Introduction to Machine Learning in Bioinformatics

Peter Sykacek[1]

Vienna Science Chair of Bioinformatics

Department of Biotechnology

BOKU University

peter.sykacek@boku.ac.at

Overview

- Fundamental Problems of Analysing Data
- Concepts in Data Analysis
- Supervised Learning
- Unsupervised Learning
- Model Fitting, Diagnosis and Evaluation
- William of Occam and Karl R. Popper
- Further Elective Courses on Data Analysis

Nature of Data

- Data type (discrete vs. continuous)
- Observation is different from ground truth.

Discrete data: phenotype, genotype, age group,... Ordering among labels can be exploited. Continuous data: length, temperature, weight, pressure, mRNA expression,...

Measurement and ground truth: Measuring pencil length of 15.3cm $\neq$ true length of 15.3cm! Why? Repeated measurements differ (15.2cm, 15.4cm, etc.)

$->$ measurement errors!

Origin of Noise

Measurement processes involve errors which arise from noise (fluctuations) that are or can not be captured:

- Measurement noise.
- Misclassifications (e.g. wrong phenotype).
- Simplified Models.

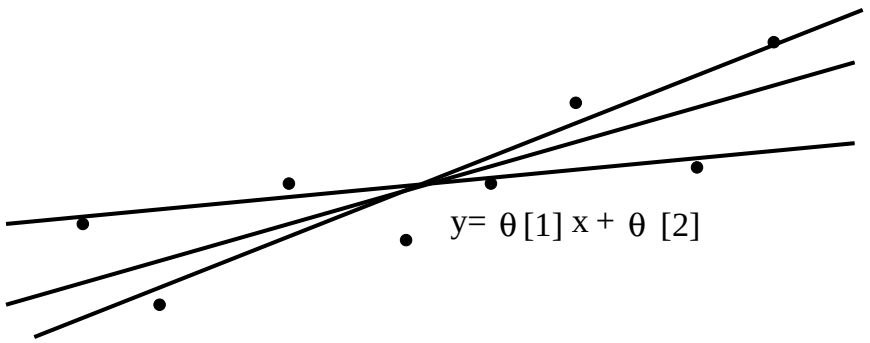
Data analysis uses replicates to remove the noise and model the remaining aspects as good as possible.

Refresher: Scalar Product

K measurements $\boldsymbol{x}^T = [\boldsymbol{x}[1], ..., \boldsymbol{x}[K]]$ (row vector) represent variable y as linear function (parameter $\boldsymbol{\theta}$). $->$ *linear regression* Express y:

$y = \sum_k \boldsymbol{x}[k]\boldsymbol{\theta}[k]$, or $y = \boldsymbol{x}^T\boldsymbol{\theta}$ and equivalently $y = \boldsymbol{\theta}^T\boldsymbol{x}$

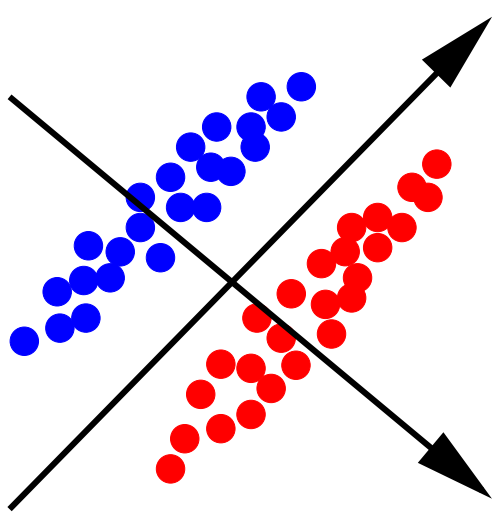
$->$ vector dot product or scalar product



Why Understand Data Analysis?

Result = Data + Model!

Linear discriminant (LDA) and principle component analysis (PCA) give different projections of the same data.



