Royal Institute of Technology

DD2447 STAT. METH. IN CS HT 2014

- Lecture 1-Intro

"Chapter 1"

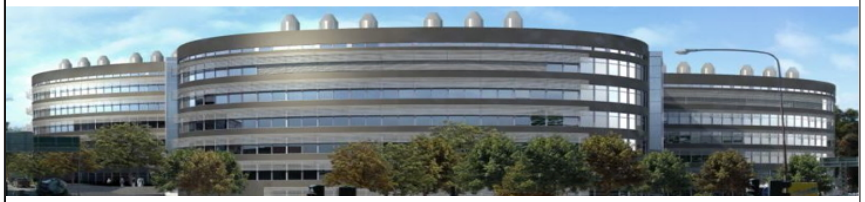
Royal Institute of Technology

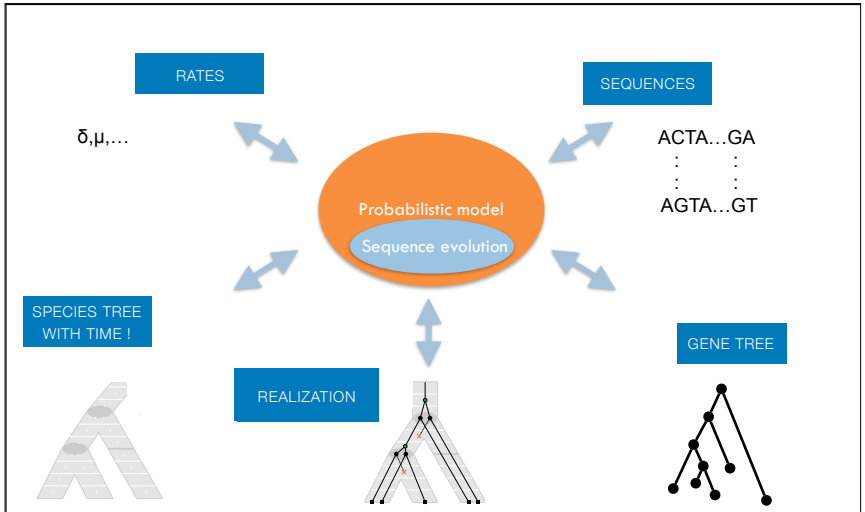
SciLifeLab

Computational Biology

Machine Learning – a main tool

Jens Lagergren





DLRS (DELIRIOUS)

THE TEXT BOOK

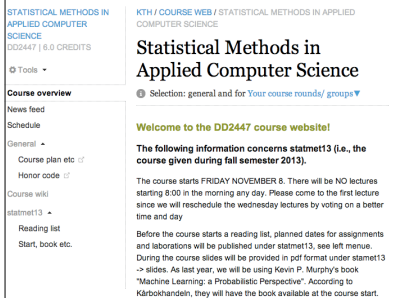
- New
- Key machine learning researcher
- Comprehensive

- Exercises – not graded
- Homework (3) – Individual, handed in and graded
  - out – nov 14, nov 28, dec 12
- Project (1) – Individual, handed in and graded
  - out – dec 12

ADMINISTRATION - EXAMINATION

1. Use latex, you will have to use it later anyway (this is a recommendation)
2. Always include your name in the file with the solutions
3. Make each step in a derivation explicit

KRISTOFFER'S COMMENTS



INTERACTION

- Lectures
- Solutions: mail, Scilife, or lectures
- KTH social www.kth.se/social/course/DD2447/

WHY MACHINE LEARNING?

- The era of big data
  - Transaction data for large corporations
    - Walmart has 2.5 petabytes ($2.5*10^{15}$) and handle 1M/hour
  - A human genome is 6 Gb
  - Meta-genomics
    - Baltic sea, hot-springs, your gut
    - A coke can can contain more microbes than there are north-americans

STATISTICS, ML, DATA MINING?

