

3F3: Signal and Pattern Processing

Lecture 1: Introduction

Zoubin Ghahramani

`zoubin@eng.cam.ac.uk`

**Department of Engineering
University of Cambridge**

Lent, 2006

Pattern Processing

Other related terms:

- Pattern Recognition
- Neural Networks
- Data Mining
- Statistical Modelling
- Machine Learning

Learning:

The view from different fields

- **Engineering:** signal processing, system identification, adaptive and optimal control, information theory, robotics, ...
- **Computer Science:** Artificial Intelligence, computer vision, information retrieval, ...
- **Statistics:** learning theory, data mining, learning and inference from data, ...
- **Cognitive Science and Psychology:** perception, movement control, reinforcement learning, mathematical psychology, computational linguistics, ...
- **Computational Neuroscience:** neuronal networks, neural information processing, ...
- **Economics:** decision theory, game theory, operational research, ...

Different fields, Convergent ideas

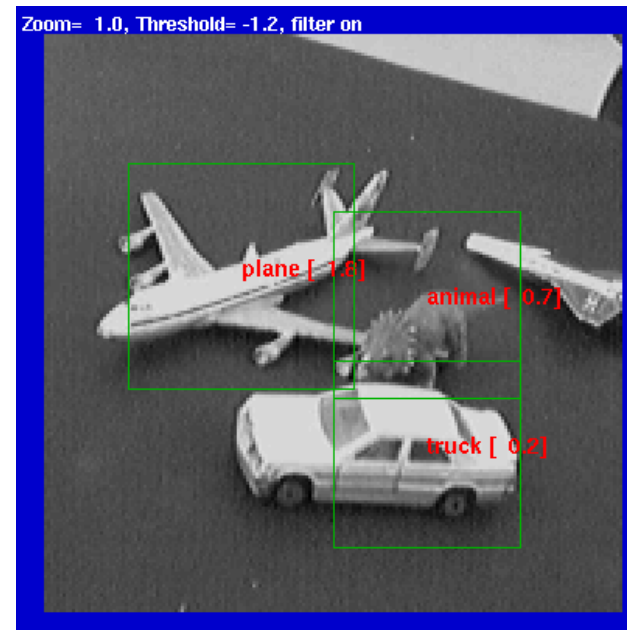
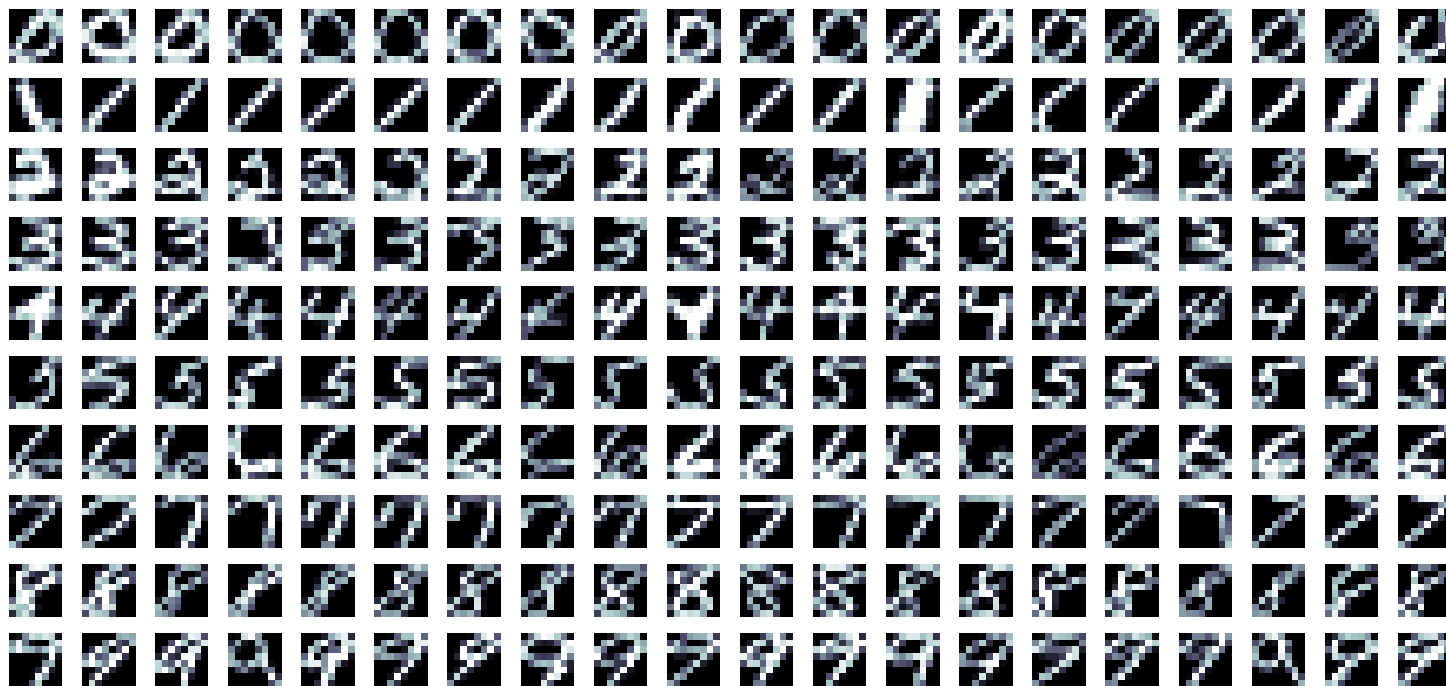
- The **same set of ideas and mathematical tools** have emerged in many of these fields, albeit with different emphases.
- *Machine learning is an interdisciplinary field focusing on both the mathematical foundations and practical applications of systems that learn, reason and act.*
- **The goal of these lectures:** to introduce very basic concepts, models and algorithms.
- **For more on this topic:** I will be teaching a 4th year module called “Machine Learning”, probably in Michaelmas 2006.

