Science is to test crazy ideas – Engineering is to put these ideas into business

Andreas Holzinger
Machine Learning & Knowledge Extraction for Health Informatics
University of Verona
Module 2 - Day 2 – April 2017

Health Data Jungle:
From the underlying physics of data to the Kullback-Leibler Divergence
a.holzinger@hci-kdd.org
http://hci-kdd.org/mini-make-machine-learning-knowledge-extraction-health

ML needs a concerted effort fostering integrated research

Interactive Knowledge Discovery

Data Mining

GDM 3 Graph-based Data Mining
TDM 4 Topological Data Mining
EDM 5 Entropy-based Data Mining

Learning Algorithms

Data Mapping

Preprocessing

Data Fusion

Privacy, Data Protection, Safety and Security

01 Data – the underlying physics of data
02 Biomedical data sources – taxonomy of data
03 Data integration, mapping, fusion
04 Probabilistic Information
05 Information Theory – Information Entropy
06 Cross-Entropy - Kullback-Leibler Divergence

Machine Learning Jungle Top-Level View

Cognition
Visualization
Data fusion
Preprocessing
Decision
Interaction
Integration

CONCEPTS
THEORIES
PARADIGMS
MODELS
METHODS
TOOLS

Dimensionality
Complexity
Unsupervised
Gaussian P.
Regularization
Python

Reinforcement
Bayesian p(x)
Supervised
Graphical M.
Scaling
Church

Representation
Entropy/KL
Semi-Superv.
Neural Nets
Aggregation
Anglican

No-free-lunch
Vapnik-Chernov.
IML
Kernel/SVM
Evolution
Julia

Multi-Task Learning
Transfer Learning
Multi-Agent-Hybrid-Systems

Data Protection, Safety and Security and Privacy Aware Machine Learning (PAML)
Application, Validation, Evaluation, Impact – Social, Economic, Acceptance, Trust
01 Reflection

Question: Where is the Biologist in this image?


Warm-up Quiz

1. Uncertainty

2. p(θ|D) = p(D|θ) * p(θ)

3. p(D) / p(θ)

4. Medical Decision Making

5. p(θ|D) = p(D|θ) * p(θ)

6. p(D) / p(θ)

7. Context

8. p(θ|D) = p(D|θ) * p(θ)

9. p(D) / p(θ)

Repetition of Bayes - on the work of Laplace

What is the simplest mathematical operation for us?

\[ p(x) = \sum_a (p(x,y)) \] (1)

How do we call repeated adding?

\[ p(x,y) = p(y|x) * p(y) \] (2)

Laplace (1773) showed that we can write:

\[ p(x,y) * p(y) = p(y|x) * p(x) \] (3)

Now we introduce a third, more complicated operation:

\[ \frac{p(x,y) * p(y)}{p(y)} = \frac{p(y|x) * p(x)}{p(y)} \] (4)

We can reduce this fraction by \( p(y) \) and we receive what is called Bayes rule:

\[ p(x,y) = \frac{p(y|x) * p(x)}{p(y)} \]

\[ p(l|h) * p(h) \]

\[ p(l|d) = \frac{p(d|h) * p(h)}{p(d)} \] (5)
Practical Example: Diagnoses

- Your MD has bad news and good news for you.
- Bad news first: You are tested positive for a serious disease, and the test is 99% accurate (T)
- Good news: It is a rare disease, striking 1 in 10,000 (D)
- How worried would you now be?

\[
p(x) = \frac{\text{likelihood } \times \text{ prior } p(x)}{\text{evidence}} \quad p(h|d) = \frac{p(d|h)p(h)}{p(d)}
\]

\[
p(T = 1|D = 1) = p(d|h) = 0.99 \quad \text{and} \quad p(D = 1) = p(h) = 0.0001
\]

\[
p(D = 1 \mid T = 1) = \frac{(0.99) \times (0.0001)}{(1-0.99) \times (1-0.0001) + 0.99 \times 0.0001} = 0.0098
\]

What is this?

Key Problems

- Heterogeneous, distributed, inconsistent data sources (need for data integration & fusion) [1]
- Complex data (high-dimensionality – challenge of dimensionality reduction and visualization) [2]
- Noisy, uncertain, missing, dirty, and imprecise, unbalanced data (challenge of pre-processing)
- The discrepancy between data-information-knowledge (various definitions)
- Big data sets in high-dimensions (manual handling of the data is often impossible) [3]

Statistical inference & Decision support: Better a good solution in time, than a perfect solution never ...

01 The underlying physics of data

What is data?
What types of data?


Example: Neonatal Screening (1/3)


Traditional Statistics versus Machine Learning

- Data in traditional Statistics
  - Low-dimensional data ($< \mathbb{R}^{100}$)
  - Problem: Much noise in the data
  - Not much structure in the data but it can be represented by a simple model

- Data in Machine Learning
  - High-dimensional data ($\gg \mathbb{R}^{100}$)
  - Problem: not noise, but complexity
  - Much structure, but the structure can not be represented by a simple model

02 Biomedical data sources: Taxonomy of data
Example: Type of Cells


Biological data is getting more complex (big sowieso ;)


To get a feeling of biological data sources (bionumbers)


Costs more decreasing than in Moor’s law (also cost!)

