A Glimpse of Machine Learning

Cheng Soon Ong

May 2017
COMP3620/6320 Artificial Intelligence .

Context

Static environment

Assume that data is independent and identically distributed

Unknown environment

The distribution generating the data is unknown

Continuous variables

The data is in $\mathbb{R}^d$ and the state space is continuous

Mathematical modeling

Focus is on expressing intuitions as mathematics

- Loss function
- Probabilistic model

Then we can "crank the handle".

What is machine 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

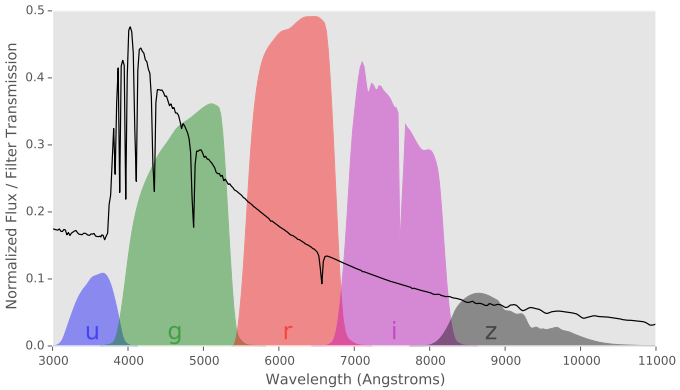
Given data $(x_1, y_1), \ldots, (x_n, y_n)$, find a predictor $f_{\mathbf{w}}(\cdot)$.

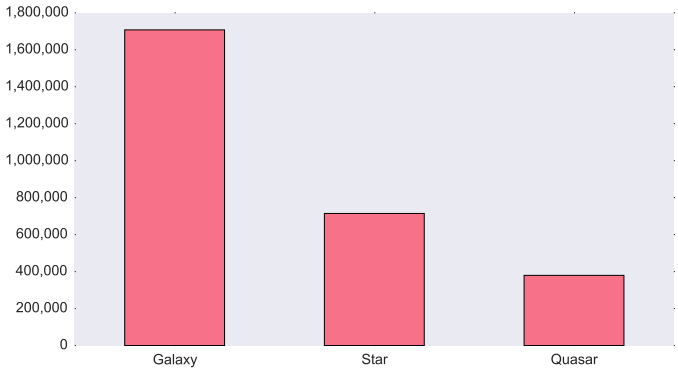
- No mechanistic model of the phenomenon
- There is relatively large amounts of data (examples, x usually $\mathbb{R}^d$)
- The outcomes (labels, y usually binary) are well defined

Prediction $\neq$ understanding

How can we use prediction to help with scientific research?

Ex: Classifying celestial objects





- Spectroscopy is expensive (3M examples), photometry is cheaper (800M examples)
- Task: Classify whether an object is a galaxy, star or quasar from photometric measurements.