- Statistics — closed formulas
- Statistical ML — computational methods
  - they share models, probability
- We will often apply a Bayesian approach
- Data mining — less mathematical

SOME STUFF I EXPECT YOU TO KNOW

- Supervised learning
- Unsupervised learning
- Training & testing

We have the answer ← means yes otherwise no

SUPERVISED LEARNING

UNSUPERVISED LEARNING

We do not have any correct answer

Find classes or groups

SOME STUFF I EXPECT YOU TO KNOW

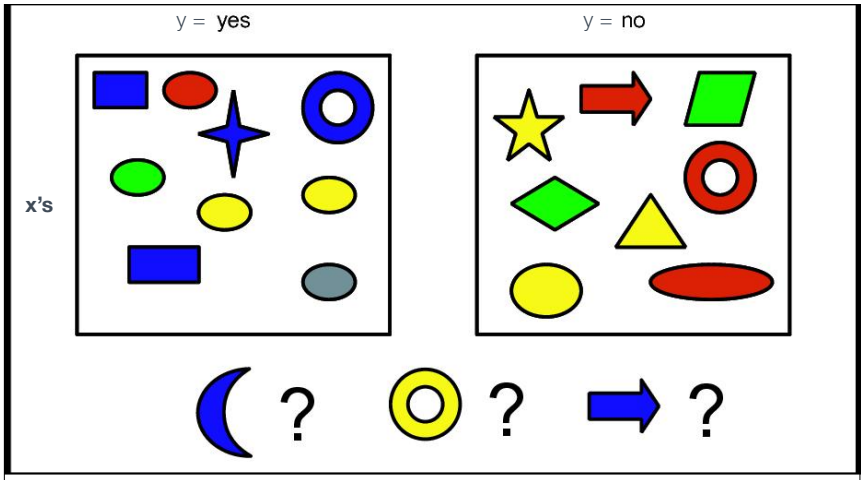
- Supervised learning
  - $D=\{(\mathbf{x}_i, y_i)\}$
  - y_i response variable (output variable)
  - $\mathbf{x}\_i$ features (input variables)
  - classification & regression
- Unsupervised learning
  - find the right y_i's, or
  - find the right dependencies between the variables of $\mathbf{x}_i$

We have the answer ← means yes otherwise no

BINARY CLASSIFICATION

- ← means apple, ← means pear, otherwise other

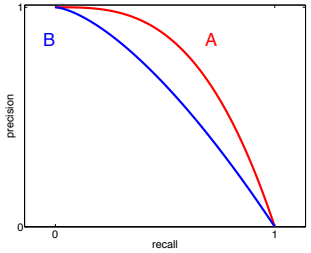
CATEGORICAL CLASSIFICATION



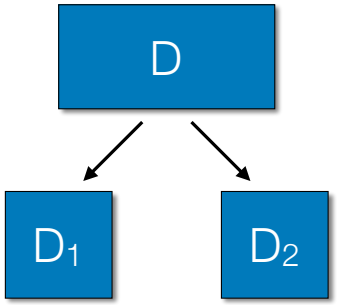
CLASSIFICATION

The probability of an answer

$$p(\text{ yes} \mid \text{blue moon}, \mathcal{D})$$

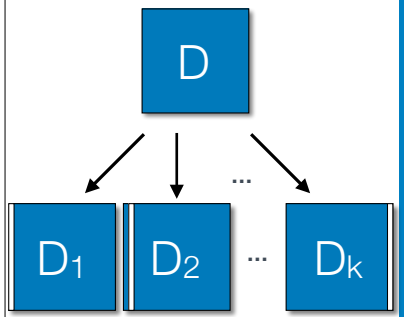


PROBABILISTIC PREDICTION



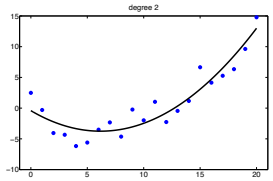
TRAINING – TEST

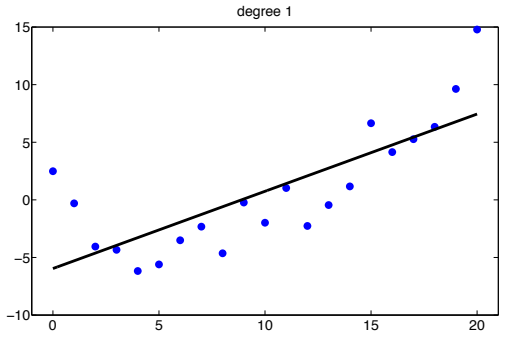
- Split data D into
  - training D_1
  - test D_2
- Use misclassification rate
- Problem: overfitting



