Week 1

Data Mining and Knowledge Discovery

Dr John Evans
j.evans8@herts.ac.uk

University of Hertfordshire UH

Plan for today

What is Data Mining?

Why Data Mining?

Data Mining Tasks

Roadmap for course

What is data mining?

- Roughly, data mining is the process of automating the procedure of discovering useful information in large data sets.
- Automation is often key since we usually deal with very large sets.

Goal: Find novel and useful patterns that might otherwise remain unknown, as well as potentially predicting the outcome of future observations.

Data mining and other information discovery tasks

We generally differentiate between data mining and other information discovery tasks.

e.g. queries such as looking up individual records in a database is not considered a data mining task.
e.g. finding web pages that contain a particular set of keywords is not considered a data mining task.

Why?

Data mining and other information discovery tasks

We generally differentiate between data mining and other information discovery tasks.

e.g. queries such as looking up individual records in a database is not considered a data mining task.
e.g. finding web pages that contain a particular set of keywords is not considered a data mining task.

Why? Because such tasks can be accomplished through simple interactions with a database management system or an information retrieval system. These rely on traditional techniques (indexing structures, query processing algorithms etc.)

Examples of data mining tasks

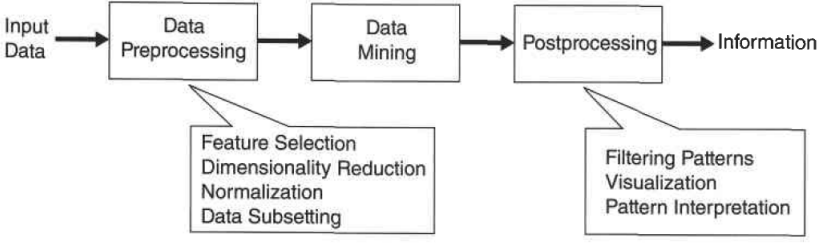
- Predictive
  - Classification
  - Prediction
  - Time-series analysis
- Descriptive
  - Association
  - Clustering
  - Summarisation

Knowledge Discovery in Databases (KDD)

- *Knowledge discovery* and *data mining* are terms often used interchangeably in the literature. However, for us we want to discriminate between the two. We see data mining as a part of the knowledge discovery process.

Knowledge Discovery in Databases (KDD)

- *Knowledge discovery* and *data mining* are terms often used interchangeably in the literature. However, for us we want to discriminate between the two. We see data mining as a part of the knowledge discovery process.
- So, knowledge discovery refers to the entire knowledge extraction process. It involves converting raw data into useful information.



Knowledge Discovery in Databases (KDD)

- Step 0 - Select the data before input. This is often domain specific.

Knowledge Discovery in Databases (KDD)

- Step 0 - Select the data before input. This is often domain specific.
- Step 1 - Input some data which can be stored in a variety of formats (flat files, spreadsheets, relational tables etc.). The data may be in a centralised data repository, or be distributed across multiple sites.

Knowledge Discovery in Databases (KDD)

- Step 0 - Select the data before input. This is often domain specific.
- Step 1 - Input some data which can be stored in a variety of formats (flat files, spreadsheets, relational tables etc.). The data may be in a centralised data repository, or be distributed across multiple sites.
- Step 2 - Transform this raw data into an appropriate format for subsequent analysis. This can include:
  - combining data from multiple sources;
  - cleaning data to remove noise;
  - selecting records and features that are relevant to the data mining task at hand;
  - transforming data into a common format (e.g. ensuring measurements are all in the same units).

  This step is generally called *preprocessing*.

Knowledge Discovery in Databases (KDD)

- Step 3 - The actual data mining takes place and results are integrated into decision support systems. This is where *postprocessing* takes place and is used to ensure that only valid and useful results are incorporated into the support system. This includes:
  - visualisation;
  - hypothesis testing.

  Key Point: Consider how the results will be interpreted.

Why data mining?

We now discuss some motivations for data mining. The first is what is often referred to as *big data* and has come from rapid advances in data collection, storage and transfer. A big player in this is, of course, the internet.

Why data mining?

We now discuss some motivations for data mining. The first is what is often referred to as *big data* and has come from rapid advances in data collection, storage and transfer. A big player in this is, of course, the internet.

Problem: Such data sets are often huge (with sizes of terabytes, petabytes, exabytes), can be highly complex (variety) and are collected at an increasingly fast pace.

