COMP3206 / COMP6229 Machine Learning

(Undergraduate and MSc classes; taught together assessed (slightly) differently)

Week 1: Introduction

Mahesan Niranjan

School of Electronics and Computer Science
University of Southampton

Slides are prompts (for me); Notes are what you make, off the white-board and from textbooks during self study.

Autumn Semester 2016/17

Overview

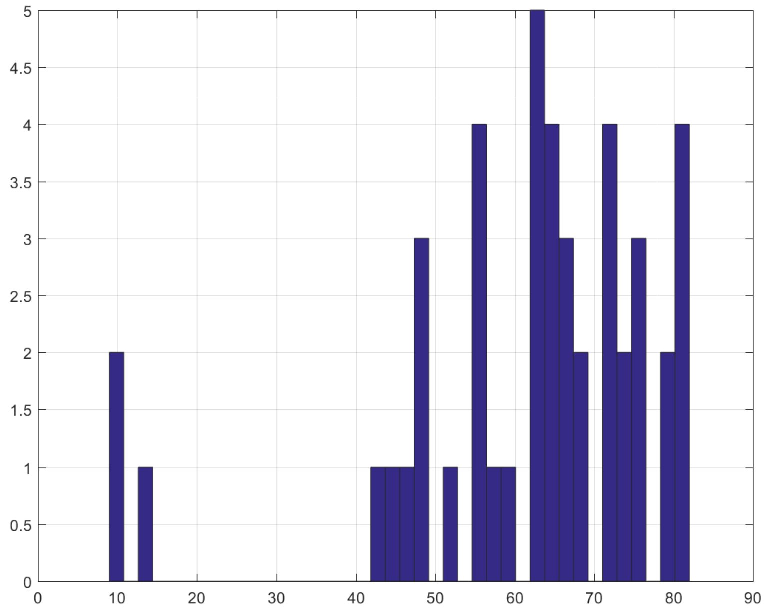
- Logistics
- Motivation
  - Some examples from my research
- Review of Mathematical Foundations
  - Linear Algebra
  - Calculus
  - Probability Theory / Statistics
  - Principles of Optimization

Emphasis is on *foundations* of the subject (mathematical and algorithmic). We will not do formal mathematics here, instead we develop an understanding of the concepts and tools.

Logistics

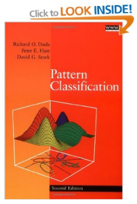
Assessment

Distribution of marks, COMP3206 2015/16

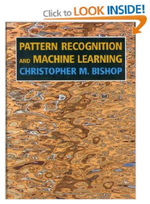


Difficult to fail this module, but please don't try!

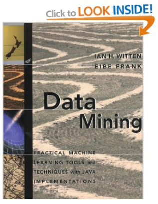
Good Books



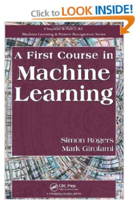
R.O.Duda, P.E.Hart & D.G.Stork
Pattern Classification