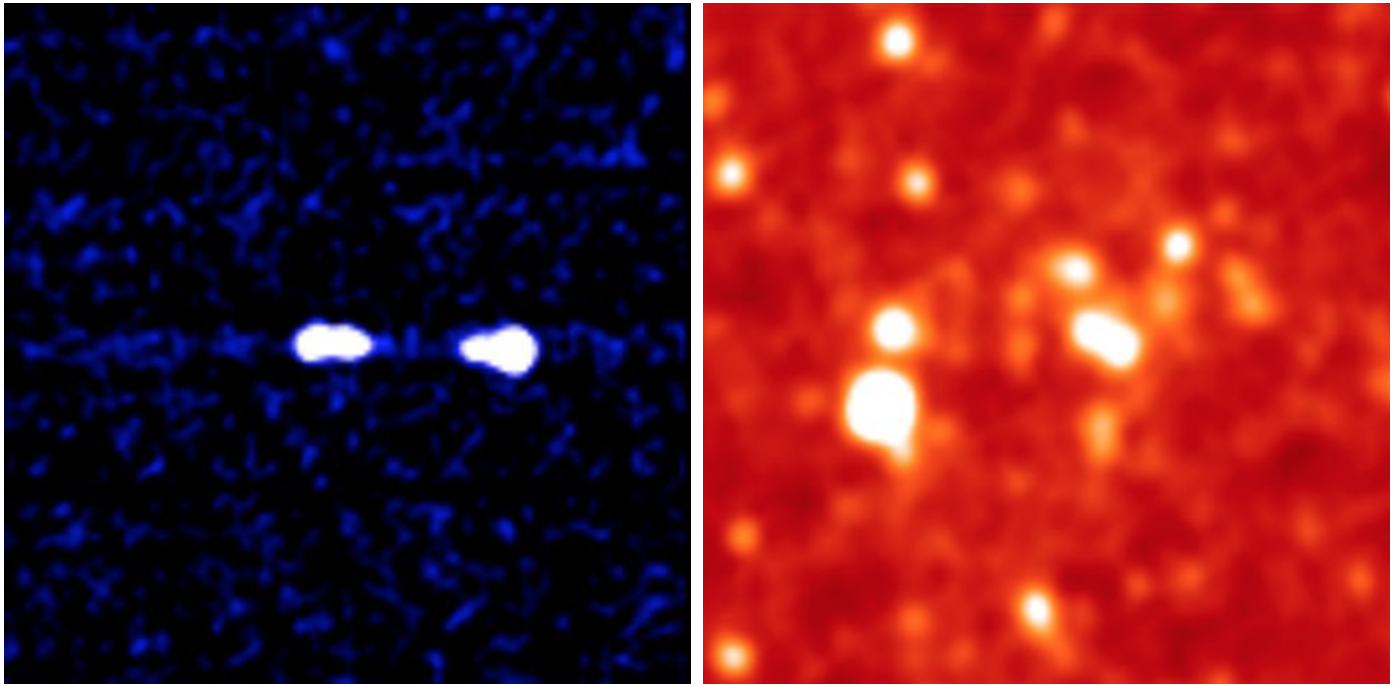
Alasdair Tran, honours at CS, ANU, 2015

Ex: Finding black holes

Optical Infrared X-ray Radio

Images of Centaurus A at different wavelengths.

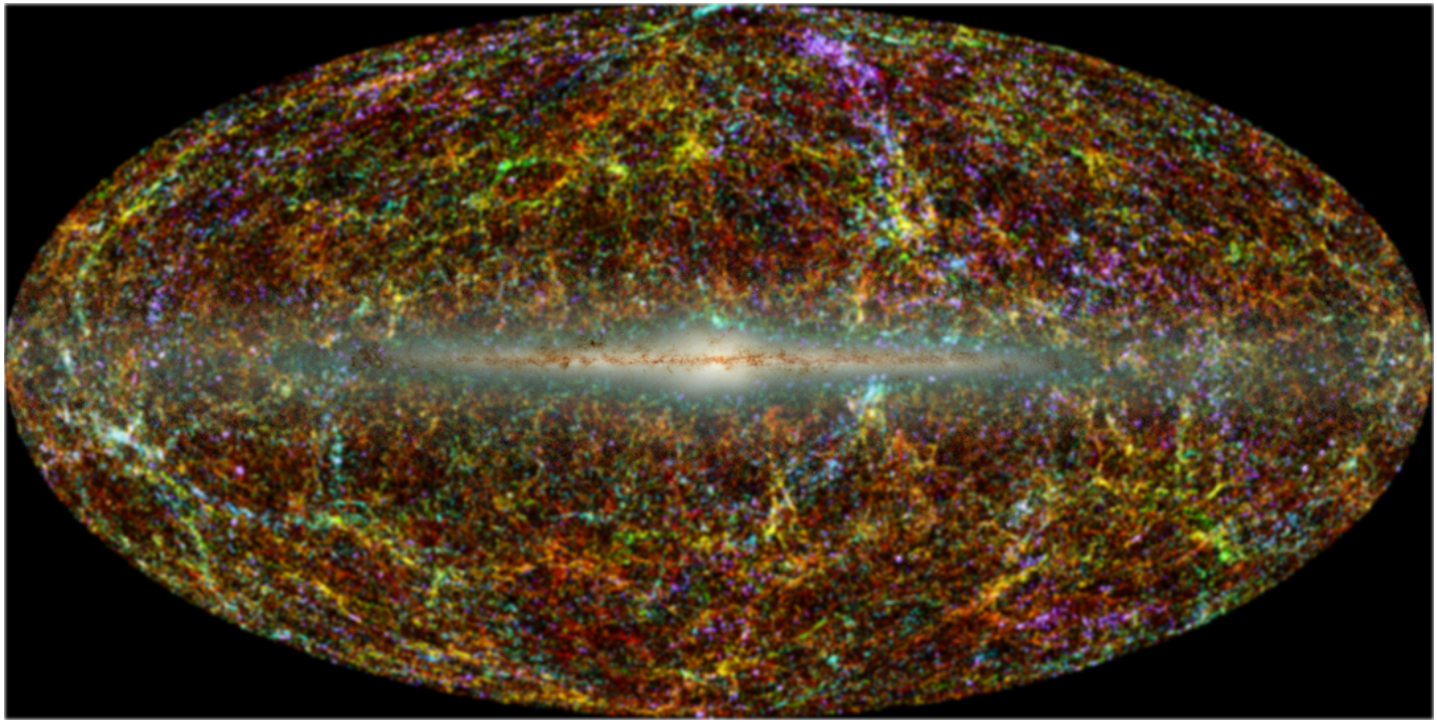
The real data



The same patch of sky in both radio (left) and infrared (right)

