Introductory Applied Machine Learning

Nigel Goddard and Victor Lavrenko
School of Informatics

Semester 1

The **primary aim** of the course is to provide the student with a set of practical tools that can be applied to solve real-world problems in machine learning.

Machine learning is the study of computer algorithms that improve automatically through experience [Mitchell, 1997].

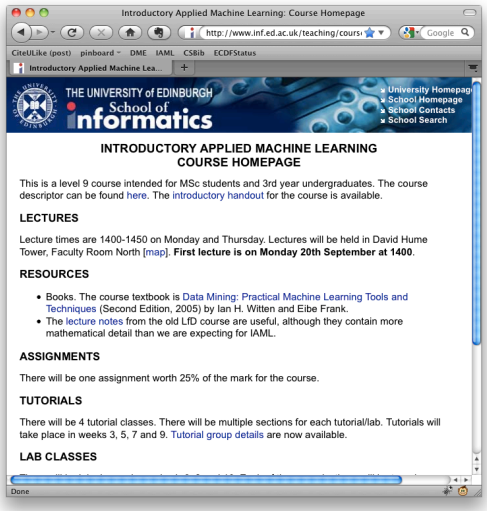
In many of today's problems it is

very hard to write a correct program

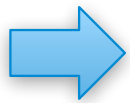
but very easy to collect examples

Idea behind machine learning:
from the examples, generate the program

Spam Classification



Web page



learning	13
lectures	3
Paris Hilton	0
assignments	7
...	

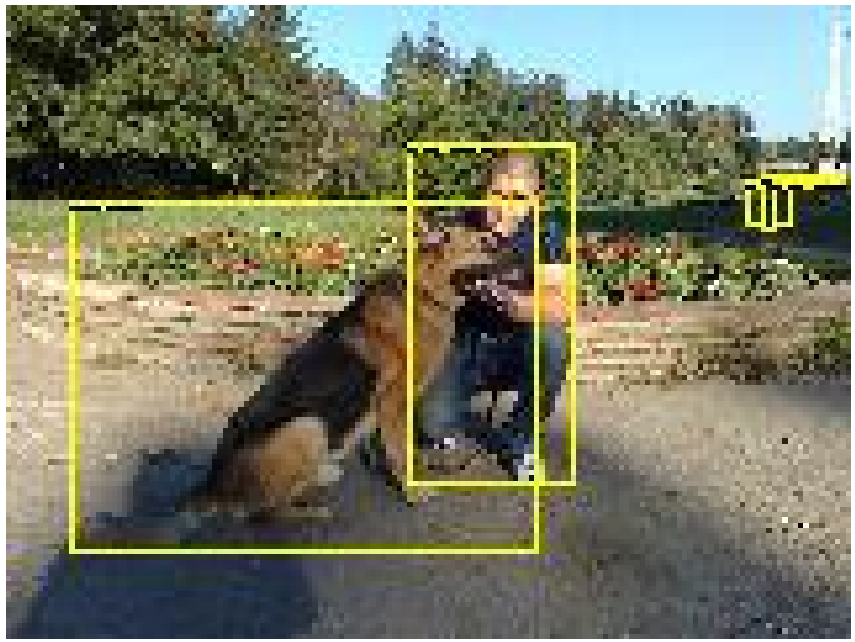
Feature vector

Classifier

SPAM

NONSPAM

Image Processing



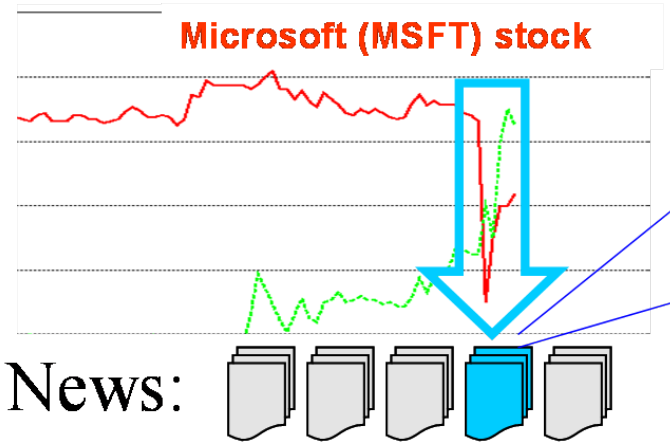
- Classification: Is there are dog in this image?
- Localization: If there is a dog in this image, draw its bounding box
- http://pascallin.ecs.soton.ac.uk/challenges/VOC/