C.M. Bishop
Pattern Recognition and Machine Learning



I.H. Witten & E. Frank
Data Mining

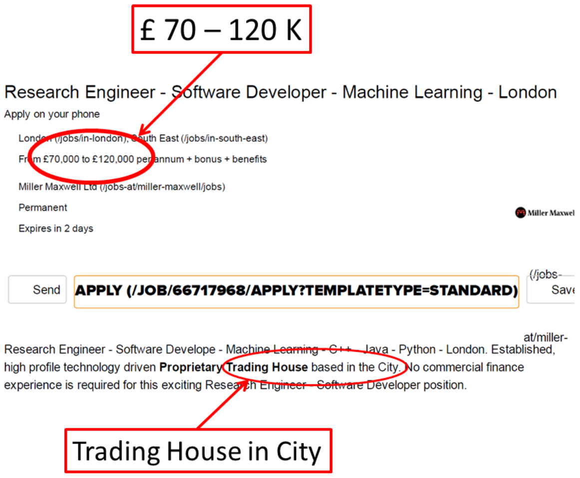


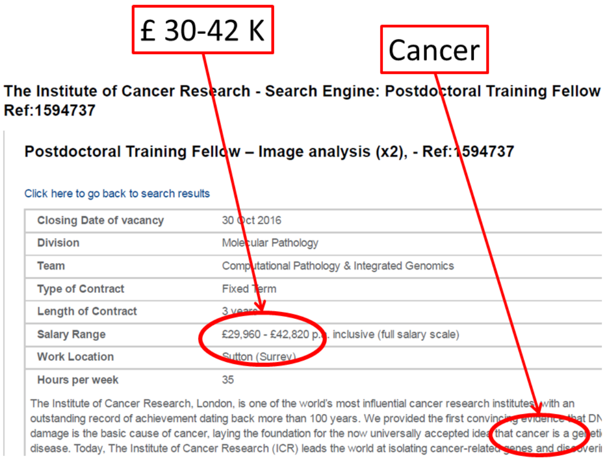
S. Rogers & M. Girolami
A First Course in Machine Learning

> *"There is nothing to be learnt from a professor, which is not to be met with in books"*
> *- David Hume (1711-1776)*

(**WikiPedia:** "Hume had little respect for the professors of his time [...] He did not graduate")

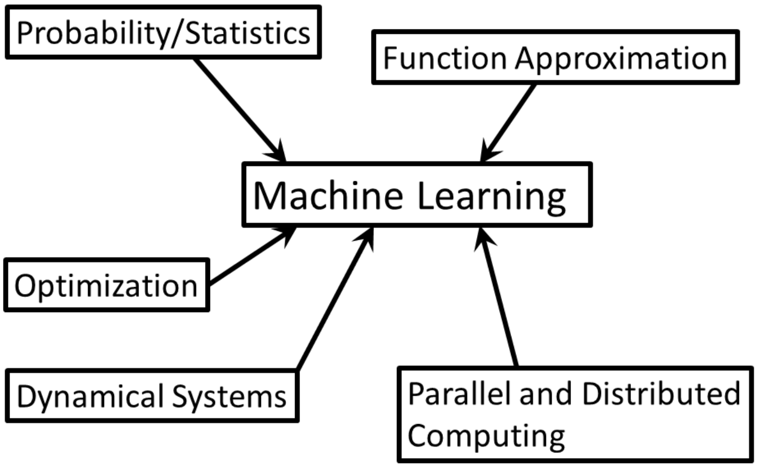
Machine Learning: Good employment prospects!





Standard disclaimers apply!

Machine Learning: Intellectually Enriching



- Mathematical / Statistical side of Artificial Intelligence
- Machine Learning draws from many fields

Machine Learning as Data-driven Modelling

Single-slide overview of the subject and challenging questions

Data	$\{\boldsymbol{x}_n, \; \boldsymbol{y}_n\}_{n=1}^{N} \qquad \{\boldsymbol{x}_n\}_{n=1}^{N}$
Function Approximator	$\boldsymbol{y} \; = \; f\,(\boldsymbol{x}, \, \boldsymbol{\theta}) \; + \; v$
Parameter Estimation	$E_0 \; = \; \sum_{n=1}^{N} \{\vert\vert \, \boldsymbol{y}_n \; - \; f\,(\boldsymbol{x}_n; \; \boldsymbol{\theta}) \; \vert\vert\}^2$
Prediction	$\hat{\boldsymbol{y}}_{N+1} \; = \; f\left(\boldsymbol{x}_{N+1}, \; \hat{\boldsymbol{\theta}}\right)$
Regularization	$E_1 \; = \; \sum_{n=1}^{N} \{\vert\vert \, \boldsymbol{y}_n \; - \; f\,(\boldsymbol{x}_n) \; \vert\vert\}^2 \; + \; g\,(\vert\vert\boldsymbol{\theta}\vert\vert)$
Modelling Uncertainty	$p\left(\boldsymbol{\theta}\vert \, \{\boldsymbol{x}_n, \boldsymbol{y}_n\}_{n=1}^{N}\right)$
Probabilistic Inference	$\boldsymbol{E}\,[g\,(\boldsymbol{\theta})] = \int g\,(\boldsymbol{\theta})\, p\,(\boldsymbol{\theta})\, d\boldsymbol{\theta} \; = \frac{1}{N_s} \sum_{n=1}^{N_s} g\,\left(\boldsymbol{\theta}^{(n)}\right)$
Sequential Estimation	$\boldsymbol{\theta}\,(n-1\vert n-1) \longrightarrow \boldsymbol{\theta}\,(n\vert n-1) \longrightarrow \boldsymbol{\theta}\,(n\vert n)$ Kalman & Particle Filters; Reinforcement Learning

Machine Learning

Many Interesting Problems (to me)

- Visual Scene Recognition
- Machine Translation
- Computational Biology
- Computational Finance
- Recommender Systems
- Physiological Signal Modelling

- “Big Data: ” Buzzword causing even more excitement!