Both use linear projections!

$$t_{\mathrm{PCA}} = \theta_{\mathrm{PCA}}^T x$$

$$t_{\mathrm{LDA}} = \theta_{\mathrm{LDA}}^T x$$

Good Analysis Practice Ensures

Technical sufficiency of data analysis

Compatibility with biological question

Technical Sufficiency

Avoid ad-hoc ideas

Verify hypothetical genes by hybridising tissue mix on N microarrays. Verified gene: on $n < N$ arrays expression above threshold δ. Problems?

1)Impossible to justify a particular n.

2)Impossible to justify a particular δ

3)Verification of low expressed (e.g. regulators) and rarely functional genes difficult.

Such “data analysis” does not fit the objective.

Biological Relevance

Avoid “hammer and nail” syndrome

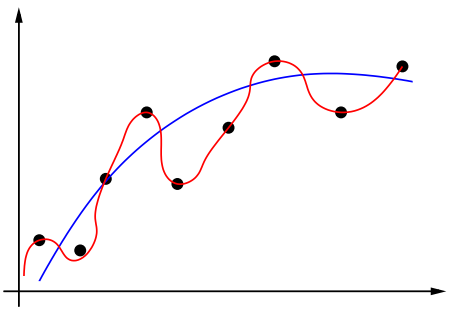
Determine “cancer genes” by combining SVM (support vector machine, a classifier) with greedy search over inputs. Problems?

1) Greedy search provides an arbitrary set of genes which separate the data. $->$ neither optimal nor complete set of “cancer genes”.

2) Gene set lacks rank information about functional importance.

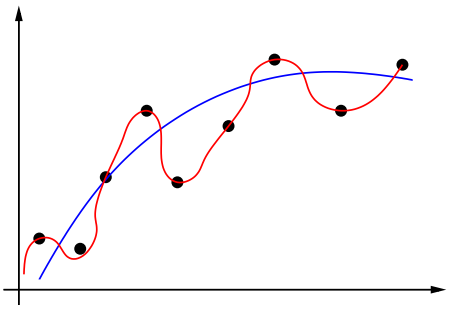
The otherwise useful approach (as a diagnostic tool) fails answering the biological question.

Fundamental Principle in Data Analysis

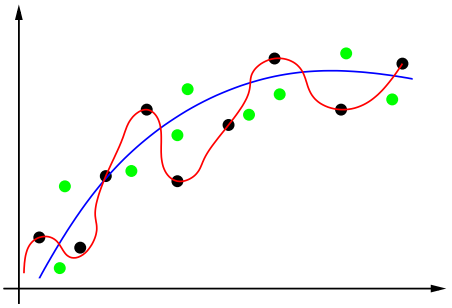


Which is the better model? Why is that the case?

Fundamental Principle in Data Analysis



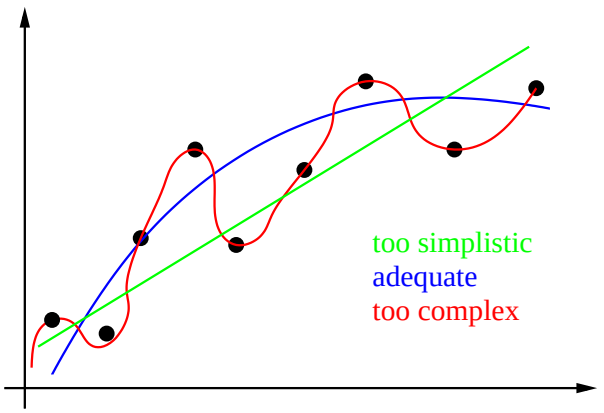
Which is the better model? Why is that the case?



Find the underlying data generating model.

Adequate models

Capture underlying structure and avoid overfitting. Adjust "fiddle parameters" $->$ avoid too simple and too complex.



Overfitting memorises training data including uninteresting noise. To "learn something useful from data" we have to get complexity right.

Why Bother With Data Analysis?