Applications of Pattern Processing and Machine Learning

Automatic speech recognition



Computer vision: object, face and handwriting recognition



(NORB image from Yann LeCun)

Information retrieval

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

 **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/~kebler/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](#)

 **Result Page:** [1](#) [2](#) [3](#) [4](#) [5](#) [6](#) [7](#) [8](#) [9](#) [10](#) [Next](#)

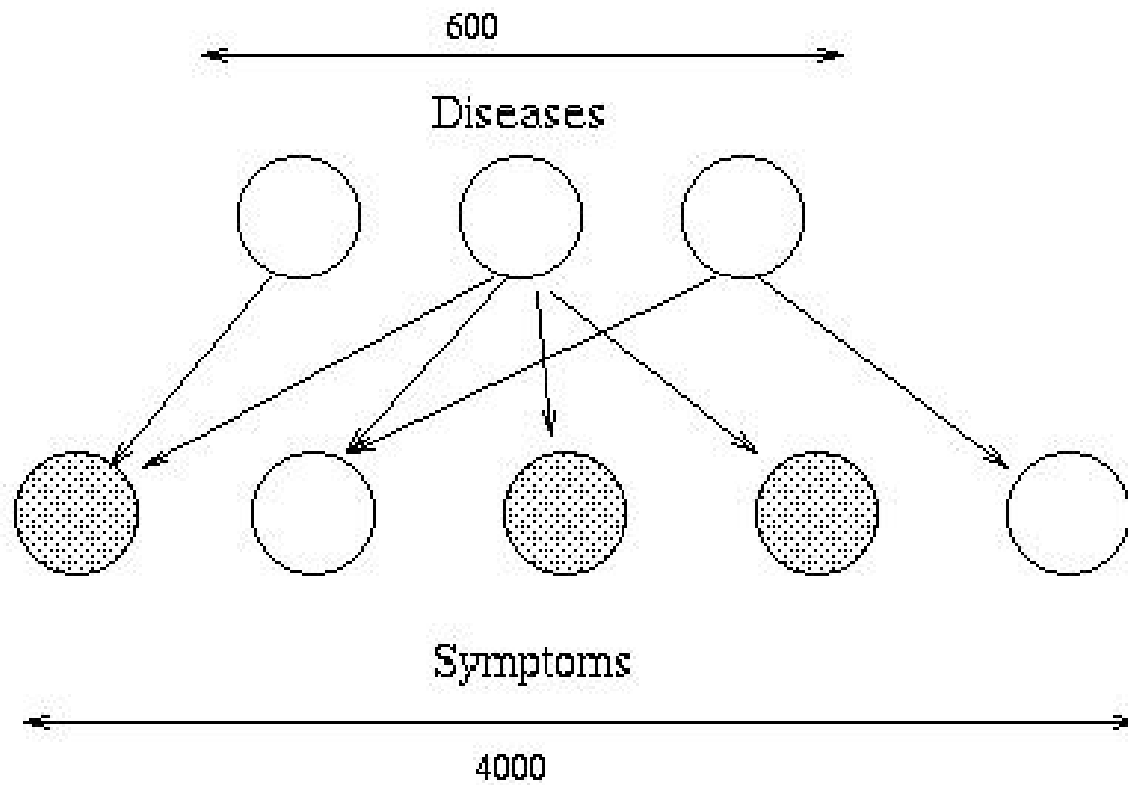
Web Pages

Retrieval
Categorisation
Clustering
Relations between pages

Financial Prediction

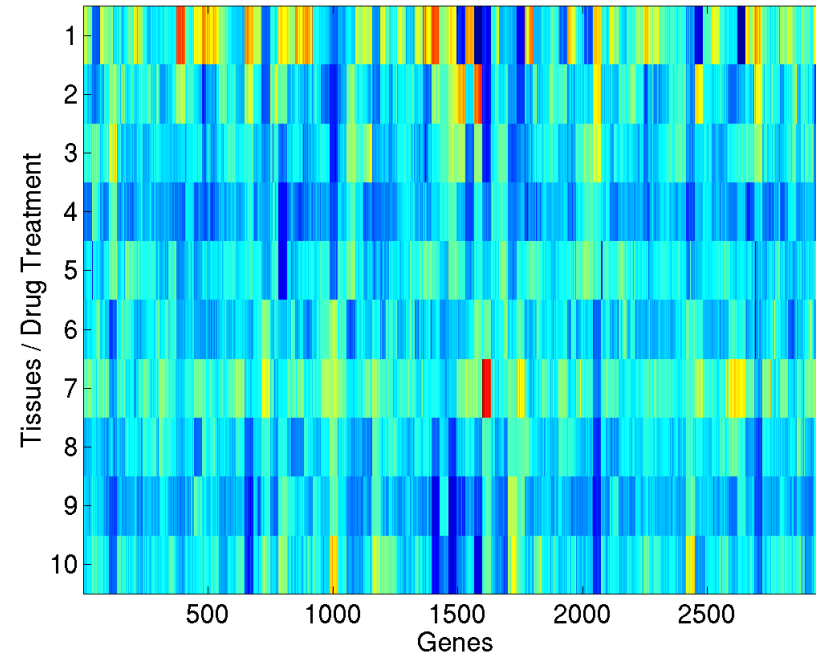
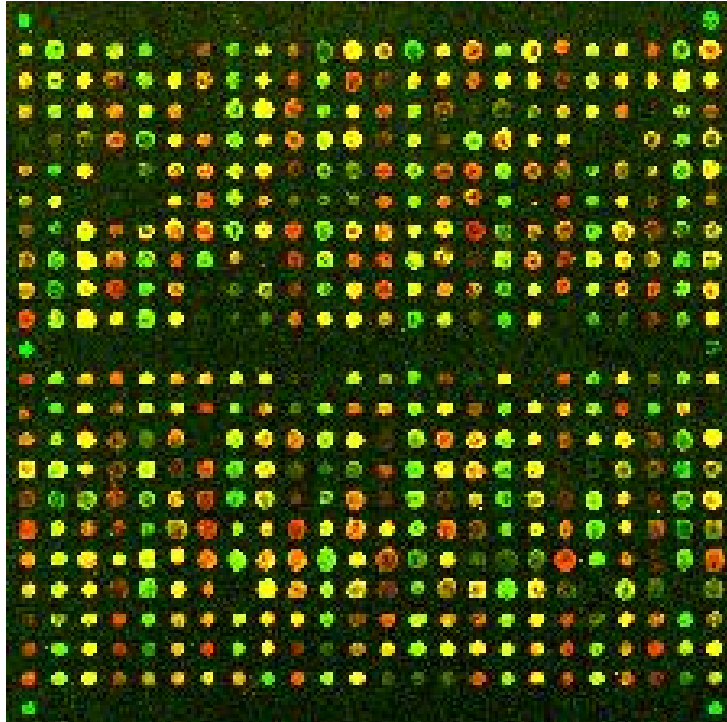


