Introduction to Data Mining and Machine Learning

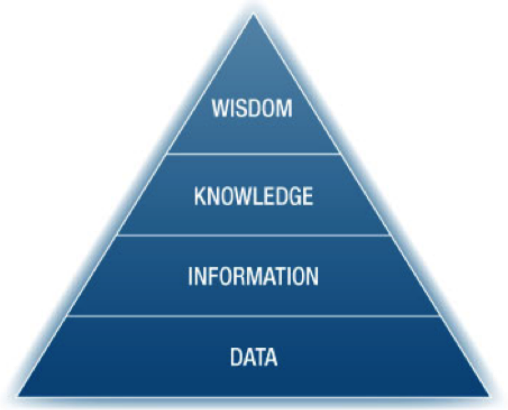
COMP90049
Knowledge Technologies

Sarah Erfani and Karin Verspoor, CIS

Semester 2, 2018

THE UNIVERSITY OF
MELBOURNE

From Data to Wisdom



WISDOM

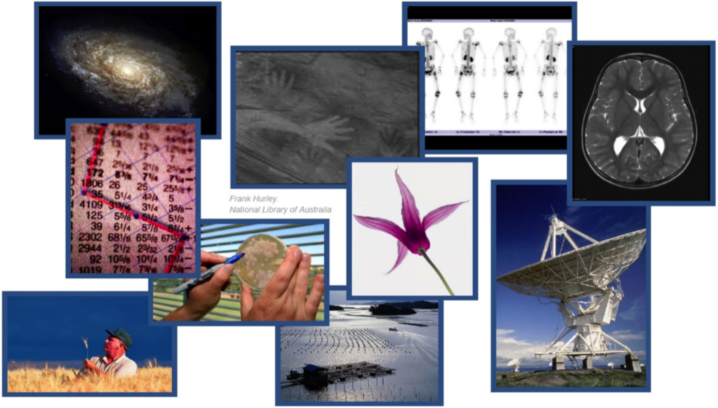
KNOWLEDGE

INFORMATION

DATA

`http://www.innovation.gov.au/Science/PMSEIC/Documents/DataForScience.pdf`

Remember: Data is everywhere



Frank Hurley.
National Library of Australia

Reminder: What is Knowledge?

Reminder: What is Knowledge?

Information interpreted with respect to a user's context to extend human understanding in a given area.

. . . In the context of data, perhaps:

Increasing insight into data, based on a user's information needs in a given context.

"Big Data"

Tackling the challenge of knowledge management and discovery at a massive scale

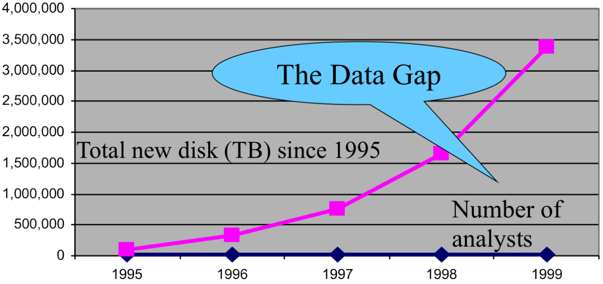
- Database modelling and integration has long been a focus of Information Technology research and development. Classic example being the application of RDBMs for commercial apps.
- A major and accelerating trend is the focus of data integration from business and enterprise applications to scientific and personal applications.
- Exponential growth of data with the spread of the Internet, Web and the multitudes of automatic data generation and collection devices.

This trend is expected to continue in the foreseeable future.

The significance of Data

Importance of Problem

- Current computational methods cannot handle magnitude and dimensionality of the data
- Decision makers and Scientists need techniques to help form hypotheses and make evidence based decisions



Tools are required to integrate, distill, and make sense of data.

What is Data Mining?

What is Data Mining?

Extracting

- implicit,
- *previously unknown*,
- potentially useful

information from data

- Needed: programs that detect patterns and regularities in the data
- Strong patterns $\rightarrow$ good predictions
  - Problem 1: most patterns are not interesting
  - Problem 2: patterns may be inexact (or spurious)
  - Problem 3: data may be garbled or missing

Machine learning definitions

Arthur Samuel (1959)

- “Field of study that gives computers the ability to learn without being explicitly programmed”

Machine learning definitions

Arthur Samuel (1959)

- “Field of study that gives computers the ability to learn without being explicitly programmed”

Tom Mitchell (1999)

- “A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.”

What is Machine Learning?

