UNIVERSITY OF CAMBRIDGE

4F10: Deep Learning and Structured Data - Introduction

Mark Gales: mjfg@eng.cam.ac.uk

Michaelmas 2021

Decision Making

In this world nothing can be said to be certain, except death and taxes.

- Benjamin Franklin

- We make decisions under uncertainty all the time

  gambling (not recommended),
  weather forecasting (not very successfully)
  insurance (risk assessment), stock market
  fourth year modules ..

  Need to formalise "intuitive decisions" mathematically

- Basically, how to quantify and manipulate uncertainty.

Machine Learning

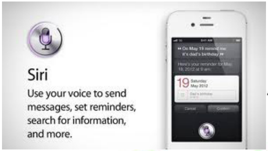
- One definition is (Mitchell):

  "A computer program is said to learn from experience (E) with some class of tasks (T) and a performance measure (P) if its performance at tasks in T as measured by P improves with E"

  alternatively

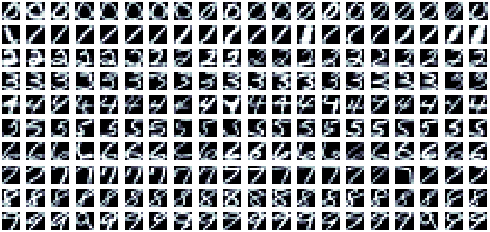
  "Systems built by analysing data sets rather than by using the intuition of experts"

- Multiple conferences in the area:
  - {International,European} Conference on Machine Learning;
  - Neural Information Processing Systems;
  - International Conference on Pattern Recognition etc etc;
- As well as many companies
  - Google, Facebook, Microsoft, Amazon, IBM, Tencent, Baidu, ...



- Interesting application area
  - automatic speech recognition
  - machine translation
  - sentiment analysis
  - summarisation

Computer Vision



Information Retrieval and Web Search

Google Search: Unsupervised Learning   http://www.google.com/search?q=Unsupervised+Learning&sourceid=fir...

Google

Web Images Groups News Froogle more »

Unsupervised Learning Search Advanced Search Preferences

Web Results **1** - **10** of about **150,000** for **Unsupervised Learning**. (**0.27** seconds)

Mixture modelling, Clustering, Intrinsic classification ...
Mixture Modelling page. Welcome to David Dowe's clustering, mixture modelling and **unsupervised learning** page. Mixture modelling (or ...
www.csse.monash.edu.au/~dld/mixture.modelling.page.html - 26k - 4 Oct 2004 - Cached - Similar pages

ACL'99 Workshop -- **Unsupervised Learning** in Natural Language ...
PROGRAM. ACL'99 Workshop **Unsupervised Learning** in Natural Language Processing. University of Maryland June 21, 1999. Endorsed by SIGNLL ...
www.ai.sri.com/~kehler/unsup-acl-99.html - 5k - Cached - Similar pages

Unsupervised learning and Clustering
cgm.cs.mcgill.ca/~soss/cs644/projects/wijhe/ - 1k - Cached - Similar pages

NIPS*98 Workshop - Integrating Supervised and **Unsupervised** ...
NIPS*98 Workshop "Integrating Supervised and **Unsupervised Learning**" Friday, December 4, 1998. ... 4:45-5:30, Theories of **Unsupervised Learning** and Missing Values. ...
www-2.cs.cmu.edu/~mccallum/supunsup/ - 7k - Cached - Similar pages

NIPS Tutorial 1999
Probabilistic Models for **Unsupervised Learning** Tutorial presented at the 1999 NIPS Conference by Zoubin Ghahramani and Sam Roweis. ...
www.gatsby.ucl.ac.uk/~zoubin/NIPStutorial.html - 4k - Cached - Similar pages

Gatsby Course: **Unsupervised Learning** : Homepage
Unsupervised Learning (Fall 2000). ... Syllabus (resources page): 10/10 1 - Introduction to **Unsupervised Learning** Geoff project: (ps, pdf). ...
www.gatsby.ucl.ac.uk/~quaid/course/ - 15k - Cached - Similar pages
[More results from www.gatsby.ucl.ac.uk]