Moore's Law:

PC 1984 5 MB Hard Drive

PC 2007 2 TB Hard Drive (4*500 GB) $\approx$ 400 Euro

How much paper on one PC in 2007 assuming 10.000 (single byte) characters per page ?

Why Bother With Data Analysis?

Moore's Law:

PC 1984 5 MB Hard Drive

PC 2007 2 TB Hard Drive (4*500 GB) $\approx$ 400 Euro

How much paper on one PC in 2007 assuming 10.000 (single byte) characters per page ?

It is actually a stack of paper 20 km high!

2 TB $\approx$ $2 * 10^{12}$ byte

$= 2 * 10^{8}$ pages, assuming 1000 pages = 10 cm

a stack $2 * 10^{5} * 10$ cm = $2 * 10^{4}$ m = 20 km

PC 2010 8 TB Hard Drive (4*2 TB) $\approx$ 440 Euro

Stack height in 2010?

How Much Data?

Medical monitoring (sleep): 20 channels, 8 hours at 200 Hz and 16 Bit: ≈ 250 MB. A lab with 8 recording units (nights only): about one TB per year.

How Much Data?

Medical monitoring (sleep): 20 channels, 8 hours at 200 Hz and 16 Bit: ≈ 250 MB. A lab with 8 recording units (nights only): about one TB per year.

Medical monitoring (cognitive neuroscience): FMRI scanner, 1dm^3 volume, 10s temporal and 1mm^3 spatial resolution, 16 bit, generates $10^6 * 360 * 2$ byte ≈ 720 MB per hour: about 1 TB every two months.

How Much Data?

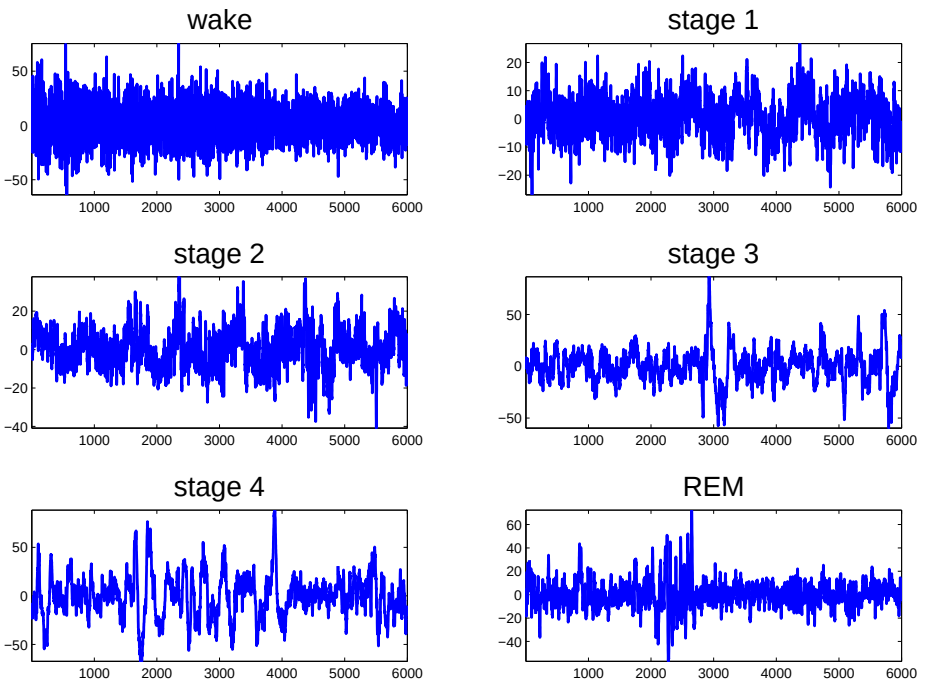
Medical monitoring (sleep): 20 channels, 8 hours at 200 Hz and 16 Bit: ≈ 250 MB. A lab with 8 recording units (nights only): about one TB per year.