Medical diagnosis



(image from Kevin Murphy)

Gene microarrays and bioinformatics



Three Types of Learning

Imagine an organism or machine which experiences a series of sensory inputs:

$$x_1, x_2, x_3, x_4, \dots$$

Supervised learning: The machine is also given **desired outputs** y_1, y_2, \dots , and its goal is to learn to **produce the correct output** given a new input.

Unsupervised learning: The goal of the machine is to **build a model** of x that can be used for reasoning, decision making, predicting things, communicating etc.

Reinforcement learning: The machine can also produce **actions** a_1, a_2, \dots which affect the state of the world, and receives **rewards (or punishments)** r_1, r_2, \dots . Its goal is to learn to act in a way that **maximises rewards** in the long term.

Three Problems

Over the next three lectures we will cover these three topics:

- Classification
- Regression
- Clustering

We will make extensive use of probability, statistics, calculus and linear algebra.

Classification

We will represent data by vectors in some vector space.

Let \mathbf{x} denote a **data point** with elements $\mathbf{x} = (x_1, x_2, \dots, x_D)$

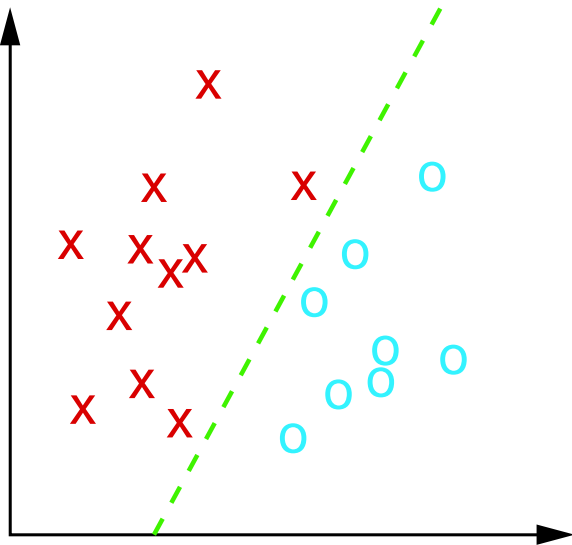
The elements of \mathbf{x} , e.g. x_d , represent measured (observed) **features** of the data point; D denotes the number of measured features of each point.

The **data set** \mathcal{D} consists of N pairs of data points and corresponding discrete class labels:

$$\mathcal{D} = \{(\mathbf{x}^{(1)}, y^{(1)}) \dots, (\mathbf{x}^{(N)}, y^{(N)})\}$$

where $y^{(n)} \in \{1, \dots, C\}$ and C is the number of classes.

The goal is to classify new inputs correctly (i.e. to *generalize*).



Examples:

