Week1: Introduction to Data Mining

CSCI 442-542 Lecture Notes, Fall 2018

Charleston Southern University, Charleston, SC, USA
By
Oluleye (Hezekiah) Babatunde, Ph.D
Email:obabatunde@csuniv.edu

August 20, 2018

Course outlines

- week1: Introduction to data mining.
- week2: Basic Mathematical Concepts I.
- week3: Basic Mathematical Concepts II.
- week4: Measurement and Data.
- week5: Visualizing and Exploring Data.
- week6: Systematic Overview of Data Mining Algorithms.
- week7 Midterm.
- week8 Bayes Theorem.
- week9, 10: Classification Algorithms/Clustering.
- week11: Artificial Neural Networks.
- week12: Genetic Algorithms.
- week13 Text Retrieval
- week14 Project Presentation.

Instructor's Educational Backgrounds

- **Postdoctoral Research**:
  Quantitative Analysis and Modelling of Chemical and Biological Systems (Systems Biology) with BioInformatics [USA]
- **PhD Thesis (Computer Science)**:
  A Neuro-Genetic Approach To Automatic Identification of Plant Species. [ECU Australia]
- **M.Sc Mathematics Thesis**:
  On Numerical simulation of a class of stochastic differential equation with boundary values [UI Ibadan].
- **M.Sc Computer Science Thesis**:
  A Cellular Neural Networks-Based Model for Edge Detection. [FUNAAB]
- **B.Sc Mathematical Sciences (Computer Major)** [FUNAAB].

Data Mining (Definition and History)

Definition (Data Mining)

The science of extracting useful information from large data sets or databases is known as data mining. It is a new discipline, lying at the intersection of statistics, machine learning, data management and databases, pattern recognition, artificial intelligence, and other areas.

Data mining is the ***exploration*** *and* ***analysis*** *of large quantities of data in order to discover valid, novel, potentially useful, and ultimately understandable patterns in data.*

Understandable*: We can interpret and comprehend the patterns.*

Data Mining (Definition and History) (contd.)

History [1], [2]

Thomas Bayes published a paper in 1763 regarding a theorem for relating current probability to prior probability called the Bayes' theorem. It is fundamental to data mining and probability, since it allows understanding of complex realities based on estimated probabilities.

Data Mining (Definition and History) (contd.)

History

In 1805 **Adrien-Marie Legendre and Carl Friedrich Gauss** applied regression to determine the orbits of bodies about the Sun (comets and planets). The goal of **regression analysis** was to estimate the relationships among variables, and the specific method they used in this case is the method of **least squares. Regression** is one of the main techniques used in data mining.

Data Mining (Definition and History) (contd.)

In 1936 **Alan Turing** introduced the idea of a Universal Machine capable of performing computations like the modern day computers. The modern day computers are built on the concepts pioneered by Turing.

Data Mining (Definition and History) (contd.)

Warren McCulloch and **Walter Pitts** created a conceptual model of a neural network in 1943. They published a paper titled "**A logical calculus of the ideas immanent in nervous activity**" in which they described the idea of a neuron in a network. It was stated that each of these neurons can do 3 things: receive inputs, process inputs and generate output.

Data Mining (Definition and History) (contd.)

In 1965 **Lawrence J. Fogel** formed a new company called **Decision Science, Inc**. for applications of evolutionary programming. It has been said to be the first company that applied evolutionary computation to solve real-world problems.

Data Mining (Definition and History) (contd.)

Sophisticated database management systems emerged in 1970s which made it possible to store and query terabytes and petabytes of data. In addition, data warehouses allow users to move from a transaction-oriented way of thinking to a more analytical way of viewing the data. However, extracting sophisticated insights from these data warehouses of multidimensional models was very limited.

Data Mining (Definition and History) (contd.)

John Henry Holland wrote Adaptation in Natural and Artificial Systems, the ground-breaking book on genetic algorithms in 1975. It's the book that initiated this field of study, presenting the theoretical foundations and exploring applications.

