IOT607U Data Mining

Week 1: Introduction

Dr Lin Wang

School of EECS, Queen Mary University of London

Table of contents

1. Module Overview
2. Why data mining?
3. What is data mining?
4. Data mining tasks

Module Overview

Teaching team

Dr Lin Wang (Module organizer)

Main research interests:
Audio and visual signal processing, machine learning, robotic perception

Demonstrators

Yazhou Li

Benjamin James Hayes

Structure

- One lecture per week (2 hours)

  Time: Friday 10:00 - 12:00

  Delivery mode: on Campus

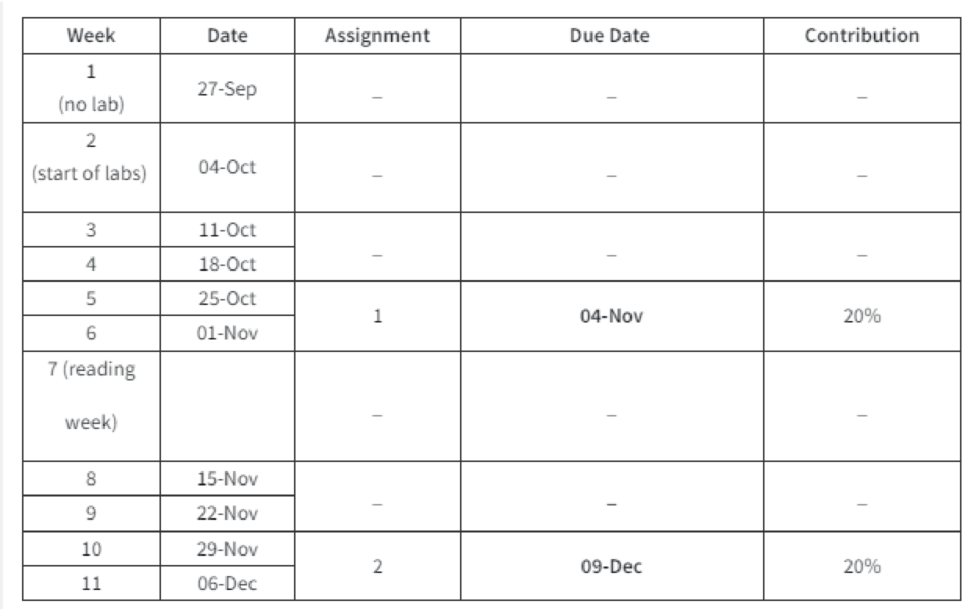
- One lab per week (2 hours)

  Time: Friday 14:00 - 16:00 (w2-w6, w8-w11)

  Delivery mode: on Campus

  Exercise and assignment

Lab & Assessment Schedule



Week	Date	Assignment	Due Date	Contribution
1 (no lab)	27-Sep	–	–	–
2 (start of labs)	04-Oct	–	–	–
3	11-Oct	–	–	–
4	18-Oct			
5	25-Oct	1	**04-Nov**	20%
6	01-Nov			
7 (reading week)		–	–	–
8	15-Nov	–	–	–
9	22-Nov			
10	29-Nov	2	**09-Dec**	20%
11	06-Dec			

Assessment and labs

Assessment:

- Final exam: 60%
- 2 assignments: 40% (20% each assignment)

  assignment 1: 20%, due by week 6; report and code

  assignment 2: 20%, due by week 12; report and code

Programming language:

- Lab exercises implemented in Python;
- "Colab" will be used, which is a free cloud service from Google, hosting Jupyter notebooks with free access to hardware acceleration tools and resources

- **PLAGIARISM: zero-tolerance, non-reversable**

Communication

- In the lecture and lab sessions

- Student Forum on QM+: primary means, questions might have been answered already and answers might be useful to others

  General module specific questions (content, labs, logistics etc) should be posted to the student forum on QM+

- Email: You should email if you want to discuss any personal issue(s).
  - Use your QMUL email account to email.
  - The subject line should start with the string: ECS607U.
  - Include your student id number in the body of the email.

Learning Outcomes

On completing this module, you will be able to:

1. Select an appropriate data representation for a given problem;
2. Apply appropriate data pre-processing and data cleaning methods for both numerical and categorical data;
3. Use data summarisation and data visualisation methods to obtain insights on a given dataset;
4. Explain the distinctions between data mining tasks (classification, clustering, association rules, outlier detection) and select the appropriate method to solve a specific problem;
5. Use appropriate performance metrics and validation techniques and explain the results;
6. Solve practical data mining problems using python and common data mining packages in python;
7. Understand the specific issues relating to ethics in data mining.

