Introduction to Statistical Machine Learning

Cheng Soon Ong & Christian Walder

Machine Learning Research Group
Data61 | CSIRO
and
College of Engineering and Computer Science
The Australian National University

Canberra
February – June 2019

(Many figures from C. M. Bishop, "Pattern Recognition and Machine Learning")

Outlines

Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

Part I

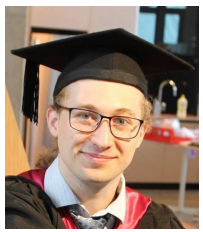
Overview

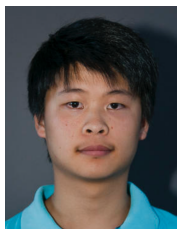
Lecturers and Tutors

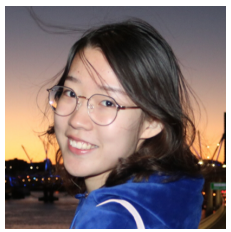
Lecturers

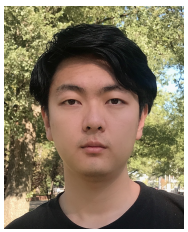
Cheng Soon Ong Christian Walder

Tutors









Chamin Hewa Koneputugodage, David Quarel, Guyver Fu, Mengyan Zhang, Tianyu Wang

Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

Course information

Lectures
- Tuesday, 12.00-13.30, Haydon Allen Tank
- Wednesday, 13.00-14.30, Science and Teaching Building, S1 (3.07)

Tutorial CSIT (building 108)
- Tutorials start this week, choose one (on Wattle)
- Thursday, 11-13, N111
- Thursday, 17-19, N115/N116
- Friday, 09-11, N111
- Friday, 11-13, N112
- Friday, 16-18, N115/116

Assignment 1 assignment, due 15 May (18%)

Exams
- In lab exam (week 3, 2% hurdle)
- Mid semester exam (week 7, 20%)
- Final Written exam, (60%)

Info `https://machlearn.gitlab.io/sml2019/`

Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

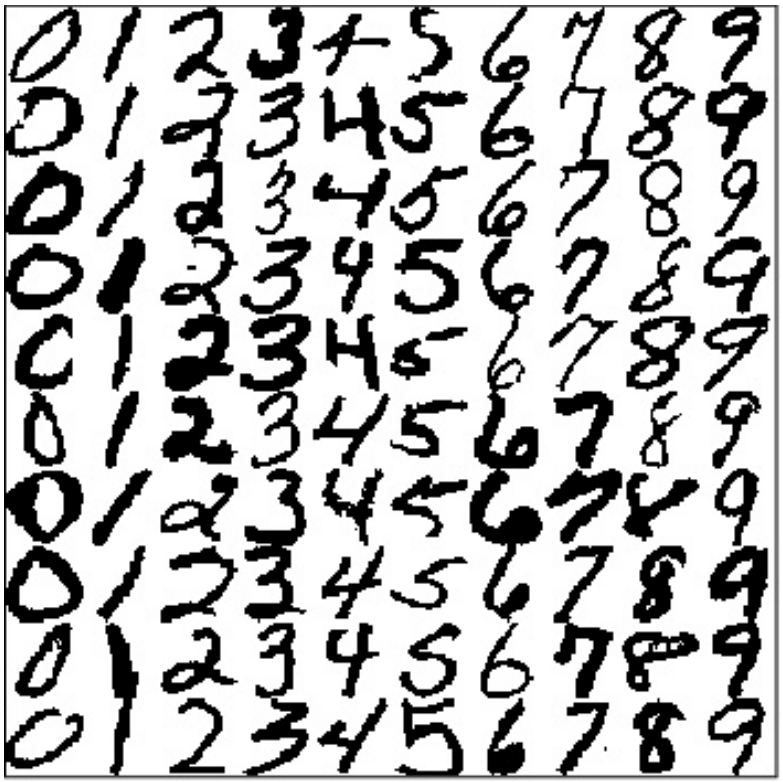
gmail - Priority inbox

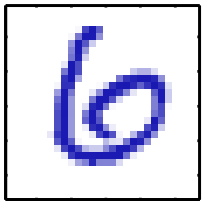
https://www.youtube.com/watch?v=5nt3gE9dGHQ

- Given examples what the user defined as important mail.
- From these examples, learn to identify new important mail.