[PDF] **Unsupervised Learning** of the Morphology of a Natural Language
File Format: PDF/Adobe Acrobat - View as HTML
Page 1. Page 2. Page 3. Page 4. Page 5. Page 6. Page 7. Page 8. Page 9. Page 10. Page 11. Page 12. Page 13. Page 14. Page 15. Page 16. Page 17. Page 18. Page 19 ...
acl.ldc.upenn.edu/J/J01/J01-2001.pdf - Similar pages

Unsupervised Learning - The MIT Press
... From Bradford Books: **Unsupervised Learning** Foundations of Neural Computation Edited by Geoffrey Hinton and Terrence J. Sejnowski Since its founding in 1989 by ...
mitpress.mit.edu/book-home.tcl?isbn=026258168X - 13k - Cached - Similar pages

[PS] **Unsupervised Learning** of Disambiguation Rules for Part of
File Format: Adobe PostScript - View as Text
Unsupervised Learning of Disambiguation Rules for Part of. Speech Tagging. Eric Brill. 1. ... It is possible to use **unsupervised learning** to train stochastic. ...
www.cs.jhu.edu/~brill/acl-wkshp.ps - Similar pages

The **Unsupervised Learning** Group (ULG) at UT Austin
The **Unsupervised Learning** Group (ULG). What ? The **Unsupervised Learning** Group (ULG) is a group of graduate students from the Computer ...
www.lans.ece.utexas.edu/ulg/ - 14k - Cached - Similar pages

Gooooooooooogle ►
Result Page: 1 2 3 4 5 6 7 8 9 10 Next

1 of 2   06/10/04 15:44

- Retrieval
- Categorisation
- Clustering
- Relations between pages
- Personalised search
- Targeted advertising
- Spam detection

Financial Prediction and Automated Trading

Protein Structure

DNA sequence

TACCGAACGCTGCTTAAACCG
ATGGCTTGCGACGAATTTGGC

mRNA sequence

AUGGCUUGCGACGAAUUUGGC

M A C D E F G

protein sequence

MACDEFG

Machine Learning

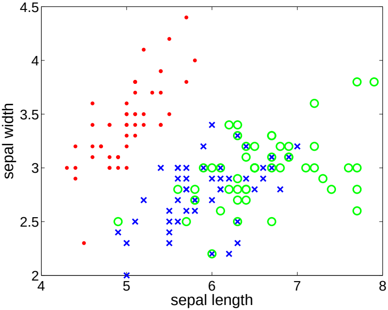
- Wide range of applications of machine-learning
- Example topics - previously seen from 3F8
    - classification
    - regression
    - clustering
    - dimensionality reduction
- This course will primarily look at classification
    - also examine low-dimensional representations
    - as a reminder of the various tasks ...

Classification: Example Iris Dataset

- Fisher (1936) used for linear discriminant techniques
  - data from 3 iris species setosa, versicolor, and virginica
  - 3 classes, 4 numeric attributes, 150 instances
  - sepal length, sepal width, petal length, and petal width

Data:

```
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
...
7.0,3.2,4.7,1.4,Iris-versicolor
6.4,3.2,4.5,1.5,Iris-versicolor
6.9,3.1,4.9,1.5,Iris-versicolor
...
6.3,3.3,6.0,2.5,Iris-virginica
5.8,2.7,5.1,1.9,Iris-virginica
7.1,3.0,5.9,2.1,Iris-virginica
```



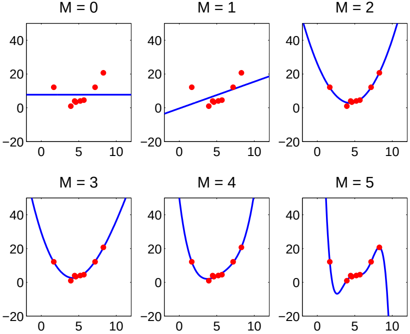
Regression

Let $\boldsymbol{x}$ denote an input point with elements $\boldsymbol{x} = [x_1, \ldots, x_d]^\mathsf{T}$. The elements of $\boldsymbol{x}$, e.g. x_j, represent measured (observed) features of the data point; d denotes the number of measured features of each point.
The data set $\mathcal{D}$ consists of N pairs of inputs and corresponding real-valued outputs:

$$\mathcal{D} = \{\{\boldsymbol{x}_1, y_1\} \ldots, \{\boldsymbol{x}_N, y_N\}\}$$

where y_i is the target value.
The goal is to predict with accuracy the output given a new input (i.e. to *generalize*).