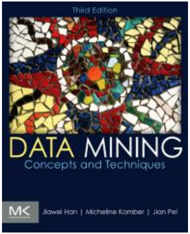
Module Contents

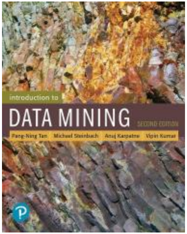
- Introduction to data mining (weeks 1)
- Data (week 2)
- Data exploration (week3)
- Data preprocessing (week 4)
- Classification (week 5-6)

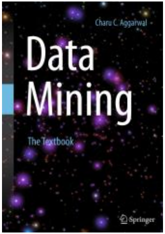
- Clustering (week 8)
- Association analysis (week 9)
- Outlier detection (week 10)
- Data mining applications (weeks 11)
- Data warehouse, Data ethics (week 12)

Reading

- Material uploaded onto QM+
- J.Han, M. Kamber, J.Pei, “Data Mining: Concepts and Techniques”, 3rd edition, Elsevier/Morgan Kaufmann, 2012
- P.-N. Tan, M. Steinbach, A. Karpatne, V. Kumar, “Introduction to Data Mining”, 2nd edition, Pearson, 2019
- C. C. Aggarwal, “Data Mining: The Textbook”, Springer, 2015



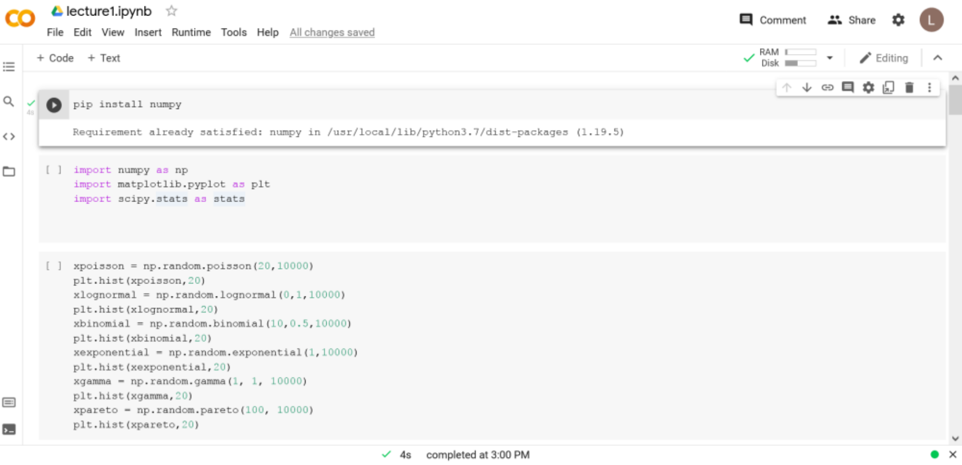




PROGRAMMING IN THE MODULE

- We will be using Python 3!
- And a number of packages like numpy and matplotlib.

Google Colab



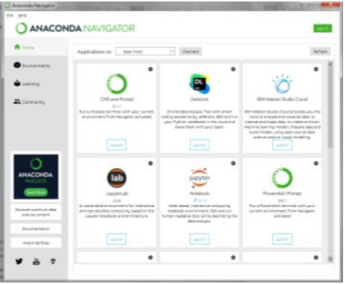
```
pip install numpy
```

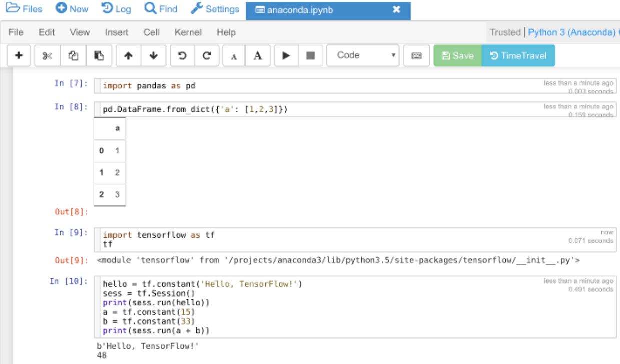
```
Requirement already satisfied: numpy in /usr/local/lib/python3.7/dist-packages (1.19.5)
```

```
[ ] import numpy as np
    import matplotlib.pyplot as plt
    import scipy.stats as stats
```

```
[ ] xpoisson = np.random.poisson(20,10000)
    plt.hist(xpoisson,20)
    xlognormal = np.random.lognormal(0,1,10000)
    plt.hist(xlognormal,20)
    xbinomial = np.random.binomial(10,0.5,10000)
    plt.hist(xbinomial,20)
    xexponential = np.random.exponential(1,10000)
    plt.hist(xexponential,20)
    xgamma = np.random.gamma(1, 1, 10000)
    plt.hist(xgamma,20)
    xpareto = np.random.pareto(100, 10000)
    plt.hist(xpareto,20)
```

