Machine Learning 2.01: Introduction

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

UNIVERSITY OF BATH

Schedule I

- Per week:
  - 2 hour lab session
  - 2 lectures
- Panopticon
- Slides may be last minute/late (sorry)
- Everything on Moodle, including a forum
  (best to ask questions there!)

UNIVERSITY OF BATH

Schedule II

Week 1	Introduction / ML triangle (Tom) Hyperparameter Optimisation (Tom)	Week 6	Natural Language Processing (Tom) Multi-armed Bandits (Tom)
Week 2	Advanced Gradient Descent (Tom) Deep Learning Introduction (Mo)	Week 7	Time Series (Tom) Active Learning (Tom)
Week 3	Multilayer Perceptrons (Mo) Back-propagation (Mo)	Week 8	*Consolidation* *(no lectures)*
Week 4	Convolutional Neural Networks (Mo) Practical Guides & Modern Architectures (Mo)	Week 9	Causality (Tom) Compressed Sensing Part 1 (Mo)
Week 5	Density Estimation (Tom) Ensembles (Tom)	Week 10	Compressed Sensing Part 2 (Mo) ML applications: Medical Imaging (Mo)
		Week 11	Guest lecture (TBD) Revision (Both)

(might change)

UNIVERSITY OF BATH

Coursework

Lab	Name	Percentage	Issued	Due
1	Kaggle group project	30%	2019-2-04	2019-5-12 12:00
2	Neural networks	40%	2019-2-18	2019-3-22 12:00
3	Natural language processing	30%	2019-3-18	2019-4-18 12:00

(Marking anonymous – shouldn't include your name!)

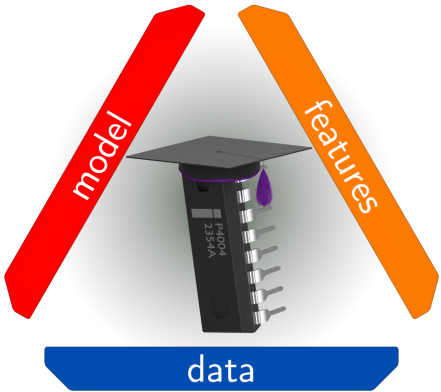
UNIVERSITY OF BATH

Now...

- Customisation to a specific problem
  (whole system – not just the ML algorithm)
- Lots of examples!

Machine learning triangle

- Trade offs:
  - Lots of data $\implies$ Simple model/features
    (e.g. k-means with raw data)
  - Smart model $\implies$ Small data, simple features
    (e.g. Gaussian process with raw data)
  - Smart features $\implies$ Small data, simple model
    (e.g. features include target function!)
- Opportunity cost
- All can be customised to problem...



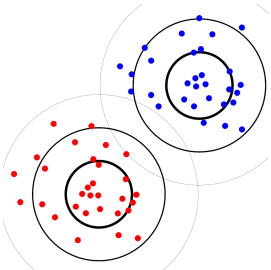
UNIVERSITY OF BATH

Feature engineering redux

- Mentioned in *ML 1.10 – The Curse of Dimensionality*
- Summary:
  - Domain dependent – find or become a *domain expert*
  - Be creative
  - Experiment
- Examples...