- spam vs non-spam
- normal vs disease
- 0 vs 1 vs 2 vs 3 ... vs 9

Classification: Example Iris Dataset

3 classes, 4 numeric attributes, 150 instances

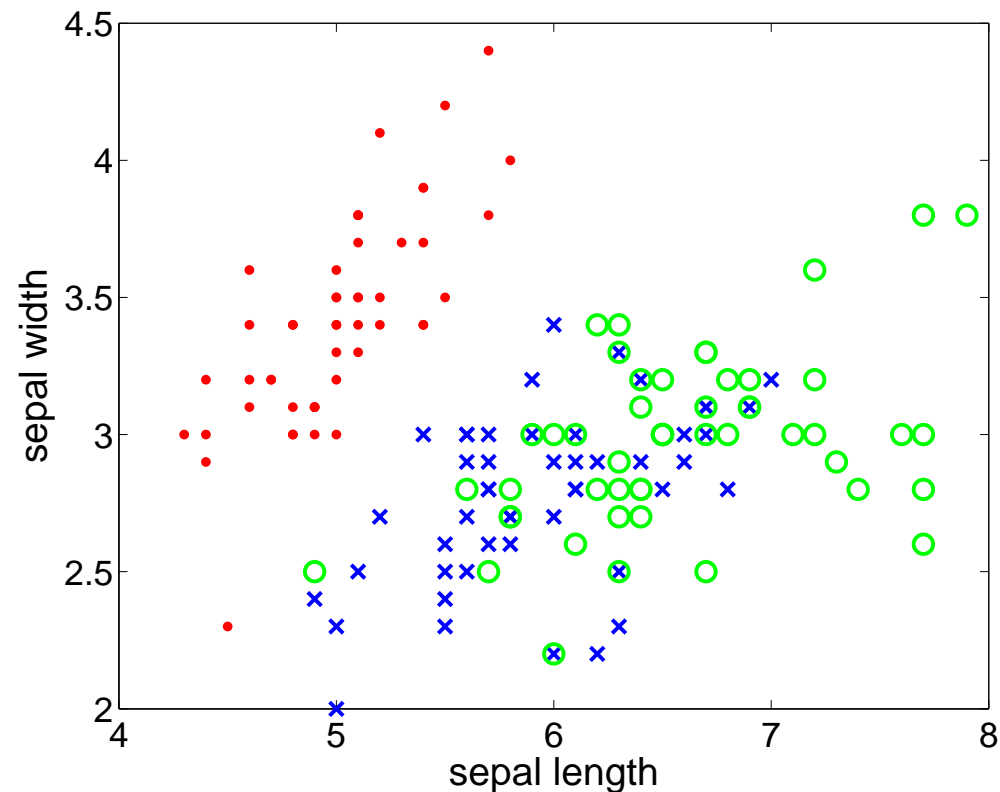
A data set with 150 points and 3 classes. Each point is a random sample of measurements of flowers from one of three iris species—setosa, versicolor, and virginica—collected by Anderson (1935). Used by Fisher (1936) for linear discriminant function technique.



The measurements are sepal length, sepal width, petal length, and petal width in cm.

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 \mathbf{x} denote an input point with elements $\mathbf{x} = (x_1, x_2, \dots, x_D)$. The elements of \mathbf{x} , e.g. x_d , represent measured (observed) features of the data point; D denotes the number of measured features of each point.

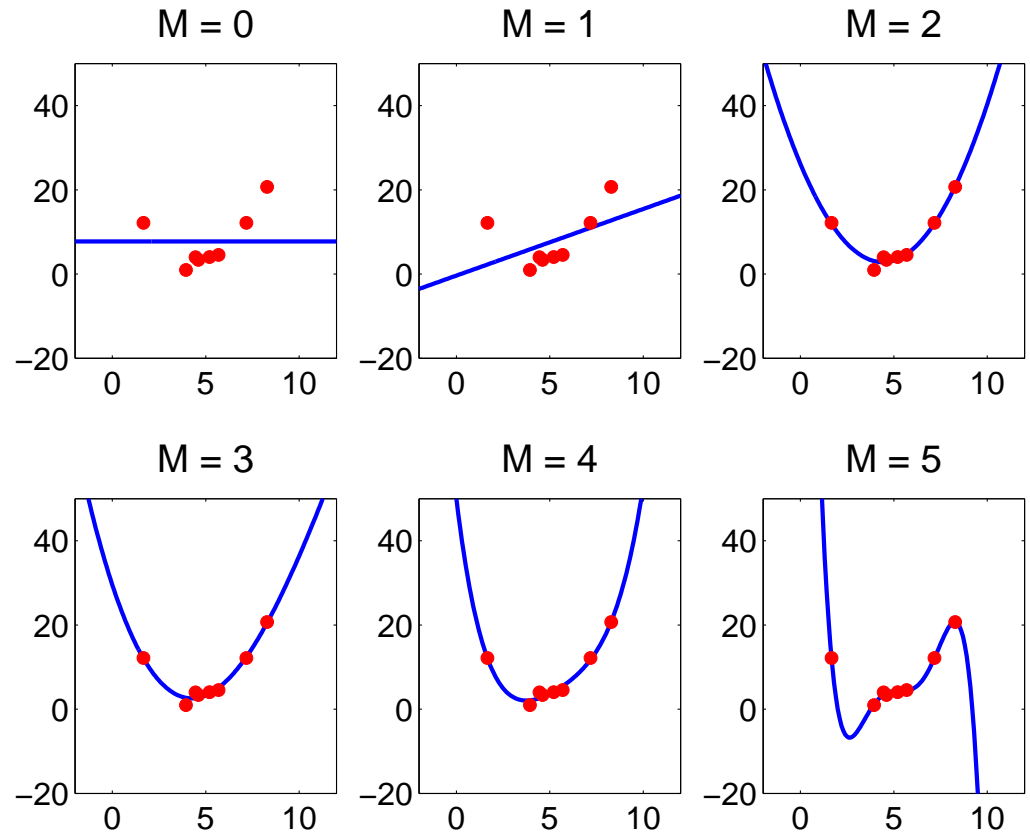
The data set \mathcal{D} consists of N pairs of inputs and corresponding real-valued outputs:

$$\mathcal{D} = \{(\mathbf{x}^{(1)}, y^{(1)}) \dots, (\mathbf{x}^{(N)}, y^{(N)})\}$$

where $y^{(n)} \in \mathfrak{R}$.

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

Linear and Nonlinear Regression



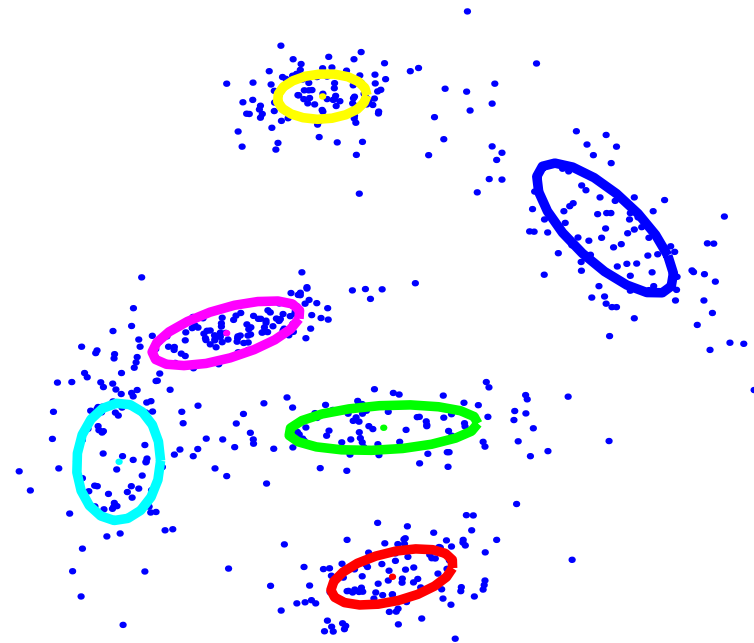
Clustering

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



Basic Rules of Probability

Let X be a random variable taking values x in some set \mathcal{X} .

Probabilities are non-negative $P(X = x) \geq 0 \forall x$.

Probabilities normalise: $\sum_{x \in \mathcal{X}} P(X = x) = 1$ for distributions if x is a discrete variable and $\int_{-\infty}^{+\infty} p(x) dx = 1$ for probability densities over continuous variables

The **joint probability** of $X = x$ and $Y = y$ is: $P(X = x, Y = y)$.

The **marginal probability** of $X = x$ is: $P(X = x) = \sum_y P(X = x, y)$, assuming y is discrete. I will generally write $P(x)$ to mean $P(X = x)$.

The **conditional probability** of x given y is: $P(x|y) = P(x, y)/P(y)$

Bayes Rule:

$$P(x, y) = P(x)P(y|x) = P(y)P(x|y) \quad \Rightarrow$$

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)}$$

Warning: I will not be obsessively careful in my use of p and P for probability density and probability distribution. Should be obvious from context.

Information, Probability and Entropy

Information is the **reduction of uncertainty**. How do we measure uncertainty?

Some axioms (informal):

- if something is certain its uncertainty = 0
- uncertainty should be maximum if all choices are equally probable
- uncertainty (information) should add for independent sources

This leads to a discrete random variable X having uncertainty equal to the **entropy** function:

$$H(X) = - \sum_{x \in \mathcal{X}} P(X = x) \log P(X = x)$$

measured in *bits* (**binary digits**) if the base 2 logarithm is used or *nats* (**natural digits**) if the natural (base e) logarithm is used.