Primate splice-junction gene sequences (DNA)

```
CCAGCTGCATCACAGGAGGCCAGCGAGCAGGTCTGTTCCAAGGGCCTTCGAGCCAGTCTG EI
GAGGTGAAGGACGTCCTTCCCCAGGAGCCGGTGAGAAGCGCAGTCGGGGGCACGGGGATG EI
TAAATTCTTCTGTTTGTTAACACCTTTCAGACTTATGTGTATGAAGGAGTAGAAGCCAAA IE
AAACTAAAGAATTATTCTTTTACATTTCAGTTTTTCTTGATCATGAAAACGCCAACAAAA IE
AAAGCAGATCAGCTGTATAAACAGAAAATTATTCGTGGTTTCTGTCACTTGTGTGATGGT N
TTGCCCTCAGCATCACCATGAACGGAGAGGCCATCGCCTGCGCTGAGGGCTGCCAGGCCA N
```

- Task is to predict if there is an IE, EI or N (neither) junction in the centre of the string
- Data from http://mlearn.ics.uci.edu/

Financial Modeling



Software giant Microsoft saw its shares dip a few percentage points this morning after U.S. District Judge Thomas Penfield Jackson issued his "findings of fact" in the government's ongoing antitrust case against the Seattle wealth-creation machine...

Words like **Jackson** and **antitrust** are more likely in the stories preceding the plunge.

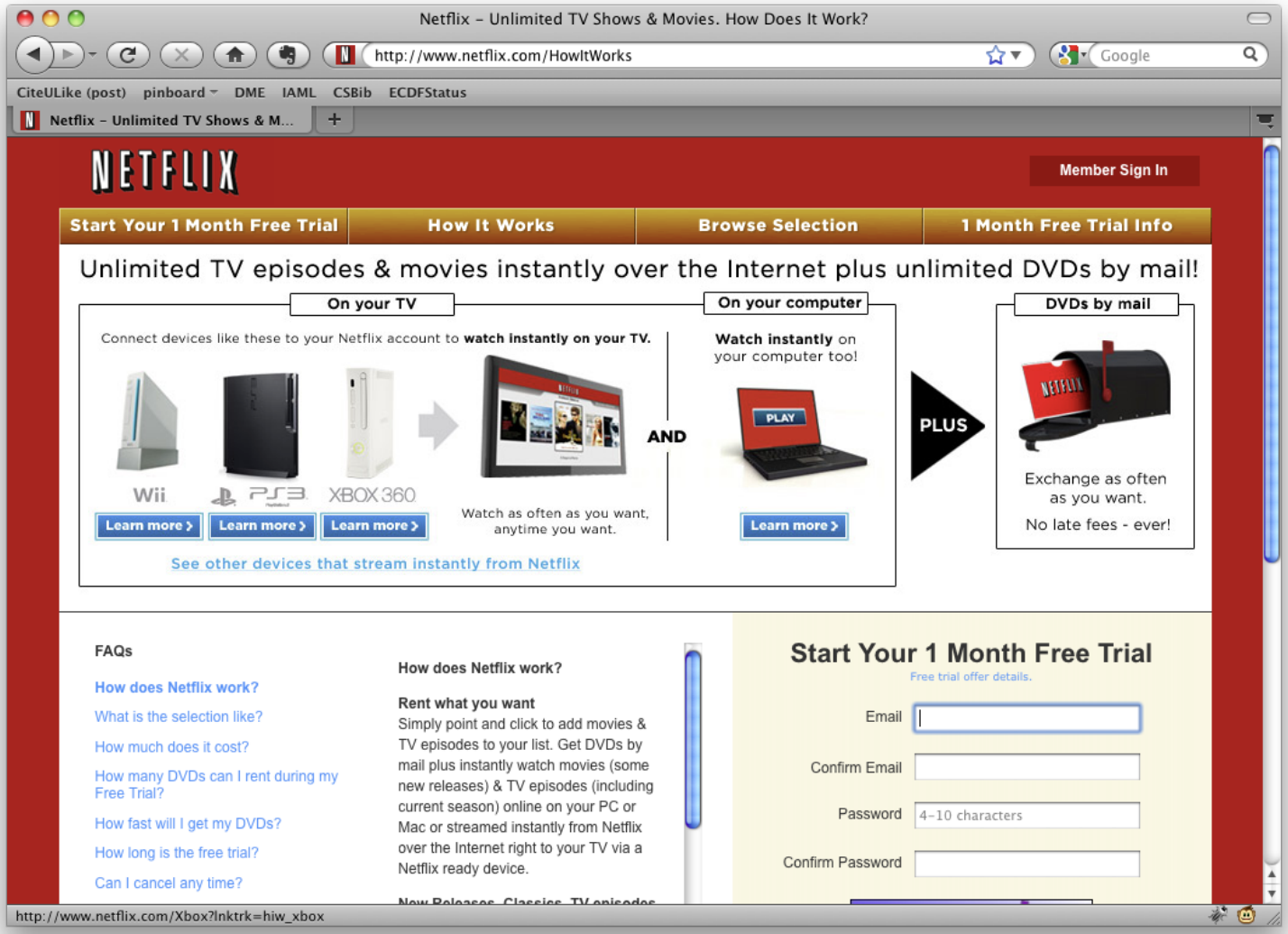
P (**shares**) = 0.074
P (**antitrust**) = 0.009
P (**judge**) = 0.006
P (**trading**) = 0.032
P (**against**) = 0.025
P (**Jackson**) = 0.001