- Make accurate predictions and make money!
- Make statements about the problem domain and become famous!

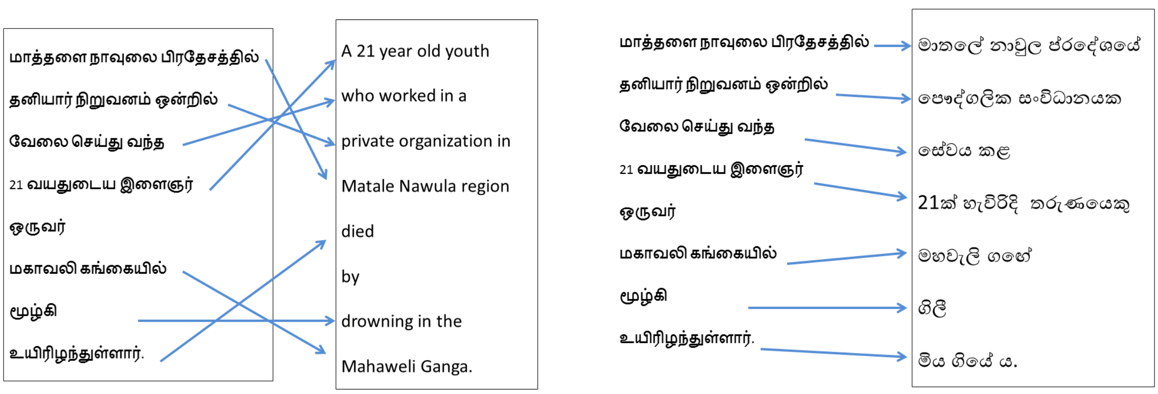
ECS: Advanced courses building on the foundations you will learn here:

- Advanced Machine Learning
- Computational Biology
- Computational Finance

Examples from my research

Example 1: Machine Translation

- Phrases move due to grammatical differences.
- Variability due to context of phrase.



- Not data rich (electronically available parallel corpora);
- Solution from active learning.

Examples from my research

Example 2: Computational Finance

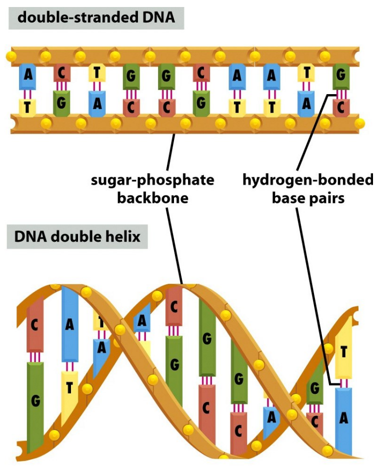
Constructing Sparse Portfolios

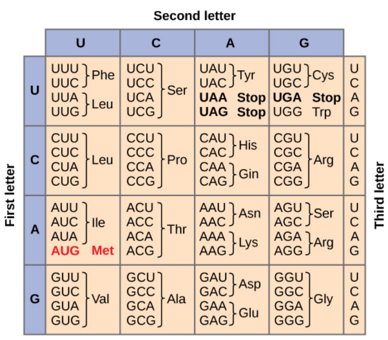
A. Takeda, M. Niranjan, J. Gotoh & Y. Kawahara (2013) "Simultaneous pursuit of out-of-sample performance and sparsity in index tracking portfolios", Computational Management Science **10**(1): 21-49.