Anaconda and Jupter notebook





```
In [7]:  import pandas as pd
```

less than a minute ago
0.003 seconds

```
In [8]:  pd.DataFrame.from_dict({'a': [1,2,3]})
```

less than a minute ago
0.159 seconds

Out[8]:

	a
0	1
1	2
2	3

```
In [9]:  import tensorflow as tf
         tf
```

now
0.071 seconds

```
Out[9]: <module 'tensorflow' from '/projects/anaconda3/lib/python3.5/site-packages/tensorflow/__init__.py'>
```

```
In [10]: hello = tf.constant('Hello, TensorFlow!')
         sess = tf.Session()
         print(sess.run(hello))
         a = tf.constant(15)
         b = tf.constant(33)
         print(sess.run(a + b))
```

less than a minute ago
0.491 seconds

```
b'Hello, TensorFlow!'
48
```

INSTALLING PYTHON PACKAGES

- Using pip to install Python packages.
  - **pip install numpy**
  - **import numpy as np**

- To install the **latest version** of SomeProject:
  - pip install SomeProject

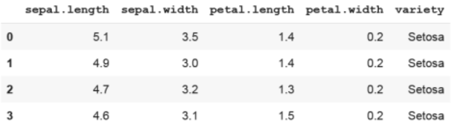
- To install a **specific version**:
  - pip install SomeProject==1.4

- To install on a shared computer, i.e. **only for your user** account:
  - pip install --user SomeProject

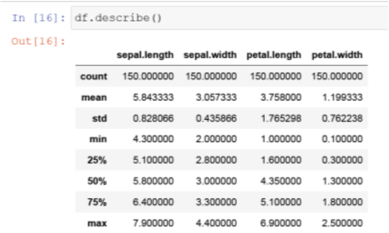
- To **upgrade** a previous installed package:
  - pip install --upgrade SomeProject

Example: load data with pandas

```python
import pandas as pd
df = pd.read_csv('http://www.eecs.qmul.ac.uk/~linwang/download/ecs764/iris.csv')
df.head(4)
```



	sepal.length	sepal.width	petal.length	petal.width	variety
0	5.1	3.5	1.4	0.2	Setosa
1	4.9	3.0	1.4	0.2	Setosa
2	4.7	3.2	1.3	0.2	Setosa
3	4.6	3.1	1.5	0.2	Setosa



In [16]: df.describe()

Out[16]:

	sepal.length	sepal.width	petal.length	petal.width
count	150.000000	150.000000	150.000000	150.000000
mean	5.843333	3.057333	3.758000	1.199333
std	0.828066	0.435866	1.765298	0.762238
min	4.300000	2.000000	1.000000	0.100000
25%	5.100000	2.800000	1.600000	0.300000
50%	5.800000	3.000000	4.350000	1.300000
75%	6.400000	3.300000	5.100000	1.800000
max	7.900000	4.400000	6.900000	2.500000

Background Survey

Go back to menti.com and input the code: **2301 5829**

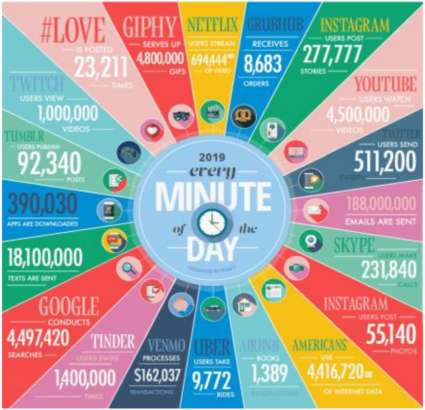
or

go this link:

https://www.menti.com/al7nqkhwg8jk

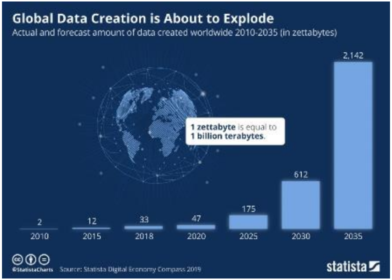
Why data mining?

WHAT IS BIG DATA?



HOW MUCH DATA IS BIG DATA?

- 1 gigabyte (GB)?
- 1 terabyte (TB)?
- 1 petabyte (PB)?
- 1 exabyte (EB)?
- 1 zettabyte (ZB)?
- 1 yottabyte (YB)?



HOW MUCH DATA IS BIG DATA?

- 1 gigabyte (GB)?
- 1 terabyte (TB)?
- 1 petabyte (PB)?
- 1 exabyte (EB)?
- 1 zettabyte (ZB)?
- 1 yottabyte (YB)?

As of 2021, we're normally beginning to talk about big data somewhere in between these.

We may encounter limitations of:
* Memory
* CPU
* Disk space

There is no formal definition though. Widely understood as **data that cannot be handled by a single computer and needs to be distributed across several computers**.

USES OF DATA MINING: RECOMMENDATIONS

- NETFLIX

“We are not recommending a movie because it suits our business needs, but because it matches the information we have from you: your explicit taste preferences and ratings, your viewing history, or even your friends’ recommendations.”

