Lecture 3

Introduction To Data Science

Dr. Faisal Kamiran

Award winning Data Scientist and Professor

What is today’s agenda?

Today we are going to learn following things :

- Introduction to Data Mining
- Basics of
  - Classification
  - Clustering
  - Association Rule Mining
  - Sequential Pattern Mining

What is (not) Data Mining

What is not Data Mining?

- Look up phone number in phone directory

What is Data Mining?

- Certain names are more prevalent in certain US locations (O’Brien, O’Rurke, O’Reilly… in Boston area)

Origins of Data Mining

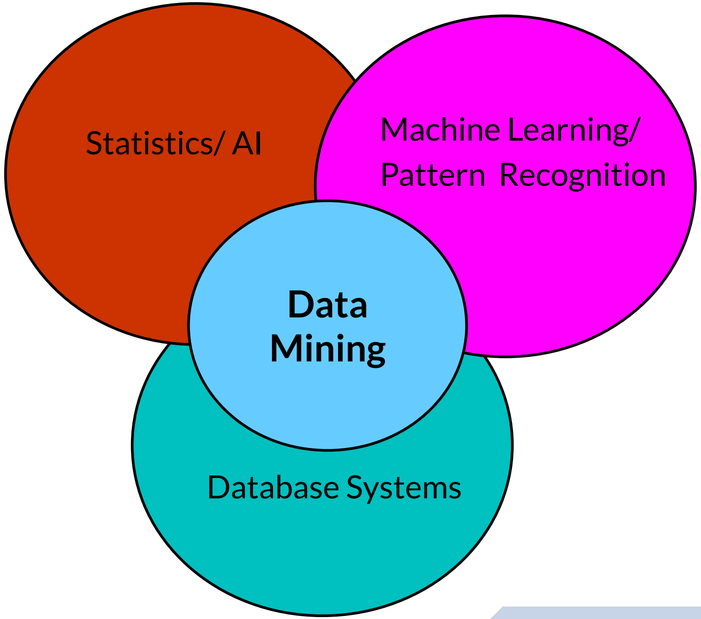
- Draws ideas from machine learning/AI, pattern recognition, statistics, and database systems.

- Traditional Techniques may be unsuitable due to

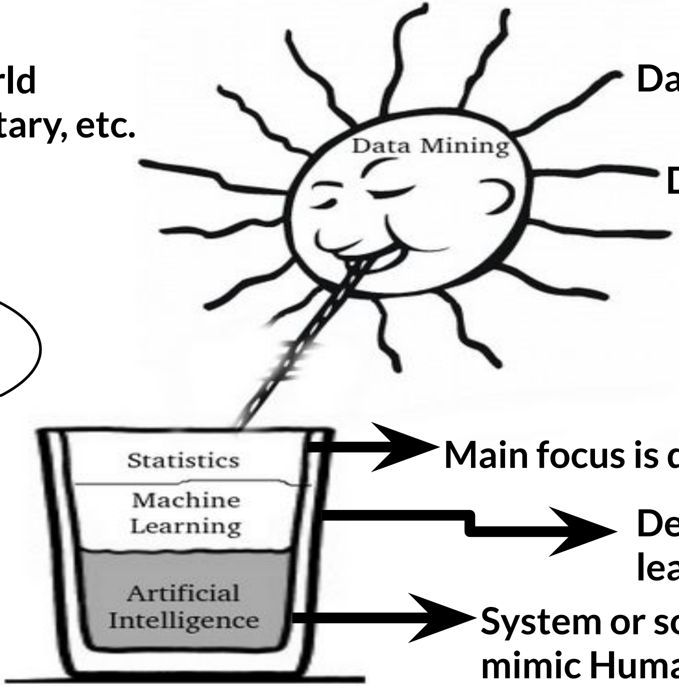
  - Enormity of data

  - High dimensionality of data

  - Heterogeneous, distributed nature of data



Origins of Data Mining



Solve real world business, Military, etc. problems

Data preprocessing

Data analysis

Practical application and Development of algorithms

Descriptive & Inferential

Main focus is data

Development of self learning algorithms

System or software to mimic Human

- NLP
- Computer Vision
- Robotics
- Reasoning

Data Mining Tasks

- Prediction Methods
  - Use some variables to predict unknown or future values of other variables.
- Description Methods
  - Find human-interpretable patterns that describe the data.

Data Mining Tasks

- Classification [Predictive]
- Clustering [Descriptive]
- Association Rule Discovery [Descriptive]
- Sequential Pattern Discovery [Descriptive]
- Regression [Predictive]
- Deviation 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.