Solution: We need to automate information extraction processes.

Business and Industry

Data can come from any area of interest. For instance, in business and industry, data can arise from the following:

- Point-of-sale data collection (barcode scanners, smart card technology, and so on).

Business and Industry

Data can come from any area of interest. For instance, in business and industry, data can arise from the following:

- Point-of-sale data collection (barcode scanners, smart card technology, and so on).
  This allows retailers to collect up-to-the-minute data about customer purchases.
  This can help with
  - customer profiling;
  - targeted marketing;
  - store layout;
  - fraud detection;
  - automated buying and selling;
  - cross-selling/up-selling.

  Techniques such as *association analysis* can help here.

Internet-based services

- Data collection from the internet
  - Data about online viewing/shopping preferences can be used to provide personalised recommendations of products.

Internet-based services

- Data collection from the internet
  - Data about online viewing/shopping preferences can be used to provide personalised recommendations of products.
  - Such methods can also be used to spread disinformation.

Internet-based services

- Data collection from the internet
  - Data about online viewing/shopping preferences can be used to provide personalised recommendations of products.
  - Such methods can also be used to spread disinformation.
  - Data mining also plays a prominent role in supporting other internet-based services, such as filtering spam messages, answering search queries, social update suggestions etc.

  Techniques such as *deep learning* can help here.

Medicine, Science and Engineering

Researchers are rapidly accumulating data that is key to significant new discoveries.

- NASA's deployment of a series of Earth orbiting satellites has led to a wealth of data that helps improve our understanding of Earth's climate system. These satellites continuously generate global observations of the land surface, oceans and atmosphere.
  - Such datasets are generally vast.

[1]This week, Svante Pääbo won the Nobel Prize for his work on Neanderthals & Denisovans.

Medicine, Science and Engineering

Researchers are rapidly accumulating data that is key to significant new discoveries.

- NASA's deployment of a series of Earth orbiting satellites has led to a wealth of data that helps improve our understanding of Earth's climate system. These satellites continuously generate global observations of the land surface, oceans and atmosphere.
  - Such datasets are generally vast.
- Molecular biology has benefited from large amounts of genomic data.[1]
  - Traditional methods allowed scientists to study a few genes at a time. Recent breakthroughs in microarray technology enable scientists to compare the behaviour of thousands of genes under various situations.
  - Such comparisons help determine the function of each gene and isolate genes responsible for certain diseases.

[1] This week, Svante Pääbo won the Nobel Prize for his work on Neanderthals & Denisovans.

Medicine, Science and Engineering

Researchers are rapidly accumulating data that is key to significant new discoveries.

- NASA's deployment of a series of Earth orbiting satellites has led to a wealth of data that helps improve our understanding of Earth's climate system. These satellites continuously generate global observations of the land surface, oceans and atmosphere.
  - Such datasets are generally vast.
- Molecular biology has benefited from large amounts of genomic data.[1]
  - Traditional methods allowed scientists to study a few genes at a time. Recent breakthroughs in microarray technology enable scientists to compare the behaviour of thousands of genes under various situations.
  - Such comparisons help determine the function of each gene and isolate genes responsible for certain diseases.
  - Data mining can also address challenges such as protein structure prediction, multiple sequence alignment, modelling of biochemical pathways, phylogenetics.

[1] This week, Svante Pääbo won the Nobel Prize for his work on Neanderthals & Denisovans.

Medicine, Science and Engineering

Researchers are rapidly accumulating data that is key to significant new discoveries.

- NASA's deployment of a series of Earth orbiting satellites has led to a wealth of data that helps improve our understanding of Earth's climate system. These satellites continuously generate global observations of the land surface, oceans and atmosphere.
  - Such datasets are generally vast.
- Molecular biology has benefited from large amounts of genomic data.[1]
  - Traditional methods allowed scientists to study a few genes at a time. Recent breakthroughs in microarray technology enable scientists to compare the behaviour of thousands of genes under various situations.
  - Such comparisons help determine the function of each gene and isolate genes responsible for certain diseases.
  - Data mining can also address challenges such as protein structure prediction, multiple sequence alignment, modelling of biochemical pathways, phylogenetics.
  - Such datasets are often noisy and high-dimensional. This means we need new methods of analysis.

[1] This week, Svante Pääbo won the Nobel Prize for his work on Neanderthals & Denisovans.