CROSS-VALIDATION

- Leave-one-out
  - let $D_i = D \setminus x_i$
  - test on x_i
- Use misclassification rate
- Redundancy, overlap





- Size
- Floor
- Location

REGRESSION

- Googles smartass (ad selection system)
  - personalisation
- Mail filter
- Handwriting recognition
  - MNIST a dataset with 60000 training and 6.000 test images (of digits 0,.., 9)
- Face recognition
- Differentiate between setosa, versicolor, and virginica ???

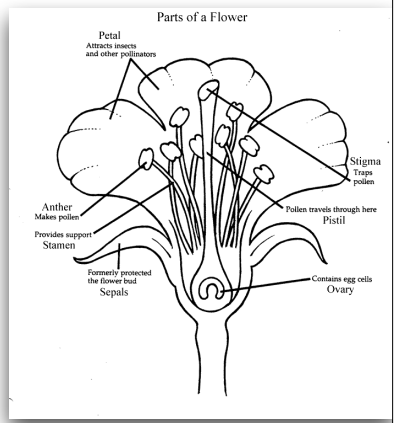
REAL WORLD APPLICATIONS

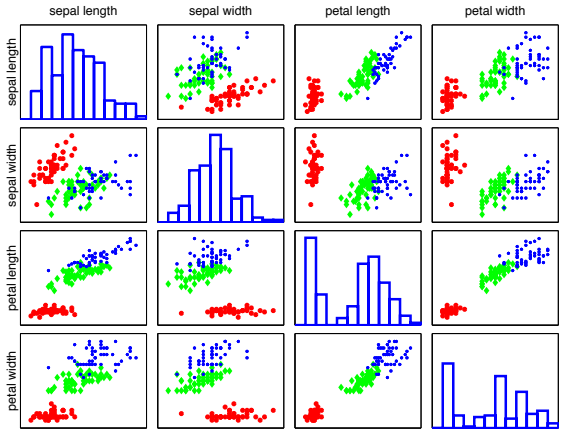
SETOSA, VERSICOLOR, AND VIRGINICA

HIGH SCHOOL BIOLOGY

Petal – attracting insects and pollinators

Sepal – formerly protected the bud



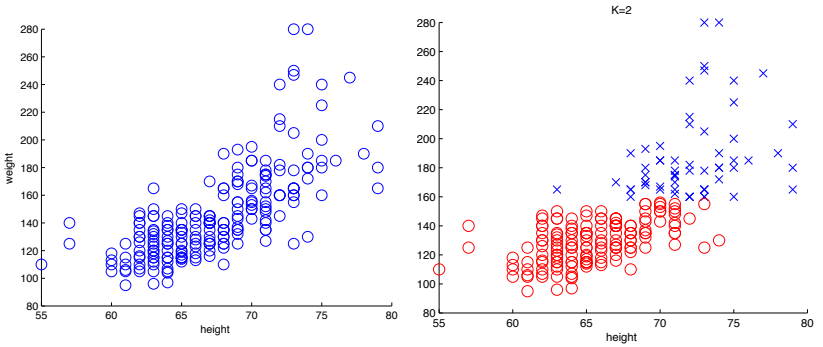


SETOSA, VERSICOLOR, AND VIRGINICA

★ sepal length, sepal width, petal length, and petal width.