See White Board

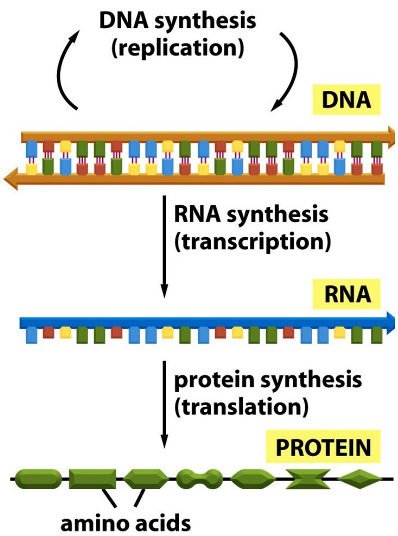
Molecular Biology

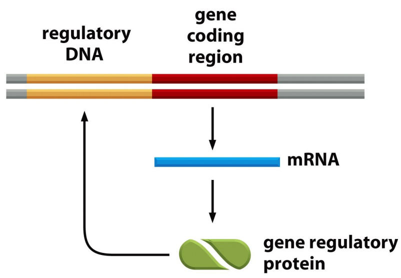
(Figures from: Alberts *et al.* Molecular Biology of the Cell)

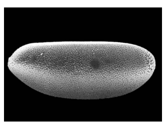


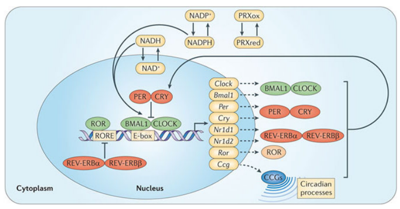


First letter	U	C	A	G	Third letter
U	UUU, UUC Phe; UUA, UUG Leu	UCU, UCC, UCA, UCG Ser	UAU, UAC Tyr; UAA Stop; UAG Stop	UGU, UGC Cys; UGA Stop; UGG Trp	U C A G
C	CUU, CUC, CUA, CUG Leu	CCU, CCC, CCA, CCG Pro	CAU, CAC His; CAA, CAG Gln	CGU, CGC, CGA, CGG Arg	U C A G
A	AUU, AUC, AUA Ile; AUG Met	ACU, ACC, ACA, ACG Thr	AAU, AAC Asn; AAA, AAG Lys	AGU, AGC Ser; AGA, AGG Arg	U C A G
G	GUU, GUC, GUA, GUG Val	GCU, GCC, GCA, GCG Ala	GAU, GAC Asp; GAA, GAG Glu	GGU, GGC, GGA, GGG Gly	U C A G



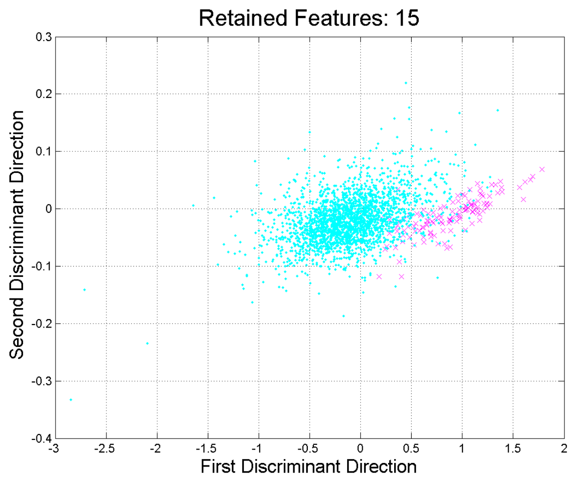


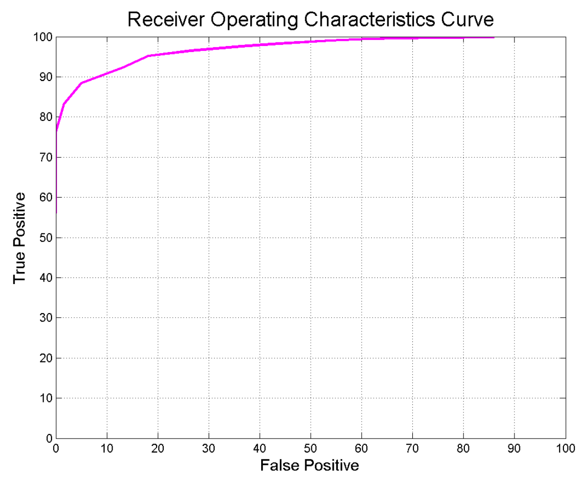




Examples from my research
Example 3: Classifying Gene Function

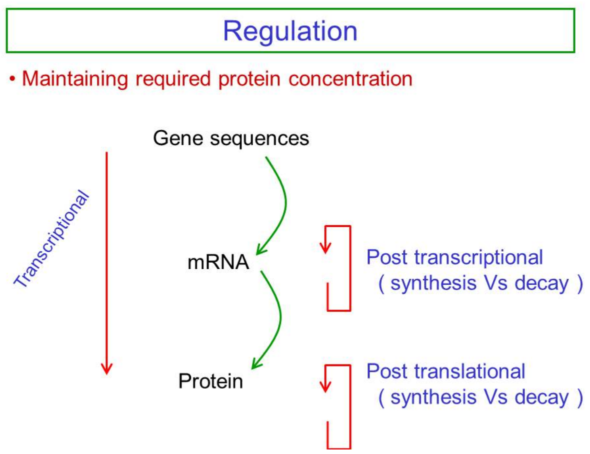
- 2000 yeast genes
- Observed (simultaneously) under 78 conditions
- Some have a specific function; others not