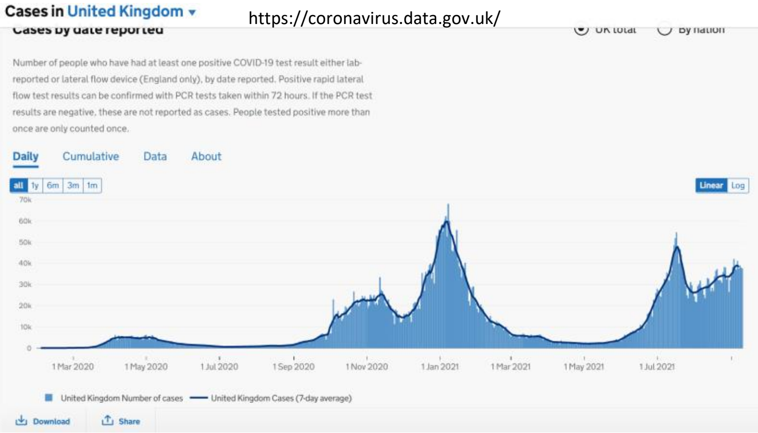
- http://techblog.netflix.com/2012/04/netflix-recommendations-beyond-5-stars.html

USES OF DATA MINING: ADVERTISING

- Forecast search trends and buying patterns
- Create effective ads without trial and error: which words/ideas will sell well with a given audience?

- http://searchengineland.com/putting-big-data-work-building-better-search-ads-191432

USES OF DATA MINING: HEALTH



https://coronavirus.data.gov.uk/

Cases in United Kingdom ▾

Cases by date reported

● UK total ○ By nation

Number of people who have had at least one positive COVID-19 test result either lab-reported or lateral flow device (England only), by date reported. Positive rapid lateral flow test results can be confirmed with PCR tests taken within 72 hours. If the PCR test results are negative, these are not reported as cases. People tested positive more than once are only counted once.

PUBLICLY AVAILABLE DATA

- OECD data, e.g. economics, demographics, agriculture, health
  - http://stats.oecd.org/

- United Nations data, e.g. various development indicators, crime, health, trade
  - http://data.un.org/

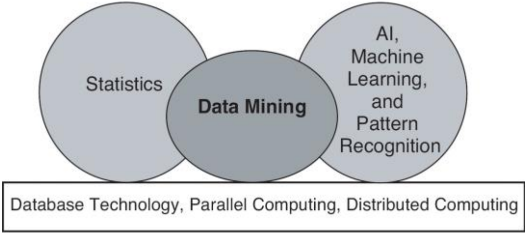
- Gov.uk open data, e.g. government, transportation, education, health
  - https://data.gov.uk/

What is data mining?

Definition of data mining

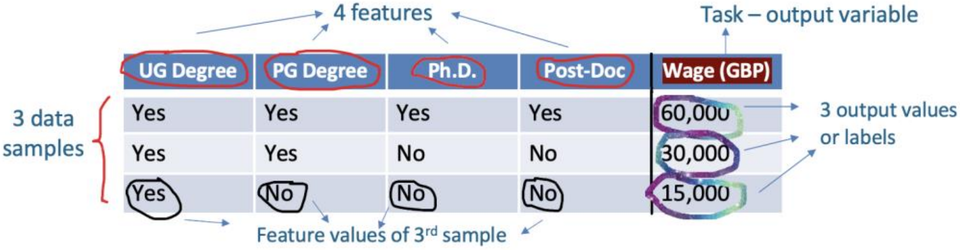
Data mining is the process of extracting and discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems.

- Input: data/database
- Output: knowledge/pattern
- Method: statistics/ML/etc.


Statistics
Data Mining
AI, Machine Learning, and Pattern Recognition
Database Technology, Parallel Computing, Distributed Computing

Data: Example 1

Data refer to characteristics, numerical or categorical, that are collected through observation.



4 features

Task – output variable

UG Degree	PG Degree	Ph.D.	Post-Doc	Wage (GBP)
Yes	Yes	Yes	Yes	60,000
Yes	Yes	No	No	30,000
Yes	No	No	No	15,000