Example: Genetic Data

For further reading this is recommended:
Buffalo, V. 2015. Bioinformatics Data Skills: Reproducible and Robust Research with Open Source Tools, Sebastopol (CA), O'Reilly.

What can ML do with such data?

Features are key to learning and understanding!

Example Species: Bacterium E. coli

% total dry weight:
DNA 3.1
RNA 20.5
Protein 55.0
Lipid 9.1

Algorithms are used to understand these important components.

Where do we get the data sets from?

- Billions of biological data sets are openly available, here only some examples:
- General Repositories:
  - GenBank, EMBL, HMCA, ...
- Specialized by data types:
  - UniProt/SwissProt, MMMP, KEGG, PDB, ...
- Specialized by organism:
  - WormBase, FlyBase, NeuroMorpho, ...
- Details: http://hci-kdd.org/open-data-sets
Example Data Set from an High-Throughput Experiment

- this figure depicts one yeast gene-expression data set
  - each row represents a gene
  - each column represents a measurement of gene expression (mRNA abundance) at some time point
- red indicates that a gene is being expressed more than some baseline; green means less

Figure from Spellman et al., Molecular Biology of the Cell. 0:3273-3287, 1998

Taxonomy of data

- Physical level -> bit = binary digit = basic indissoluble unit (= Shannon, Sh), ≠ Bit (!) in Quantum Systems -> qubit
- Logical Level -> integers, booleans, characters, floating-point numbers, alphanumeric strings, ...
- Conceptual (Abstract) Level -> data-structures, e.g. lists, arrays, trees, graphs, ...
- Technical Level -> Application data, e.g. text, graphics, images, audio, video, multimedia, ...
- "Hospital Level" -> Narrative (textual) data, numerical measurements (physiological data, lab results, vital signs, ...), recorded signals (ECG, EEG, ...), Images (x-ray, MR, CT, PET, ...) ; -omics

Taxonomy of data at Hospital Level

- Clinical workplace data sources
  - Med docs: text (non-standardized free-text), semi-structured, standard terminologies (ICD, SNOMED-CT)
  - Measurements: lab results, ECG, EEG, EOG, ...
  - Surveys, Clinical studies, trials
- Image data sources
  - Radiology: MRI (256x256, 200 slices, 16 bit per pixel, uncompressed, ~26 MB); CT (512x512, 60 slices, 16 bit per pixel, uncompressed ~32 MB; MR, US;
  - Digital Microscopy: WSI (15mm slide, 20x magn., 24 bits per pixel, uncompressed, 2,5 GB, WSI 10 GB; confocal laser scanning, etc.
- -omics data sources
  - Sanger sequencing, NGS whole genome sequencing (3 billion reads, read length of 36) ~ 200 GB; NGS exome sequencing ("only" 110,000,000 reads, read length of 75) ~7GB;
  - Microarray, mass-spectrometry, gas chromatography, ...

Example Data Structures (1/3): List

Example Data Structures (2/3): Graph

Evolutionary dynamics act on populations. Neither genes, nor cells, nor individuals evolve; only populations evolve.

[Graph diagrams]


Example Data Structures (3/3) Tree


Algorithms in nature: Shared principles

[Diagram showing distributed computing, network processes, reusable components, randomness, and stochasticity]


Translational Health Informatics Continuum

[Diagram showing translational medicine continuum: Bench (T1) → Bedside (T2) → Community (T3) → Policy]

Biomedical Research needs Open Data Access

- Grand Challenges in this area:
  - Production of Open Data Sets
  - Synthetic data sets for learning algorithm testing
  - Privacy preserving machine learning
  - Data leak detection
  - Data citation
  - Differential privacy
  - Anonymization and pseudonymization
  - Evaluation and benchmarking

Please visit:

---

03 Data Integration, mapping, fusion

---

In medicine we have two different worlds ...

---

Unsolved Problem:
Data Integration and Data Fusion in the Life Sciences

How to combine these different data types together to obtain a unified view of the activity in the cell is one of the major challenges of systems biology


---

Our central hypothesis:
Information may bridge this gap

Example Data Integration Architecture

- Genomics (sequence annotation)
- Transcriptomics (microarray)
- Proteomics (Proteome Databases)
- Metabolomics (enzyme annotation)
- Fluxomics (isotopic tracing, metabolic pathways)
- Phenomics (biomarkers)
- Epigenomics (epigenetic modifications)
- Microbiomics (microorganisms)
- Lipidomics (pathways of cellular lipids)
Omics-data integration