©2019 Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

Handwritten Digit Recognition



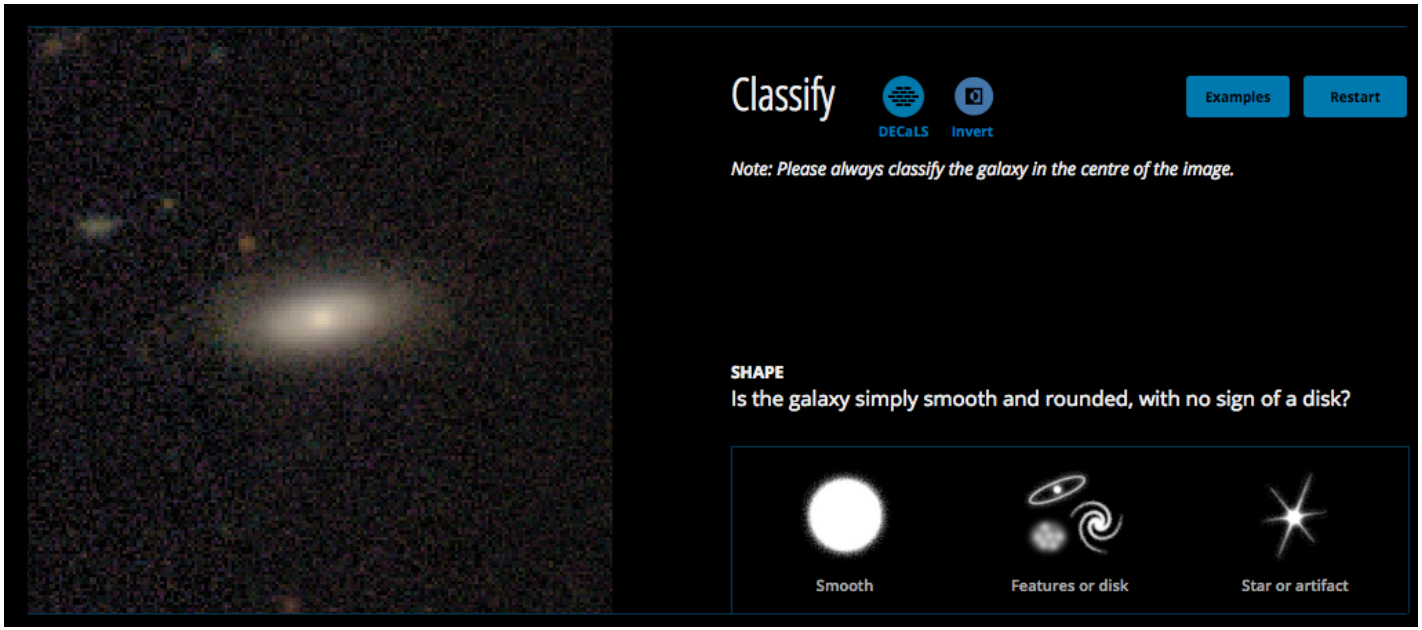


- Given handwritten ZIP codes on letters, money amounts on cheques etc.
- Learn to recognise the correct digit written by hand.

Predicting Solar Panel Output

- Photovoltaics now very close to grid electricity in price
- Distributed system of generators
- Energy market
- Great Machine Learning Problem: Predict the solar energy output (variability primarily due to clouds) for Australia
- Pilot project in Canberra : Use cheap cameras to take 360° sky photos in several location.
- Learn to predict 3-D model of cloud movement.
- Learn orientation and efficiency of solar panels for each house from time series of energy output.
- Predict output of each solar panel for 15 min to 1 hour from current snapshots.