Medical monitoring (cognitive neuroscience): FMRI scanner, 1dm^3 volume, 10s temporal and 1mm^3 spatial resolution, 16 bit, generates $10^6 * 360 * 2$ byte ≈ 720 MB per hour: about 1 TB every two months.

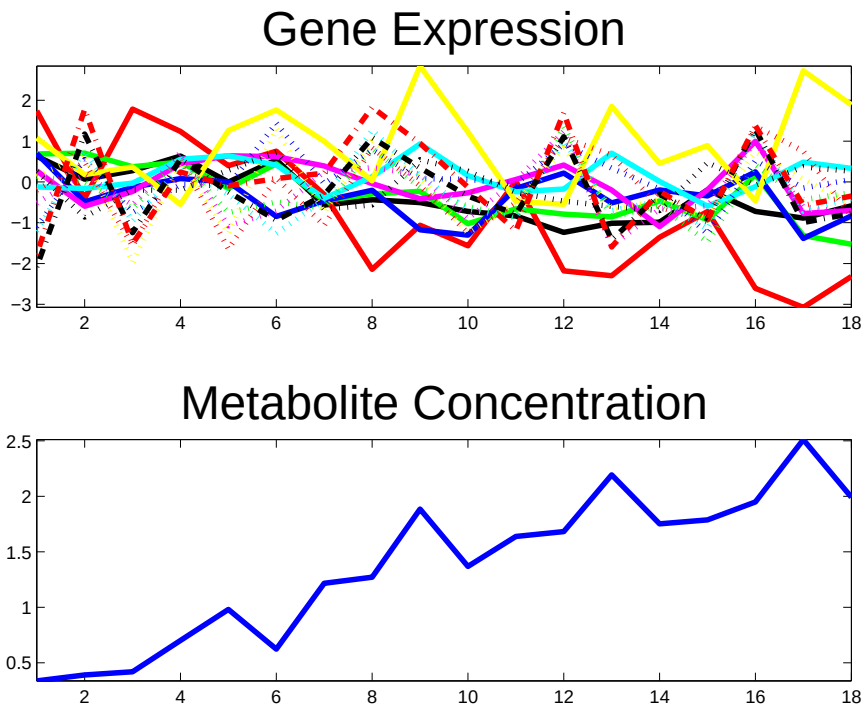
High throughput molecular biology: Small microarray facility, 12 slides per 24 hours: about 240 MB image files per day. Deep sequencing (ABI Solid+) about 300 GB per week (two flow cells, aligned reads).

Amount and noise prevent manual analysis

Example: Sleep EEG



Example: Metabolomics



Analysis Strategies

All data analysis problems can be grouped into two categories:

1. Supervised Learning methods are used for regression problems.
2. Unsupervised Learning methods are used for exploratory data analysis.

Regression Problems

Noisy Data from life science experiment $\mathcal{Z} = \{(y_1, \boldsymbol{x}_1), ..., (y_N, \boldsymbol{x}_N)\}$ with $\boldsymbol{x}_n$ denoting vectors.

Regression fits based on $\mathcal{Z}$ an "optimal" function relating $\boldsymbol{x}$ and y:

$$y = f(\boldsymbol{x}; \boldsymbol{\theta}) + \epsilon(\lambda)$$

Regression Problems

Noisy Data from life science experiment $\mathcal{Z} = \{(y_1, \boldsymbol{x}_1), ..., (y_N, \boldsymbol{x}_N)\}$ with $\boldsymbol{x}_n$ denoting vectors.

Regression fits based on $\mathcal{Z}$ an "optimal" function relating $\boldsymbol{x}$ and y:

$$y = f(\boldsymbol{x}; \boldsymbol{\theta}) + \epsilon(\lambda)$$

Noise requires a deterministic and a random component.