Example I: Ordering I

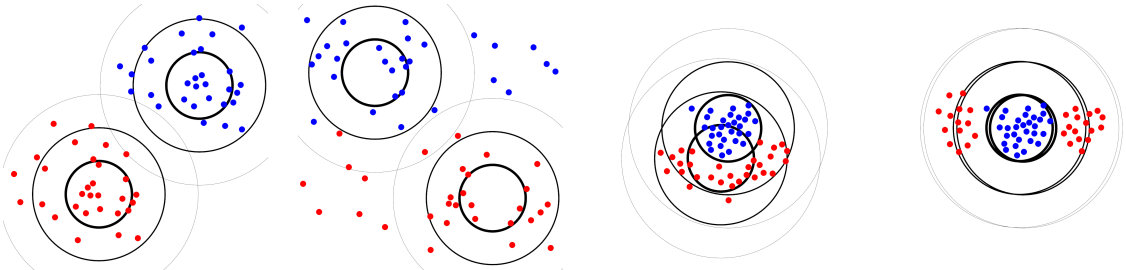
- Clustering,



UNIVERSITY OF BATH

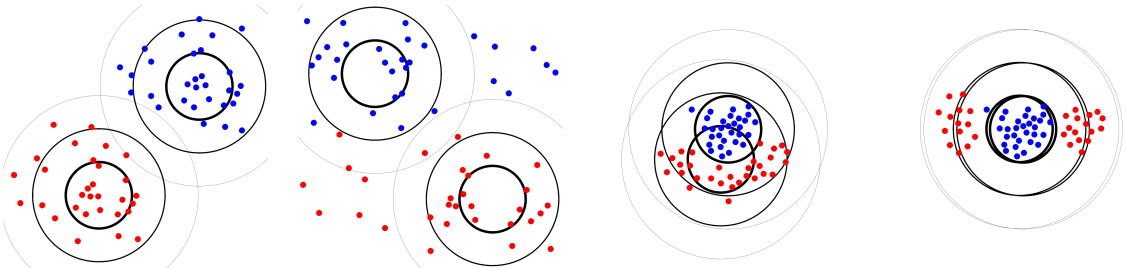
Example I: Ordering I

- Clustering, is not always easy! (gets worse with more dimensions)



Example I: Ordering I

- Clustering, is not always easy! (gets worse with more dimensions)



- Solutions:
  - Advanced clustering algorithm (not today)
  - Learn new distance function...

Example I: Ordering II

- Learn new distance function from *training data* e.g. ask some humans (collect more data)
- Works even if distances are crude, e.g. $\in \{0, 1\}$

UNIVERSITY OF BATH

Example I: Ordering II

- Learn new distance function from *training data*
  e.g. ask some humans (collect more data)
- Works even if distances are crude, e.g. $\in \{0, 1\}$

- Learn $y_{ij} = f([\mathbf{x}_i, \mathbf{x}_j])$ (regression)
  - y_{ij} = distance between items i and j
  - $[\mathbf{x}_i, \mathbf{x}_j]$ = concatenation of feature vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ for items i and j

Example I: Ordering II

- Learn new distance function from *training data*
  e.g. ask some humans (collect more data)
- Works even if distances are crude, e.g. $\in \{0, 1\}$

- Learn $y_{ij} = f([\mathbf{x}_i, \mathbf{x}_j])$ (regression)
  - y_{ij} = distance between items i and j
  - $[\mathbf{x}_i, \mathbf{x}_j]$ = concatenation of feature vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ for items i and j

- Called:
  - Similarity learning
  - Metric learning, but no guarantee its a metric...

UNIVERSITY OF BATH

Example I: Ordering III

- Metric definition:
  1. $f([\mathbf{x}_i, \mathbf{x}_j]) \geq 0$
  2. $f([\mathbf{x}_i, \mathbf{x}_j]) = 0 \Leftrightarrow \mathbf{x}_i = \mathbf{x}_j$
  3. $f([\mathbf{x}_i, \mathbf{x}_j]) = f([\mathbf{x}_j, \mathbf{x}_i])$
  4. $f([\mathbf{x}_i, \mathbf{x}_j]) \leq f([\mathbf{x}_i, \mathbf{x}_k])f([\mathbf{x}_k, \mathbf{x}_j])$

UNIVERSITY OF BATH

Example I: Ordering III

- Metric definition:
  1. $f([\mathbf{x}_i, \mathbf{x}_j]) \geq 0$
  2. $f([\mathbf{x}_i, \mathbf{x}_j]) = 0 \Leftrightarrow \mathbf{x}_i = \mathbf{x}_j$
  3. $f([\mathbf{x}_i, \mathbf{x}_j]) = f([\mathbf{x}_j, \mathbf{x}_i])$
  4. $f([\mathbf{x}_i, \mathbf{x}_j]) \leq f([\mathbf{x}_i, \mathbf{x}_k])f([\mathbf{x}_k, \mathbf{x}_j])$

- 1 and 2: Augment data set with $\forall i; 0 = f([\mathbf{x}_i, \mathbf{x}_i])$