Matthew Alger, honours at CS, ANU, 2016

Ex: Estimating redshift of galaxies



Atlas Image [or Atlas Image mosaic] courtesy of 2MASS/UMass/IPAC-Caltech/NASA/NSF

Jakub Nabaglo, honours at CS, ANU, 2017

COMP4670/8600

- 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

Strategy for machine learning

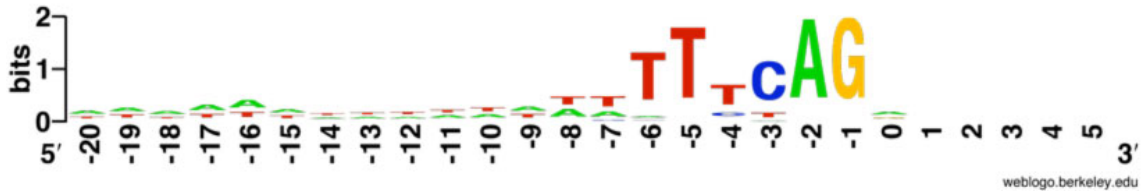
Estimate best predictor = training = learning

Given data $(x_1, y_1), \ldots, (x_n, y_n)$, find a predictor $f_{\mathbf{w}}(\cdot)$.

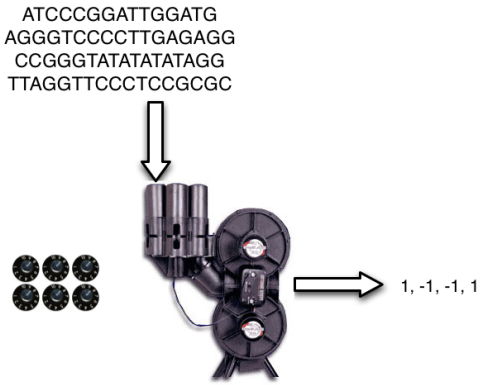
1. Identify the type of input x and output y data
2. Propose a (linear) mathematical model for $f_{\mathbf{w}}$
3. Design an objective function or likelihood
4. Calculate the optimal parameter $(\mathbf{w})$
5. Model uncertainty using the Bayesian approach
6. Implement and compute (the algorithm in python)
7. Interpret and diagnose results

Classification of Sequences

Example: Recognition of splice sites



- Every 'AG' is a possible acceptor splice site
- Computer has to learn what splice sites look like
  - given some known genes/splice sites . . .
- Prediction on unknown DNA



From Sequences to Features

- Many algorithms depend on numerical representations.
  - Each example is a vector of values (features).
- Use background knowledge to design good features.

	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$	$\mathbf{x}_6$	$\mathbf{x}_7$	$\mathbf{x}_8$	...
GC before	0.6	0.2	0.4	0.3	0.2	0.4	0.5	0.5	...
GC after	0.7	0.7	0.3	0.6	0.3	0.4	0.7	0.6	...
AG**AG**AAG	0	0	0	1	1	0	0	1	...
TTT**AG**	1	1	1	0	0	1	0	0	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱
Label	+1	+1	+1	−1	−1	+1	−1	−1	...

Comparing Example / Linear Predictor

How similar are two examples?

$$\langle \mathbf{x}_1, \mathbf{x}_2 \rangle = 0.6 \times 0.2 + 0.7 \times 0.7 + 0 \times 0 + 1 \times 1 = 1.61$$
$$\langle \mathbf{x}_1, \mathbf{x}_5 \rangle = 0.6 \times 0.2 + 0.7 \times 0.3 + 0 \times 1 + 1 \times 0 = 0.33$$

Is $\mathbf{x}_1$ more similar to $\mathbf{x}_2$ or $\mathbf{x}_5$?

	$\mathbf{w}$	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$	$\mathbf{x}_6$	$\mathbf{x}_7$	$\mathbf{x}_8$	...
GC before	2.7	0.6	0.2	0.4	0.3	0.2	0.4	0.5	0.5	...
GC after	8.3	0.7	0.7	0.3	0.6	0.3	0.4	0.7	0.6	...
AG**AG**AAG	−1.2	0	0	0	1	1	0	0	1	...
TTT**AG**	0.88	1	1	1	0	0	1	0	0	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱	
Label		+1	+1	+1	−1	−1	+1	−1	−1	...