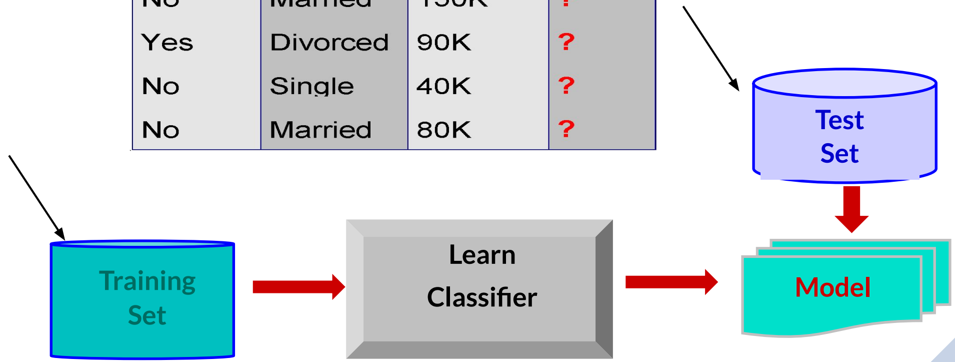
- Goal: <u>previously unseen</u> records should be assigned a class as accurately as possible.
  - A *test set* is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

Classification : Example

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

Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?



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.

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.

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:

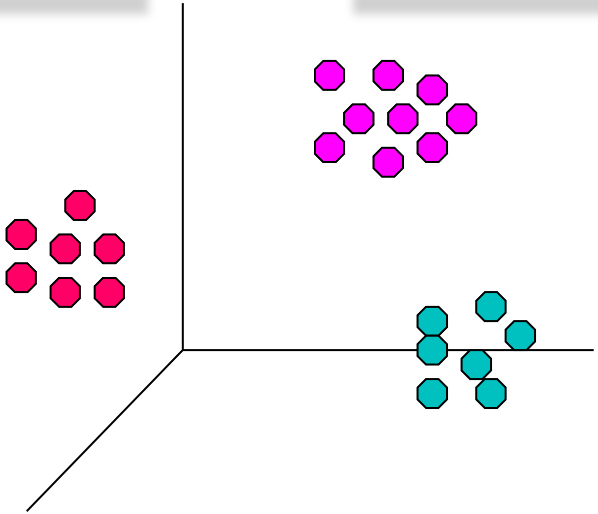
  - Euclidean Distance if attributes are continuous.
  - Other Problem-specific Measures.

Illustrating Clustering

- Euclidean Distance Based Clustering in 3-D space.