P (**shares** | MSFT↓) = 0.071
P (**antitrust** | MSFT↓) = 0.044
P (**judge** | MSFT↓) = 0.039
P (**trading** | MSFT↓) = 0.029
P (**against** | MSFT↓) = 0.027
P (**Jackson** | MSFT↓) = 0.025

$$P(\text{MSFT}\downarrow \mid \textbf{Jackson}) = P(\textbf{Jackson} \mid \text{MSFT}\downarrow)\ P(\text{MSFT}\downarrow) / P(\textbf{Jackson})$$

[Victor Lavrenko]

Collaborative Filtering



More applications

- Science (Astronomy, neuroscience, medical imaging, bio-informatics)
- Environment (energy, climate, weather, resources)
- Retail (Intelligent stock control, demographic store placement)
- Manufacturing (Intelligent control, automated monitoring, detection methods)
- Security (Intelligent smoke alarms, fraud detection)
- Marketing (targetting promotions, ...)
- Management (Scheduling, timetabling)
- Finance (credit scoring, risk analysis...)
- Web data (information retrieval, information extraction, ...)

Overview

- What is ML? Who uses it?
- Course structure / Assessment
- Relationships between ML courses
- Overview of Machine Learning
- Overview of the Course
- Maths Level
- Reading: W & F chapter 1

Acknowledgements: We thank Amos Storkey, David Barber, Chris Williams, and Charles Sutton for permission to use course material from previous years. Additionally, inspiration has been obtained from Geoff Hinton's slides for CSC 2515 in Toronto

Administration

- Course text: Data Mining: Practical Machine Learning Tools and Techniques (Third Edition, 2011) by Ian H. Witten and Eibe Frank
- All material in course accessible to 3rd year undergraduates. Postgraduates also welcome.
- Assessment:
  - Assignments (4) (25% of mark)
  - Exam (75% of mark)
- 4 Tutorials and 4 Labs
- Maths surgeries
- Course rep
- Plagiarism

`http://www.inf.ed.ac.uk/teaching/plagiarism.html`

Machine Learning Courses

IAML Basic introductory course on supervised and unsupervised learning

MLPR More advanced course on machine learning, including coverage of Bayesian methods (Semester 2)

RL Reinforcement Learning.

PMR Probabilistic modelling and reasoning. Focus on learning and inference for probabilistic models, e.g. probabilistic expert systems, latent variable models, Hidden Markov models

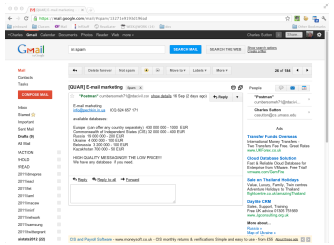
DME Data mining and Exploration. Using methods from PMR to deal with practical issues in learning from large datasets. (Semester 2)

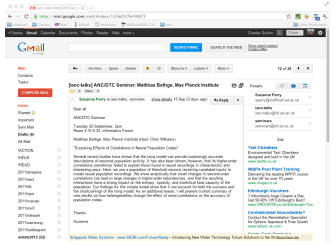
- Basically, IAML: Users of ML; MLPR: Developers of new ML techniques.