3 data samples

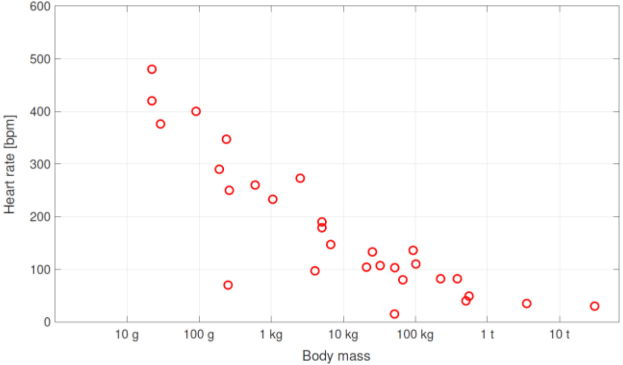
3 output values or labels

Feature values of 3rd sample

Data: Example 2

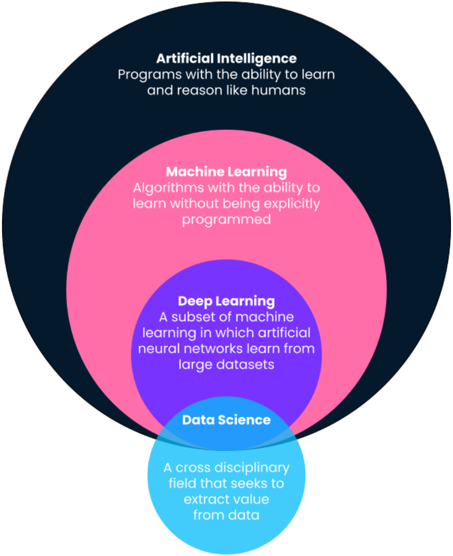
Animal	Body mass [g]	Heart rate [bpm]
Wild mouse	22	480
Rabbit	2.5×10^3	250
Humpback whale	30×10^6	30
...	...	...

Example 2: animal body mass vs. heart rate



Machine Learning

Machine Learning (ML), is a subset of artificial intelligence (AI) that focuses on the development of computer algorithms that improve automatically through experience and by the use of data.



Artificial Intelligence
Programs with the ability to learn and reason like humans

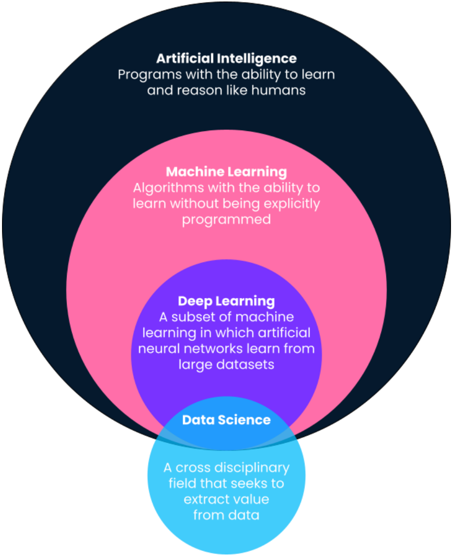
Machine Learning
Algorithms with the ability to learn without being explicitly programmed

Deep Learning
A subset of machine learning in which artificial neural networks learn from large datasets

Data Science
A cross disciplinary field that seeks to extract value from data

Machine Learning

Machine Learning (ML), is a subset of artificial intelligence (AI) that focuses on the development of computer algorithms that improve automatically through experience and by the use of data.



Artificial Intelligence
Programs with the ability to learn and reason like humans

Machine Learning
Algorithms with the ability to learn without being explicitly programmed

Deep Learning
A subset of machine learning in which artificial neural networks learn from large datasets

Data Science
A cross disciplinary field that seeks to extract value from data

Data mining definitions

Data mining is the process of extracting and discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems.

Remark 1

Non-trivial extraction of implicit, previously unknown and potentially useful information from data.

Remark 2

The human activity consisting in extracting knowledge from data.

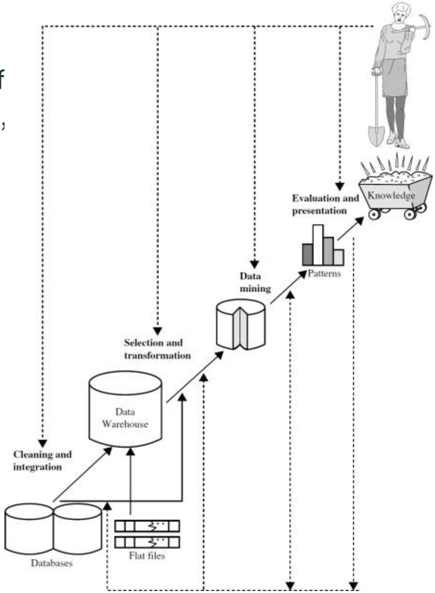
Remark 3

Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns.

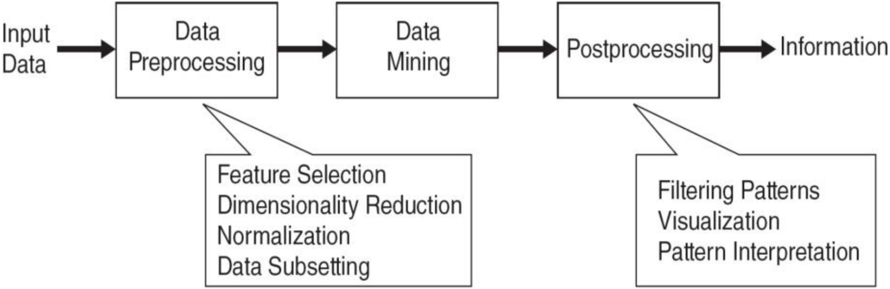
Knowledge discovery from data (KDD)