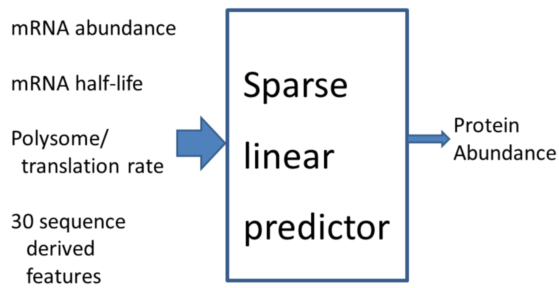




See `MATLAB` Demo

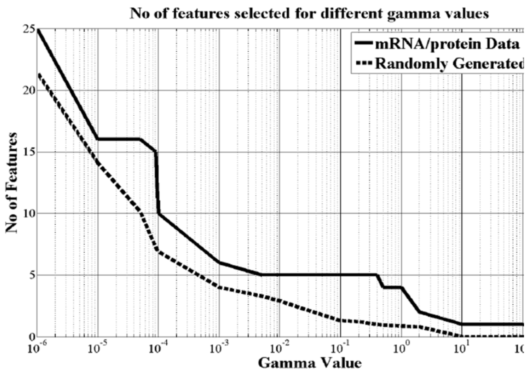
Example 4: Regulation of Protein Concentrations [Yawwani Gunawardana]

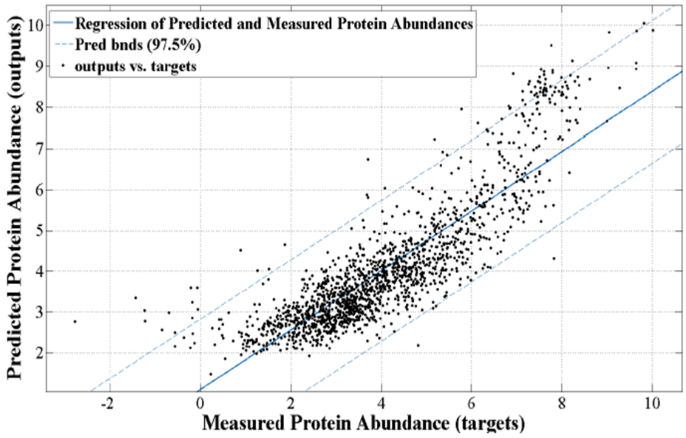


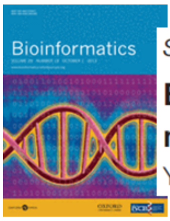


- Set up a predictor of protein concentration
- Sparse model selects relevant features
- Outliers $\Longrightarrow$ post-translationally regulated proteins

Transcriptome $\longrightarrow$ Proteome [Yawwani Gunawardana]





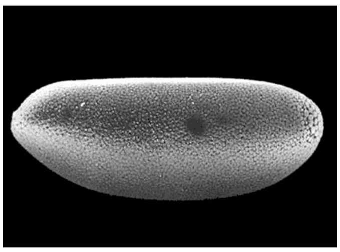


Systems biology *doi:10.1093/bioinformatics/btt537*

Bridging the gap between transcriptome and proteome measurements identifies post-translationally regulated genes

Yawwani Gunawardana and Mahesan Niranjan*

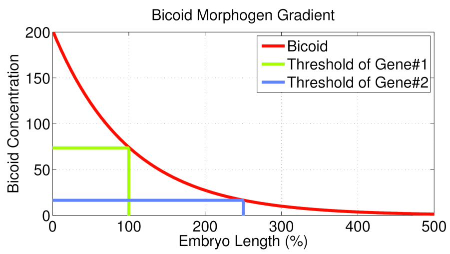
Example 5: Morphogen Propagation in Development