Linear and Nonlinear Regression



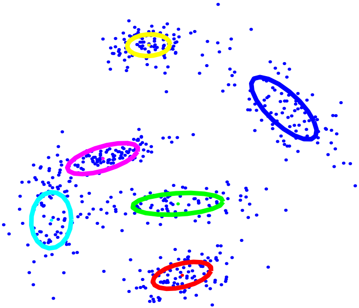
Clustering

Given some data, the goal is to discover "clusters" of points

Roughly speaking, two points belonging to the same cluster are generally more similar to each other or closer to each other than two points belonging to different clusters.

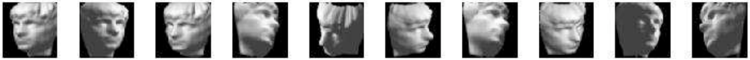
Examples:

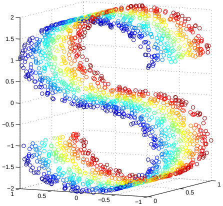
- cluster news stories into topics
- cluster genes by similar function
- cluster movies into categories
- cluster astronomical objects

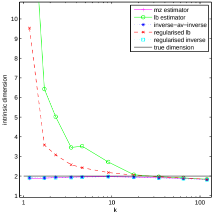


Dimensionality Reduction

Given some data, the goal is to discover and model the intrinsic dimensions of the data, and/or to project high dimensional data onto a lower number of dimensions that preserve the relevant information.







Course Syllabus

- Introduction [1 lecture] - Mark Gales (MJFG)
  - Decision Boundaries/Probability of Error [1 lecture] - MJFG
  - Conditional Independence [1 lecture] - MJFG/MHL
  - Latent Variable & Sequence Models [3 lectures] - MJFG
  - Deep Learning [2 lectures] - MJFG
- Example Class 1
  - Deep Learning for Structured Data [2 lectures] - MJFG
  - Ensemble Methods [1 lecture] - MJFG
  - Support Vector Machines [2 lectures] - MJFG/MHL
  - Kernels for Structured Data [1 lecture] - MJFG/MHL
- Example Class 2
- See web-page for more details

Course Structure

- Total 14 lectures and 2 Examples Classes (2 examples papers)
- Assessment by Examination (1.5 hours): 3 questions from 4
- A number of books cover course material:
  - Christopher M. Bishop: *Pattern Recognition and Machine Learning*, Springer, 2006. ISBN 0-38-731073-8
  - Richard Duda, Peter Hart and David Stork: *Pattern Classification*, Second Edition, John Wiley & Sons Inc, 2000. ISBN 0471056693
  - David J.C. MacKay: *Information Theory, Inference and Learning Algorithms*, Cambridge University Press, 2003. ISBN 0521642981.
  - Kevin P. Murphy: *Machine Learning: a Probabilistic Perspective*, MIT Press, 2012. ISBN 0262018020.
- Specific deep-learning book
  - Ian Goodfellow, Yoshua Bengio and Aarn Courville: *Deep Learning*, MIT Press, 2016.

What is Deep Learning?

From Wikipedia:

> *Deep learning is a branch of machine learning based on a set of algorithms that attempt to model high-level abstractions in data by using multiple processing layers, with complex structures or otherwise, composed of multiple non-linear transformations.*

- Specific area within Machine Learning
  - yields state-of-the-art in a range of tasks
    speech processing, image processing, Go ...
- Highly active research area
  - range of standard tools: pyTorch, TensorFlow, CNTK, ...

What is Structured Data?

- Structured data is a general term for example

  Structured data refers to any data that resides in a fixed field within a record or file.

- In this course structured data will refer to:
  - high dimensional data
  - sequence data

  with (unknown) dependencies between sections of the data

- Examples already shown of this
  - speech and language processing
  - image (video) processing
  - biological sequences (DNA/RNA)

Assessing System Performance

(a) Siberian husky

(b) Eskimo dog

- Images above are from the ImageNet corpus
- Simple to quantify system performance:
  - the dog is a Siberian Husky or Eskimo Dog
  - the system either predicts the class correctly or not
- The loss from making a mistake is normally 1:wrong 0:correct
  - the aim is to minimise the loss of the system on the task

Structured Data "Loss"

- Structured data loss is more interesting
  - e.g. meaning of right/wrong (the "loss") for sequence data?