Linear predictor

$$f_{\mathbf{w}}(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + b = \mathbf{w}^\top \mathbf{x} + b$$

Inner product vs distance

$$||\mathbf{x}_i - \mathbf{x}_j||^2 = \langle \mathbf{x}_i, \mathbf{x}_i \rangle - 2\langle \mathbf{x}_i, \mathbf{x}_j \rangle + \langle \mathbf{x}_j, \mathbf{x}_j \rangle$$

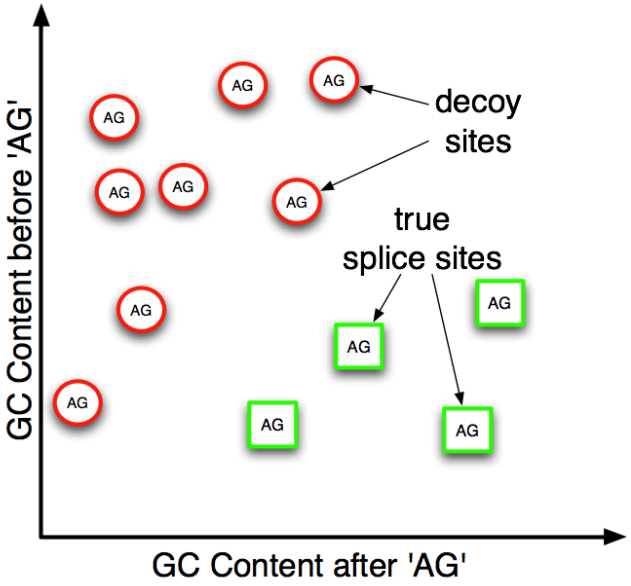
Recognition of Splice Sites

- Given: Potential acceptor splice sites

```
AAACAAATAAGTAACTAATCTTTTAGGAAGAACGTTTCAACCATTTTGAG
AAGATTAAAAAAAAACAAATTTTTAGCATTACAGATATAATAATCTAATT
CACTCCCCAAATCAACGATATTTTAGTTCACTAACACATCCGTCTGTGCC
TTAATTTCACTTCCACATACTTCCAGATCATCAATCTCCAAAACCAACAC
```

