COMP30027 Machine Learning
Welcome and Introduction

Semester 1, 2019

Jeremy Nicholson & Tim Baldwin & Karin Verspoor

Lecture Outline

1. Introduction to Machine Learning
2. Welcome to COMP30027
3. Overview of Machine Learning (why this subject is for me)

What is Machine Learning? I

Define “Machine Learning”:

What is Machine Learning? II

Define “Machine Learning”:

Attempted definition of ML (i):

... the computer learns something ...

But what does it mean to “learn”?

What is Machine Learning? III

Define "Machine Learning":

Attempted definition of ML (ii):

... how to construct computer programs that automatically improve with experience .... A computer program is said to learn from experience ... if its performance ... improves with experience...

Compare: "Artificial Intelligence"

Source(s): Mitchell [1997, pp xv–17]

What is Machine Learning? IV

Define "Machine Learning":

Attempted definition of ML (iii):

Data Mining is ... the process of discovering patterns in data.... Machine Learning [comprises] techniques for finding and describing structural patterns in data, as a tool for helping to explain that data and make predictions from it...

More on "predictions" later!

Source(s): Witten et al. [2011, pp3–9]

What is Machine Learning? V

Define "Machine Learning":

Attempted definition of ML (iv):
Statistics, plus marketing...

Cynical, but a certain degree of truth!

Defining Machine Learning I

Motivation:

- *We are drowning in data, but starving for knowledge!*
- Data = raw information
- Knowledge = set of patterns or models underlying the data

Defining Machine Learning II

Hypothesis: pre-existing data repositories contain a lot of potentially important information

Mission of ML: find it

Definition of ML:

automatic extraction of **valid**, **novel**, **useful** and **comprehensible** knowledge (rules, regularities, patterns, constraints, models, ...) from arbitrary sets of data

Defining Machine Learning III

Our main tasks (depending on data set & problem)

- Classification
- Clustering
- Regression
- Sequence Discovery
- Association Rule Mining
- Outlier Detection
- ...

Lecture Outline

1. Introduction to Machine Learning
2. Welcome to COMP30027
3. Overview of Machine Learning (why this subject is for me)

Who? I

- **Lecturers:**
    - Jeremy Nicholson (subject co-ordinator)
        - `nj@unimelb.edu.au`
    - Afshin Rahimi
        - `rahimia@unimelb.edu.au`

Who? II

- **Tutors:**
  - Evan Cranney
  - Yunzhe Jia
  - Masoud Moravej Khorasani
  - Prashan Madumal
  - Muhammad Ali Qadar
  - Shima Rashidi
  - Hasti Samadi
  - Anubhav Singh
  - Pei-Yun Sun

Where and When? I

- Lectures:

Wed	3:15pm–4:15pm	(Copland Theatre)
Thu	4:15pm–5:15pm	(Wright Theatre)

Where and When? II

- Tutorials (begin in Week 2):

#05	Mon	9–10am	John Medley WG03
#14	Mon	10–11am	David Caro Podium 201
#07	Mon	4.15–5.15pm	John Medley WG05
#09	Mon	5.15–6.15pm	Alan Gilbert 101
#08	Tue	10–11am	John Medley WG05
#04	Thu	8–9am	Sidney Myer Asia Centre G06
#15	Thu	9–10am	Arts West North Wing 355
#13	Thu	11am–12pm	Alice Hoy 101
#10	Fri	11am–12pm	Old Arts 254
#11	Fri	12pm–1pm	Old Arts 254
#12	Fri	3.15–4.15pm	Alice Hoy 101
#06	Fri	4.15–5.15pm	Alan Gilbert G03

Where and When? III

- Practicals (begin in Week 2):

#06	Tue	9–10am	207 Bouverie St B113
#10	Tue	10–11am	207 Bouverie St B116
#02	Tue	4.15–5.15pm	Doug McDonnell 502
#08	Wed	11am–12pm	207 Bouverie St B113
#09	Wed	2.15–3.15pm	207 Bouverie St B114
#03	Wed	4.15–5.15pm	207 Bouverie St B116
#12	Thu	11am–12pm	207 Bouverie St B113
#11	Thu	12–1pm	207 Bouverie St B113
#13	Thu	12–1pm	207 Bouverie St B116
#04	Fri	11am–12pm	Alan Gilbert 111
#05	Fri	2.15–3.15pm	207 Bouverie St B113
#07	Fri	3.15–4.15pm	207 Bouverie St B117

Lecture Schedule (for the first couple of weeks)

Week	Day	Time	Content
1	Wed	3.15pm	Subject introduction
	Thu	4.15pm	Machine learning basics
2	Wed	3.15pm	Probability and Models
	Thu	4.15pm	Naive Bayes
3	Wed	3.15pm	Evaluation (Part I)
	Thu	4.15pm	Feature Selection (?)