UNSUPERVISED LEARNING

- We do not have any correct answer
- Find classes or groups

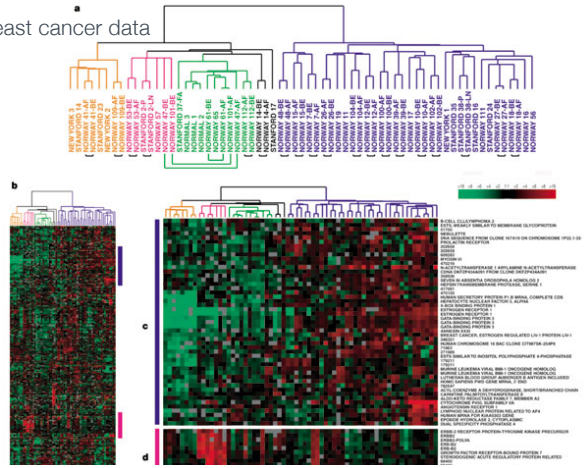


- Each subset should contain similar points
- Pairs of subsets should have dissimilar points.

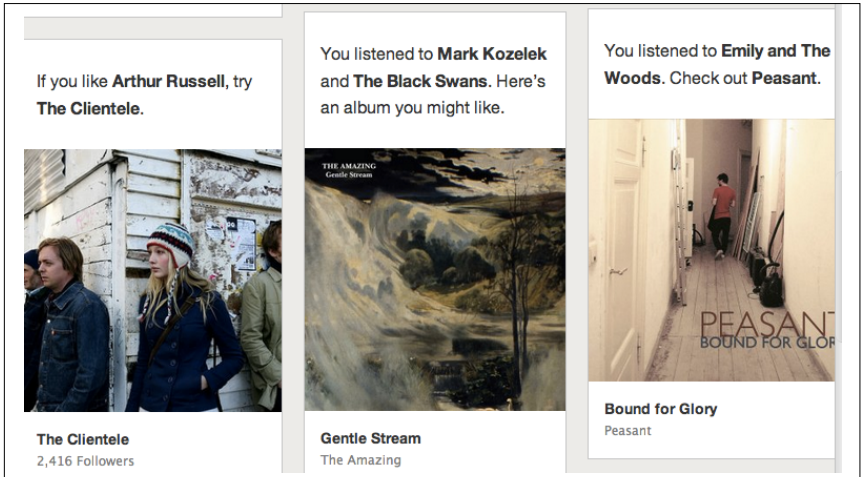
CLUSTERING

Molecular breast cancer data

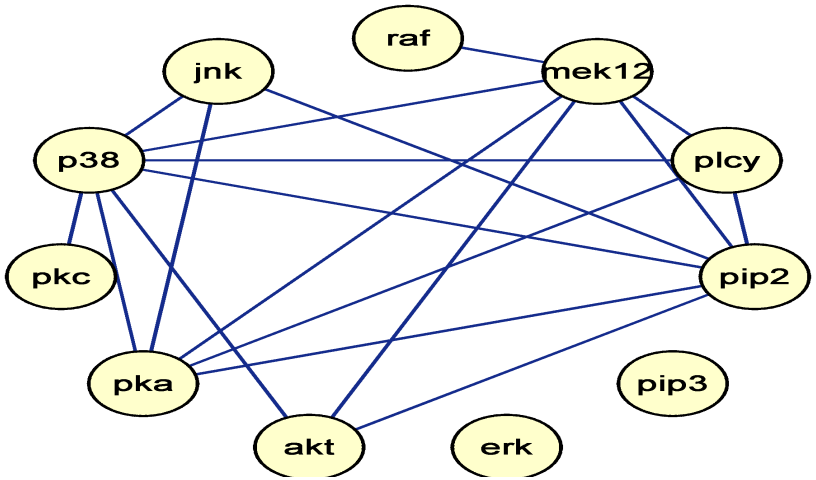
5 subtypes



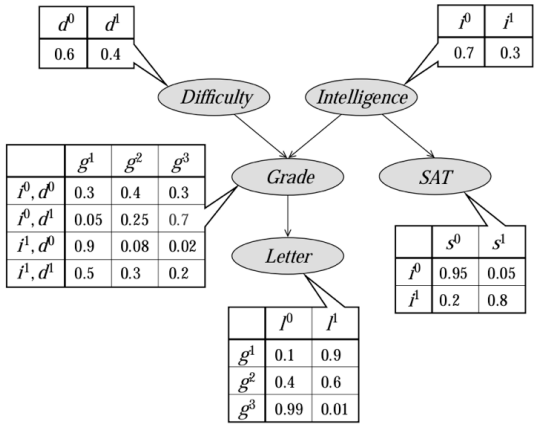
HIERARCHICAL CLUSTERING



COLLECTIVE FILTERING