Galaxy classification



http://www.galaxyzoo.org/#/classify

- Given images of the sky from SDSS and CTIO, and crowdsourced labels
- From these examples, learn to identify galaxies.

Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

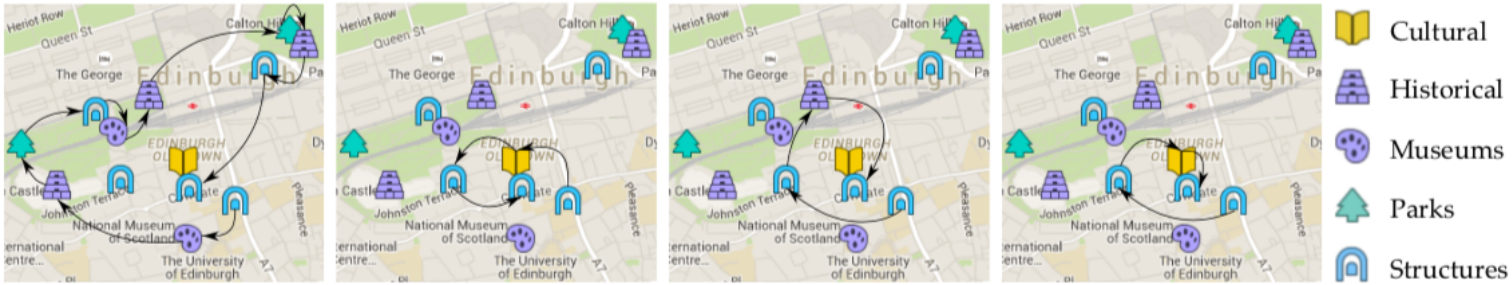
Generating Captions with Sentiments

http://cm.cecs.anu.edu.au/post/senticap/

- Given an image, generate text to describe it
- Additionally tune positive or negative sentiment

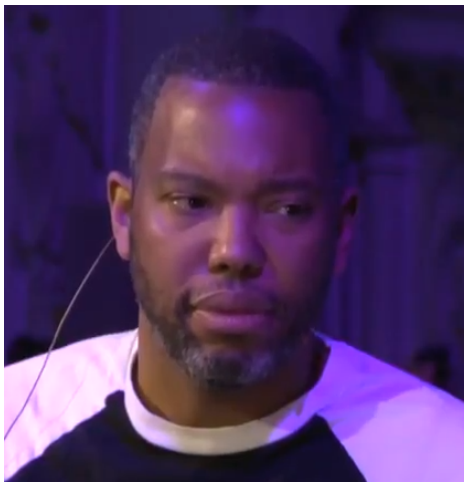
Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

Recommending Tours in a City



- Given start and end points in the city, suggest a tour
- Try to learn the preferences of individual users

It will affect society...



Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

Other applications of Machine Learning

- autonomous robotics,
- bioinformatics,
- detecting network intrusion,
- neuroscience,
- medical diagnosis,
- stock market analysis,
- social network analysis,
- traffic and infrastructure planning,
- ⋮
- See more at `en.wikipedia.org/wiki/Machine_learning#Applications`...

Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

What is common to these examples?

1. Given some data (e.g. hand written digits).
2. Possibly some extra information (e.g. which digit does this number represent)
3. Goal: Build a machine which can learn from the given data utilising the extra information (if available).

Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

Flavour of this course

- Formalise intuitions about problems
- Use language of mathematics to express models
- Geometry, vectors, linear algebra for reasoning
- Probabilistic models to capture uncertainty
- Calculus to identify good parameters
- Design and analysis of algorithms
- Numerical algorithms in python
- Understand the choices when designing machine learning methods

Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

What is Machine Learning?

Definition (First Try)

Machine learning is concerned with the design and development of algorithms that allow machines to learn.

Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

What is Machine Learning?

Definition (First Try)

Machine learning is concerned with the design and development of algorithms that allow machines to learn.