Motivating challenges

1. Scalability - If data mining algorithms are to handle massive data sets, they must be scalable.

Motivating challenges

1. Scalability - If data mining algorithms are to handle massive data sets, they must be scalable.
2. High dimensionality - data may now contain hundreds or thousands of attributes
   - e.g. In bioinformatics, gene expression data often involves thousands of features.
   - e.g. Healthcare datasets can involve a large number of features, such as blood pressure, resting heart rate, immune system status, surgery history, height, weight, existing conditions etc.
   - Data with temporal and/or spatial components also tend to have high dimensionality, e.g. a dataset that contains measurements of temperature at various locations and repeated over an extended period of time.

Motivating challenges

1. Scalability - If data mining algorithms are to handle massive data sets, they must be scalable.
2. High dimensionality - data may now contain hundreds or thousands of attributes
   - e.g. In bioinformatics, gene expression data often involves thousands of features.
   - e.g. Healthcare datasets can involve a large number of features, such as blood pressure, resting heart rate, immune system status, surgery history, height, weight, existing conditions etc.
   - Data with temporal and/or spatial components also tend to have high dimensionality, e.g. a dataset that contains measurements of temperature at various locations and repeated over an extended period of time.
   - Traditional data analysis techniques that were developed for low-dimensional data often do not work well for high dimensional data. This is often referred to as the *curse of dimensionality*

Curse of dimensionality

As a simple example, consider the scenario where you want to create a model to predict the location of a large bacteria in a $25cm^2$ petri dish.

Curse of dimensionality

As a simple example, consider the scenario where you want to create a model to predict the location of a large bacteria in a $25cm^2$ petri dish. For two dimensions, our model may work well at pinning the bacteria down to the nearest square cm. However, if instead of a two-dimensional petri dish we use a three-dimensional beaker, then the predictive space increase from $25cm^2$ to $125cm^3$.

Statistical curse of dimensionality: A required sample size n will grow exponentially with data that has d dimensions, i.e. adding more dimensions generally means that the sample size we need quickly becomes unmanageable.

Data mining tasks

Roughly, we can divide data mining into two major categories:

1. Predictive tasks
   The objective here is to predict the value of a particular attribute based on the values of other attributes.
   - The *target/dependent variable* is what we call the attribute to be predicted.
   - The *explanatory or independent variables* are what we call the attributes used for making the prediction.

Data mining tasks

Roughly, we can divide data mining into two major categories:

1. Predictive tasks
   The objective here is to predict the value of a particular attribute based on the values of other attributes.
   - The *target/dependent variable* is what we call the attribute to be predicted.
   - The *explanatory or independent variables* are what we call the attributes used for making the prediction.
2. Descriptive tasks
   This time, the objective is to derive patterns (correlations, trends, clusters, trajectories and anomalies) that summarise the underlying relationships in the data. These tasks are often exploratory in nature and frequently require postprocessing techniques to validate and explain the results.

Four core data mining tasks

Within these two categories exist four core data mining tasks:

- Predictive modelling
  - Classification
  - Regression
- Association analysis;
- Cluster analysis;
- Anomaly detection.

Goal for this course: Predictive Modelling (mainly Classification) and Cluster Analysis.

Predictive modelling

Idea: Create a model for the target variable as a function of the explanatory variables.

Two main types:

- Classification (used for disccrete target variables)
- Regression (used for continuous target variables)

Predictive modelling: Examples

e.g. Predicting whether a web user will make a purchase at an online bookstore.

Predictive modelling: Examples

e.g. Predicting whether a web user will make a purchase at an online bookstore. This is a classification task since the target variable is binary-valued.

Predictive modelling: Examples

e.g. Predicting whether a web user will make a purchase at an online bookstore. This is a classification task since the target variable is binary-valued.

e.g. Forecasting the future price of a stock.

Predictive modelling: Examples

e.g. Predicting whether a web user will make a purchase at an online bookstore. This is a classification task since the target variable is binary-valued.

e.g. Forecasting the future price of a stock. This is a regression task because price is a continuous valued attribute.

In both cases, the goal is to learn a model that minimises error between the predicted and true values of the target variable.

Predictive modelling: Iris example

Consider the task of predicting a species of flower based on the characteristics of the flower. In particular, we consider classifying an Iris flower as one of the following three species: *Setosa, Versicolour* or *Virginica*.