intron **exon**

- Goal: Rule that distinguishes true from false ones

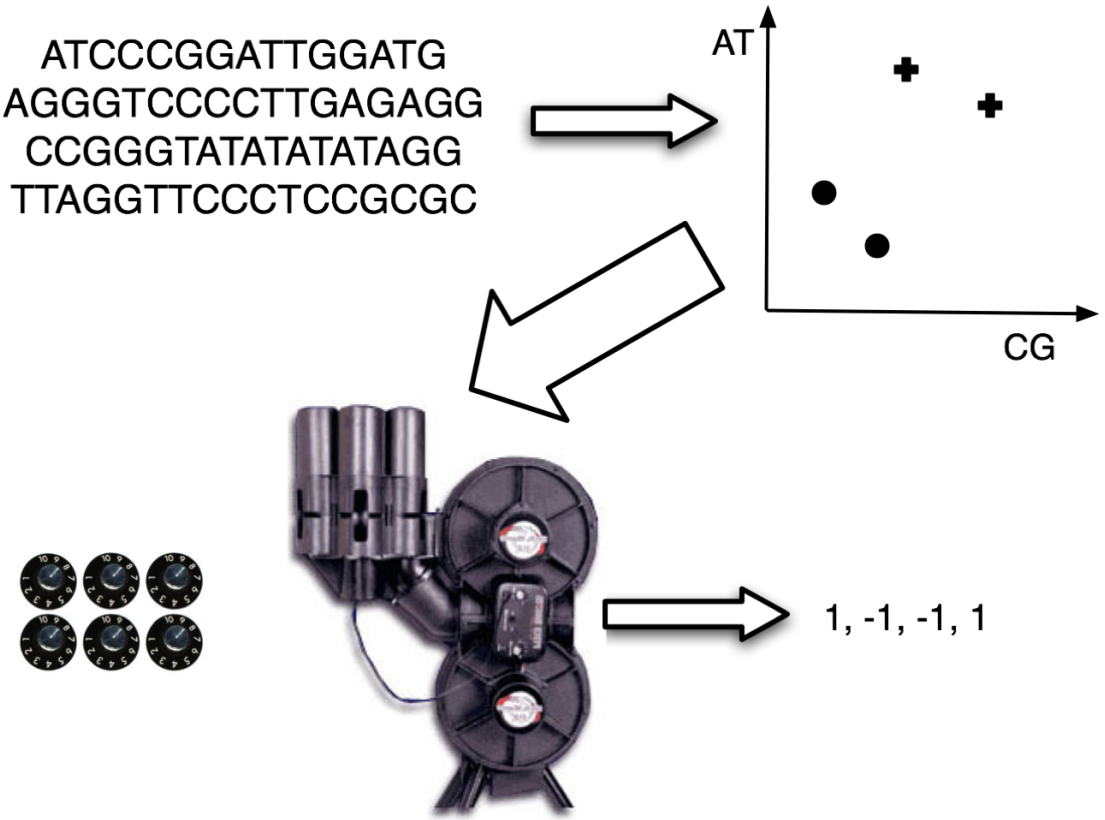


e.g. exploit that exons have higher GC content

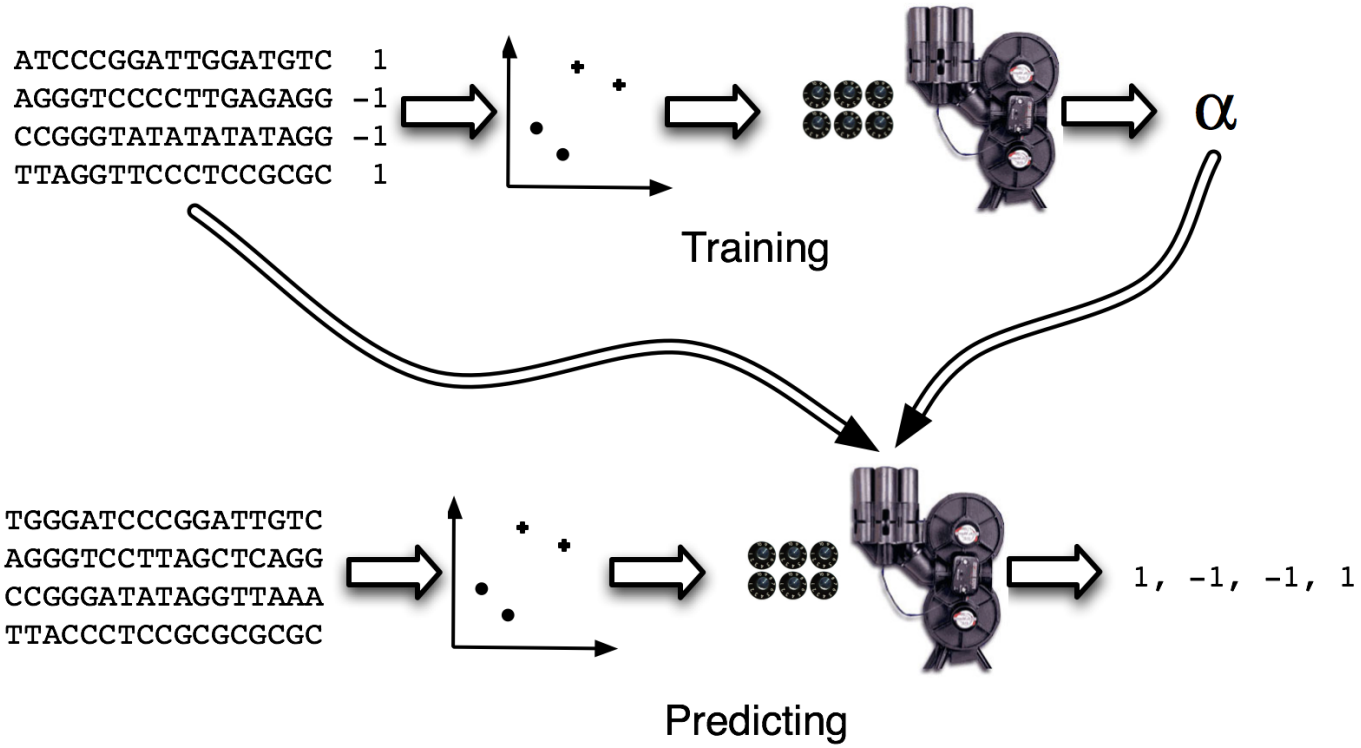
or

that certain motifs are located nearby

Numerical Representation



Training vs Predicting



Training a linear regressor

Data

Given $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$.