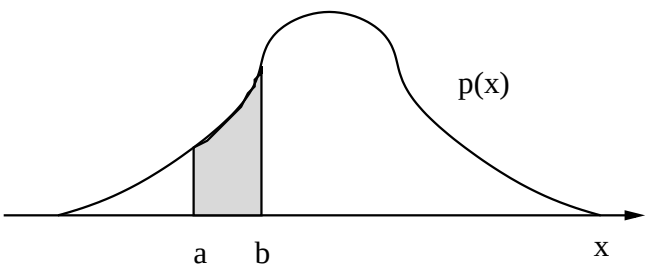
$->$ Inherent uncertainty, y is a random variable!

Variables vs. Random Variables

Variables: x represents deterministic value.

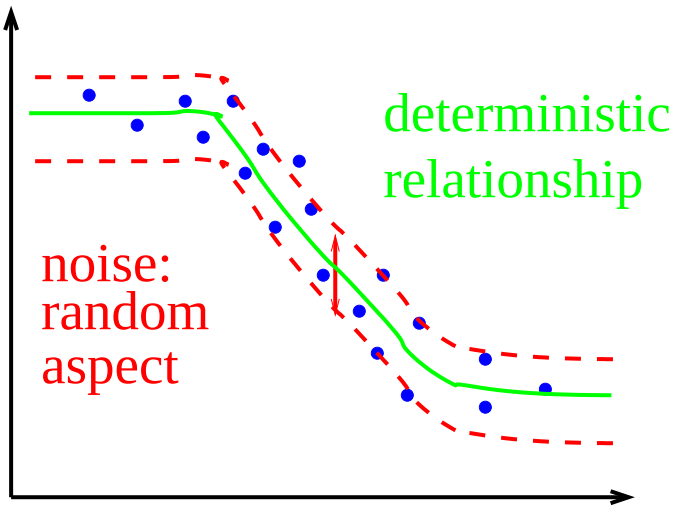
Random Variables: x represents collection of values. Density function $p(x)$ describes relative occurrence of values. Like sand heap specifying occurrence of grain positions.

Area of shaded region: probability observing sample in interval $P(x \in [a,b]) = \int_{x=a}^{b} p(x)dx$.



Implication of Randomness

Best predicting expected y values from x (local average). Complete description includes noise characteristics.

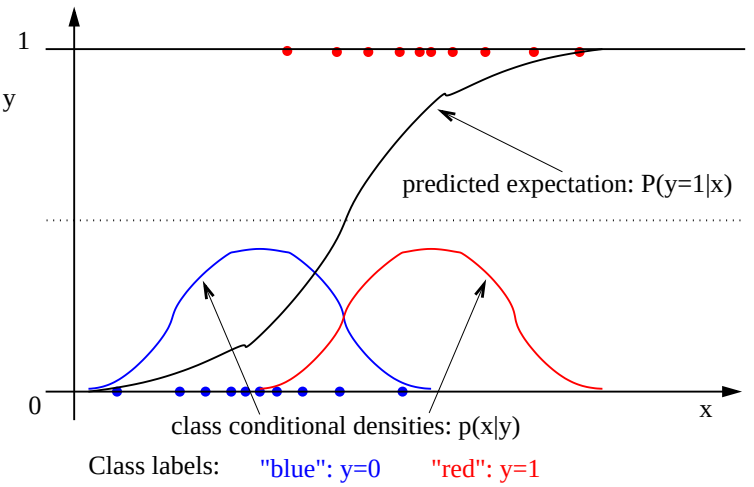


Red error bars represent the standard deviation which is complete description of Gaussian noise.

Classification

Classification is instance of regression, with predicted values (i.e. the y) being discrete.

Two classes: $y = \{0, 1\}$. Predicted expectations are class probabilities $P(y = 1|x)$.



Exploratory Data Analysis

Search of unknown structure in a data set $\mathcal{X} = \{x_1, x_2, ..., x_N\}$, x_n distributed according to unknown pdf $p(x)$.

Learning task: summarise x by an unobserved variable t.