Predictive modelling: Iris example

Consider the task of predicting a species of flower based on the characteristics of the flower. In particular, we consider classifying an Iris flower as one of the following three species: *Setosa, Versicolour* or *Virginica*.

- Step 1 - we need a data set containing the characteristics of various flowers of these three species.
  We can get such a dataset from the UCI Machine Learning Repository at `https://archive.ics.uci.edu/ml/datasets/iris`.
  - This data set contains four other attributes: sepal width, sepal length, petal length, petal width.

Predictive modelling: Iris example II

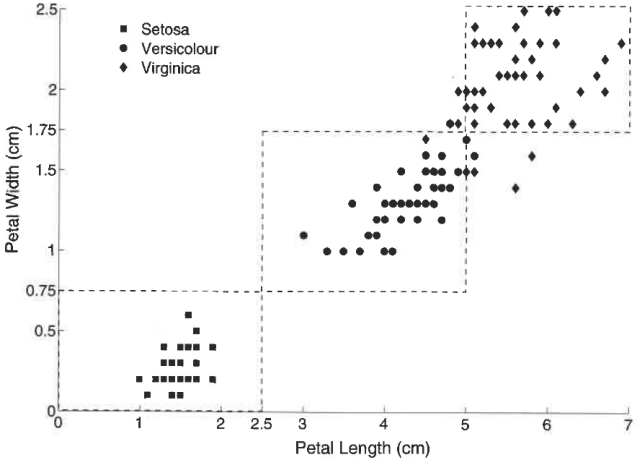
- Step 2 - decide upon categories
  - Petal width is broken into the categories *low, medium* and *high*, which corresponds to the intervals $[0, 0.75)$, $[0.75, 1.75)$, $[1.75, \infty)$, respectively.
  - Likewise, petal length is broken into the categories *low, medium* and *high*, which correspond to the intervals $[0, 2.5)$, $[2.5, 5)$, $[5, \infty)$, respectively.

Predictive modelling: Iris example II

- Step 2 - decide upon categories
  - Petal width is broken into the categories *low, medium* and *high*, which corresponds to the intervals $[0, 0.75)$, $[0.75, 1.75)$, $[1.75, \infty)$, respectively.
  - Likewise, petal length is broken into the categories *low, medium* and *high*, which correspond to the intervals $[0, 2.5)$, $[2.5, 5)$, $[5, \infty)$, respectively.
- Step 3 - Formulate rules
  - Petal width low and petal length low implies Setosa.
  - Petal width medium and petal length medium implies Versicolour.
  - Petal width high and petal length high implies Virginica.

N.B. While these rules do not classify all of the flowers, they do a good (but not perfect) job of classifying most of the flowers.

Petal width versus petal length for 150 Iris flowers



Association analysis

- Used to discover patterns that describe strongly associated features in the data.
- Discovered patterns are typically represented in the form of implication rules or feature subsets.

Association analysis

- Used to discover patterns that describe strongly associated features in the data.
- Discovered patterns are typically represented in the form of implication rules or feature subsets.
- Because of the exponential size of its search space, the goal is to extract the most interesting patterns in an efficient manner.
  - e.g. Find groups of genes that have related functionality.
  - e.g. Identify web pages that are accessed together.
  - e.g. Understand the relationships between different elements in Earth's climate system.

Cluster analysis

- Seeks to find groups of closely related observations so that observations which belong to the same cluster are ‘more similar’ to each other than observations which belong to other clusters.
- Can be used to group sets of related customers, find areas of the ocean that have a significant impact on the Earth’s climate, and compress data.

Cluster analysis: Example

Consider the following collection of news articles:

Collection of News Articles	
Article	Word-frequency pairs
1	dollar: 1, industry: 4, country: 2, loan: 3, deal: 2, government: 2
2	machinery: 2, labour: 3, market: 4, industry: 2, work: 3, country: 1
3	job: 5, inflation: 3, rise: 2, jobless: 2, market: 3, country: 2, index: 3
4	domestic: 3, forecast: 2, gain: 1, market: 2, sale: 3, price: 2
5	patient: 4, symptom: 2, drug: 3, health: 2, clinic: 2, doctor: 2
6	pharmaceutical: 2, company: 3, drug: 2, vaccine: 1, flu: 3
7	death: 2, cancer: 4, drug: 3, public: 4, health: 3, director: 2
8	medical: 2, cost: 3, increase: 2, patient: 2, health: 3, care: 1

Anomaly detection

- Identify observations whose characteristics are significantly different from the rest of the data. Such observations are known as anomalies or outliers.