Objective function

Choose to minimise the the squared error

$$\min_{\mathbf{w},b} \frac{1}{2}\sum_{i=1}^{N}(f_{\mathbf{w}}(x_i) - y_i)^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

Calculus and linear algebra

- Compute the gradient of the objective with respect to $\mathbf{w}$ and b.
- Set gradient to zero and solve linear system of equations
- This gives $\mathbf{w}^*$ and b^*

Use for prediction

A new data point $\mathbf{x}_{test}$ has prediction

$$\mathbf{w}^{*\top}\mathbf{x}_{test} + b^*$$

Generalisation Error

Estimating the performance of a predictor

Performance on unseen examples

- Theoretical ideal: expected risk
- The accuracy on all (infinite) data

$$\mathbb{E}_{x\sim\mathcal{X},y\sim\mathcal{Y}} \quad \mathrm{perf}(f_{\mathbf{w}}(x), y)$$

Estimate using empirical sample

- Approximate expected risk with empirical risk
- Cross validation or bootstrap $(x_{n+1}, y_{n+1}), \cdots, (x_{n+m}, y_{n+m})$

$$\sum_{i=n+1}^{n+m} \mathrm{perf}(f_{\mathbf{w}}(x_i), y_i)$$

Be careful about splitting data

- Use a representative set of examples and labels
- Be careful of batch effects and correlations

Two types of parameters

Integration vs optimisation

- We generally refer to training as optimising an objective
- Some machine learning problems are better solved by expressing as integration problems.
- Computationally challenging to integrate $\Rightarrow$ optimise

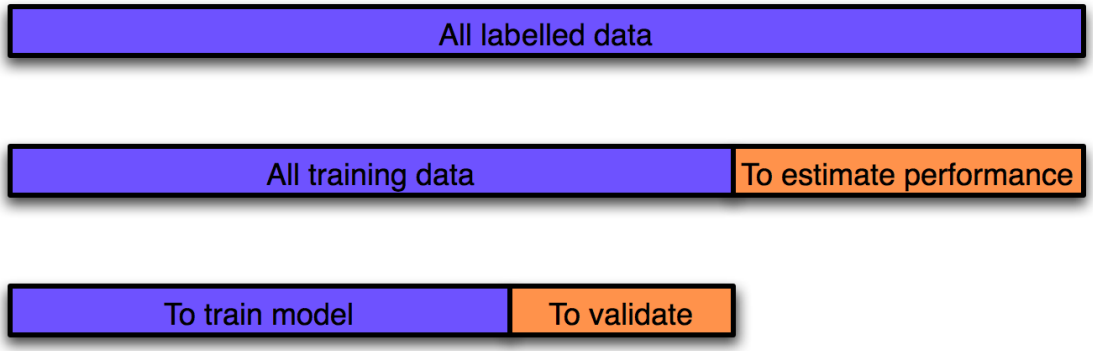
Inner loop – Training – parameter

- Optimize an objective function
- Use gradient information or other efficient method
- Many third party libraries

Outer loop – Model selection – hyperparameter

- Select over a small set of options
- Hard to directly optimise
- Cross validation or bootstrap

No Peeking!



Do not train on the test set!

- Use subset of data for training
- From subset, further split to select model.

Model Selection = Find best hyperparameters

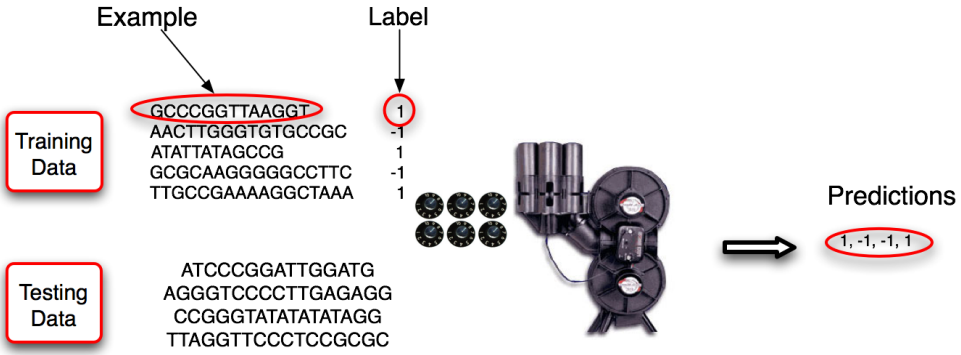
- Regularization parameter
- Feature construction parameters
- Learning rate

Anatomy of a ML workflow

Feature engineering

- Represent sequences and structures numerically
- Capture domain knowledge and scale
- Correlated and higher order features

Training and evaluation



Sanity checks and deployment

- Resources: memory, computation, disk
- Software engineering: unit tests, APIs, support

Summary so far ...

What is machine learning?

- Use data to construct a predictor
- Expert knowledge used in design

Takeaway messages

- Let the data do the talking
- Feature engineering is crucial
- Estimate generalisation performance
- Do not train on the test set!