Data Mining (Definition and History) (contd.)

HNC trademarks the phrased the term “**database mining**.” in 1980s. The trademark was meant to protect a product called **DataBase Mining Workstation**. It’s during this period that sophisticated algorithms can “learn” relationships from data that allow subject matter experts to reason about what the relationships mean.

Data Mining (Definition and History) (contd.)

Gregory Piatetsky-Shapiro coined the term "**Knowledge Discovery in Databases**" (KDD) in 1989. It also at this time that he co-founded the first workshop also named KDD.

Data Mining (Definition and History) (contd.)

The term "**data mining**" appeared in the database community in 1990s. Retail companies and the financial community started using data mining to analyze data and recognize trends to increase their customer base, predict fluctuations in interest rates, stock prices, customer demand.

Data Mining (Definition and History) (contd.)

In 1992 **Bernhard E. Boser, Isabelle M. Guyon and Vladimir N. Vapnik** suggested an improvement on the original **support vector machine** which allows for the creation of nonlinear classifiers. Support vector machines are a supervised learning approach that analyzes data and recognizes patterns used for classification and regression analysis.

Data Mining (Definition and History) (contd.)

In 1993 **Gregory Piatetsky-Shapiro** starts the newsletter Knowledge Discovery Nuggets (KDnuggets). It was originally meant to connect researchers who attended the KDD workshop. However, www.KDnuggets.com seems to have a much wider audience now.

Data Mining (Definition and History) (contd.)

In 2001 **William S. Cleveland** introduced it as an independent discipline (although the term data science has existed since 1960s,). As per Build Data Science Teams, DJ Patil and Jeff Hammerbacher then used the term to describe their roles at **LinkedIn** and **Facebook**

Data Mining (Definition and History) (contd.)

Michael Lewis published Moneyball in 2003 and that changed the way many major league front offices do business. The Oakland Athletics used a statistical, data-driven approach to select for qualities in players that were undervalued and cheaper to obtain. In this manner, they successfully assembled a team that brought them to the 2002 and 2003 playoffs with 1/3 the payroll.

CHARLESTON SOUTHERN UNIVERSITY

Data Mining (Definition and History) (contd.)

DJ Patil became the first Chief Data Scientist at the White House in February 2015. Today, data mining is widespread in business, science, engineering and medicine just to name a few. Mining of credit card transactions, stock market movements, national security, genome sequencing and clinical trials are just the tip of the iceberg for data mining applications.Terms like Big Data are now commonplace with the collection of data becoming cheaper and the proliferation of devices capable of collecting data.

CHARLESTON SOUTHERN UNIVERSITY

Data Mining (Definition and History) (contd.)

One of the most active techniques being explored today is Deep Learning. Capable of capturing dependencies and complex patterns far beyond other techniques, it is reigniting some of the biggest challenges in the world of data mining, data science and artificial intelligence.

Models in Data Mining

Models in Data Mining

- The **relationship and summaries** derived through the process are called models or patterns. Examples are:

- : linear equations
- : rules
- : graphs
- : trees
- : clusters

- Usually data mining is referred to as **secondary** data analysis because the data were collected for another purpose (sell product, health care, etc)

Data Mining and other disciplines

Data Mining and other disciplines

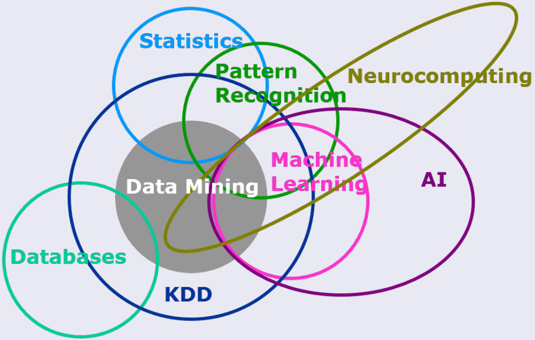


Figure 1: Data mining and other related areas

(Data Matrix

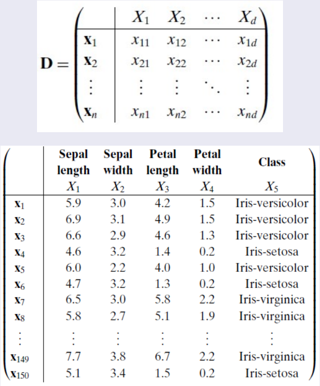
Dataset (Data Matrix)

data set: set of measurements taken from some environment or process.