Many people treat data mining as part of knowledge discovery from data (KDD), which includes the following steps:

1. Data cleaning
2. Data integration
3. Data selection
4. Data transformation
5. Data mining
6. Pattern evaluation
7. Knowledge presentation



Data mining Pipeline



Input Data

Data Preprocessing

Data Mining

Postprocessing

Information

Feature Selection
Dimensionality Reduction
Normalization
Data Subsetting

Filtering Patterns
Visualization
Pattern Interpretation

Go back to menti.com and input the code: **2301 5829**
or

go this link:

https://www.menti.com/al7nqkhwg8jk

Data mining tasks

Types of Machine Learning Models

- Supervised learning
  - Classification
  - Regression
- Unsupervised learning
  - Clustering
  - Dimensionality Reduction
  - Association Analysis

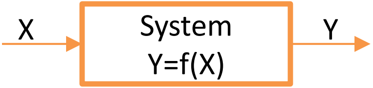
Types of Machine Learning: Supervised learning

In *Supervised Learning* we have:

- input variables (X)

and

- output variables (Y)



We use a model to learn the mapping function, f, from the inputs to the outputs:

$$Y = f(X)$$

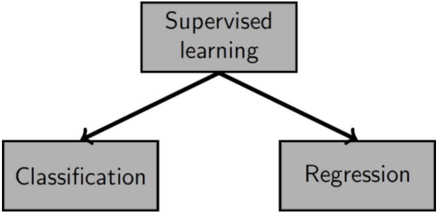
based on input-output pairs, i.e., the *labelled/annotated training dataset*

Supervised learning subcategories

Supervised learning can be split into two subcategories:

1. Classification:
   when the output variable is a category/discrete variable, such as “red” or “blue” ; “disease” or “no disease”.

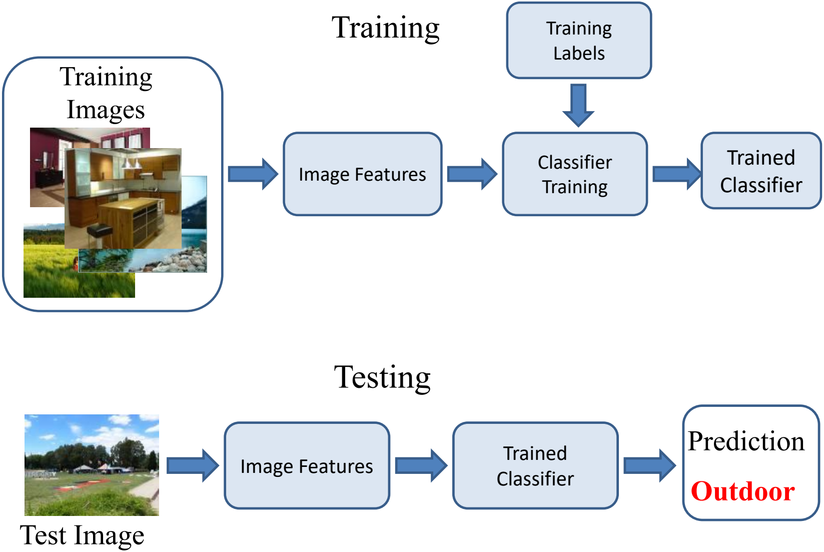
2. Regression:
   when the output variable is a real value/continuous variable, such as “amount of pounds” or “weight”.


Supervised learning

Classification

Regression

Classification example: image categorization



Training

Training Images

Image Features

Training Labels

Classifier Training

Trained Classifier

Testing

Test Image

Image Features

Trained Classifier

Prediction

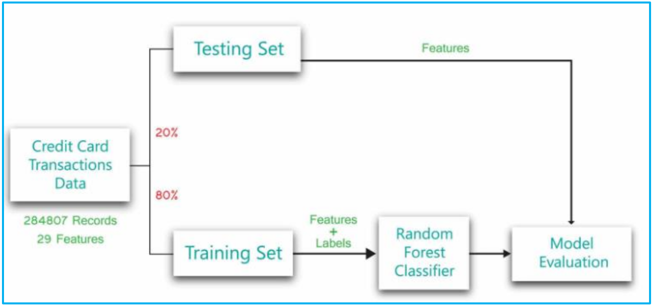
Outdoor

Classification example: fraud detection

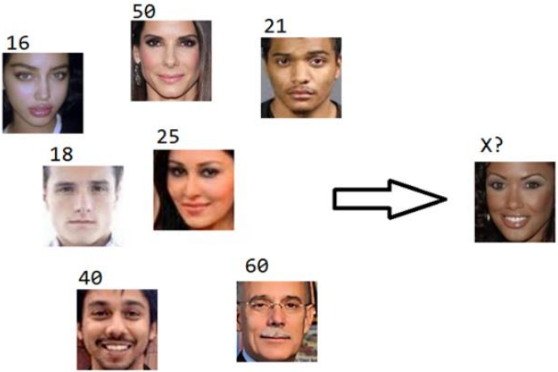
Goal: Predict fraudulent cases in credit card transactions.