Plug and Pray

Machine Learning Open Source Software

mloss.org mldata.org

Do We Need Hundreds of Classifiers to Solve Real World Classification Problems?

jmlr.org/papers/v15/delgado14a.html

Spoiler: No

Usability and Reproducibility

- (too much) focus on new algorithms
- Documentation, modularity issues
- Literate programming

  rmarkdown.rstudio.com yihui.name/knitr jupyter.org
- Scientific computing workflows

  galaxyproject.org www.taverna.org.uk

Dream: App Bazaar for data science

Practical tips (I)

Version control

- Learn how to use `git` or `hg`
- Do not add automatically generated files.
- Software repository: github.com, bitbucket.org, gitlab.com
- Data repository: figshare.com, zenodo.org
- Document repository: overleaf.com, authorea.com

Literate programming

- Live code, equations, visualisation and explanatory text
- jupyter.org, rmarkdown.rstudio.com, yihui.name/knitr
- Refactor code into a package as project matures

Practical tips (II)

Do not use email to manage projects!

Automate, automate, automate

Use scripts for manipulating data, and avoid manual editing if possible. Includes copying data, renaming files, etc.

Issue tracker

Explicitly name the version of software that solves a particular problem

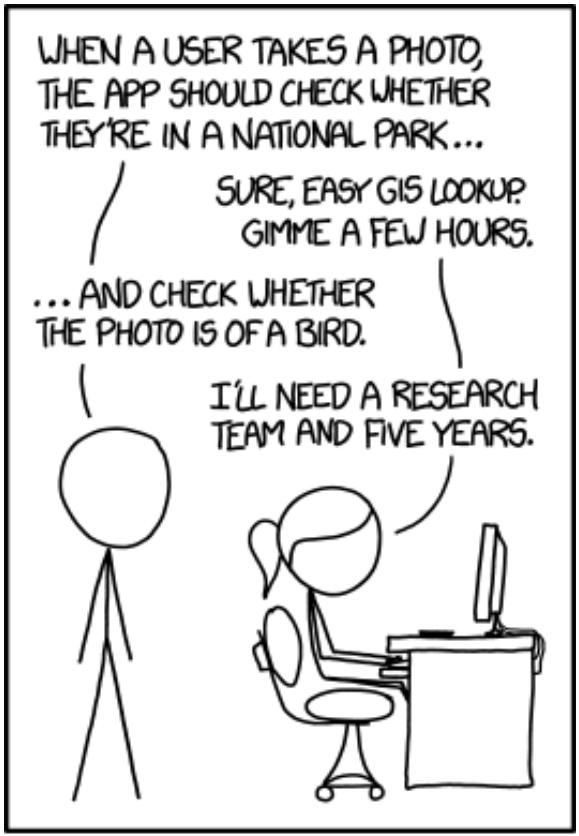
Use common structure

Write a README, Follow common directory formats in your field.

Continuous integration

Trigger tests upon commit, travis-ci.org

When do you talk to an expert?



http://xkcd.com/1425/

Leaving the textbook scenario...

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

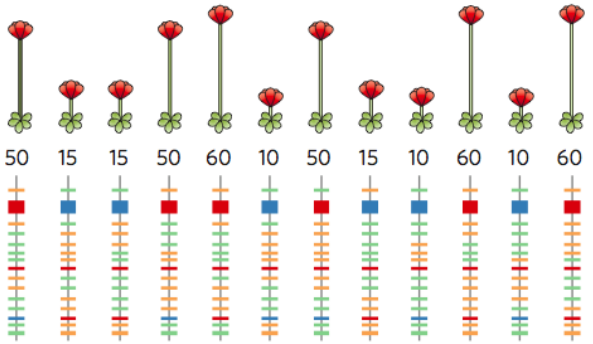
Given data $(x_1, y_1), \ldots, (x_n, y_n)$, find a predictor $f_{\mathbf{w}}(\cdot)$.

- 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

What are good features?

$$f_{\mathbf{w}}(x) : \mathcal{X} \rightarrow \mathcal{Y}$$

What are good biomarkers?