A. Turing **C. Nüsslein-Volhard**

$$\frac{\partial}{\partial t} M(x,t) = D \frac{\partial^2}{\partial x^2} M(x,t) - \tau_p^{-1} M(x,t) + \mathbf{S}(\mathbf{x},\mathbf{t})$$

B. Houchmandzadeh *et al.* (2007), *Nature*

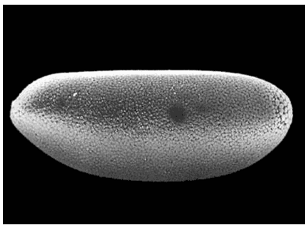


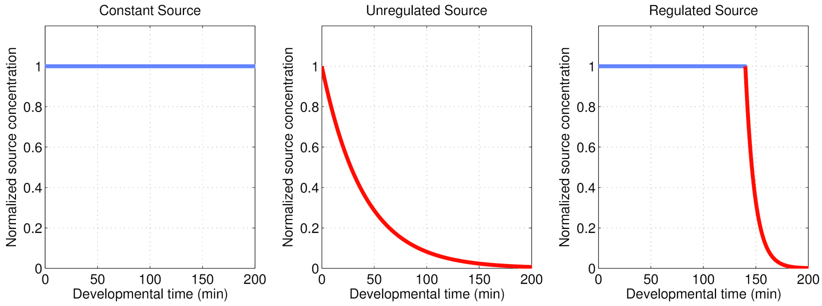
Establishment of developmental precision and proportions in the early *Drosophila* embryo

Bahram Houchmandzadeh*†, Eric Wieschaus* & Stanislas Leibler*‡§

** Howard Hughes Medical Institute, Department of Molecular Biology, Princeton University, Princeton, New Jersey 08544, USA*
† CNRS, Laboratoire de Spectrometrie Physique, BP87, 38402, St-Martin D'Heres Cedex, France

Is Maternal mRNA Stability Regulated? Wei Liu





OPEN ACCESS Freely available online PLoS one

The Role of Regulated mRNA Stability in Establishing Bicoid Morphogen Gradient in *Drosophila* Embryonic Development

Wei Liu*, Mahesan Niranjan

School of Electronics and Computer Science, University of Southampton, Southampton, United Kingdom

Abstract

The Bicoid morphogen is amongst the earliest triggers of differential spatial pattern of gene expression and subsequent cell

BIOINFORMATICS Vol. 00 no. 00 2005 Pages 1–6

Gaussian process modelling for *bicoid* mRNA regulation in spatio-temporal Bicoid profile

Wei Liu and Mahesan Niranjan*

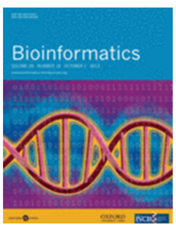
School of Electronics and Computer Science, University of Southampton, Southampton, UK

Example 6: Systems Level Modelling [Xin Liu]

$$\dot{D}_t = K_d \frac{S_t}{1 + \frac{K_s D_t}{1 + K_u U_f}} - \alpha_d D_t$$

$$\dot{S}_t = \eta(t) - \alpha_0 S_t - \alpha_s \frac{\frac{K_s D_t}{1 + K_u U_f}}{1 + \frac{K_s D_t}{1 + K_u U_f}} S_t$$

$$\dot{U}_f = K(t)[P_t - U_f] - [K(t) + K_{fold}] D_t$$



Systems biology *Vol. 28 no. 11 2012, pages 1501–1507*

State and parameter estimation of the heat shock response system using Kalman and particle filters

Xin Liu and Mahesan Niranjan*

School of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, UK

Associate Editor: Trey Ideker

… and now for more serious matters!

Rapid Review of Foundations

- Linear Algebra
- Calculus
- Optimization
- Probabilities

- This is not a course on any of the above!
- We need tools from these topics.
- Quickly review what we need today, and will return to each topic as and when we need them (in just about enough depth) to understand machine learning.