- machines? computers? HAL?
- to learn?

Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

What is Machine Learning?

Definition (First Try)

Machine learning is concerned with the design and development of algorithms that allow machines to learn.

- machines? computers? HAL?
- to learn?
- need to quantify "learning"
- to improve their performance over time

Definition (Second Try)

Machine learning is concerned with the design and development of algorithms that allow computers (machines) to improve their performance over time.

Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

What is Machine Learning?

- What is the source of the improved performance?
- New insights by the algorithm designer?

Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

What is Machine Learning?

- What is the source of the improved performance?
- New insights by the algorithm designer?

Definition (Final Version)

Machine learning is concerned with the design and development of algorithms that allow computers (machines) to improve their performance over time based on data.

Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

What is Machine Learning?

Definition (Mitchell, 1998)

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.

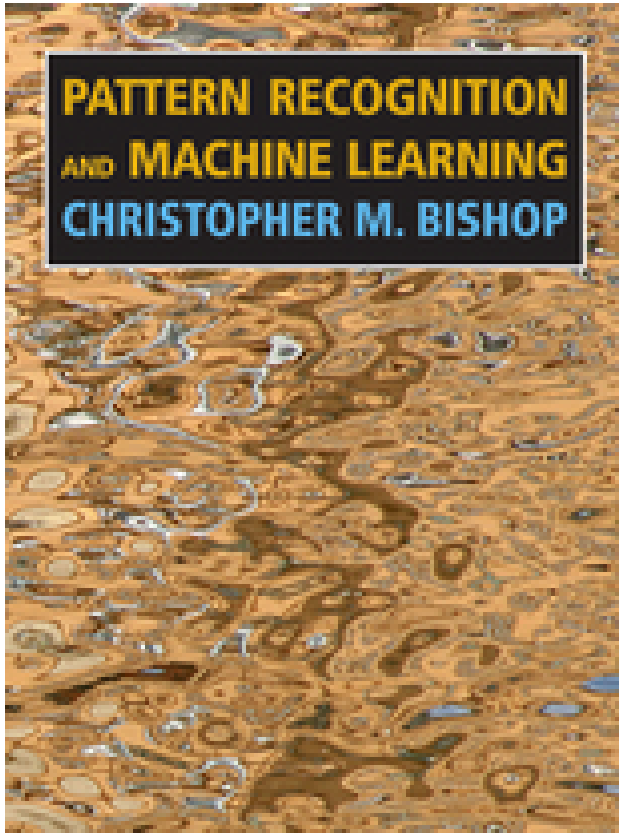
Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

What is the challenge?

- Given only some examples.
- Need to derive a relation for many more (possibly infinite) unseen examples.
- Occam's Razor (also written as Ockham's razor, William Ockham, circa 1285 – 1349): "Plurality must never be posited without necessity"
- "The simplest explanation or strategy tends to be the best one."

Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

Textbook



Christopher Bishop
Pattern Recognition and Machine Learning, 2006
`https://aka.ms/prml`

Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

Statistical Machine Learning - Some History

- 1960's : symbolic AI; computers learn rules from data; analysis of the underlying statistics is seldom done.
- Perceptron (Rosenblatt, 1957), "Perceptrons" (Minsky and Papert, 1969)
-
- 1980's : artificial neural networks
- 1990's - 2000's : statistical machine learning (kernel methods, decision trees, graphical models)
- Why Statistical Machine Learning not earlier?
  - faster computers with larger memory to represent statistical models have become available
  - numerical methods on the desktop computer (BLAS, LAPACK, Optimisation)
  - found new interesting classes of algorithms (e.g. on graphs)
  - large amounts of data available which can be tapped into (flickr, social networks)
  - many data sets with partial/incomplete data (e.g. netflix)

Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

Supervised learning

- Machine learning is about prediction

Examples/features	$x_1, \ldots, x_n \sim \mathcal{X}$
Labels/annotations	$y_1, \ldots, y_n \sim \mathcal{Y}$
Predictor	$f_{\mathbf{w}}(x) : \mathcal{X} \to \mathcal{Y}$