Some Definitions and Intuitions

- **Surprise** (for event $X = x$): $-\log P(X = x)$
- **Entropy** = average surprise: $H(X) = -\sum_{x \in \mathcal{X}} P(X = x) \log_2 P(X = x)$
- **Conditional entropy**

$$H(X|Y) = -\sum_x \sum_y P(x, y) \log_2 P(x|y)$$

- **Mutual information**

$$I(X; Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X, Y)$$

- **Independent random variables**: $P(x, y) = P(x)P(y) \forall x \forall y$

Shannon's Source Coding Theorem

A discrete random variable X , distributed according to $P(X)$ has **entropy** equal to:

$$H(X) = - \sum_{x \in \mathcal{X}} P(x) \log P(x)$$

Shannon's source coding theorem: Consider a random variable X , with entropy $H(X)$. A sequence of n independent draws from X can be losslessly compressed into a minimum expected code of length $n\mathcal{L}$ bits, where $H(X) \leq \mathcal{L} < H(X) + \frac{1}{n}$.

If each symbol is given a code length $l(x) = -\log_2 Q(x)$ then the expected per-symbol length \mathcal{L}_Q of the code is

$$H(X) + KL(P\|Q) \leq \mathcal{L}_Q < H(X) + KL(P\|Q) + \frac{1}{n},$$

where the **relative-entropy** or **Kullback-Leibler divergence** is

$$KL(P\|Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)} \geq 0$$

Take home message: better probabilistic models \equiv more efficient codes

Some distributions

Univariate Gaussian density ($x \in \mathfrak{R}$):

$$p(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{(x - \mu)^2}{2\sigma^2} \right\}$$

Multivariate Gaussian density ($\mathbf{x} \in \mathfrak{R}^D$):

$$p(\mathbf{x}|\mu, \Sigma) = |2\pi\Sigma|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2}(\mathbf{x} - \mu)^\top \Sigma^{-1}(\mathbf{x} - \mu) \right\}$$

Bernoulli distribution ($x \in \{0, 1\}$):

$$p(x|\theta) = \theta^x (1 - \theta)^{1-x}$$

Discrete distribution ($x \in \{1, \dots, L\}$):

$$p(x|\theta) = \prod_{\ell=1}^L \theta_\ell^{\delta(x,\ell)}$$

where $\delta(a, b) = 1$ iff $a = b$, and $\sum_{\ell=1}^L \theta_\ell = 1$ and $\theta_\ell \geq 0 \forall \ell$.

Some distributions (cont)

Uniform ($x \in [a, b]$):

$$p(x|a, b) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

Gamma ($x \geq 0$):

$$p(x|a, b) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp\{-bx\}$$

Beta ($x \in [0, 1]$):