Linear Algebra: Vectors and Matrices

- Vectors and matrices as collections of numbers

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \qquad A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1d} \\ a_{11} & a_{12} & \dots & a_{1d} \\ \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nd} \end{bmatrix}$$

- Operations on collections on numbers
  - Scalar product

$$\boldsymbol{w} \bullet \boldsymbol{x} = \sum_{i=1}^{n} w_i \, x_i$$

- With useful geometric insights
  - Angle between vectors in n dimensional space

$$\boldsymbol{w} \bullet \boldsymbol{x} = |\boldsymbol{w}| \, |\boldsymbol{x}| \, \cos(\theta)$$

Vectors

- Linear independence
  ... set of p vectors $\boldsymbol{x}_j$, $j = 1, ..., p$

$$\sum_{i=j}^{p} \alpha_j \, \boldsymbol{x}_j \; = \; \boldsymbol{0}$$

  ... only solution is all $\alpha_j \; = \; 0$
  ... no vector in the set can be expressed as a linear combination of the others.
- Scalar product as projection: projection of vector $\boldsymbol{x}$ on a direction specified by vector $\boldsymbol{u}$

$$\frac{\boldsymbol{x} \bullet \boldsymbol{u}}{|\, \boldsymbol{u} \,|} \boldsymbol{u}$$

  ... we will also write this as

$$\frac{\boldsymbol{x}^T \, \boldsymbol{u}}{|\, \boldsymbol{u} \,|} \boldsymbol{u}$$

Matrices

- Simple operations *e.g.* addition: $[A + B]_{ij} \; = \; [A]_{ij} + [B]_{ij}$; transpose: $[A]_{ij}^T \; = \; [A]_{ji}$; multiplication by a scalar: $[\alpha A]_{ij} \; = \; \alpha \, [A]_{ij}$
- Matrix multiplication:

$$[A \, B]_{ij} \; = \; \sum_{k=1}^{n} [A]_{ik} \, [B]_{kj}$$

- $(AB)^T \; = \; A^T \, B^T$
- Square: number of rows = number of columns
- Symmetric: $A^T \; = \; A$
- Identity matrix: I diagonal elements 1, off diagonals 0.
- Determinant: $\det \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \; = \; a_{11} \, a_{22} - a_{21} \, a_{12}$
- Trace: $\mathrm{trace}(A) \; = \; \sum_{i=1}^{n} a_{ii}$

Linear transformation

- $\boldsymbol{y} = A\,\boldsymbol{x}$
- Rotation: $R = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$
  $R\,\boldsymbol{x}$ rotates $\boldsymbol{x}$ by angle θ radians.
  Magnitude of $\boldsymbol{x}$ does not change.
- A special relationship between a square matrix A and vector $\boldsymbol{x}$

$$A\,\boldsymbol{x} = \lambda\,\boldsymbol{x}$$

  Magnitude scales, but no rotation... have you come across this?
  Eigenvalues, eigenvectors
  Found by

$$\det(A - \lambda I) = 0$$

- **Homework:** Look up if the following are true and how they are proved.
  - $\det(A) = \prod_{i=1}^{n} \lambda_i$
  - $\mathrm{trace}(A) = \sum_{i=1}^{n} \lambda_i$
- Real symmetric matrix $\boldsymbol{A} = \boldsymbol{U}\,\boldsymbol{D}\,\boldsymbol{U}^T$
  Columns of $\boldsymbol{U}$ orthogonal.
- More advanced (very powerful) topic: Singular value decomposition (SVD)

Rapid Review of Foundations II: Calculus

- Function $y = f(\boldsymbol{x})$
  - Derivative $\frac{dy}{dx}$ is gradient/slope;
  - Integral $\int_{x=a}^{x=b} f(x)dx$ is area under the curve.
- Function of several variables $y = f(x_1, x_2 \ldots, x_p)$
  - Partial derivatives $\frac{\partial f}{\partial x_i}$: Differentiate with respect to x_i pretending all other variables remain constant.
  - Gradient vector

$$\boldsymbol{\nabla f} = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_p} \end{pmatrix}$$

Homework: Consider $f = \boldsymbol{x}^t \boldsymbol{A} \boldsymbol{x}$, $A = A^T$; $\boldsymbol{\nabla f} = 2\boldsymbol{A}\boldsymbol{x}$. Using scalars in two dimensions, *i.e.* $\boldsymbol{x} = [x_1\ x_2]^T$ and $\boldsymbol{A}$ to contain elements $\begin{bmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{bmatrix}$, verify the claim. Writing out the algebra helps in learning!