Typical models:

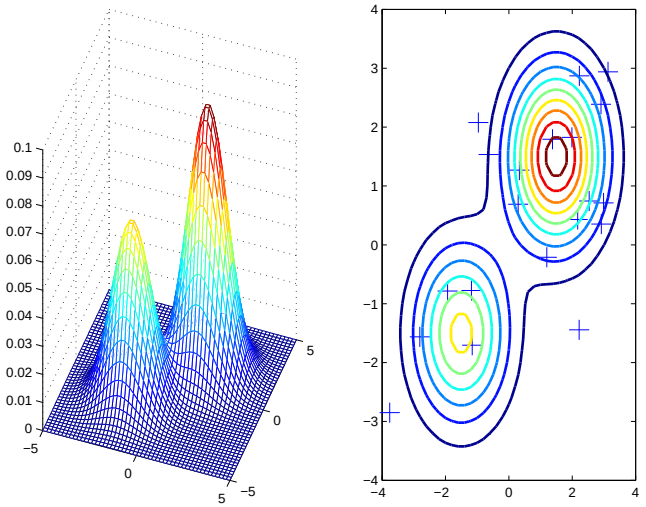
Mixture density models:
$p(x) = \sum_k P(t = k)p(x|t = k)$, and $t \in \{1, .., K\}$.

Continuous latent variable models:
$p(x) = \int_t p(t)p(x|t)dt$, $x \in \Re^k$, $t \in \Re^d$ and $k > d$.

Mixture Density Model

Example: Gaussian mixture model

$p(x|t = k) = \mathcal{N}(x; \mu_k, \lambda_k)$ - a Gaussian density function.



Summary: the k which generated x $->$ Clustering.

Continuous Latent Variable Model

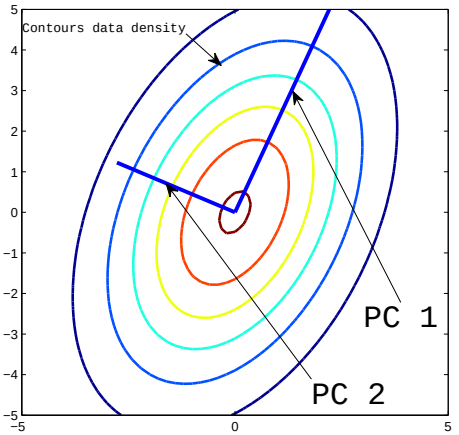
Example - PCA (principle component analysis):

$x = m + W_1 t_1 + ... + W_d t_d$, $x \in \Re^k$, $t = [t_1, ..., t_d] \in \Re^d$

$W_d : [k \times 1]$ d-th eigenvector of sample covariance matrix

$t \sim \mathcal{N}(t; 0, \Lambda)$, with $\Lambda : [d \times d]$ diagonal cov. matrix

Summary: lower dimensional continuous representation $->$ dimensionality reduction



Analysis Tasks and Methods

Task	$->$	Method
predict continuous y from input data	$->$	Regression
predict discrete y from input data	$->$	Classification
find unknown groups in input data	$->$	Clustering (e.g. k-means, mixture models)
find low dimensional representation for input data	$->$	Dimensionality reduction PCA, independent component analysis (ICA)

Model Fitting

- Choose appropriate analysis methodology
- Adapt model parameters to data (“learning”, inference)
- Apply model diagnostics (performance assessment, model selection)

Assessing Model Parameters

Goal: tune $\boldsymbol{\theta}$ such that $f(\boldsymbol{x}_n; \boldsymbol{\theta})$ represents all $(y_n \boldsymbol{x}_n)$ pairs well.

Need expression we may optimise (maximise, minimise) for good fit of all n "training" samples.

Assessing Model Parameters

Goal: tune $\boldsymbol{\theta}$ such that $f(\boldsymbol{x}_n; \boldsymbol{\theta})$ represents all $(y_n \boldsymbol{x}_n)$ pairs well.

Need expression we may optimise (maximise, minimise) for good fit of all n "training" samples.

Possible choice: sum of squared errors (SSE).

