Data Mining: An Overview

Professor Mo Saraee
University of Salford

Overview

- Business Intelligence
- What is Data Mining?
- What is Machine Learning?
- Data Mining vs Machine Learning: Key Differences
- Data Mining: On what kind of data?
- Data mining functionality
- Are all the patterns interesting?
- Data Mining Task Primitives
- Integration of data mining system with a DB and DW System

Business Intelligence

- In today's business environment, three things are certain
  - ★ Competition is more intense than ever
  - ★ The quantity of information is increasing proportionally
  - ★ Markets and products evolve faster than ever
- Businesses need
  - To have the appropriate information
  - To make informed decisions
  - To take timely action

Business Intelligence

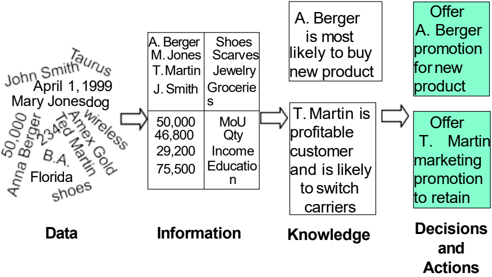
- According to the International Data Corporation (IDC), in 65 organisations, the average return on business intelligence investments was over 400% over a period of 2.3 years.

- Implies services and products in the areas of
  - Data Mining
  - Data Warehousing

Business Intelligence: Process

- Business Plan
- Architecture
- Project Planning
- Data Acquisition
- Implementation of the business intelligence solution
    - Data Warehouse
    - Intelligence tools e.g., data mining
- Evaluate the use of business intelligence

Business Intelligence
Data-Information-Knowledge-Decision



Taurus
John Smith
April 1, 1999
Mary Jones
dog
wireless
Amex Gold
Ted Martin
234
Anna Berger
50,000
B.A.
Florida
shoes

A. Berger	Shoes
M. Jones	Scarves
T.Martin	Jewelry
J. Smith	Groceries
50,000	MoU
46,800	Qty
29,200	Income
75,500	Education

A. Berger is most likely to buy new product

T. Martin is profitable customer and is likely to switch carriers

Offer A. Berger promotion for new product

Offer T. Martin marketing promotion to retain

Data

Information

Knowledge

Decisions and Actions

What is data mining?

- After years of data mining there is still no unique answer to this question.

- A tentative definition:

Data mining is the use of efficient techniques for the analysis of very large collections of data and the extraction of useful and possibly unexpected patterns indata.

What is Machine Learning?

- Machine Learning is a sub-branch of Artificial Intelligence. It is the scientific study of intelligent algorithms that can be used by machines (computers) to perform human-like tasks without being explicitly programmed or trained for it.

- A unique aspect of Machine Learning algorithms is that they can learn from data

Data Mining vs Machine Learning: Key Differences (i)

- Both Data Mining and Machine Learning are sub-domains of Data Science. So, naturally, they are inter-related.

- Data Mining and Machine Learning both employ advanced algorithms to uncover relevant data patterns. However, even though Data Mining and Machine Learning intersect each other, they have a fair share of differences as to how they are used.

- Data mining is the process of discovering hidden patterns, relationships, or insights from large datasets. Machine learning, on the other hand, focuses on building algorithms that learn from data to make predictions or decisions automatically.

Data Mining Application : Beer & Nappies

- There is a very beautiful story, which is always told in many data mining courses, about a big warehouse chain in the United States, Wal-Mart, that made a market research about its clients' purchase habits at the end of the 90's. Surprisingly, they discovered a notable statistical correlation between the purchase of nappies and drink: Friday afternoon, men between 25 and 35 year-old use to bought both products**.**

- After a detailed analysis, this result is explained in a very curious way. As nappies are quite bulky, women usually sent their husbands to buy them. Husbands and fathers, youths between 25 and 35 years-old (average age to have so little children), use to go shopping on Friday, with reluctant air, at the last possible moment.

Data Mining Application: Beer & Nappies

- These poor fathers, with no happy social life, took advantage to buy drink at the same time they bought nappies for their babies, because they couldn't go to the pub.
- It is also said that Wal-Mart used that analysis to reorganize these products in strategic places: they put Beer near nappies.

Result

- Fathers who usually bought beer started to buymore, because of it convenient situation.