Algorithms for acquiring structural descriptions from examples

- Structural descriptions represent patterns explicitly
- Can be used to predict outcome in new situation
- Can be used to understand and *explain* how prediction is derived (may be even more important)

Methods originate from artificial intelligence, statistics, and research on databases

Can machines really learn?

Dictionary definitions of "learning":

- To get knowledge of by study, experience, or being taught
- To become aware by information or from observation
- To commit to memory
- To be informed of, ascertain; to receive instruction

$\rightarrow$ Difficult to measure; Trivial for computers

Operational definition:

- Things learn when they change their behaviour in a way that makes them perform better in the future.
- Does learning imply intention?

Types of machine learning algorithms

Supervised learning

- *Teach* the computer how to do something *(by example)*, then let it use its new-found knowledge to do it
- Labeled data: for given inputs, provide the expected output ("the answer")
- Infer a function mapping from inputs to outputs

Unsupervised learning

- Let the computer *learn how to do something*
- Determine structure and patterns in data
- Unlabeled data: Don't give the computer "the answer"

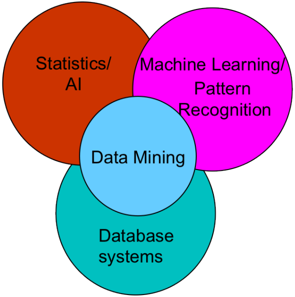
Data Mining vs Machine Learning

The distinctions between Data Mining and Machine Learning are not cut-and-dried.

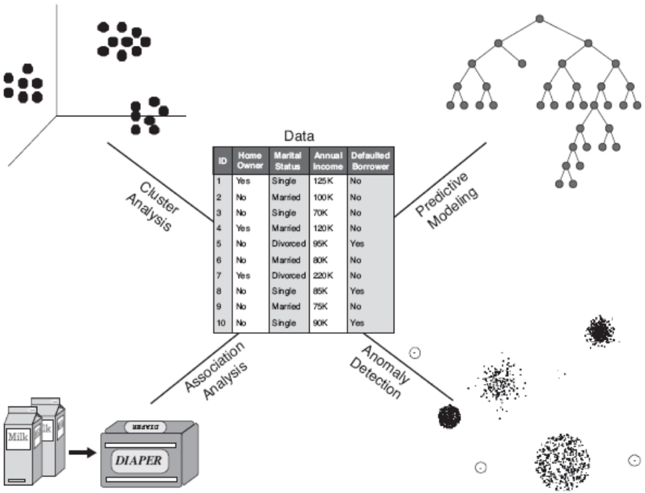
Data mining is primarily about discovering something hidden in your data, that you did not know before, as "new" as possible. *Knowledge obtained from data.*

Machine learning emphasises algorithms used to generalise existing knowledge to new data, as accurately as possible. *Techniques used to learn from data.*

Data mining applications typically use a lot of machine learning techniques. For example a pattern in a data set that is useful for generalisation might represent new knowledge.



Core tasks



ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
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	80K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Figure 1.3. Four of the core data mining tasks.

From: Tan, Steinbach, Kumar (2006) Introduction to Data Mining.

Types of applications

Supervised learning

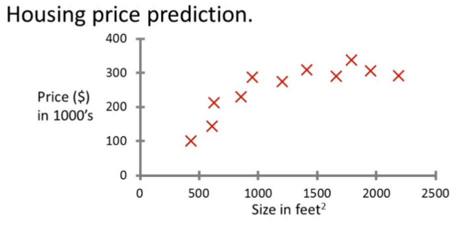
- Classification
  predicting a discrete class
- Regression
  predicting a numeric quantity

Unsupervised learning

- Association
  detecting associations between features
- Information organisation; Clustering
  grouping similar instances into clusters
- Reinforcement learning
- Recommender systems
- Anomaly/outlier detection

Example: Supervised Learning (Regression)

Can we predict housing prices?

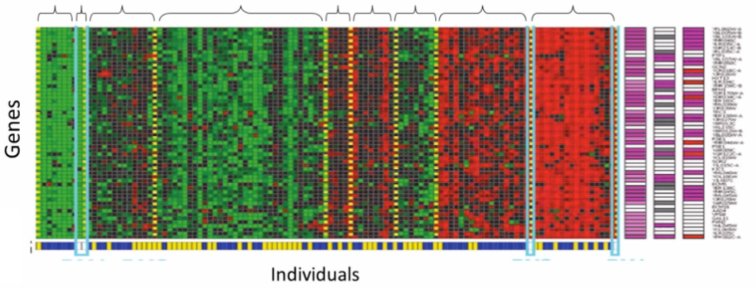


A friend has a house which is 750 square feet – how much can he expect to get?