$$p(x|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}$$

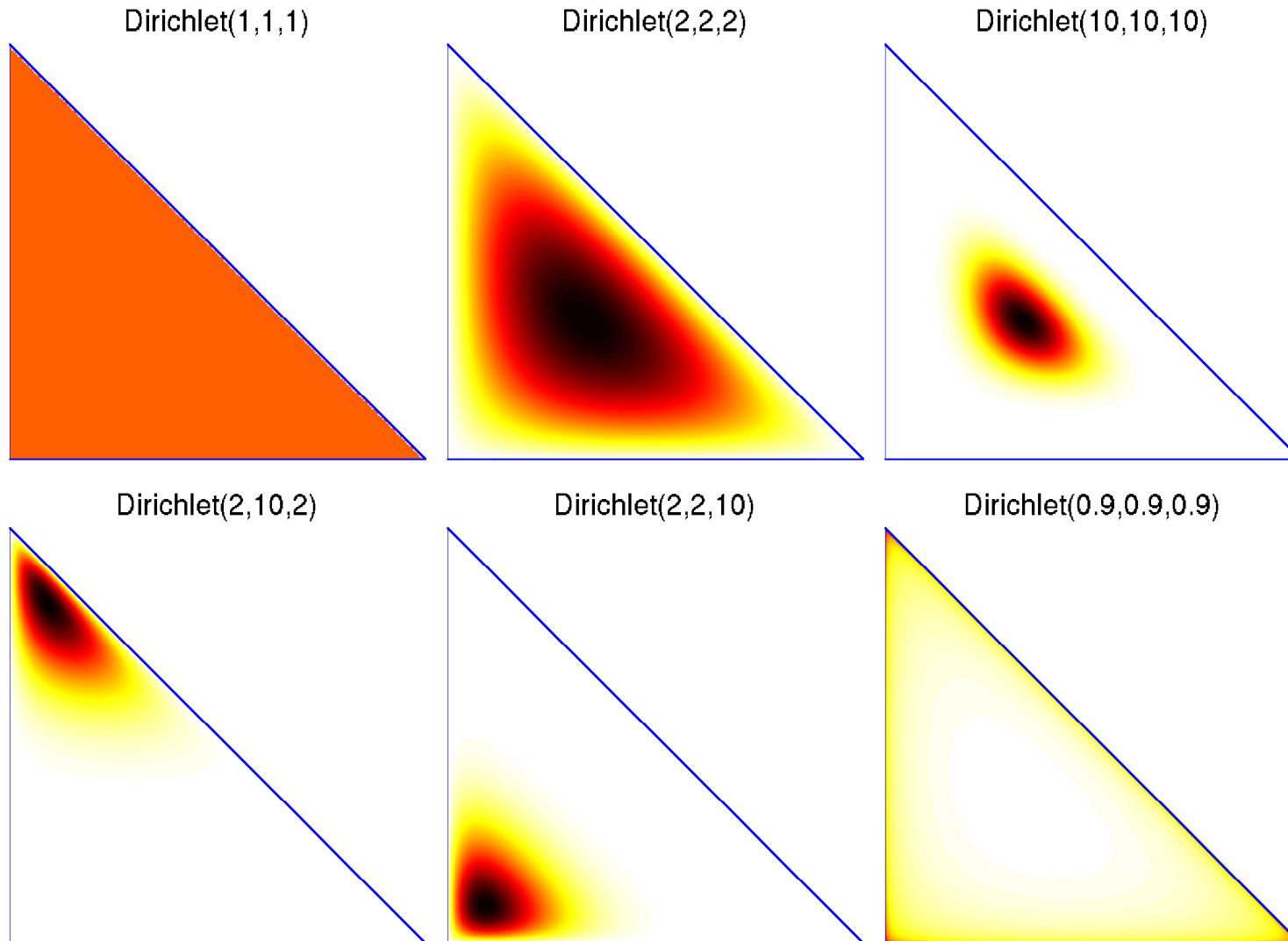
where $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt$ is the gamma function, a generalisation of the factorial:
 $\Gamma(n) = (n-1)!$.

Dirichlet ($\mathbf{p} \in \mathbb{R}^D$, $p_d \geq 0$, $\sum_{d=1}^D p_d = 1$):

$$p(\mathbf{p}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{d=1}^D \alpha_d)}{\prod_{d=1}^D \Gamma(\alpha_d)} \prod_{d=1}^D p_d^{\alpha_d-1}$$

Dirichlet Distributions

Examples of Dirichlet distributions over $\mathbf{p} = (p_1, p_2, p_3)$ which can be plotted in 2D since $p_3 = 1 - p_1 - p_2$:



Other distributions you should know about...

Exponential family of distributions:

$$P(\mathbf{x}|\boldsymbol{\theta}) = f(\mathbf{x}) g(\boldsymbol{\theta}) \exp \{ \boldsymbol{\phi}(\boldsymbol{\theta})^\top \mathbf{u}(\mathbf{x}) \}$$

where $\boldsymbol{\phi}(\boldsymbol{\theta})$ is the vector of *natural parameters*, \mathbf{u} are *sufficient statistics*

- Binomial
- Multinomial
- Poisson
- Student t distribution
- ...

End Notes

It is very important that you *understand* all the material in the following cribsheet:

<http://www.gatsby.ucl.ac.uk/~zoubin/course05/cribsheet.pdf>

Here is a useful statistics / pattern recognition glossary:

<http://research.microsoft.com/~minka/statlearn/glossary/>