Idea: subtract deterministic part from y_n:

$\epsilon_n = y_n - f(\boldsymbol{x}_n; \boldsymbol{\theta})$ + summation

$$\mathrm{SSE} = \sum_n \epsilon_n^2 = \sum_n (y_n - f(\boldsymbol{x}_n; \boldsymbol{\theta}))^2$$

Several objective functions e.g. (log)-likelihood

Major Problem

True model - linear regression:

$$y_n = \boldsymbol{x}_n^T \boldsymbol{\theta} + \epsilon_n$$

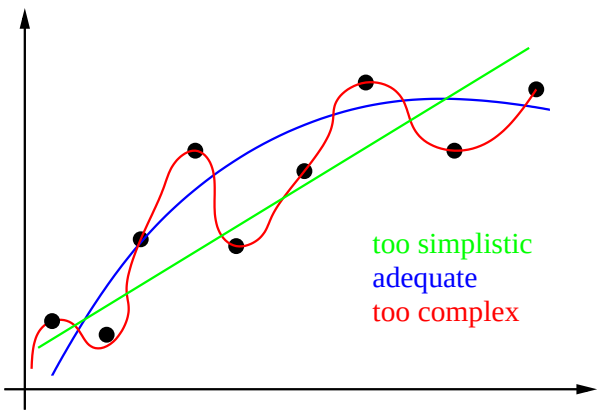
Finite sample size and arbitrarily complex models: What is the minimum of the SSE?

Think "phone book": Perfect memorising of all y_n, modelling error 0, SSE $-> 0$

$->$ SSE unsuitable for model selection! (likelihood likewise!)

Adequacy of models

Optimise model structure and avoid overfitting.



Wrong model class does not capture "truth" and performs worse in applications.

Some model classes:

$y = kx + d + \epsilon$

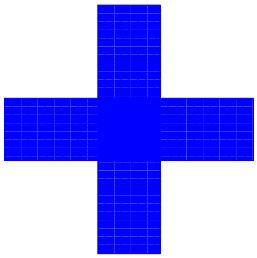
$y = lx^2 + kx + d + \epsilon$

$y = \sum_{j=0}^{J} (x^j k_j) + \epsilon$

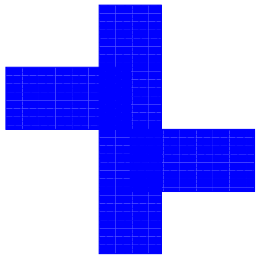
How getting complexity right ?
See the ideas by Karl R. Popper!

Human Intuition and Complexity

How many components?



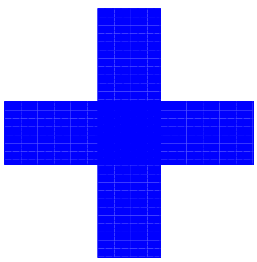
Object A



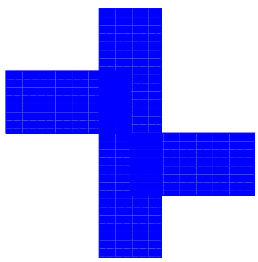
Object B

Human Intuition and Complexity

How many components?

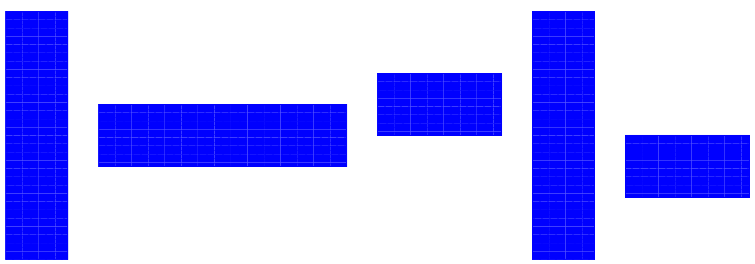


Object A



Object B

Most likely answer:



Object A

Object B

Occam's Razor

We implicitly apply Occam's Razor

William of Occam (or Ockham) (1288 - 1348)

Entia non sunt multiplicanda sine necessitate: Entities are not to be multiplied without necessity.