(draw a straight line vs. fit a curve)

Example: Unsupervised Learning

- Given gene expression data for individuals, cluster based on expression profiles
- Group newspaper articles into cohesive groups
- Credit card fraud
- Network intrusion behaviour



Some basic Machine Learning concepts

- The input to a machine learning system consists of:
  - *Instances*: the individual, independent examples of a concept

    also known as exemplars
  - *Attributes*: measuring aspects of an instance

    also known as features
  - *Concepts*: things that we aim to learn

Example: Supervised Learning (Classification)

Outlook	Temperature	Humidity	Windy	Play
sunny	hot	high	FALSE	no
sunny	hot	high	TRUE	no
overcast	hot	high	FALSE	yes
rainy	mild	high	FALSE	yes
rainy	cool	normal	FALSE	yes
rainy	cool	normal	TRUE	no
⋮	⋮	⋮	⋮	⋮

Given information about current weather conditions and the forecast, can we determine whether we will go out to play?

Classification (Instances/Training examples)

	Outlook	Temperature	Humidity	Windy	Play
INSTANCE$_1$	sunny	hot	high	FALSE	no
INSTANCE$_2$	sunny	hot	high	TRUE	no
	overcast	hot	high	FALSE	yes
	rainy	mild	high	FALSE	yes
	rainy	cool	normal	FALSE	yes
	rainy	cool	normal	TRUE	no
	⋮	⋮	⋮	⋮	⋮

Classification (Attributes/Features)

Outlook	Temperature	Humidity	Windy	Play
sunny	hot	high	FALSE	no
sunny	hot	high	TRUE	no
overcast	hot	high	FALSE	yes
rainy	mild	high	FALSE	yes
rainy	cool	normal	FALSE	yes
rainy	cool	normal	TRUE	no
⋮	⋮	⋮	⋮	⋮

ATTRIBUTE$_1$

ATTRIBUTE$_2$

Classification (Classes/Labels)

Outlook	Temperature	Humidity	Windy	Play
sunny	hot	high	FALSE	no
sunny	hot	high	TRUE	no
overcast	hot	high	FALSE	yes
rainy	mild	high	FALSE	yes
rainy	cool	normal	FALSE	yes
rainy	cool	normal	TRUE	no
⋮	⋮	⋮	⋮	⋮

CLASS

Attributes

Each instance is described by a fixed feature vector

Possible attribute types (levels of measurement):

- nominal
- ordinal
- continuous

Nominal attributes

- Values are distinct symbols (e.g. {`sunny`,`overcast`,`rainy`})
  - values themselves serve only as labels or names
- Also called *categorical*, *enumerated*, or *discrete* (NB. “enumerated” and “discrete” imply an order which tends not to exist)
- Special case: dichotomy (“boolean” attribute)
- No relation is implied among nominal values (no ordering or distance measure), and only equality tests can be performed

Ordinal attributes

- An explicit order is imposed on the values (e.g. $\{\texttt{hot},\texttt{mild},\texttt{cool}\}$ where $\texttt{hot} > \texttt{mild} > \texttt{cool}$)
- No distance between values defined; addition and subtraction don't make sense
- Example rule: $\texttt{temperature} < \texttt{hot} \rightarrow \texttt{play} = \texttt{yes}$
- Distinction between nominal and ordinal not always clear (e.g. `outlook`)

Continuous attributes

- Continuous features are real-valued with a well-defined zero point and no explicit upper bound
- Also called *numeric*
- Example: attribute `distance`

  Distance between an object and itself is zero
- All mathematical operations are allowed

Thought experiment

How might you approach data mining the Weather dataset?

- Methods
  - Using Supervised methods?
  - Using Unsupervised methods?
- Attributes
  - Are there regularities among the attributes?
  - Are there different ways you could make use of the attributes (e.g. different combinations? different thresholds?)?

Books and Websites

Introduction to Data Mining
Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. 2006. Addison Wesley.
`http://www-users.cs.umn.edu/~kumar/dmbook/index.php`

Data Mining: Practical Machine Learning Tools and Techniques
Ian Witten, Eibe Frank, Mark Hall
`http://www.cs.waikato.ac.nz/ml/weka/book.html`

Tools and Resources

WEKA Toolkit
`http://www.cs.waikato.ac.nz/ml/weka/index.html`

List of more specific tools
`http://www-users.cs.umn.edu/~kumar/dmbook/resources.htm`

Data sets
UC Irvine Machine Learning Data Repository
`http://archive.ics.uci.edu/ml/datasets.html`