UNIVERSITY OF BATH

Example I: Ordering III

- Metric definition:
  1. $f([\mathbf{x}_i, \mathbf{x}_j]) \geq 0$
  2. $f([\mathbf{x}_i, \mathbf{x}_j]) = 0 \Leftrightarrow \mathbf{x}_i = \mathbf{x}_j$
  3. $f([\mathbf{x}_i, \mathbf{x}_j]) = f([\mathbf{x}_j, \mathbf{x}_i])$
  4. $f([\mathbf{x}_i, \mathbf{x}_j]) \leq f([\mathbf{x}_i, \mathbf{x}_k])f([\mathbf{x}_k, \mathbf{x}_j])$

- 1 and 2: Augment data set with $\forall i; 0 = f([\mathbf{x}_i, \mathbf{x}_i])$

- 3: Make feature **invariant** to order!
  (feature engineering)

  Replace $[\mathbf{x}_i, \mathbf{x}_j]$ with $\left[\frac{1}{2}(\mathbf{x}_i + \mathbf{x}_j), |\mathbf{x}_i - \mathbf{x}_j|\right]$
  (flawed – also invariant to swapping feature values between i and j, but in practise rarely a problem)

Example I: Ordering III

- Metric definition:
  1. $f([\mathbf{x}_i, \mathbf{x}_j]) \geq 0$
  2. $f([\mathbf{x}_i, \mathbf{x}_j]) = 0 \Leftrightarrow \mathbf{x}_i = \mathbf{x}_j$
  3. $f([\mathbf{x}_i, \mathbf{x}_j]) = f([\mathbf{x}_j, \mathbf{x}_i])$
  4. $f([\mathbf{x}_i, \mathbf{x}_j]) \leq f([\mathbf{x}_i, \mathbf{x}_k]) f([\mathbf{x}_k, \mathbf{x}_j])$

- 1 and 2: Augment data set with $\forall i; 0 = f([\mathbf{x}_i, \mathbf{x}_i])$

- 3: Make feature **invariant** to order! (feature engineering)

  Replace $[\mathbf{x}_i, \mathbf{x}_j]$ with $\left[\frac{1}{2}(\mathbf{x}_i + \mathbf{x}_j), |\mathbf{x}_i - \mathbf{x}_j|\right]$ (flawed – also invariant to swapping feature values between i and j, but in practise rarely a problem)

- 4: Ignore – don't need it and really hard
  - Not metric!
  - Regression algorithm will violate other conditions anyway
  - *Mahalanobis distance learning* is metric

UNIVERSITY OF BATH

Example II: Bag of words

- Sentence to feature vector? (variable length)

  "*I'm sorry, Dave. I'm afraid I can't do that.*"

  "*Look Dave, I can see you're really upset about this. I honestly think you ought to sit down calmly, take a stress pill, and think things over.*"

UNIVERSITY OF BATH

Example II: Bag of words

- Sentence to feature vector? (variable length)

  "*I'm sorry, Dave. I'm afraid I can't do that.*"

  "*Look Dave, I can see you're really upset about this. I honestly think you ought to sit down calmly, take a stress pill, and think things over.*"

- One approach: Ignore order!
  (Another **invariance**)
  - Clearly stupid
    – language is all about context
  - But works for many problems!
    e.g. sentiment analysis,
    topic identification, spam filtering

Example II: Bag of words

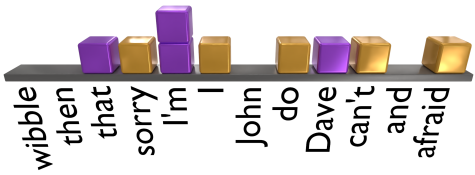
- Sentence to feature vector? (variable length)

  "*I'm sorry, Dave. I'm afraid I can't do that.*"

  "*Look Dave, I can see you're really upset about this. I honestly think you ought to sit down calmly, take a stress pill, and think things over.*"

- One approach: Ignore order!
  (Another **invariance**)
  - Clearly stupid
    – language is all about context
  - But works for many problems!
    e.g. sentiment analysis,
    topic identification, spam filtering

- Feature vector = word count histogram:



- Use sparse encoding (very long, mostly zeros)
- Can now train $y = f(x)$

UNIVERSITY OF BATH

Ablation study

- Examples: Feature engineering that works
- In practise: Try many ideas to find good approach

UNIVERSITY OF BATH

Ablation study

- Examples: Feature engineering that works
- In practise: Try many ideas to find good approach

- Need to measure value of adding feature?
- Some algorithms indicate **abstract** feature importance
  (e.g. logistic regression, random forest)

UNIVERSITY OF BATH

Ablation study

- Examples: Feature engineering that works
- In practise: Try many ideas to find good approach

- Need to measure value of adding feature?
- Some algorithms indicate **abstract** feature importance
  (e.g. logistic regression, random forest)

- Better: Ablation study
  1. Train model with all features
  2. Train model with all but one feature; every combination
  3. Report improvement of "adding" each feature
     (using problem specific cost function)

  (not perfect – imagine same feature included twice)

Data

- Data often feels like a constraint
- i.e. you have what you have, nothing more

- But you can go beyond ‘collect and label more”...

UNIVERSITY OF BATH

Data calibration I

- Netflix prize:
  - 2006, $1 million prize
  - Predict film ratings
  - 10% improvement to win
  - Won after 3 years

Data calibration I

- Netflix prize:
  - 2006, $1 million prize
  - Predict film ratings
  - 10% improvement to win
  - Won after 3 years

(image stolen from Wired article)

- Gavin Potter:
  - Top 10 in final year
  - Psychologist
  - Just him – no team
  - One old desktop
  - Weak maths
    (his college-age daughter did it all)
- Winner (BellKor's Pragmatic Chaos) used his approach...
- ...their biggest final year improvement
  (everyone shared their tricks)

UNIVERSITY OF BATH

Data calibration II

- Ideas from behavioural economics
- *Kahneman–Tversky anchoring effect*
  - Expensive product makes cheaper, but still expensive, products seem reasonable
  - A sale does this with only one product!
  - People think relative to "anchor" (expensive product)

UNIVERSITY OF BATH

Data calibration II

- Ideas from behavioural economics
- *Kahneman–Tversky anchoring effect*
  - Expensive product makes cheaper, but still expensive, products seem reasonable
  - A sale does this with only one product!
  - People think relative to "anchor" (expensive product)
- Applies to films:
  - Good film $\implies$ rate next film higher
  - Bad film $\implies$ rate next film lower

Data calibration II

- Ideas from behavioural economics
- *Kahneman–Tversky anchoring effect*
  - Expensive product makes cheaper, but still expensive, products seem reasonable
  - A sale does this with only one product!
  - People think relative to "anchor" (expensive product)
- Applies to films:
  - Good film $\implies$ rate next film higher
  - Bad film $\implies$ rate next film lower

- Algorithm:
  - $r(f_t)$ = predicted rating for film f_t
  - $a(f_t)$ = actual rating for film f_t
  - Calculate two averages over n recent films:

$$\mu_r = \frac{1}{n} \sum_{i=t-n}^{t-1} r(f_i)$$

$$\mu_a = \frac{1}{n} \sum_{i=t-n}^{t-1} a(f_i)$$

  - Predict:

$$a(f_t) \approx r(f_t) \frac{\mu_a}{\mu_r}$$

- Can factor in time (learn falloff function)

UNIVERSITY OF BATH

Missing data

- Missing values in data set:
  - Smart algorithm – not always practical
  - Fill in with data set mean – stupid
  - Dud value – may work, often doesn't
  - Use machine learning?

UNIVERSITY OF BATH

Missing data

- Missing values in data set:
  - Smart algorithm – not always practical
  - Fill in with data set mean – stupid
  - Dud value – may work, often doesn't
  - Use machine learning?
- Train model to predict feature when missing
  - Inputs: features that are almost always there (calculate probability for each)
  - Need train/test from complete examples
  - Fallback to mean when model won't work (or another model)

UNIVERSITY OF BATH

Unseen data I

- What about a “missing value” you never see?