Overview of Machine Learning

- Supervised learning
  - Predict an output y when given an input $\mathbf{x}$
  - For categorical y : *classification*.
  - For real-valued y : *regression*.
- Unsupervised learning
  - Create an internal representation of the input, e.g. clustering, dimensionality
  - This is important in machine learning as getting labels is often difficult and expensive
- Other areas of ML
  - Learning to predict structured objects (e.g., graphs, trees)
  - Reinforcement learning (learning from “rewards”)
  - Semi-supervised learning (combines supervised + unsupervised)
  - We will not cover these at all in the course

Supervised Learning (Classification)





Feature processing

Training data

$\mathbf{x_1} = (1, 0, 0, 3, \ldots.)$ $y_1 = \mathbf{SPAM}$

$\mathbf{x_2} = (-1, 4, 0, 3, \ldots.)$ $y_2 = \mathbf{NOTSPAM}$

Learning algorithm

Classifier

Prediction on new example

$\mathbf{x_{1000}} = (1, 0, 1, 2, \ldots.)$ $y_{1000} = \mathbf{???}$

Supervised Learning (Regression)

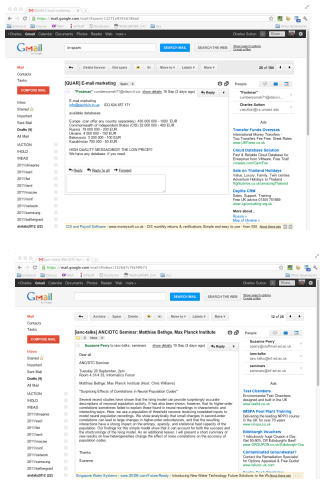
In this class we will talk about linear regression

$$f(\mathbf{x}) = w_0 + w_1 x_1 + \ldots + w_D x_D$$

- $\mathbf{x} = (x_1, \ldots, x_D)^T$
- Here the assumption B is that $f(\mathbf{x})$ is a linear function in $\mathbf{x}$
- The specific setting of the parameters $w_0, w_1, \ldots, w_D$ is done by minimizing a score function
- Usual score function is $\sum_{i=1}^{n} (y^i - f(\mathbf{x}^i))^2$ where the sum runs over all training cases
- Linear regression is discussed in W & F §4.6, and we will cover it later in the course

Unsupervised Learning

In this class we will focus on one kind of unsupervised learning, clustering.



Feature processing

Training data

$\mathbf{x_1} = (1, 0, 0, 3, \ldots.)$

$\mathbf{x_2} = (-1, 4, 0, 3, \ldots.)$

$\ldots.$

$\mathbf{x_{1000}} = (1, 0, 1, 2, \ldots.)$

Learning algorithm

Cluster labels

$c_1 = 4$

$c_2 = 1$

$\ldots.$

$c_2 = 4$

General structure of supervised learning algorithms

Hand, Mannila, Smyth (2001)

- Define the **task**
- Decide on the **model structure** (choice of inductive bias)
- Decide on the **score function** (judge quality of fitted model)
- Decide on **optimization/search method** to optimize the score function

Inductive bias

- Supervised learning is inductive, i.e. we make generalizations about the form of $f(\mathbf{x})$ based on instances $\mathcal{D}$
- Let $f(\mathbf{x}; L, \mathcal{D})$ be the function learned by algorithm L with data $\mathcal{D}$
- Learning is impossible without making assumptions about f !!

The futility of bias-free learning

1

0

???

The futility of bias-free learning

- *A learner that makes no a priori assumptions regarding the target concept has no rational basis for classifying any unseen examples* (Mitchell, 1997, p 42)
- The *inductive bias* of a learner is the set of prior assumptions that it makes (we will not define this formally)
- We will consider a number of different supervised learning methods in the IAML; these correspond to different inductive biases

Machine Learning and Statistics

- A lot of work in machine learning can be seen as a rediscovery of things that were known in statistics; but there are also flows in the other direction
- The emphasis is rather different. One difference is a focus on *prediction* in machine learning vs *interpretation* of the model in statistics
- Machine learning often refers to tasks associated with artificial intelligence (AI) such as recognition, diagnosis, planning, robot control, prediction, etc. These provide rich and interesting tasks
- Goals can be autonomous machine performance, or enabling humans to learn from data (data mining)

Provisional Course Outline

- Introduction (NG)
- Basic probability (NG)
- Thinking about data (VL)
- Naïve Bayes classification (VL)
- Decision trees (VL)
- Linear regression (NG)
- Generalization and Overfitting (NG)
- Linear classification: logistic regression, perceptrons (NG)
- Kernel classifiers: support vector machines (NG)
- Dimensionality reduction (PCA etc) (VL)
- Instance-based methods (VL)
- Performance evaluation (VL)
- Clustering (k-means, hierarchical) (VL)
- Further topics as time permits ...

Maths Level

- Machine learning generally involves a significant number of mathematical ideas and a significant amount of mathematical manipulation
- IAML aims to keep the maths level to a minimum, explaining things more in terms of higher-level concepts, and developing understanding in a procedural way (e.g. how to program an algorithm)
- For those wanting to pursue research in any of the areas covered you will need courses like PMR, MLPR