Approach:

- Use credit card transactions and the information on its account-holder as attributes-features (e.g. when and what a customer buy).
- Label past transactions as fraud or fair transactions.
- Learn a model for the class of the transactions.
- Use this model to detect fraud by observing credit card transactions on an account.



Regression example: age estimation



Types of Machine Learning: Unsupervised learning

In *Unsupervised Learning* we only have input data (X) and neither corresponding output variables nor labels.

The goal of unsupervised learning is to model the underlying structure or distribution in the data in order to search for interesting/useful characteristics in the data, e.g.,

- find groups of samples that exhibit similarity in some sense
- find subset(s) of features that behave similarly
- find combinations of features with the greatest variation

Unsupervised learning subcategories

Unsupervised learning problems can be further grouped into:

- Clustering:
We use clustering algorithms to discover the inherent groupings in the data, such as grouping customers by purchasing behaviour.

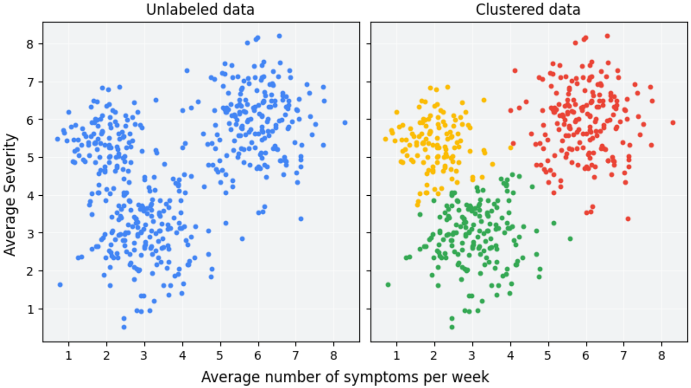
- Dimensionality Reduction:
We use dimensionality reduction algorithms when the number of input variables becomes very/quite large.

- Association:
We use association rule learning so as to discover rules that describe large portions of data, such as people that buy X also tend to buy Y.

Unsupervised learning: Clustering

Clustering

Finding groups of objects such that the objects in a group will be similar to one another and different from the objects in other groups.

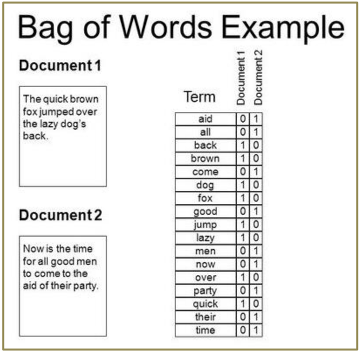


Clustering Application

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 the documents.



Bag of Words Example

Document 1

The quick brown fox jumped over the lazy dog's back.

Document 2

Now is the time for all good men to come to the aid of their party.

Term	Document 1	Document 2
aid	0	1
all	0	1
back	1	0
brown	1	0
come	0	1
dog	1	0
fox	1	0
good	0	1
jump	1	0
lazy	1	0
men	0	1
now	0	1
over	1	0
party	0	1
quick	1	0
their	0	1
time	0	1

Unsupervised learning: Association Analysis

Given a set of records each of which contain some number of items from a given collection:

- Produce dependency rules which will predict occurrence of an item based on occurrences of other items.

TID	Items
1	Bread, eggs, milk
2	Juice, bread
3	Juice, eggs, butter, milk
4	Juice, bread, butter, milk
5	Eggs, butter, milk

Rules discovered:
$\{\text{Milk}\} \Rightarrow \{\text{Eggs}\}$
$\{\text{butter,milk}\} \Rightarrow \{\text{Juice}\}$

Association Analysis Applications

Market-basket analysis:
Rules are used for sales promotion, shelf management, and inventory management.

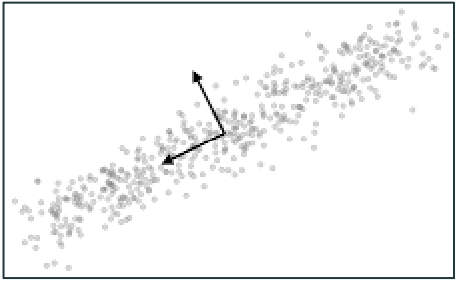
Medical Informatics:
Rules are used to find combination of patient symptoms and test results associated with certain diseases.