Interpretation: always opt for an explanation with as few as possible causes, factors, or variables.

Material from `http://en.wikipedia.org/wiki/William_of_Ockham`.

Occams Razor in ML

Model selection can be done by the following approaches:

- Add complexity penalty to objective function. Many choices: AIC (Akaikes information criterion), BIC (Bayesian information criterion), MDL (minimum description length), etc.
- Use learning methods like Bayesian inference with Occam's razor built in.
- Use empirical approaches comparing model classes by validation testing (computer simulation using independent data).

Instances of model selection

- Variable selection: search for input subsets which improve predictive performance or identify important variables (e.g. differentially expressed genes).
- Change point detection: Separating data into groups which show similar statistical properties.
- Clustering: (see above)
- Determining optimal model orders (applies to most ML methods!)
- Determining suitable noise characteristics.
- ...

Validation: Diagnostic Measures

Mean square generalisation error (MSE_{test}, average SSE!) for assessing regression models.

$$\mathrm{MSE}_{test} = \frac{1}{N} \sum_n \left(y_n^{test} - f(\boldsymbol{x}_n^{test}, \boldsymbol{\theta}^{train}) \right)^2$$

Classification is tested by the generalisation accuracy acc_{test}.

$$\forall n \; \hat{y}_n^{test} = \mathrm{argmax}_k (P(y = k | \boldsymbol{x}_n^{test}, \boldsymbol{\theta}^{train}))$$

$$\mathrm{acc}_{test} = \frac{1}{N} \sum_n \delta(y_n^{test}, \hat{y}_n^{test})$$

Classify such that most probable class wins. Estimate fraction of correctly classified test cases.

Validation: Estimation Procedures

Trade-off:

Reliable model fit requires large “training sets”

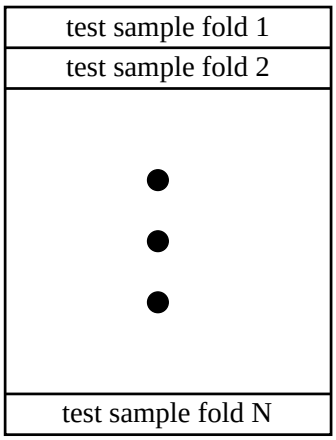
Unbiased diagnostics require large “test sets”

Diagnostic quantities are only unbiased if we leave test samples untouched! Test samples must not be used for any modelling decisions.

$->$ Solution: reuse samples by iterating over model fitting and testing.

N-Fold Cross Testing

Sketch and MatLab like pseudo code




```
allres=[]; allpred=[];
for n=1:n_folds
    % split into training and test data
    [train, test]=foldsplit(orig_data, n_folds, n);
    % model inference
    [model]=trainfunc(train, fiddleparams);
    % store this folds true targets and predictions
    [res]=truetarg(test);
    [pred]=predtarg(test, model);
    allres=[allres; res];
    allpred=[allpred; pred];
end
```

Leave one out has as many folds as samples. An alternative by resampling with replacement is called Bootstrapping.

In Depth Courses

Data Analysis is an important topic in modern life sciences. Three elective courses provide more advanced topics (In English, providing theoretical concepts and practical experience in the computer lab).

- *Efficient Microarray Data Analysis using R and FSPMA (793.403)* 1.0 HRS, winter term, 1.5 ETCS, A two day blocked lecture held entirely in computer lab.
- *Machine Learning and Pattern Recognition for Bioinformatics (793.404)* 3.0 HRS, winter term, catalogued elective course with 4.5 ETCS - theoretical part and MatLab based practical in the computer lab.
- *Bayesian Data Analysis in the Life Sciences (793.402)* 3.0 HRS, summer term, 4.5 ETCS - theoretical part and 3 days blocked MatLab practical in the computer lab.

Further details at
`http://www.sykacek.net/teaching.html`