Tutorials vs Practicals

- The tutorial will focus on revising theoretical concepts/methods covered in classes, with a strong focus on working through numeric examples
  - Aside from aiming to provide you with a basis for understanding how various ML elements work, these will also be in the style of the questions on the Final Exam
- In the practical sessions, you will gain hands-on experience with applying ML methods
  - This will particularly focus on becoming familiar with `scikit-learn` — which will be helpful for the Project work — but will also allow you to see some practical consequences on various (small) data sets

Subject Materials

- The LMS will be the primary portal for the subject:
- Lecture slides, Tutorial/Prac exercises, Project materials, Project submission will be accessible through the LMS — you will require access
- The Discussion Forum will be our first port-of-call for answering subject-related questions. (This is most salient for the Projects!)

Prerequisites I

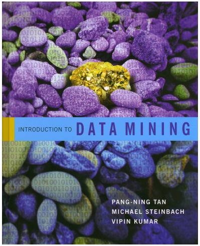
- Programming skills:
  - we will use Python predominantly in the labs and project work (`numpy`, `scipy`, `scikit-learn`)
  - this subject will have implementation as a secondary focus

Prerequisites II

- Mathematical skills:
    - familiarity with formal mathematical notation
    - basic familiarity with:
        - Probability
        - Statistics
        - Univariate/multivariate differential calculus
        - Geometry
        - Linear algebra
    - (Why?)

Recommended Textbook

- Pang-Ning Tan, Michael Steinbach and Vipin Kumar (2005) *Introduction to Data Mining*, Addison-Wesley.



Other Recommended References

- Tom M. Mitchell (1997) *Machine Learning*, WCB/McGraw-Hill.
- Trevor Hastie, Robert Tibshirani and Jerome Friedman (2001) *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer.
- Ian Witten, Eibe Frank, and Mark A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 3rd edition, 2011.

Assessment

- Final marks will be calculated as follows:

20%	*Project 1 (Week 5-ish)*
20%	*Project 2 (Week 10-ish)*
60%	*Final exam (Exam period)*

Student Representatives

- We need 2 student reps for the subject
- Responsibilities
  - solicit feedback from fellow students
  - attend a Staff–Student Liaison Committee meeting
- Benefits
  - experience at public speaking/chance to get to know CIS staff
  - something to put on your CV
  - free food!
- Volunteers?

Lecture Outline

1. Introduction to Machine Learning
2. Welcome to COMP30027
3. Overview of Machine Learning (why this subject is for me)

Subject Structure (again)

- Lectures (most theoretical)
- Tutorials
- Practicals
- Projects (most practical)

Subject Content

- The subject will cover topics including:
  - **classification**: instance-based learning, decision tree induction, supervised Bayesian learning, logistic regression, support vector machines, deep learning, ensemble learning
  - **data processing**: discretisation, feature selection, regularisation, visualisation
  - **regression**, **clustering**, **evaluation**
- You will gain both theoretical and hands-on experience with all of this, and have the experience to get your hands dirty with real-world data analytics in the projects

Subject Objectives

- Recognise real-world problems as amenable to machine learning
- Apply machine learning algorithms and end-to-end statistical processes correctly
- Interpret the results of machine learning run on real data
- Compare benefits/drawbacks of competing models and algorithms, relevant to real problems
- Derive machine learning algorithms from statistical first principles.

Relevant Disciplines

- Points of contact between machine learning and:
  - Artificial intelligence
  - Statistics
  - Computational complexity theory
  - Information theory
  - Philosophy
  - Psychology and neurobiology

⋮

ML Example: the *Cool/Cute* Classifier

- According to ~~my~~ Tim's 2 y.o. son:

Entity	Class
self	cute
self as baby	???
big brother (4 y.o.)	cool
big sister (6 y.o.)	cute
Mummy	cute

Entity	Class
sports car	cool
tiger	cool
Hello Kitty	cute
spoon	???
water	???

- What would we predict the class for the following to be:

koala, book on ML, train

Yeah yeah, but what's in it for me?

- Scenario 1:

  You are an archaeologist in charge of classifying a mountain of fossilised bones, and want to quickly identify any "finds of the century" before sending the bones off to a museum

- Solution:

  Identify bones which are of different size/dimensions/characteristics to others in the sample and/or pre-identified bones

Yeah yeah, but what's in it for me?

- Scenario 1:

  You are an archaeologist in charge of classifying a mountain of fossilised bones, and want to quickly identify any "finds of the century" before sending the bones off to a museum

- Solution:

  Identify bones which are of different size/dimensions/characteristics to others in the sample and/or pre-identified bones

CLUSTERING/OUTLIER DETECTION

Yeah yeah, but what's in it for me?

- Scenario 2:

  You are an archaeologist in charge of classifying a mountain of fossilised bones, and want to come up with a consistent way of determining the species and type of each bone which doesn't require specialist skills

- Solution:

  Identify some easily measurable properties of bones (size, shape, number of "lumps", ...) and compare any new bones to a pre-classified DB of bones

Yeah yeah, but what's in it for me?

- Scenario 2:

  You are an archaeologist in charge of classifying a mountain of fossilised bones, and want to come up with a consistent way of determining the species and type of each bone which doesn't require specialist skills

- Solution:

  Identify some easily measurable properties of bones (size, shape, number of "lumps", ...) and compare any new bones to a pre-classified DB of bones

SUPERVISED CLASSIFICATION

Yeah yeah, but what's in it for me?