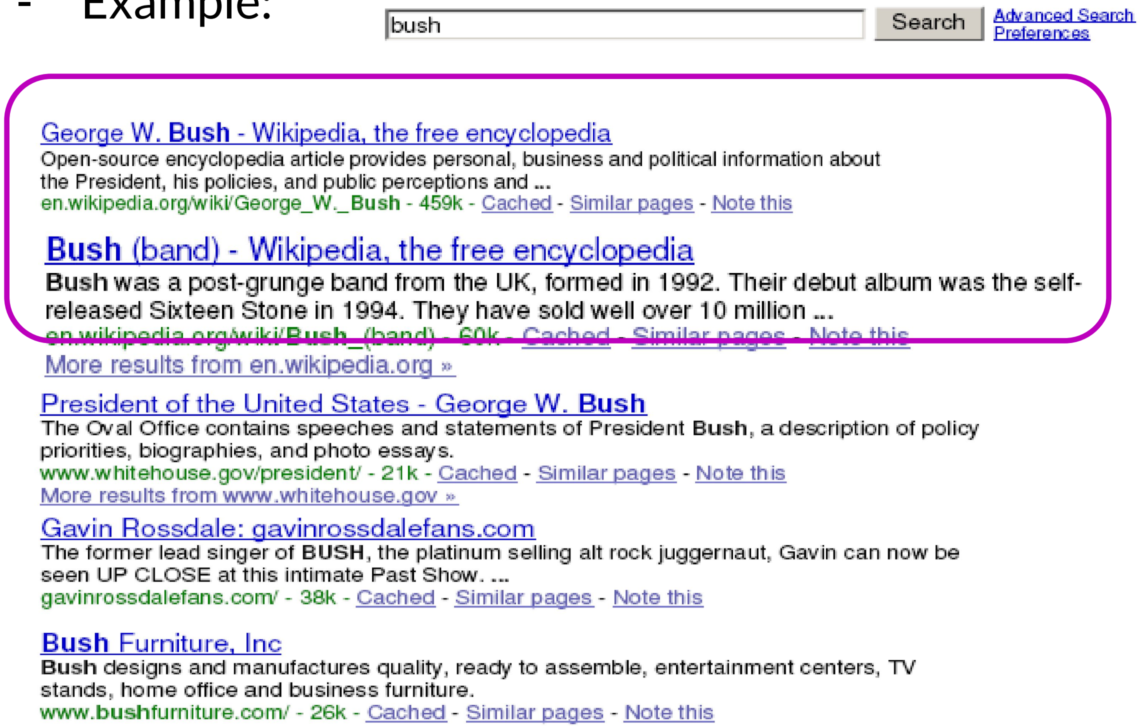
Intracluster distances are minimized

Intercluster distances are maximized



Data Mining Techniques : Clustering

- Example:



George W. Bush - Wikipedia, the free encyclopedia
Open-source encyclopedia article provides personal, business and political information about the President, his policies, and public perceptions and ...
en.wikipedia.org/wiki/George_W._Bush - 459k - Cached - Similar pages - Note this

Bush (band) - Wikipedia, the free encyclopedia
Bush was a post-grunge band from the UK, formed in 1992. Their debut album was the self-released Sixteen Stone in 1994. They have sold well over 10 million ...
en.wikipedia.org/wiki/Bush_(band) - 60k - Cached - Similar pages - Note this
More results from en.wikipedia.org »

President of the United States - George W. Bush
The Oval Office contains speeches and statements of President Bush, a description of policy priorities, biographies, and photo essays.
www.whitehouse.gov/president/ - 21k - Cached - Similar pages - Note this
More results from www.whitehouse.gov »

Gavin Rossdale: gavinrossdalefans.com
The former lead singer of BUSH, the platinum selling alt rock juggernaut, Gavin can now be seen UP CLOSE at this intimate Past Show. ...
gavinrossdalefans.com/ - 38k - Cached - Similar pages - Note this

Bush Furniture, Inc
Bush designs and manufactures quality, ready to assemble, entertainment centers, TV stands, home office and business furniture.
www.bushfurniture.com/ - 26k - Cached - Similar pages - Note this

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.

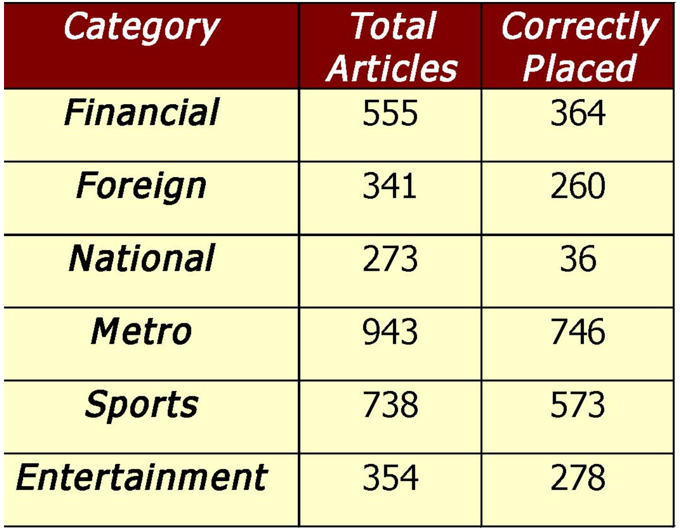
Clustering : Application 2

- 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.

Illustrating Document Clustering

- Clustering Points: 3204 Articles of Los Angeles Times.

- Similarity Measure: How many words are common in these documents (after some word filtering).



Category	*Total Articles*	*Correctly Placed*
Financial	555	364
Foreign	341	260
National	273	36
Metro	943	746
Sports	738	573
Entertainment	354	278

Classification vs Clustering

Classification

- **Input:** We have a Training set containing data that have been previously categorized

- **Task:** Based on this training set, the algorithms finds the category that the new data points belong to

- Since a Training set exists, we describe this technique as **Supervised learning**

Clustering

- **Input:** We do not know the characteristics of similarity of data in advance

- **Task:** Using statistical concepts, we split the datasets into sub-datasets such that the Sub-datasets have “Similar” data

- Since Training set is not used, we describe this technique as **Unsupervised learning**

Supervised vs Unsupervised Learning

Supervised Learning

- Correct results/labels during the training are given.
- Resultant models are generalized ones, usually fast and accurate

Unsupervised Learning

- Correct results/labels are **NOT** given in input data
- Usually computationally expensive
- Grouping of input data w.r.t. its statistical properties