Unsupervised learning: Dimensionality reduction

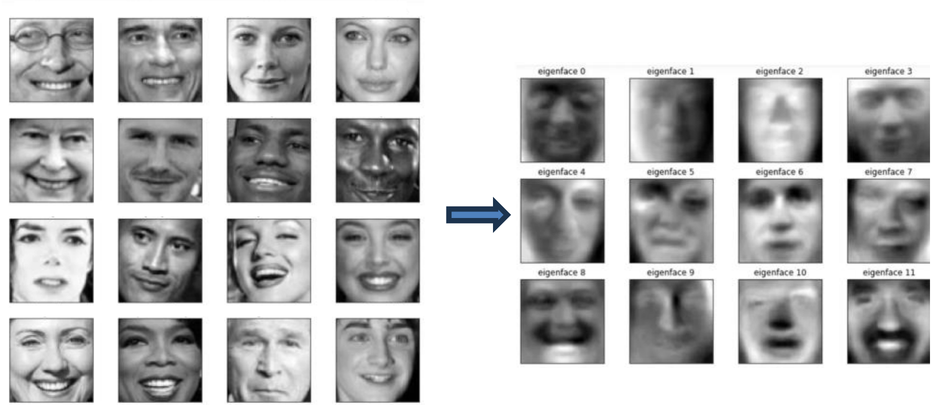
Dimensionality reduction is an unsupervised learning technique that reduces the number of features, or dimensions, in a dataset.

- Data compression
- Data visualization
- Feature selection

Principal component analysis (PCA)



PCA Applications: eigenface for face recognition



Outlier Analysis / Anomaly Detection

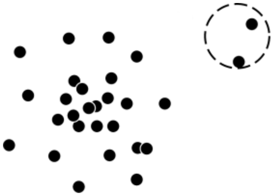
A dataset may contain objects that do not comply with the general behavior or model of the data. These data objects are outliers.

The analysis of outlier data is referred to as outlier analysis or anomaly detection.

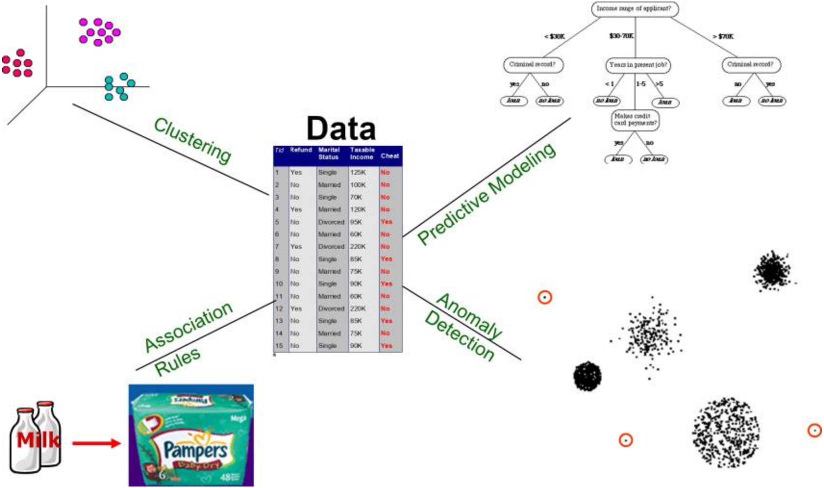
Can use both supervised (e.g. classification) and unsupervised (e.g. clustering) approaches

Applications:

- Credit card fraud detection
- Network intrusion detection
- Monitoring and surveillance in sensor networks
- Detecting changes in the global forest cover



Data mining tasks



Data

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
11	No	Married	60K	No
12	Yes	Divorced	220K	No
13	No	Single	85K	Yes
14	No	Married	75K	No
15	No	Single	90K	Yes

Clustering

Predictive Modeling

Association Rules

Anomaly Detection

Data mining tasks

In general, data mining tasks can be classified into two categories: descriptive and predictive.

Descriptive mining tasks characterise properties of the data in a target data set.

Predictive mining tasks perform induction on the current data in order to make predictions.

Go back to menti.com and input the code: **2301 5829**
or

go this link:

https://www.menti.com/al7nqkhwg8jk

Challenges in data mining

Challenges in data mining

Mining methodology

- Researchers have been vigorously developing new data mining methodologies.
- Current topics: investigation of new kinds of knowledge, mining in multidimensional space, integrating methods from other disciplines...
- Mining methodologies should consider issues such as data uncertainty, noise, and incompleteness.

User Interaction-Human in the Loop

- How to interact with a data mining system
- How to incorporate a user's background knowledge in mining
- How to visualize and comprehend data mining results

Challenges in data mining

Efficiency and Scalability

- Algorithms must be efficient and scalable in order to effectively extract information from huge amounts of data.
- The wide distribution of data, and the computational complexity of some data mining methods motivate the development of parallel and distributed algorithms.

Diversity of Database Types

- Handling complex types of data
- Mining dynamic, networked, and global data repositories

Challenges in data mining

Data Mining and Society

- Social impacts of data mining: How can we use data mining technology to benefit society? How can we guard against its misuse?
- Privacy-preserving data mining
- Invisible data mining

Data Ethics

An emerging branch of applied ethics which describes the value judgments and approaches we make when generating, analysing and disseminating data.

Questions?

also please use the forum on QM+