Genome Wide Association Studies

- Which mutations are associated with tall poppies?
- Identify biomarkers with hypothesis tests

Finding stable biomarkers

- Split cohort into two (cross validation)
- Use p-value as a score
- Investigate rank correlation between scores

bioinformatics.research.nicta.com.au/software/gwis/

Not standard classifcation

$$f_{\mathbf{w}}(x) : \mathcal{X} \to \mathcal{Y}$$

Unknown objects

During training	
After deployment	

Identifying wheel defects in trains

- Wheel defects destroy infrastructure
- Classify type of defect from time series

Collaboration with Swiss National Railway

Classifying celestial objects

- Skymapper southern sky survey
- Rare objects not available at training

Discussion with Christian Wolf, RSAA, ANU

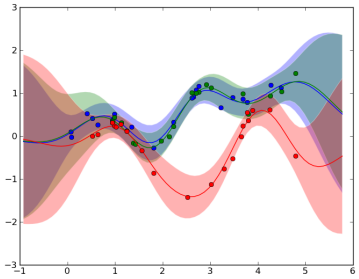
What to measure?

$$f_{\mathbf{w}}(x) : \mathcal{X} \to \mathcal{Y}$$

Active Learning / Expt. Design

Use predictor to identify good candidates

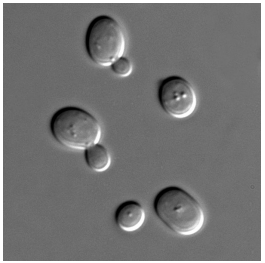
- Annotate top-k items
- Confidence interval improves performance
- Explore - exploit tradeoff



Glucose metabolism in Yeast

- Multiple possible models
- Design biological experiments that maximise information gain

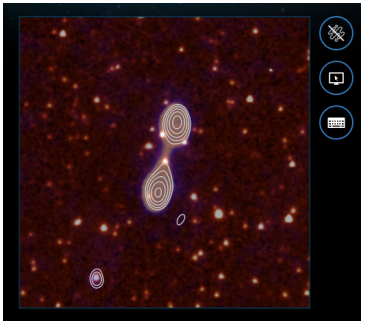
Collaboration with SystemsX Switzerland



Finding host galaxies

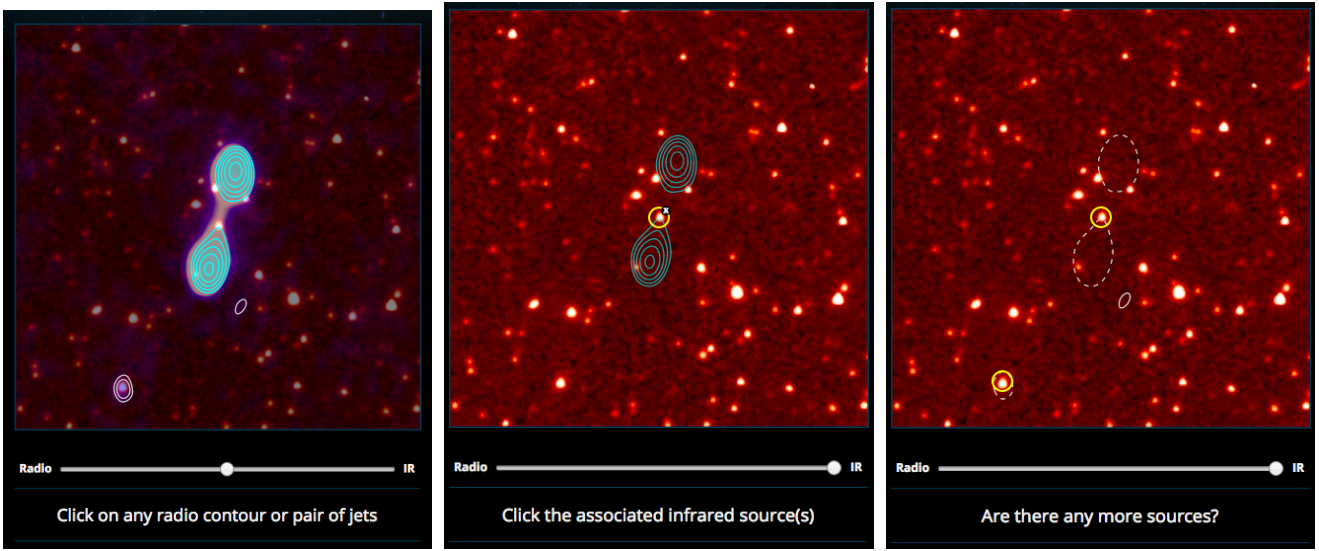
- Machine learning to locate objects
- Show 10 candidates to expert daily

Discussion with Julie Banfield at RSAA, ANU



Which example to ask for a label?

Radio Galaxy Zoo:
citizen science project to cross identify radio galaxies



Summary

What is machine learning?

- Use data to construct a predictor
- Expert knowledge used in design

Takeaway messages

- Let the data do the talking
- Feature engineering is crucial
- Estimate generalisation performance
- Do not train on the test set!

Plug and pray

- Use version control
- Literate programming for designing analysis workflow
- Many open problems in machine learning

Please make your research open

www.ong-home.my