DISCOVERING GRAPH STRUCTURE

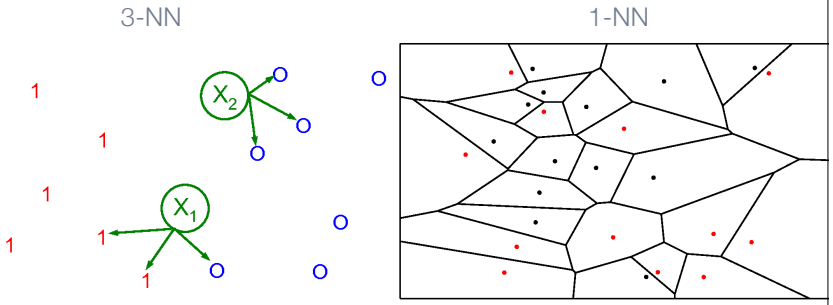


DIRECTED GRAPHICAL MODELS

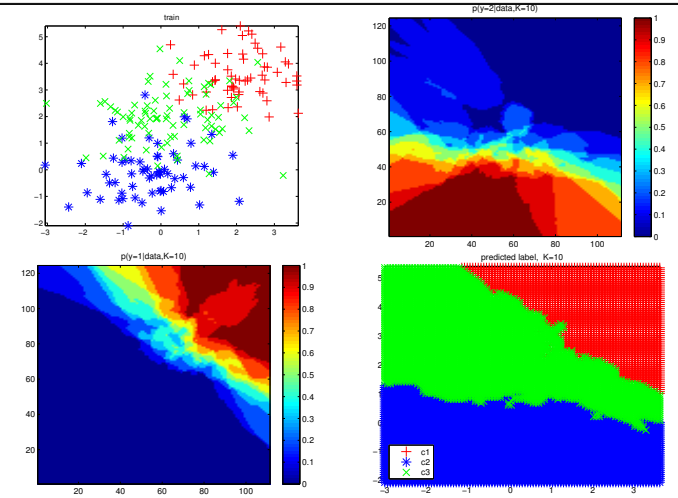
PARAMETRIC VS NON-PARAMETRIC

- Constant # parameters – parametric model (any distribution)
- Representation grows with data – non-parametric model

K-NN $$p(y = c|\boldsymbol{x}, \mathcal{D}, K) = \frac{1}{K} \sum_{n \in N_K(\boldsymbol{x}, \mathcal{D})} I(y_n = c)$$



K-NEAREST NEIGHBOUR (K-NN)