Goal: Discover the real anomalies and avoid falsely labelling normal objects as anomalous. In other words, a good anomaly detector must have a high detection rate and a low false alarm rate.

Anomaly detection: Applications

Applications of anomaly detection include the detection of fraud, network intrusions, unusual patterns of disease and ecosystem disturbances, such as droughts, floods, fires, hurricanes, etc.

[2] 'Using persistent homology as preprocessing of early warning signals for critical transition in flood', Scientific Reports, 11, 2021

Anomaly detection: Applications

Applications of anomaly detection include the detection of fraud, network intrusions, unusual patterns of disease and ecosystem disturbances, such as droughts, floods, fires, hurricanes, etc.

Often, practices such as preprocessing can impact the effectiveness of an anomaly detection algorithm. For example, in a 2021 paper by Musa et al.[2] the authors developed early warning systems for flooding and tested these at the Guillemard Bridge station at the Kelantan River in Malaysia. By performing a preprocessing step using a novel use of algebraic topology (called persistent homology), the number of false alarms reduced from six to four. This was a reduction in rate from 33.33% to 25%.

[2] 'Using persistent homology as preprocessing of early warning signals for critical transition in flood', Scientific Reports, 11, 2021

(Rough) Roadmap for course

Week	Lecture (1-hour)	Lab (2-hour)
1	Introduction and Overview	
2	Data Storage & Databases	SQL
3	Data Storage & Databases	SQL
4	Types of Data & Data Cleaning	Importing Data and Data Exploration
5	Data Preprocessing	Data Preprocessing
6	Measures of Similarity and Dissimilarity	Data Wrangling
7	Classification	Regression: Lab
8	Classification	Classification: Lab
9	Classification & Cluster Analysis	Classification: Lab
10	Cluster Analysis	Cluster Analysis: Lab
11	Cluster Analysis	Cluster Analysis: Lab
12	Submission of report	

Assessment

There will be three pieces of assessment:

1. ‘Weekly’ Quiz (combined worth 20%).

Assessment

There will be three pieces of assessment:

1. ‘Weekly’ Quiz (combined worth 20%).
2. SQL and Database Assessment (worth 25%).

Assessment

There will be three pieces of assessment:

1. ‘Weekly’ Quiz (combined worth 20%).
2. SQL and Database Assessment (worth 25%).
3. Report (worth 55%).
   - 5% will be for a short Q & A sheet.
   - The remaining 50% will be for the report.

The report

- You will be split into study groups of 5.
- Each study group will be assigned a topic.
- Your task is to learn this topic and apply this (along with anything/everything else we do in class) to a dataset of your choice (from a selection of datasets).
- You then submit a 2page report (not including code) explaining what you have done and your results.

The report

- You will be split into study groups of 5.
- Each study group will be assigned a topic.
- Your task is to learn this topic and apply this (along with anything/everything else we do in class) to a dataset of your choice (from a selection of datasets).
- You then submit a 2page report (not including code) explaining what you have done and your results.
- People in the same study group do not have to work on the same dataset. The study group is simply there to help you throughout the semester.

Reading list: Data mining

The main lectures will follow the book “Introduction to Data Mining” by Pang-Ning Tan, Michael Steinbach, Anuj Karpatne and Vipin Kumar.
Some other books which may be of use are as follows:

- *Data Mining: Introductory and Advanced Topics*, by M.H. Dunham (Pearson)
- *Data Mining: Concepts and Techniques*, Third Edition, by J. Han, M. Kamber, J. Pei (MK Publishers)
- *Principles of Data Mining*, by D. Hand, H. Mannila, P. Smyth (MIT Press)

Reading list: Python

The main python sessions will follow the book “Python for Data Analysis: Data Wrangling with Pandas, Numpy and iPython” by Wes McKinney.
Some other books which may be of use are as follows:

- *Python Data Science Handbook*, by Jake VanderPlas (O’Reilly)
- *Python Cookbook*, Third Edition, by David Beazley and Brian K. Jones (O’Reilly)
- *Fluent Python*, by Luciano Ramalho (O’Reilly)
- *Effective Python*, by Brett Slatkin (Pearson)
- *Python Crash Course*, Second Edition, by Eric Matthes (No Starch Press)