- Estimate best predictor = training = learning
  Given data $(x_1, y_1), \ldots, (x_n, y_n)$, find a predictor $f_{\mathbf{w}}(\cdot)$.
  - No mechanistic model of the phenomenon
  - There is relatively large amounts of data (examples, x usually $\mathbb{R}^d$)
  - The outcomes (labels, y usually binary) are well defined

Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

Strategy in this course

- Estimate best predictor = training = learning
  Given data $(x_1, y_1), \ldots, (x_n, y_n)$, find a predictor $f_{\mathbf{w}}(\cdot)$.
  1. Identify the type of input x and output y data
  2. Propose a (linear) mathematical model for $f_{\mathbf{w}}$
  3. Design an objective function or likelihood
  4. Calculate the optimal parameter ($\mathbf{w}$)
  5. Model uncertainty using the Bayesian approach
  6. Implement and compute (the algorithm in python)
  7. Interpret and diagnose results

Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

Some Basic Notation - Data

- The set of all input data is denoted as $\mathcal{X}$. For instance, $\mathcal{X} = \{ x \mid x \text{ is an image containing a handwritten digit} \}$.
- One data point with D elements :

$$\mathrm{x} = \begin{bmatrix} x_1 \\ \dots \\ x_D \end{bmatrix} = (x_1, \dots, x_D)^T.$$

- Data matrix : A set of N data points $\mathrm{x_i}$, where $i = 1 \dots N$,

$$\mathrm{X} = \begin{bmatrix} x_{1,1} \dots x_{1,D} \\ x_{2,1} \dots x_{2,D} \\ \dots \\ x_{N,1} \dots x_{N,D} \end{bmatrix} = \begin{bmatrix} \mathrm{x}_1^{\mathrm{T}} \\ \dots \\ \mathrm{x}_{\mathrm{N}}^{\mathrm{T}} \end{bmatrix}.$$

  (Note : Each data point $\mathrm{x_i}$ is a column vector, but appears as a row vector in X.)
- If $D = 1$, X is a vector of N scalar data points. We write

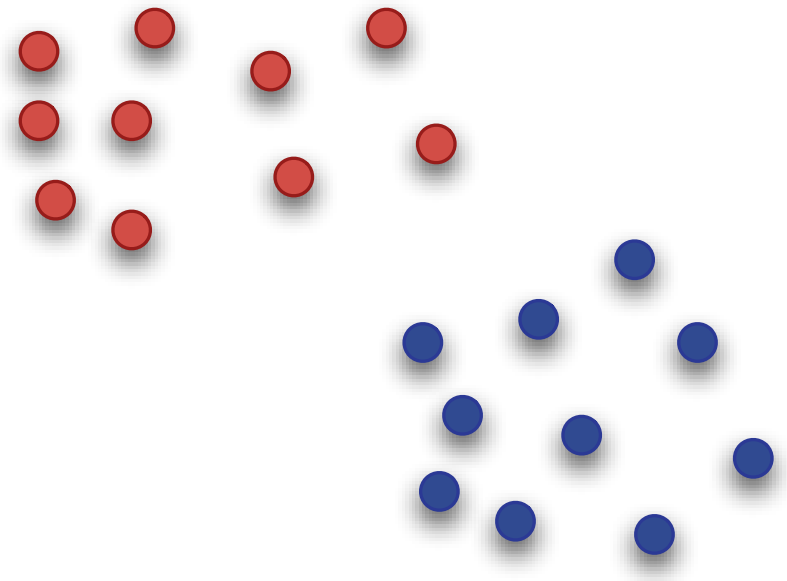
$$\mathbf{x} = \begin{bmatrix} \mathrm{x}_1 \\ \dots \\ \mathrm{x}_{\mathrm{N}} \end{bmatrix}.$$

Some Basic Notation - Targets

- A target can be from a finite discrete set ('labels') or from $\mathbb{R}$.
  (Note: Can extend this idea to m-dimensional labels and $\mathbb{R}^m$.)
- Set of Targets $\mathcal{T}$, e.g.
  $\mathcal{T} = \{\text{one}, \text{two}, \text{three}, \text{four}, \text{five}, \text{six}, \text{seven}, \text{eight}, \text{nine}, \text{zero}\}$.
- An ordered set of N scalar labels $\mathbf{t} = \begin{bmatrix} t_1 \\ \dots \\ t_N \end{bmatrix} = (t_1, \dots, t_N)^T$.