- If we have a collection of n objects and d measurements on those objects, we can think of our data as an $n \times d$ data matrix
- the n rows are normally called individuals, entities, cases, objects or records
- the d columns are often called variables, features, attributes or fields

(Data Matrix (contd.)

Data Matrix



$$\mathbf{D} = \left(\begin{array}{c|cccc} & X_1 & X_2 & \cdots & X_d \\ \hline \mathbf{x}_1 & x_{11} & x_{12} & \cdots & x_{1d} \\ \mathbf{x}_2 & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_n & x_{n1} & x_{n2} & \cdots & x_{nd} \end{array}\right)$$

	Sepal	**Sepal**	**Petal**	**Petal**	**Class**
	length	**width**	**length**	**width**	
	X_1	X_2	X_3	X_4	X_5
$\mathbf{x}_1$	5.9	3.0	4.2	1.5	Iris-versicolor
$\mathbf{x}_2$	6.9	3.1	4.9	1.5	Iris-versicolor
$\mathbf{x}_3$	6.6	2.9	4.6	1.3	Iris-versicolor
$\mathbf{x}_4$	4.6	3.2	1.4	0.2	Iris-setosa
$\mathbf{x}_5$	6.0	2.2	4.0	1.0	Iris-versicolor
$\mathbf{x}_6$	4.7	3.2	1.3	0.2	Iris-setosa
$\mathbf{x}_7$	6.5	3.0	5.8	2.2	Iris-virginica
$\mathbf{x}_8$	5.8	2.7	5.1	1.9	Iris-virginica
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathbf{x}_{149}$	7.7	3.8	6.7	2.2	Iris-virginica
$\mathbf{x}_{150}$	5.1	3.4	1.5	0.2	Iris-setosa

Figure 2: Data matrix

(Data Matrix (contd.)

Depending on the application domain, rows may also be referred to as ***entities, instances, examples, records, transactions, objects, points, feature-vectors, tuples***, and so on.

Likewise, columns may also be called ***attributes, properties, features, dimensions, variables, fields***, and so on.

The number of instances n is referred to as the size of the data, whereas the number of attributes d is called the dimensionality of the data.

The analysis of a single attribute is referred to as **univariate analysis** (e.g $f(x) = x^2 + 5$), whereas the simultaneous analysis of two attributes is called **bivariate analysis** (e.g $f(x, y) = x^2 + y^2$) and the simultaneous analysis of more than two attributes is called **multivariate analysis** e.g (e.g $f(x, y, z) = x^2 + y^2 + z^2$).

(Data Matrix (contd.)

Example data sets

- US Census Bureau Public Use Microdata Samples (PUMS).
- UCI Machine Learning Repository `www.ics.uci.edu/~mlearn/mlsummary.html`
- Textual data for mining “Reuters-21578, Distribution 1.0”, `www.research.att.com/~lewis`
- Transaction data from businesses.

Main types of data

Main types of data

- **Quantitative**: measured on a numerical scale and in principle can take any value examples:age, income, weight, etc.
- **Categorical variables**: can only be certain discrete variables examples: gender, race, etc.
  - **Ordinal** – possessing a natural order – example: grade in school – 1st, 2nd, 3rd
  - **nominal** – simply naming categories – example: marital status (married, single, divorsed, etc)

References

- Principles of Data Mining. Hand, David, Heikki Manilla, Padhraic Smyth. ISBN: 9780262082907

- DataMining .
  `https://hackerbits.com/data/history-of-data-mining/`