Unseen data I

- What about a "missing value" you never see?
- Another data set (with overlapping features)
- Example:
  - Goal: *Predict university performance*
  - Research tells you: *Parents going to uni correlated with children doing better*
  - But it's not included in dataset!

UNIVERSITY OF BATH

Unseen data II

- One solution:
  - Data includes: *Home address*, *race*, *religion*
  - Second data set: *National census*
  - Overlaps, and includes *highest qualification level*
  - Model 1: $\text{highest qualification} = f_1(\text{overlaping data})$
    (only summary statistics are available, but still learnable)
  - Augment data with model prediction
    (assume parents have same race & religion as student)
  - Model 2: $\text{performance} = f_2(\text{augmented data})$

UNIVERSITY OF BATH

Unseen data II

- One solution:
  - Data includes: *Home address*, *race*, *religion*
  - Second data set: *National census*
  - Overlaps, and includes *highest qualification level*
  - Model 1: $\text{highest qualification} = f_1(\text{overlaping data})$
    (only summary statistics are available, but still learnable)
  - Augment data with model prediction
    (assume parents have same race & religion as student)
  - Model 2: $\text{performance} = f_2(\text{augmented data})$
- Have mistakes from both models – doesn't always work
- Best if probabilistic
- Often unethical...

Systems approach

- Unseen data: Example of **systems approach**
- Multiple machine learning models, with outputs attached to inputs
- Another example...

Example: Word vectors I

- Bag of words weaknesses:
  - Large feature vector
  - Doesn't *share statistical strength*
    i.e. similar words have to be learned independently

UNIVERSITY OF BATH

Example: Word vectors I

- Bag of words weaknesses:
  - Large feature vector
  - Doesn't *share statistical strength*
    i.e. similar words have to be learned independently
- Avoided by *word vectors*:
  - Assign fixed length vector to each word
  - Similar words nearby, dissimilar words far away

UNIVERSITY OF BATH

Example: Word vectors II

- One such algorithm: **GloVe**
- Objective:

$$\log P(i|j) = \log P(j|i) = \mathbf{x}_i^T \mathbf{x}_j$$

  - $P(i|j)$ = Probability of finding word i in context of word j
  - Context = Length 21 window centred on each occurrence (has $\frac{1}{d}$ weighting, where d is how far apart the words are)
  - $\mathbf{x}_i$ = Length 300 column vector for word i
  - Also downscales rare words
  - Random initialisation and (fancy) gradient descent

UNIVERSITY OF BATH

Example: Word vectors III

- As part of a system:
  - Model 1: Word vector
    (train yourself or download from internet)
  - Model 2: $y = f(\text{word vector(s)})$
    (combine word vectors by taking their mean)
- Very few natural language processing (NLP) tasks for which not useful!
- Has been adapted to *not-words*,
  e.g. AirBnB adapted concept to identify "similar listings" from property features

Model

- Customising model to problem – the obvious approach
- Other lectures/units...

Aside: Beware the central limit theorem

UNIVERSITY OF BATH

Nothing is Normal

- Central limit theorem:
  It's normal to be Normal (Gaussian distribution)

UNIVERSITY OF BATH

Nothing is Normal

- Central limit theorem:
  It's normal to be Normal (Gaussian distribution)
- Reality:
  Most things are not

UNIVERSITY OF BATH

Nothing is Normal

- Central limit theorem:
  It's normal to be Normal (Gaussian distribution)
- Reality:
  Most things are not
- Aside:
  - *Gaussian* is the original name, after Gauss
    (de Moivre got there first, then Laplace, plus Adrain also discovered it in parallel)
  - Pearson popularised *Normal*
  - Regretted doing so

UNIVERSITY OF BATH

Nothing is Normal

- Central limit theorem:
  It's normal to be Normal (Gaussian distribution)
- Reality:
  Most things are not
- Aside:
  - *Gaussian* is the original name, after Gauss
    (de Moivre got there first, then Laplace, plus Adrain also discovered it in parallel)
  - Pearson popularised *Normal*
  - Regretted doing so
- Sometimes it's correct/doesn't matter
- Sometimes it really, really matters

2008 financial crash

- Value at risk – one of the key contributors
  e.g. you might say you have a 1% one year VaR of 1000 units
  Means you expect to lose 1000 units every 100 years

