Machine Learning 1: Assorted Ramblings

Tom S. F. Haines
T.S.F.Haines@bath.ac.uk

UNIVERSITY OF BATH

Purpose of talk?

UNIVERSITY OF BATH

What is data science?

- I don't know! (all the definitions suck)

UNIVERSITY OF BATH

What is data science?

- I don't know! (all the definitions suck)
- Wikipedia claims:
  "Data science [...] is an interdisciplinary field about scientific methods, processes, and systems to extract knowledge or insights from data in various forms, either structured or unstructured, similar to data mining."
  Isn't that just **science**?

UNIVERSITY OF BATH

What is data science?

- I don't know! (all the definitions suck)
- Wikipedia claims:
  "Data science [...] is an interdisciplinary field about scientific methods, processes, and systems to extract knowledge or insights from data in various forms, either structured or unstructured, similar to data mining."
  Isn't that just **science**?
- Wikipedia also says:
  "When Harvard Business Review called it "The Sexiest Job of the 21st Century" the term became a buzzword, and is now often applied to business analytics, or even arbitrary use of data, or used as a sexed-up term for statistics."

UNIVERSITY OF BATH

What is data science?

- I don't know! (all the definitions suck)
- Wikipedia claims:
  "Data science [...] is an interdisciplinary field about scientific methods, processes, and systems to extract knowledge or insights from data in various forms, either structured or unstructured, similar to data mining."
  Isn't that just **science**?
- Wikipedia also says:
  "When Harvard Business Review called it "The Sexiest Job of the 21st Century" the term became a buzzword, and is now often applied to business analytics, or even arbitrary use of data, or used as a sexed-up term for statistics."

- Bad news: It's a vacuous term.
- Good news: It's whatever we want.

UNIVERSITY OF BATH

Our interpretation

- **End-to-end machine learning**. It's really a **Machine Learning MSc**.

UNIVERSITY OF BATH

Our interpretation

- **End-to-end machine learning**. It's really a **Machine Learning MSc**.
- Modules:
  - Statistics for Data Science
  - Software Technologies for Data Science
  - Machine Learning 1
  - Applied Data Science
  - Machine Learning 2
  - Bayesian Machine Learning*
  - Neural Computation*
  - Reinforcement Learning*
  - Research Project

  * = pick two.

UNIVERSITY OF BATH

Our interpretation

- **End-to-end machine learning**. It's really a **Machine Learning MSc**.
- Modules:
  - Statistics for Data Science
  - Software Technologies for Data Science
  - Machine Learning 1

  } Dependencies for ML (+ data management)

  – ML

  - Applied Data Science

  – Data preparation, visualisation etc.

  - Machine Learning 2
  - Bayesian Machine Learning*
  - Neural Computation*
  - Reinforcement Learning*

  } All ML

  - Research Project

  * = pick two.

UNIVERSITY OF BATH

Software Stack

- **Jupyter**
- Python 3
  - numpy
  - scipy
  - h5py
  - pandas
  - matplotlib
  - scikit-learn
  - tensorflow
- Jupyter hub.

Blank Slates

- Mathematicians don't program.
- Computer scientists don't math.

UNIVERSITY OF BATH

Blank Slates

- Mathematicians don't program.
- Computer scientists don't math.

Software Technologies for Data Science
Programming
Vectorisation
Databases etc.

Statistics for Data Science
Linear algebra
Probability
Statistics

UNIVERSITY OF BATH

Blank Slates

- Mathematicians don't program.
- Computer scientists don't math.

Software Technologies for Data Science	**Statistics for Data Science**	**Machine Learning 1**
Programming	Linear algebra	ML for babies
Vectorisation	Probability	ML without probability
Databases etc.	Statistics	Real ML

UNIVERSITY OF BATH

Lectures: Overview

- 2 per week, 22 total.
- 17 proper.
- 3 guest.
- 2 spare.

UNIVERSITY OF BATH

Lectures: Warm Up

- L01: Overview, what is ML, classical problem, categorical distribution summary building to naive Bayes, criticising naive Bayes.
- L02: Decision tree as one answer to criticisms.
- L03: Types of output (classification, regression), regressions trees, bagging to get random forests.
- L04: Other problems (unsupervised etc.), clustering, k-means, n-nearest graph.
- Minimal maths and programming required!