- Those who didn't buy beer before began to doit.

- So that, beer sales increased spectacularly

Why do we need data mining?

- Data is power!
  - Today,the collected data is one of the biggest assets of an onlinecompany
    - Query logs of Google, The friendship and updates of Facebook, Tweets and follows of Twitter, Amazon transactions
- Data for the people:
  - Using data from the people activity we can improve their individual lives but also the overall society life.
- We need a way to harness the collective intelligence

- From Data mining to Data Science

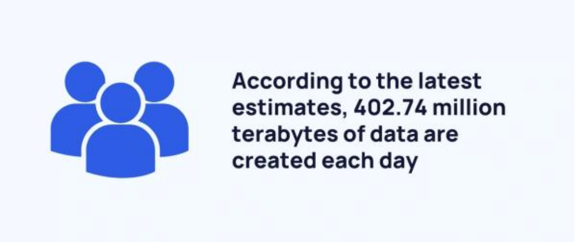
A Day of Data

- How much data is generated in a day – and what could this look like as we enter an even more data-driven future?

Here are some key daily statistics highlighted in the infographic:

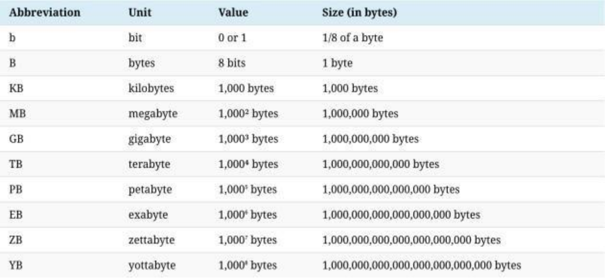
- 500 million tweets are sent
- 294 billion emails aresent
- 4 petabytes of data are created on Facebook
- 4 terabytes of data are created from each connected car
- 65 billion messages are sent on WhatsApp
- 5 billion searches are made

By 2025, it's estimated that 463 exabytes of data will be created each day globally – that's the equivalent of 212,765,957 DVDs per day!



According to the latest estimates, 402.74 million terabytes of data are created each day

https://explodingtopics.com/blog/data-generated-per-day



Abbreviation	Unit	Value	Size (in bytes)
b	bit	0 or 1	1/8 of a byte
B	bytes	8 bits	1 byte
KB	kilobytes	1,000 bytes	1,000 bytes
MB	megabyte	$1{,}000^2$ bytes	1,000,000 bytes
GB	gigabyte	$1{,}000^3$ bytes	1,000,000,000 bytes
TB	terabyte	$1{,}000^4$ bytes	1,000,000,000,000 bytes
PB	petabyte	$1{,}000^5$ bytes	1,000,000,000,000,000 bytes
EB	exabyte	$1{,}000^6$ bytes	1,000,000,000,000,000,000 bytes
ZB	zettabyte	$1{,}000^7$ bytes	1,000,000,000,000,000,000,000 bytes
YB	yottabyte	$1{,}000^8$ bytes	1,000,000,000,000,000,000,000,000 bytes

https://enewswithoutborders.com/2019/04/23/how-much-data-is-generated-each-day/

Example: transaction data

- Billions of real-life customers:
  - WALMART: 20M transactions per day
  - AT&T 300 M calls per day
  - Credit card companies: billions of transactions per day.
  - Amazon:millions of purchases per day
- The point cards allow companies to collect information about specific users

Example: document data

- Web as a document repository: estimated 50billions of web pages in Google index
  - Several trillions overall
- Wikipedia: 4.9 million articles (andcounting)

- Online news portals: steady stream of 100's of new articles every day
- Twitter: ~500 million tweets everyday

Example: genomic sequences

- http://www.1000genomes.org/page.php
- Full sequence of 1000 individuals
- 3 billion nucleotides per person 3 trillion nucleotides
- Lots more data in fact: medical history of the persons, gene expression data

Example: Medical data

- Wearable devices can measure your heart rate, blood sugar, blood pressure, and other signals about your health. Medical records are becoming available to individuals
    - Wearable computing

- Brain imaging
    - Images that monitor the activity in different areas of the brain under different stimuli
        - TB of data that need to be analyzed.

- Gene and Protein interaction networks
    - It is rare that a single gene regulates deterministically the expression of a condition.
    - There are complex networks and probabilistic models that govern the protein expression.