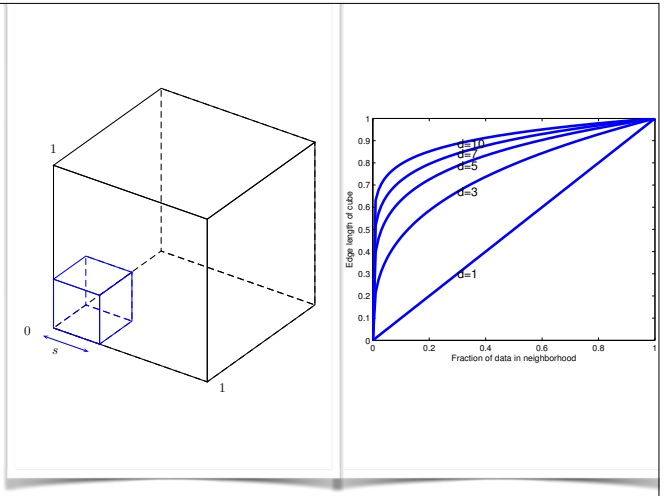


K-NEAREST NEIGHBOUR

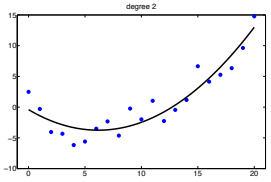
Assume small Euklidean is necessary

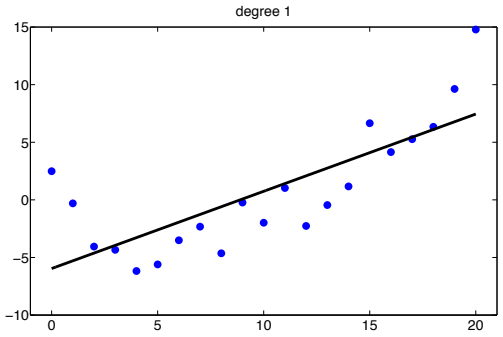
Fraction of cube with side, say, 1/2 is $(1/2)^D$

So is expected fraction of points in the smaller cube



THE CURSE





- Size
- Floor
- Location

REGRESSION

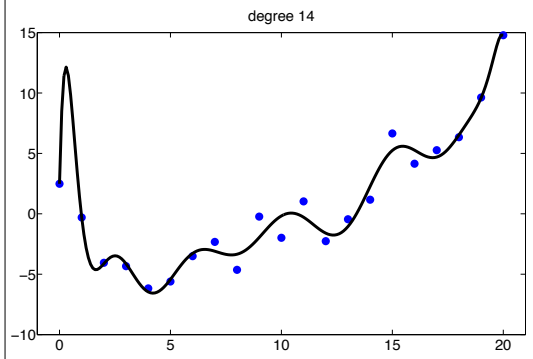
LINEAR REGRESSION

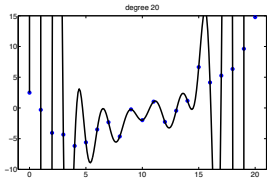
- Fit line to data, i.e., find **w** and use

$$y(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} + \epsilon = \sum_{d=1}^{D} w_d x_d + \epsilon$$

- **w** – model weight vector, β in statistiks
- ε – residual error

$$\epsilon \sim \mathcal{N}(\mu, \sigma^2)$$





OVERFITTING

BAYESIAN

Fair (F)

$P(i|F) = \frac{1}{2}, \forall i$

Used 99% FEL

Biased/loaded (B)

$p(6|B) = \frac{1}{2}$ FEL

$P(i|B) = \frac{1}{2}, \forall 1 \leq i \leq 5$ FEL

Used 1%

- A is the event 6,6,6
- Bayesian
- We get $P(M|A) = \frac{P(A|M)P(M)}{P(A)}$ The same for F & B

$$P(A|B)P(B) = \frac{1}{2}^3 * 0.01 \quad < \quad P(A|F)P(F) = \frac{1}{6}^3 * 0.99$$

SOME THOUGHTS ON MODELING

- All models are wrong, but some are useful.
  - Models are what we call the lies we are used to
- There are no model free approaches!
  - use the term assumption instead
- Using models is a way to make assumptions explicit.
- Bayesian is a non-deterministic logic.

K-MEANS

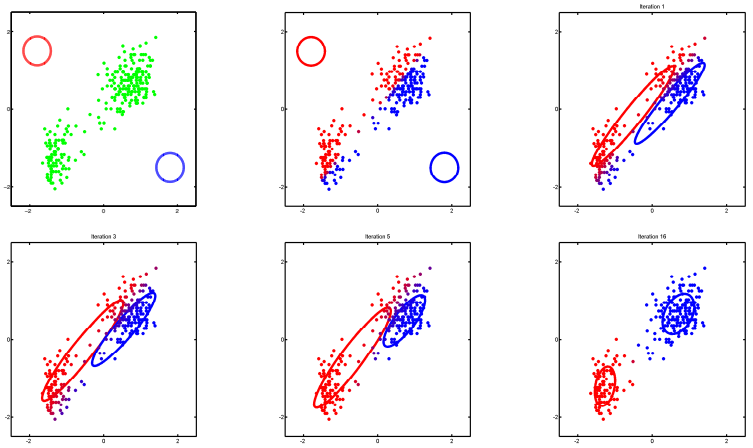
- Data vectors $D=\{x_1,\ldots,x_N\}$
- Randomly selected classes $z_1,\ldots,z_N$
- Iteratively do