UNIVERSITY OF BATH

2008 financial crash

- Value at risk – one of the key contributors
  e.g. you might say you have a 1% one year VaR of 1000 units
  Means you expect to lose 1000 units every 100 years

- Multiple ways to calculate, but two assumptions:
  - Risk has a Gaussian distribution
  - Assets are independent

UNIVERSITY OF BATH

2008 financial crash

- Value at risk – one of the key contributors
  e.g. you might say you have a 1% one year VaR of 1000 units
  Means you expect to lose 1000 units every 100 years

- Multiple ways to calculate, but two assumptions:
  - Risk has a Gaussian distribution
  - Assets are independent

- Underestimated risk:
  - Actually "fat tailed"
  - Markets are not independent!
    (don't make this mistake either)

Beware summary statistics

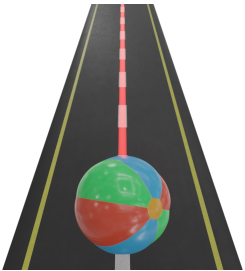
- Imagine a ball bouncing down a road. . . is it safe?

$\mu = 0$

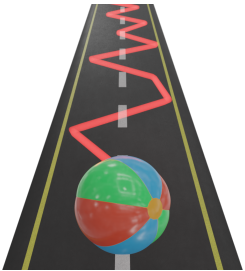
Beware summary statistics

- Imagine a ball bouncing down a road. . . is it safe?

$\mu = 0$



$\mu = 0$

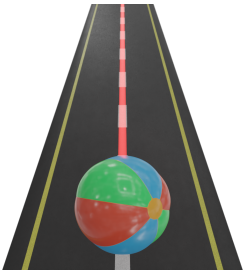


$\mu = 0$

Beware summary statistics

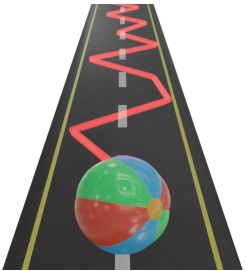
- Imagine a ball bouncing down a road. . . is it safe?

$\mu = 0$



$\mu = 0$

$\sigma^2 = 0$



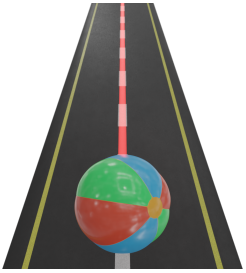
$\mu = 0$

$\sigma^2 = 1$

Beware summary statistics

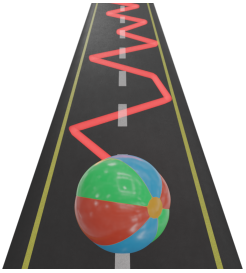
- Imagine a ball bouncing down a road. . . is it safe?

$\mu = 0$



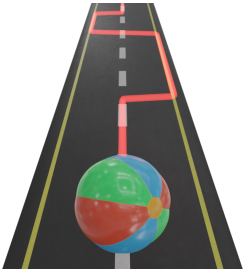
$\mu = 0$

$\sigma^2 = 0$



$\mu = 0$

$\sigma^2 = 1$



$\mu = 0$

$\sigma^2 = 1$

- Evaluate cost functions on the distribution ($\mathbb{E}[C(x)]$), never the mean ($C(\mathbb{E}[c])$)

UNIVERSITY OF BATH

Papers

- Ordering invariance:
  "*Random Forests for Metric Learning with Implicit Pairwise Position Dependence*"
  by C. Xiong, D. Johnson, R. Xu and J. J. Corso

- Wired article about Gavin Potter:
  `https://www.wired.com/2008/02/mf-netflix`

- Example of value at risk approach: (advanced)
  "*CAViaR: Conditional Autoregressive Value at Risk by Regression Quantiles*"
  by R. F. Engle and S. Manganelli

- A word vector paper: (will appear again)
  "*GloVe: Global Vectors for Word Representation*
  by J. Pennington, R. Socher and C. D. Manning

UNIVERSITY OF BATH

Summary

- Unit overview
- Examples of customisation
- An aside

- Next lecture: Hyperparameter optimisation