactions with each of the past and present customers, to find attributes.
    - Find attributes of past customers who have switched their services.
    - Label the customers as **loyal or disloyal** based on a criteria (Age, Income, Occupation, Gender, Education Levels etc.)
    - Find a model for loyalty.

From [Berry & Linoff] Data Mining Techniques, 1997

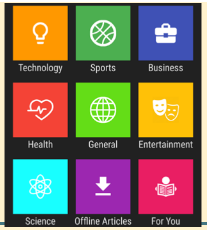
Classification: Other Use Cases

Identifying intruders in the cyberspace

Classifying land covers (water bodies, urban areas, forests, etc.) using satellite data



Predicting tumor cells as benign or malignant

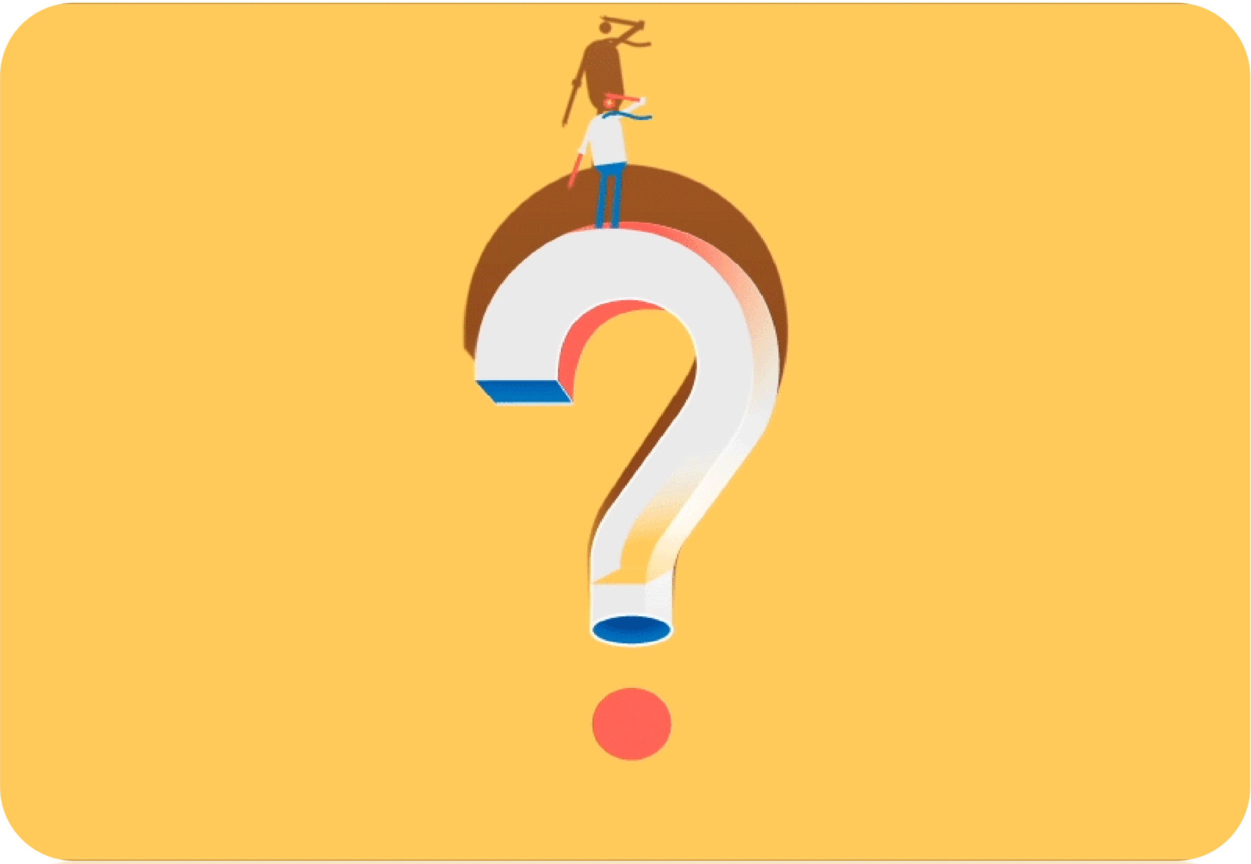


Categorizing news stories as finance, weather, entertainment, sports, etc

Summary

We discussed how data is spread out there.
We defined the meaning of data mining.
We then discussed the process of knowledge discovery in databases (KDD)
We discussed different methods of data mining and the application of those methods.

Any Questions



Join menti

https://www.menti.com/xdzes46ta9