$$\boldsymbol{\mu}_c = \frac{1}{N_c} \sum_{n:z_n=c} \boldsymbol{x}_n, \qquad \text{where } N_c = |\{n : z_n = c\}|$$

$$z_n = \text{argmin}_c ||\boldsymbol{x}_n - \boldsymbol{\mu}_c||_2$$

- One step O(NKD), can be improved

EXAMPLE



Expected complete: notation

$$
\begin{aligned}
&\log p(\boldsymbol{x}_n|\boldsymbol{\theta}') \\
&\quad = \log \sum_{\boldsymbol{z}_n} p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta}') \\
&\quad = \log \sum_{\boldsymbol{z}_n} p(\boldsymbol{z}_n|\boldsymbol{x}_n, \boldsymbol{\theta}) \frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta}')}{p(\boldsymbol{z}_n|\boldsymbol{x}_n, \boldsymbol{\theta})} \\
&\quad = \log E_{\boldsymbol{z}_n} \left(\frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta}')}{p(\boldsymbol{z}_n|\boldsymbol{x}_n, \boldsymbol{\theta})} \mid \boldsymbol{x}_n, \boldsymbol{\theta} \right) \\
&\quad \geq^{\text{Jensen}} E_{\boldsymbol{z}_n} \left(\log \frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta}')}{p(\boldsymbol{z}_n|\boldsymbol{x}_n, \boldsymbol{\theta})} \mid \boldsymbol{x}_n, \boldsymbol{\theta} \right) \\
&\quad = \sum_{\boldsymbol{z}_n} p(\boldsymbol{z}_n|\boldsymbol{x}_n, \boldsymbol{\theta}) \log \frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta}')}{p(\boldsymbol{z}_n|\boldsymbol{x}_n, \boldsymbol{\theta})} \\
&\quad = \sum_{\boldsymbol{z}_n} p(\boldsymbol{z}_n|\boldsymbol{x}_n, \boldsymbol{\theta}) \log p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta}') - \sum_{\boldsymbol{z}_n} p(\boldsymbol{z}_n|\boldsymbol{x}_n, \boldsymbol{\theta}) \log p(\boldsymbol{z}_n|\boldsymbol{x}_n, \boldsymbol{\theta}) \\
&\quad = Q_n(\boldsymbol{\theta}'; \boldsymbol{\theta}) - R_n(\boldsymbol{\theta}; \boldsymbol{\theta})
\end{aligned}
$$

COMBINING INFO FROM VARIOUS SENSORS

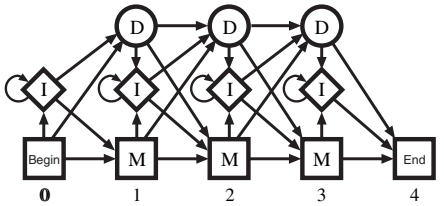
- Inertial sensors
- Camera
- Barometer

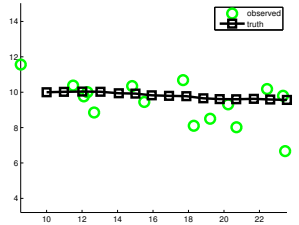
- Inertial sensors
- Radar
- Barometer
- Map

- Inertial sensors
- Cameras
- Radars
- Wheel speed sensors
- Steering wheel sensor

- Inertial sensors
- Ultra-wideband

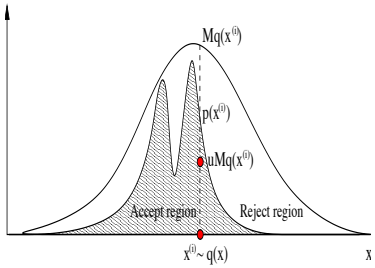
Aim: Motion capture, find the motion (position, orientation, velocity and acceleration) of a person (or object) over time.