Example: environmental data

- Climate data (just an example)

http://www.ncdc.gov/oa/climate/ghcn-monthly/index.php

- “a database of temperature, precipitation and pressure records managed by the National Climatic Data Center, Arizona State University and the Carbon Dioxide Information Analysis Center”

- “6000 temperature stations, 7500 precipitation stations, 2000 pressure stations”
  - Spatiotemporal data

Example: Behavioral data

- Mobile phones today record a large amount of information about the user behavior
    - GPS records position
    - Camera produces images
    - Communication via phone and SMS
    - Text via facebook updates
    - Association with entities via check-ins

- Amazon collects all the items that you browsed, placed into your basket, read reviews about, purchased.

- Google and Bing record all your browsing activity via toolbar plugins. They also record the queries you asked, the pages you saw and the clicks you did.

- Data collected for millions of users on a daily basis

The data is also very complex

- Multiple types of data: database tables, text, time series, images, videos, graphs, etc

- Interconnected data of different types:
  - From the mobile phone we can collect, location of the user, friendship information, check-ins to venues, opinions through twitter, status updates in FB, images though cameras, queries to searchengines

- Spatial and temporal aspects

What can you do with the data?

- Suppose that you are the owner of a supermarket and you have collected billions of market basket data. What information would you extract from it and how would you use it?

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

Productplacement

Catalog creation

Are All the “Discovered” Patterns Interesting?

- Data mining may generate thousands of patterns: Not all of them are interesting
  - Suggested approach: Human-centered, query-based, focused mining
- **Interestingness measures**
  - A pattern is interesting if it is easily understood by humans, valid on new or test data with some degree of certainty, potentially useful, novel, or validates some hypothesis that a user seeks toconfirm
- **Objective vs. subjective interestingness measures**
  - Objective: based on statistics and structures of patterns, e.g., support, confidence, etc.
  - Subjective: based on user’s belief in the data, e.g., unexpectedness, novelty, actionability, etc.

Supervised Learning /Unsupervised Learning

- Supervised Learning
  - Build a learner model using data instances of known origin.
  - Use the model to determine the outcome new instances of unknown origin
- Unsupervised Learning
  - A data mining method that builds models from data without predefined classes.

Basic Data Mining Tasks

- **Classification**
  - maps data into predefined groups or classes
  - Supervised learning
- **Clustering**
  - groups similar data together into clusters.
  - Unsupervised learning
- **Association Rules**
  - uncovers relationships amongdata.
  - Unsupervised learning

Frequent Itemsets and Association Rules

- Given a set of records each of which contain some number of items from a given collection;
  - Identify sets of items (itemsets) occurring frequently together
  - Produce dependency rules which will predict occurrence of an item based on occurrences of other items.
- Challenge: Do this efficiently for millions of records and items

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

Itemsets Discovered:

{Milk,Coke}

{Diaper, Milk}

Rules Discovered:

{Milk} --> {Coke}

{Diaper, Milk} --> {Beer}

Example Applications

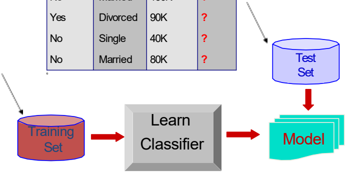
- Supermarket shelf management.
  - Goal: To identify items that are bought together by sufficiently many customers.
  - Approach: Process the point-of-sale data collected with barcode scanners to find dependencies among items.
  - A classic rule -- If a customer buys diaper and milk, then he is very likely to buy beer.
    - So, don't be surprised if you find six-packs stacked next to diapers!
- Text mining: finding associated phrases in text
  - There are lots of documents that contain the phrases "association rules", "data mining" and "efficient algorithm"
  - Can be used to define key phrases, correct spelling mistakes, associate different concepts.

Classification Example: Tax Fraud

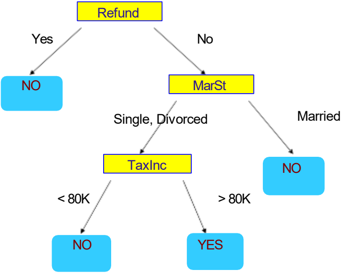
categorical categorical continuous class

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	?