Rapid Review of Foundations III: Optimization

- Unconstrained optimization: $\min f(\boldsymbol{x})$
- Constrained optimization:

$$\begin{aligned}&\min f(\boldsymbol{x})\\&\text{subject to } g_i(\boldsymbol{x}) \leq b_i, \;\; i = 1, 2, .., m\end{aligned}$$

- Gradient and Hessian:

$$\boldsymbol{\nabla f} = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_p} \end{pmatrix} \qquad \boldsymbol{H} = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_p} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial^2 x_2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_p} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial^2 f}{\partial x_p \partial x_1} & \frac{\partial^2 f}{\partial x_p \partial x_2} & \cdots & \frac{\partial^2 f}{\partial^2 x_p} \end{bmatrix}$$

Optimizations (cont'd)

- Example: Gradient descent algorithm

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \eta \, \boldsymbol{\nabla f}$$

- Newton's Method

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - H^{-1} \, \boldsymbol{\nabla f}$$

- Example: Lagrange Multipliers

$$\begin{aligned}&\min f(\boldsymbol{x})\\&\text{subject to } g_i(\boldsymbol{x}) \leq b_i, \;\; i = 1, 2, .., m\end{aligned}$$

$$F(\boldsymbol{x}, \lambda) = f(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i \, [b_i - g_i(\boldsymbol{x})]$$

- We will use various optimization algorithms in this module (later in the coursework).
- Advanced $\boxed{\textbf{Homework:}}$ Search for `CVX` Disciplined Convex Programming and have a rough read.

Rapid Review of Foundations IV: Probabilities

- Discrete probabilities $P[X]$
- Continuous densities $p(\boldsymbol{x})$
- Joint $P[X, Y]$; Marginal $P[X]$; Conditional $P[X|Y]$

$$\begin{aligned} P[Y\,|\,X] &= \frac{P[X|Y]\; P[Y]}{P[X]} \\ P[X] &= \sum_Y P[X|Y]\; P[Y] \end{aligned}$$

$$\begin{aligned} P[X, Y] &= P[X|Y]\; P[Y] \\ P[X] &= \sum_y P[X, Y] \\ &= \sum_Y P[X|Y]\; P[Y] \end{aligned}$$

Gaussian Densities: Univariate and Multivariate

- Univariate Gaussian

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

What are properties we know? $\boxed{\textbf{Homework:}}$ Draw sketches for different values of m and σ.

- Multivariate Gaussian

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}(\det \boldsymbol{C})^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{m})^t\, \boldsymbol{C}^{-1}\,(\boldsymbol{x}-\boldsymbol{m})\right\}$$

Mean $\boldsymbol{m}$ is a vector
Covariance, C, matrix: symmetric, positive semi definite!
$\boxed{\textbf{Homework:}}$ Draw sketches for different values of $\boldsymbol{m}$ and $\boldsymbol{C}$

$$\boxed{\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{m}, \boldsymbol{C}),\ \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} \implies \boldsymbol{y} \sim \mathcal{N}(\boldsymbol{A}\boldsymbol{m}, \boldsymbol{A}\boldsymbol{C}\boldsymbol{A}^T)}$$

Estimation

Univariate Mean	$\widehat{m} \;=\; \frac{1}{N}\sum_{n=1}^{N} x_n$
Univariate Covariance	$\widehat{\sigma} \;=\; \frac{1}{N}\sum_{n=1}^{N}(x_n - \widehat{m})^2$
Multivariate Mean	$\widehat{\boldsymbol{m}} \;=\; \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{x}_n$
Covariance Matrix	$\widehat{C} \;=\; \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_n - \boldsymbol{m})(\boldsymbol{x}_n - \boldsymbol{m})^T$

- These are known as maximum likelihood estimates (see later).
- **Homework:** Have you noticed there are two buttons in a calculator for estimating standard deviation, denoted σ_n and σ_{n-1}? Find out why.

What Next?

Weeks 2 and 3: Pattern Classification

- Classifying based on $P\left[\omega_j \,|\, \boldsymbol{x}\right]$
- Optimal classifier for simple distributions
- Linear classifier – when is it optimal?
- Distance based classifiers
  - Nearest Neighbour classifier
  - Mahalanobis distance
- Linear discriminant analysis
  - Fisher LDA
- Classifier Performance
  - Receiver Operating Characteristics (ROC) Curve
- Perceptron learning rule and convergence

See sketches on whiteboard – these illustrations are important