HMMS ANS SSM

REJECTION SAMPLING



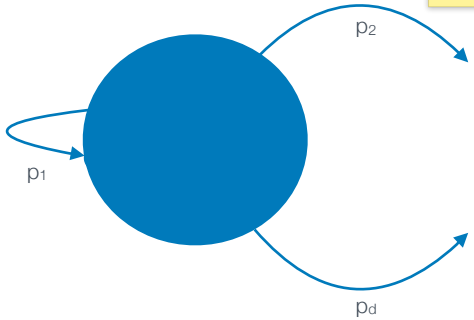
Algorithm

- sample x~q(x)
- sample u~U(0,1)
- if uMq(x) ≤ p(x), accept (and output x)

MARKOV CHAIN (DISCRETE)

importans sampling
kalman filter
particel filter
gibbs sampling

Probabilities on outgoing edges sum to one



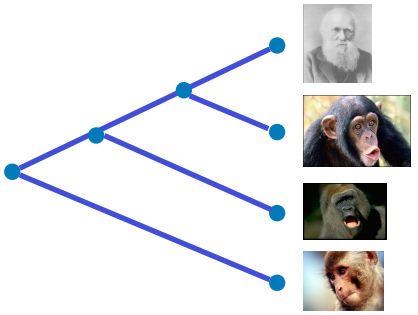
$\sum_{i \in [d]} p_i = 1$

IS THE CHIMP OUR CLOSEST RELATIVE?

PHYLOGENY

Input: species

Output: tree where proximity correlates with similarity



MR BAYES

BIOINFORMATICS APPLICATIONS NOTE Vol. 19 no. 12 2003, pages 1572–1574
DOI: 10.1093/bioinformatics/btg180

MrBayes 3: Bayesian phylogenetic inference under mixed models

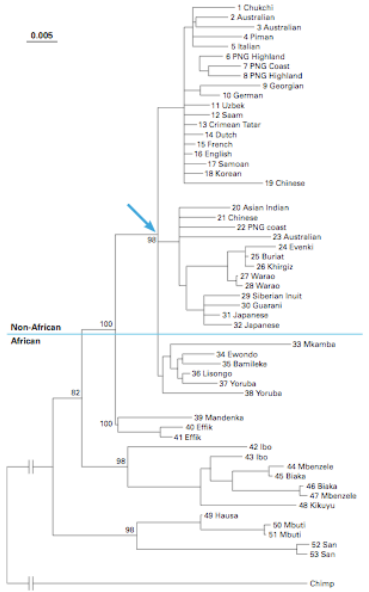
Fredrik Ronquist [1,] and John P. Huelsenbeck [2]*

[1]Department of Systematic Zoology, Evolutionary Biology Centre, Uppsala University, Norbyv. 18D, SE-752 36 Uppsala, Sweden and [2]Section of Ecology, Behavior and Evolution, Division of Biological Sciences, University of California, San Diego, La Jolla, CA 92093-0116, USA

Received on December 20, 2002; revised on February 14, 2003; accepted on February 19, 2003

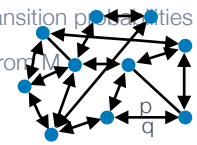
Prof. Entomology

OUR ORIGIN



MCMC

- In order to sample from p, set up a MC M
  - select transition probabilities so that p = stationary distribution
  - sample from M



CONTENT

- Chapter 1: Introduction.
- Chapter 2: Probability.
- Chapter 3: Generative models for discrete data.
- Chapter 4: Gaussian models.
- Chapter 5: Bayesian statistics.
- Chapter 10: Directed graphical models. Probably deeper.
- Chapter 18: SSM (HMMs)
- (Chapter 19. Undirected graphical models.)
- Chapter 20 Exact inference for graphical models.
- Chapter 23: Monte Carlo inference.
- Chapter 24: Markov Chain Monte Carlo.
- Chapter 25: Clustering
- Chapter 26: Graphical Model Structure Learning
- Particle MCMC