The cat sat on the mat
The rat sat on the mat
Cat settled on the green mat

- Which has the smallest loss?
- It depends on the task! For speech recognition

The	cat	sat	on	the		mat	—
The	rat	sat	on	the		mat	1
DEL	Cat	settled	on	the	green	mat	3

Nature of Learning

- Training data is used to estimate the model parameters, $\boldsymbol{\theta}$
- Three distinct forms:
  - supervised: the correct classes of the training data are known.

    $$\mathcal{D} = \{\{\boldsymbol{x}_1, y_1\}, \ldots, \{\boldsymbol{x}_N, y_N\}\}$$

    classification: y_i one of K classes, denoted by $\omega_1, \ldots, \omega_K$.
    regression: y_i is a continuous value (may be a vector)
  - unsupervised: classes of the training data are not known

    $$\mathcal{D} = \{\{\boldsymbol{x}_1\}, \ldots, \{\boldsymbol{x}_N\}\}$$

  - reinforcement learning: given an input $\boldsymbol{x}_i$ learn to produce action, a_i that maximise a reward (or penalty) r_i
- This course will primarily consider supervised training

Module Selection

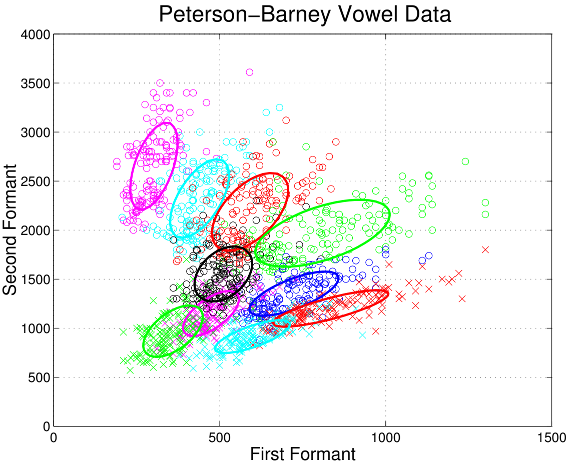
- As engineers did a risk analysis:

Module	Attribute	Probability	Weight
4A2	Presentations	0.92	1.5
	Handouts	0.86	2.0
	⋮		
4F10	Presentations	0.81	1.5
	Handouts	0.75	2.0
	"Hot" topic	1.00	0.5
	Easy course	0.32	3.0
	Job opportunity	0.93	4.0
4F12	Presentations	0.88	1.5
	⋮		

- Maximise (s.t $\sum_{\texttt{modules}} \texttt{Select}(\texttt{module}) = 8$)

$$\sum_{\texttt{modules}} \left[\texttt{Select}(\texttt{module}) \times \sum_{\texttt{attributes}} \texttt{Probability} \times \texttt{weight} \right]$$

Vowel Classification Using Formants Features



- No separation of classes given observation $\boldsymbol{x}$ (uncertainty)
  - $P(\omega|\boldsymbol{x})$ is not zero or one ... there is no "perfect" decision

Making Decisions under Uncertainty

- Need to classify an "unseen" (not in training) observation $\boldsymbol{x}^\star$

  Make a decision that minimises the expected loss (error)

- Consider a classifier, $f(\boldsymbol{x}^\star, \boldsymbol{\theta})$
  - classifier generates a "decision", $\hat{\omega} \in \{\omega_1, \dots, \omega_K\}$
  - associated with any decision is a loss (risk), $\mathcal{L}(\hat{\omega}, y^\star)$
  - $y^\star \in \{\omega_1, \dots, \omega_K\}$ is the "correct" outcome (class label)
  - because of uncertainty $y^\star \sim P(\omega|\boldsymbol{x}^\star)$
- Bayes' Decision Rule - minimise expected loss

$$\hat{\omega} = \arg\min_{\omega} \left\{ \sum_{i=1}^{K} \mathcal{L}(\omega, \omega_i) P(\omega_i|\boldsymbol{x}^\star) \right\}$$

  - decision just requires $P(\omega_i|\boldsymbol{x}^\star)$ for all K classes
  - but we don't know $P(\omega_i|\boldsymbol{x}^\star)$ - need to train a model