UNIVERSITY OF BATH

Lectures: Optimisation

- L05 Optimisation for Pirates: Analytic, brute force, random, stupid approaches; degrees of freedom, RANSAC. Linear regression as running example.
- L06 Optimisation for Ninjas: Cost functions, gradient descent, local vs. global minima, automatic differentiation.
- L07: Logistic regression, multilevel regression and poststratification (polling application), generalised linear model.

Lectures: Details

- L08 Is it working?: Train/test, hyper-opt, examples of failure, checking with visualisation.
- L09: Overfitting, regularisation (why), maximum likelihood, maximum a posteriori, Bayesian, priors, generative models.
- L10: Curse of Dimensionality, feature design, invariance/equivariance, dimensionality reduction with PCA.

UNIVERSITY OF BATH

Lectures: Graphical Models

- L11: Representations (factor graphs as primary), as factorisation, examples, Markov random chain, dynamic programming, Kalman filtering/smoothing.
- L12: Fitting to data, belief propagation, Bayesian decision theory, running medical scenario.
- L13 Do you like Terminator?: Latent variables, plate notation, Latent semantic analysis for film recommendation.
- L14: Variational, mean field procedure, LDA topic model, Reuters data set.

UNIVERSITY OF BATH

Lectures: Some Hard Stuff

- L15 Optimisation for Wizards: Overparameterizisation, multiple restarts, hierarchy of complexity, line search, momentum inc. Nestorov, (multivariate) Newtons method.
- L16: Support vector machines, solving with linear programming, kernel trick. Algorithm shootout on several data sets.
- L17: Gaussian processes.

UNIVERSITY OF BATH

What's missing?

- Density estimation.
- Particle filtering.
- Hyper-parameter optimisation.
- MCMC!

All in ML2. (CDE students attending ML1 only).

UNIVERSITY OF BATH

Coursework

- 60% – 3 or 4 lab exercises.
- 40% – final project, 3000 word limit.
- Plan to use nbgrader for semi-automatic marking.

UNIVERSITY OF BATH

Coursework

- 60% – 3 or 4 lab exercises.
- 40% – final project, 3000 word limit.
- Plan to use nbgrader for semi-automatic marking.

- No real plan as of yet, but ideas include:
  - Random forest.
  - Something that is just optimisation.
  - Linear model, using Maximum Likelihood, MAP, Bayesian.
  - Potts model, both belief propagation and variational.

UNIVERSITY OF BATH

Software Technologies for Data Science

- I'm teaching how to program (50%), Ken is teaching databases (50%).
- Most students will already know how to code! How to avoid boredom?

UNIVERSITY OF BATH

Software Technologies for Data Science

- I'm teaching how to program (50%), Ken is teaching databases (50%).
- Most students will already know how to code! How to avoid boredom?
- Stealing idea of multiple difficulties.

UNIVERSITY OF BATH

Software Technologies for Data Science

- I'm teaching how to program (50%), Ken is teaching databases (50%).
- Most students will already know how to code! How to avoid boredom?
- Stealing idea of multiple difficulties.
- Do numpy (arrays, vectorisation) in parallel.

UNIVERSITY OF BATH

Software Technologies for Data Science

- I'm teaching how to program (50%), Ken is teaching databases (50%).
- Most students will already know how to code! How to avoid boredom?
- Stealing idea of multiple difficulties.
- Do numpy (arrays, vectorisation) in parallel.
- Series of interesting projects (lectures and coursework):
  - Need to introduce new programming concepts.
  - Have multiple difficulties.
  - Can give clear pseudo code for basic version – no understanding required.
  - Not ML, but could add ML!

UNIVERSITY OF BATH

Software Technologies for Data Science

- I'm teaching how to program (50%), Ken is teaching databases (50%).
- Most students will already know how to code! How to avoid boredom?
- Stealing idea of multiple difficulties.
- Do numpy (arrays, vectorisation) in parallel.
- Series of interesting projects (lectures and coursework):
  - Need to introduce new programming concepts.
  - Have multiple difficulties.
  - Can give clear pseudo code for basic version – no understanding required.
  - Not ML, but could add ML!
- Example ideas:
  - Guitar simulation (wave guides).
  - Finding shortest route on map (A*).
  - Board games (search).
  - Stock market simulation (rule based agents).
- If you have any ideas...

Questions?
Feedback?
Suggestions?