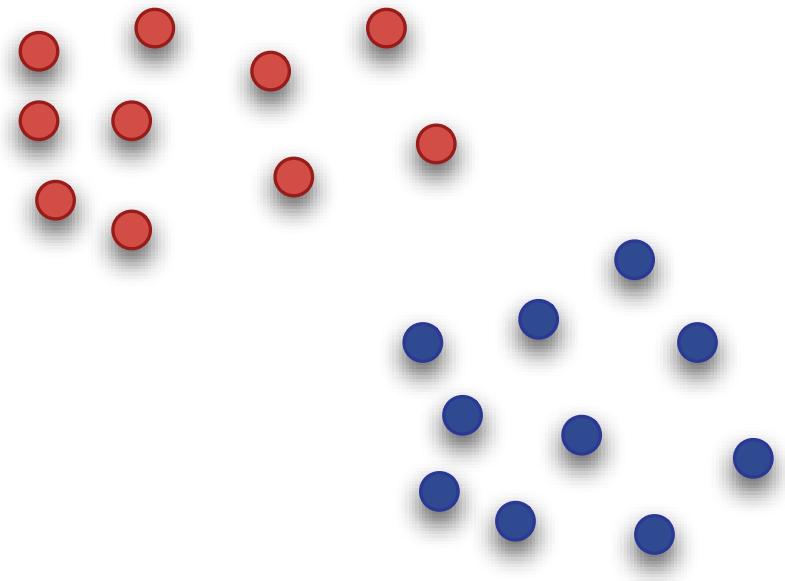
Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

Supervised Learning



Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

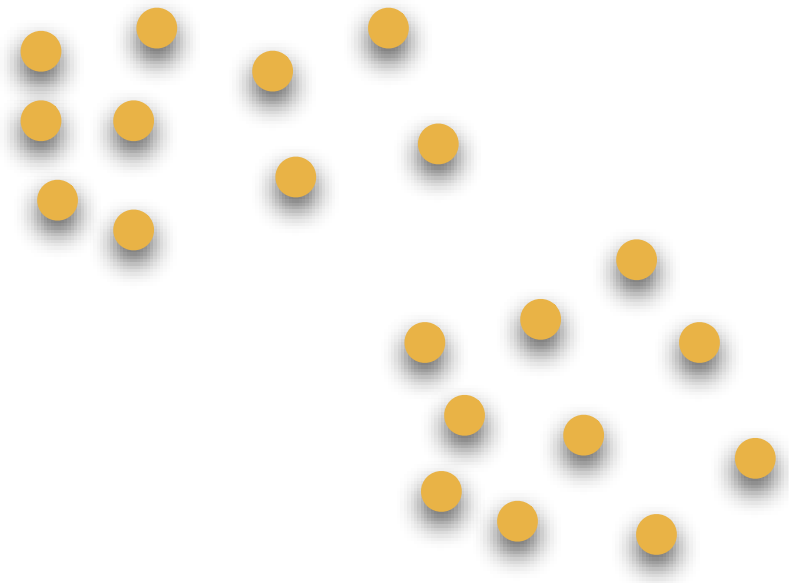
Supervised Learning



- Given are pairs of data $\mathrm{x_i} \in \mathcal{X}$ and targets $t_i \in \mathcal{T}$ in the form $(\mathrm{x_i}, t_i)$, where $i = 1..N$.
- Learn a mapping between the data X and the target $\mathbf{t}$ which generalises well to new data.

Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

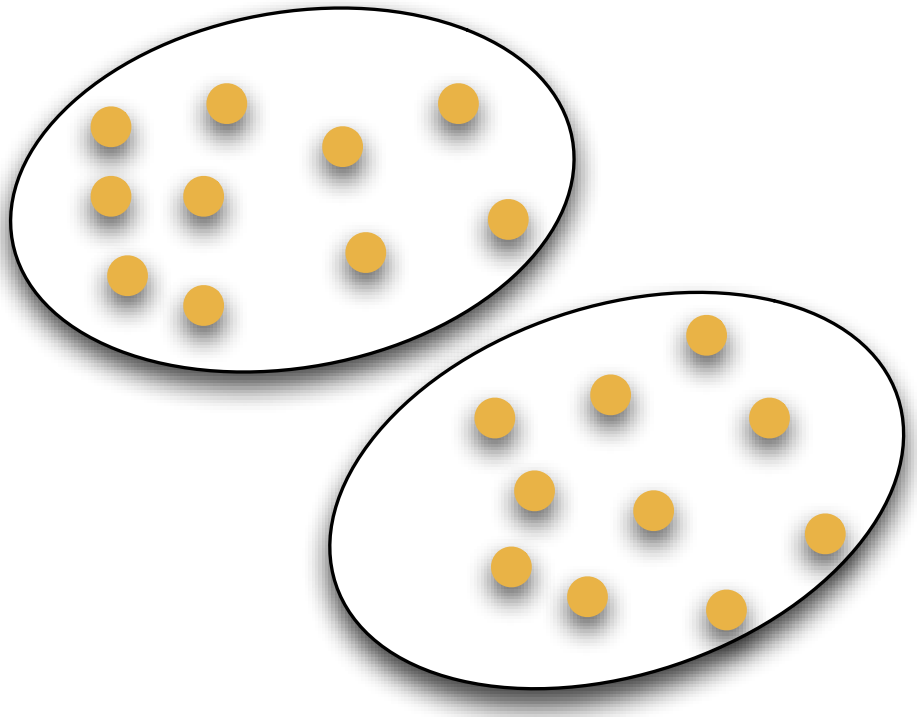
Unsupervised Learning



- Given only the data $\mathrm{x}_i \in \mathcal{X}$.
- Discover (=learn) some interesting structure inherent in the data X.

Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

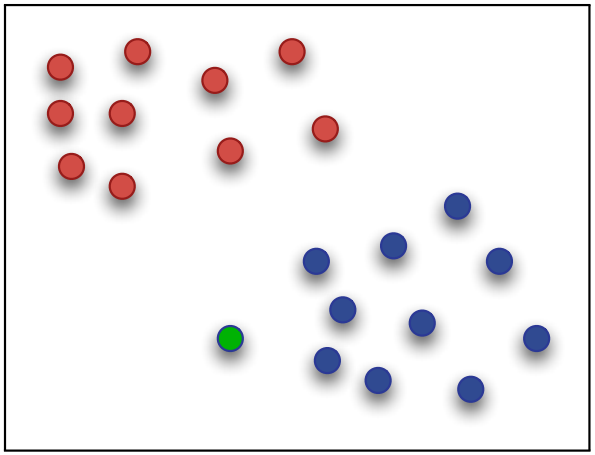
Unsupervised Learning

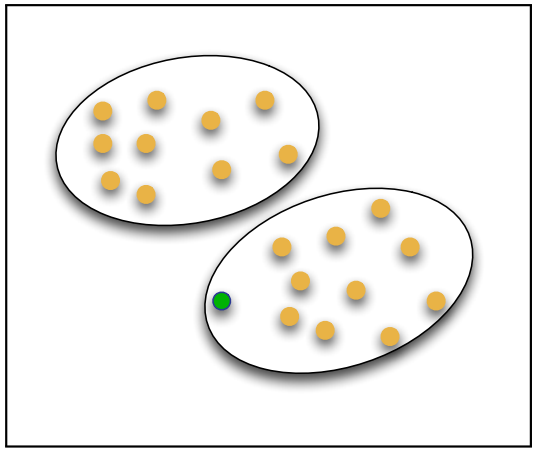


- Given only the data $x_i \in \mathcal{X}$.
- Discover (=learn) some interesting structure inherent in the data.

Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

Testing - Supervised versus Unsupervised Learning





Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

Reinforcement Learning

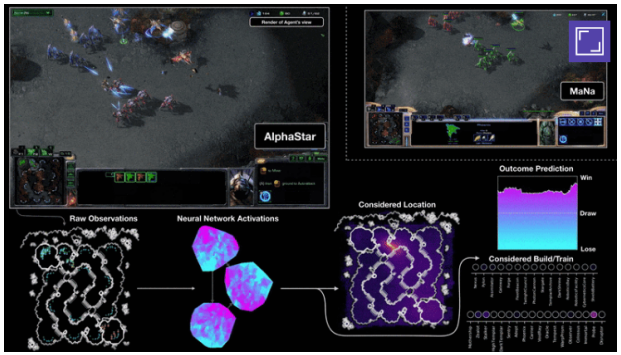
- Example: Game playing. There is one reward at the end of the game (negative or positive).
- Find suitable actions in a given environment with the goal of maximising some reward.
- correct input/output pairs never presented
- Reward might only come after many actions.
- Current action may not only influence the current reward, but future rewards too.

Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

Reinforcement Learning

go

DOTA



Starcraft II

- Exploration versus Exploitation.
- Well suited for problems with a long-term versus short-term reward trade-off.
- Naturally focusing on online performance.

Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

Other Machine Learning Types

- Active Learning
  - The algorithm may choose which data $x_i \in \mathcal{X}$ to select next when building the model.
  - The order of the data is actively chosen by the algorithm at run-time.
- Transduction
  - The algorithms is allowed to use the test data (but of course not labels!) when building a model.
- Estimation with missing variables.
- Co-training with two different but related data sets.
- ... and others.

Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

Training Regimes

- Batch Learning
  - All training data $\mathrm{X} = \{\mathrm{x}_1, \ldots, \mathrm{x}_n\}$ and targets $\mathbf{t} = \{t_1, \ldots, t_n\}$ are given.
  - Learn a mapping from $\mathrm{x_i}$ to t_i which can then be applied to yet unseen data $\mathrm{X}' = \{\mathrm{x}'_1, \ldots, \mathrm{x}'_m\}$ to find $\mathbf{t}' = \{t'_1, \ldots, t'_m\}$.
- Online Processing
  - Pairs of $(\mathrm{x_i}, t_i)$ become available one at a time.
  - At each step, learn and refine a mapping from $\mathrm{x_i}$ to t_i which can then be applied to yet unseen data x'_i.

Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

This semester

Supervised Learning

- Regression
- Classification

Unsupervised Learning

- Dimensionality Reduction
- Graphical Models

Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

Python

- dynamically typed programming language (no declarations for variables)
- supports object oriented, imperative, and functional programming style
- many built-in data types (str, tuple, list, set, dict, . . .)
- packages for scientific programming (numpy, scipy, matplotlib)
- easily extensible to use code written in C and C++ (or FORTRAN for that matter)
- Python runs on Windows, Linux/Unix, Mac OS X, Android, Raspberry Pi
- OSI-approved Open Source License

Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

Jupyter Notebooks

- Literate programming environment
- Markdown cells also understand LaTeX
- Code cells support many languages (we use Python3)
- Tutorials and assignments all use notebooks

Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

Topics to Review

- Linear Algebra
- Analytic Geometry
- Matrix Decomposition
- Vector Calculus
- Probability and Statistics
- Continuous Optimization

https://mml-book.com

Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

On the Way to Learning (in Indonesia)

Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

On the Way to Learning (in China)

Ong & Walder & Webers
Data61 | CSIRO
The Australian National University