Model Example: Decision Trees



Classification: Application 1

- Ad Click Prediction
  - Goal: Predict if a user that visits a web page will click on a displayed ad. Use it to target users with high click probability.
  - Approach:
    - Collect data for users over a period of time and record who clicks and who does not. The {click, no click} information forms the class attribute.
    - Use the history of the user (web pages browsed, queries issued) as the features.
    - Learn a classifier model and test on new users.

Clustering Definition

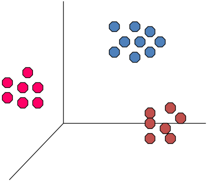
- Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that
  - Data points in one cluster are more similar to one another.
  - Data points in separate clusters are less similar to one another.
- Similarity Measures?
  - Euclidean Distance 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



Clustering: Application 1

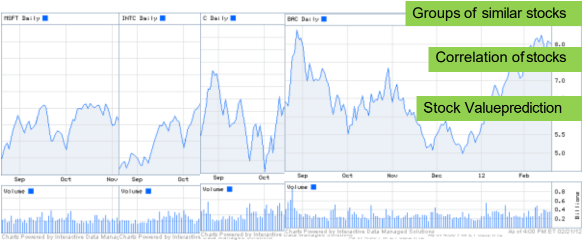
- Bioinformatics applications:
  - Goal: Group genes and tissues together such that genes are coexpressed on the same tissues

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 basedon the frequencies of different terms. Use it to cluster.
  - Gain: Information Retrieval can utilize the clustersto relate a new document or search term to clustered documents.

What can you do with the data?

- Suppose you are a stock broker and you observe the fluctuations of multiple stocks over time. What information would you like to get our of your data?



Groups of similar stocks

Correlation of stocks

Stock Valueprediction

What can you do with the data?

- You are the owner of a social network, and you have full access to the social graph, what kind of information do you want to get out of your graph?

  - Who is the most important node in the graph?
  - What is the shortest path between two nodes?
  - How many friends two nodes have in common?
  - How does information spread in the network?
  - How likely are two nodes to become friends?

Why Data Mining?

- Scientific point of view
  - Scientists are at an unprecedented position where they can collect TB of information
    - Examples: Sensor data, astronomy data, social network data, genedata
  - We need the tools to analyze such data to get a better understanding of the world and advance science and help people
- Commercial point of view
  - Data has become the key competitive advantage of companies
    - Examples: Facebook, Google, Amazon
  - Being able to extract useful information out of the data is key for exploiting them commercially.
- Scale (in data size and feature dimension)
  - Why not use traditional analytic methods?
  - Enormity of data, curse of dimensionality
  - The amount and the complexity of data does not allow for manual processing of the data. We need automated techniques.
- "Data is the new oil"

What is Data Mining again?

- "Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data analyst" (Hand, Mannila, Smyth)

- "Data mining is the discovery of models for data" (Rajaraman, Ullman)
  - We can have the following types of models
    - Models that explain the data (e.g., a single function)
    - Models that predict the future data instances.
    - Models that summarize the data
    - Models the extract the most prominent features of the data.

What is data mining again?

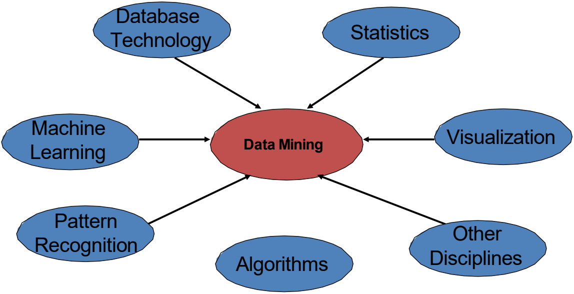
- “Data Mining is the study of collecting, processing, analyzing, and gaining useful insights from data” – Charu Aggarwal

- Essentially, anything that has to do with data is data mining

What is data mining again?

- The industry point of view: The analysis of huge amounts of data for extracting useful and actionable information, which is then integrated into production systems in the form of new features of products
  - Data Scientists should be good at data analysis, math, statistics, but also be able to code with huge amountsof data and use the extracted information to build products.

Data Mining: Confluence of Multiple Disciplines


Database Technology
Statistics
Machine Learning
Data Mining
Visualization
Pattern Recognition
Algorithms
Other Disciplines