- Scenario 3:

  You are a supermarket manager, wishing to boost sales without increasing expenditure, but with lots of historical purchase data

- Solution:

  Strategically position products to entice consumers to spend more:

  - beer next to chips?
  - beer next to bathroom cleaner?

Yeah yeah, but what's in it for me?

- Scenario 3:

  You are a supermarket manager, wishing to boost sales without increasing expenditure, but with lots of historical purchase data

- Solution:

  Strategically position products to entice consumers to spend more:

  - beer next to chips?
  - beer next to bathroom cleaner?

ASSOCIATION RULES

Yeah yeah, but what's in it for me?

- Scenario 4:

  You are in charge of developing the next "release" of Coca Cola, and want to be able to estimate how well received a given recipe will be

- Solution:

  Carry out tast tests over various "recipes" with varying proportions of sugar, caramel, caffeine, phosphoric acid, coca leaf extract, ... (and any number of "secret" new ingredients), and estimate the function which predicts customer satisfaction from these numbers

Yeah yeah, but what's in it for me?

- Scenario 4:

  You are in charge of developing the next "release" of Coca Cola, and want to be able to estimate how well received a given recipe will be

- Solution:

  Carry out tast tests over various "recipes" with varying proportions of sugar, caramel, caffeine, phosphoric acid, coca leaf extract, ... (and any number of "secret" new ingredients), and estimate the function which predicts customer satisfaction from these numbers

REGRESSION

Everyday Applications of Machine Learning

- Medical diagnosis
- Bioinformatics
- Network security
- Spam filtering
- Price prediction
- Image labelling

$\vdots$

Machine Learning vs. Data Mining vs. Data Science

- Machine learning tends to:
  - be more concerned with theory than applications
  - largely ignore questions of run time/scalability
- Data mining tends to:
  - be more concerned with (business) applications than theory
  - talk a lot about databases and run time/scalability
- Data science tends to:
  - straddle data mining and machine learning, and have more focus on applications and interpreting/communicating data insights
- Very fuzzy dividing line between the three

General Expectations for the Subject

- Mix of theory and practical applications
- Open-ended hands-on projects — to be completed in "groups" of 1 or 2
- Some maths (more and more as we go through the subject)
- Some background/supplementary readings
- (Largely) self-directed programming

What Level of Maths are we Talking? I

$$
\begin{aligned}
\ln \frac{P(y = \texttt{true}|x)}{1 - P(y = \texttt{true}|x)} &= \boldsymbol{w} \cdot \boldsymbol{f} \\
\frac{P(y = \texttt{true}|x)}{1 - P(y = \texttt{true}|x)} &= e^{\boldsymbol{w} \cdot \boldsymbol{f}} \\
P(y = \texttt{true}|x) &= e^{\boldsymbol{w} \cdot \boldsymbol{f}} - e^{\boldsymbol{w} \cdot \boldsymbol{f}} P(y = \texttt{true}|x) \\
P(y = \texttt{true}|x) + e^{\boldsymbol{w} \cdot \boldsymbol{f}} P(y = \texttt{true}|x) &= e^{\boldsymbol{w} \cdot \boldsymbol{f}} \\
P(y = \texttt{true}|x) = h(x) &= \frac{e^{\boldsymbol{w} \cdot \boldsymbol{f}}}{1 + e^{\boldsymbol{w} \cdot \boldsymbol{f}}} = \frac{1}{1 + e^{-\boldsymbol{w} \cdot \boldsymbol{f}}} \\
P(y = \texttt{false}|x) &= \frac{1}{1 + e^{\boldsymbol{w} \cdot \boldsymbol{f}}} = \frac{e^{-\boldsymbol{w} \cdot \boldsymbol{f}}}{1 + e^{-\boldsymbol{w} \cdot \boldsymbol{f}}}
\end{aligned}
$$

What Level of Maths are we Talking? II

$$P(y = 1|x; \beta) = h_\beta(x)$$
$$P(y = 0|x; \beta) = 1 - h_\beta(x)$$
$$\rightarrow \quad P(y|x; \beta) = (h_\beta(x))^y * (1 - h_\beta(x))^{1-y}$$

$$\begin{aligned} & \arg\max_\beta \textstyle\prod_{i=1}^n P(y_i|x_i; \beta) \\ = \quad & \arg\max_\beta \textstyle\prod_{i=1}^n (h_\beta(x_i))^{y_i} * (1 - h_\beta(x_i))^{1-y_i} \\ = \quad & \arg\max_\beta \textstyle\sum_{i=1}^n y_i \log h_\beta(x_i) + (1 - y_i) \log(1 - h_\beta(x_i)) \end{aligned}$$

Summary

- What is machine learning?
- What are differences between machine learning and data mining?
- What are the basic “flavours” of machine learning?

Looking ahead to Lecture 2

- Check that you can access the web site via the LMS

References I

Tom M. Mitchell. *Machine Learning*. McGraw-Hill, New York, USA, 1997.

Ian Witten, Eibe Frank, and Mark A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 3rd edition, 2011.