Bayes' Decision Rule Example - Sequences

- Consider two loss functions for word sequences
  - sentence: the loss is 1 if the sentence is incorrect
  - word: the loss is the number of word errors

- Consider 3 word sequences from vocabulary `A,B,C,X,Y`

Sentence	Model Prob	Expected Loss Sentence	Expected Loss Word
(1) `A B C`	0.4	0.6	1.2
(2) `A D X`	0.3	0.7	1.1
(3) `A D Y`	0.3	0.7	1.1
(4) `A D C`	0.0	1.0	1.0

- Detailed calculation for `A D X` - order (1) (2) (3) (4)

  sentence: $0.4 \times 1.0 + 0.3 \times 0.0 + 0.3 \times 1.0 + 0.0 \times 1.0 = 0.7$

  word: $0.4 \times 2.0 + 0.3 \times 0.0 + 0.3 \times 1.0 + 0.0 \times 1.0 = 1.1$

Training Criterion for Classification

- Classifier, $f(\boldsymbol{x}^{\star}, \boldsymbol{\theta})$, has model parameters $\boldsymbol{\theta}$
  - need to obtain the "optimal" values for the model parameters
- Train parameters, $\boldsymbol{\theta}$, to minimise the expected loss, $\mathcal{L}_{\tt act}$,

$$\mathcal{L}_{\tt act} = \int \left[\sum_{i=1}^{K} \mathcal{L}(f(\boldsymbol{x}, \boldsymbol{\theta}), \omega_i) P(\omega_i|\boldsymbol{x}) \right] p(\boldsymbol{x}) d\boldsymbol{x}$$

  - $P(\omega_i|\boldsymbol{x})$ is the "true" probability of class given observation
  - $p(\boldsymbol{x})$ is the "true" probability of an observation
  - neither of these is usually known (why we need the model!)
- BUT we may have samples drawn from $p(\omega, \boldsymbol{x}) = P(\omega|\boldsymbol{x})p(\boldsymbol{x})$

the (supervised) training data $\mathcal{D}$!

"Generating" Training Data

- Interested in supervised training data for classification

$$\mathcal{D} = \{\{\boldsymbol{x}_1, y_1\}, \dots, \{\boldsymbol{x}_N, y_N\}\}$$

  - $\boldsymbol{x}_i$: the observation, feature vector
  - $y_i \in \{\omega_1, \dots, \omega_K\}$: class label for observation $\boldsymbol{x}_i$

- Samples are "draws" from joint distribution $p(\omega, \boldsymbol{x})$
- "Standard" process for obtaining data for a task
  - from an initial deployment, or available data, obtain $\boldsymbol{x}_i$

$$\boldsymbol{x}_i \sim p(\boldsymbol{x})$$

  - manually, or from outcomes, obtain label y_i

$$y_i \sim P(\omega|\boldsymbol{x}_i)$$

Empirical Loss

- The empirical loss, $\mathcal{L}_{\texttt{emp}}$, is computed from training data

$$\mathcal{L}_{\texttt{emp}} = \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}(f(\boldsymbol{x}_i, \boldsymbol{\theta}), y_i)$$

  - in the limit, $N \rightarrow \infty$, $\mathcal{L}_{\texttt{emp}} = \mathcal{L}_{\texttt{act}}$
  - in practice N is finite(!) so usually

$$\mathcal{L}_{\texttt{emp}} \leq \mathcal{L}_{\texttt{act}}$$

- Optimising empirical risk may not yield "good" performance
  - don't care about training data performance (known labels)
  - care about heldout data performance - generalisation

Course Syllabus

- Introduction [1 lecture] - Mark Gales (MJFG)
  - Decision Boundaries/Probability of Error [1 lecture] - MJFG
  - Conditional Independence [1 lecture] - MJFG/MHL
  - Latent Variable & Sequence Models [3 lectures] - MJFG
  - Deep Learning [2 lectures] - MJFG
- Example Class 1
  - Deep Learning for Structured Data [2 lectures] - MJFG
  - Ensemble Methods [1 lecture] - MJFG
  - Support Vector Machines [2 lectures] - MJFG/MHL
  - Kernels for Structured Data [1 lecture] - MJFG/MHL
- Example Class 2
- See web-page for more details
- Questions: please ask - or

Email: mjfg@eng.cam.ac.uk