<table>
<thead>
<tr>
<th>Genomics (sequence annotation)</th>
<th>Transcriptomics</th>
<th>Proteomics</th>
<th>Metabolomics</th>
<th>Protein-DNA interactions</th>
<th>Protein-protein interactions</th>
<th>Proteomics (transcriptional modification)</th>
<th>Metabolomics (metabolite abundance)</th>
<th>Protein-DNA interactions (ChIP-chip)</th>
<th>Protein-protein interactions (PPI, coIP-MS)</th>
<th>Pharmacology</th>
<th>Pathway identification activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transcriptomics (microarray, SAGE)</td>
<td>- Protein transcript correlation</td>
<td></td>
<td>- Enzyme annotation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proteomics (identification and quantification)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Metabolomics (metabolite abundance)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Protein-DNA interactions (ChIP-chip)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Protein-protein interactions (PPI, coIP-MS)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pharmacology</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pathway identification activity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>


Example of typical clinical data sets

- 50+ Patients per day ~ 5000 data points per day ...
- Aggregated with specific scores (Disease Activity Score, DAS)
- Current patient status is related to previous data
- = convolution over time
- ⇒ time-series data


DAS28 Predicted Mean Responses


Standardization versus Structurization

Data Dimensionality examples

- 0-D data = a data point existing isolated from other data, e.g. integers, letters, Booleans, etc.
- 1-D data = consist of a string of 0-D data, e.g. Sequences representing nucleotide bases and amino acids, SMILES etc.
- 2-D data = having spatial component, such as images, NMR-spectra etc.
- 2.5-D data = can be stored as a 2-D matrix, but can represent biological entities in three or more dimensions, e.g. PDB records
- 3-D data = having 3-D spatial component, e.g. image voxels, e-density maps, etc.
- H-D Data = data having arbitrarily high dimensions

Example: 1-D data (univariate sequential data objects)

SMILES (Simplified Molecular Input Line Entry Specification)

... is a compact machine and human-readable chemical nomenclature:

e.g. Viagra: 
CCC1nn(C)c2c(=O)[nH]c(nc12)c3cc(ccc3OCC)S(=O)(=O)N4CC N(C)CC4

...is Canonicalizalizable
...is Comprehensive
...is Well Documented


Example: 2-D data (bivariate data)


Example: 2.5-D data (structural information & metadata)
Example: 3-D Voxel data (volumetric picture elements)


Example: Data structures - Classification


Note: The curse of dimensionality


http://www.iro.umontreal.ca/~bengio/yoshua_en/research.html

Categorization of Data (Classic “scales”)

Challenges in Data Integration

- Bridging the gap between natural sciences and clinical medicine (who has seen genomics and patient data integrated in routine?)
- Organizational barriers, data provenance, data ownership, privacy, accessibility, usability, fair use of data, security, safety, data protection
- Combine Ontologies with Machine Learning
- Stochastic Ontologies, Ontology learning
- Integration of data from wet-labs with in-silico experimental data (e.g. tumor growth simulation)

Please always distinguish models between ...

Boolean models
Algebraic models
Probabilistic models *)

*) Our probabilistic models describes data which we can observe from our environment – and if we use the mathematics of probability theory, in order to express the uncertainties around our model then the inverse probability allows us to infer unknown unknowns ... learning from data and making predictions – the core essence of machine learning and of vital importance for health informatics


04 Probabilistic Information p(x)

Life is complex information

What is information?

For ML and Health always remember

Probabilistic Information $p(x)$

Thomas Bayes
1701 - 1761

$$p(x) = \sum P(x, y_j)$$

Bayes’ Rule is a corollary of the Sum Rule and Product Rule:

$$p(x|y_j) = \frac{p(y_j|x_i)p(x_i)}{\sum p(x, y_j)p(x_i)}$$


Always remember:

The inverse probability allows to infer unknowns, learn from data and make predictions:

1) Maximum-Likelihood Learning

finds a parameter setting, that maximizes the $p(x)$ of the data: $P(D | \theta)$

2) Maximum a Posteriori Learning (e.g. for MCMC)

assumes a prior over the model parameters $P(\theta)$ and finds a parameter setting that maximizes the posterior: $P(\theta | D) \propto P(\theta)P(D | \theta)$

3) Bayesian Learning

assumes a prior over the model parameters and computes the posterior distribution $P(\theta | D)$
Parameter Estimation

- General setting:
  - Given a (hypothesized & probabilistic) model that governs the random experiment
  - The model gives a probability of any data \( p(D|\theta) \) that depends on the parameter \( \theta \)
  - Now, given actual sample data \( X = \{x_1, \ldots, x_n\} \), what can we say about the value of \( \theta \)?
- Intuitively, take your best guess of \( \theta \)
- “best” means “best explaining/fitting the data”
- Generally an optimization problem

Maximum Likelihood vs. Bayesian

- 1) Maximum likelihood estimation (given \( X \))
  - “Best” means “data likelihood reaches maximum”
  \[
  \hat{\theta} = \arg \max_{\theta} P(X|\theta)
  \]
  - Problem: massive amount of data necessary
- 2) Bayesian estimation (use posterior)
  \[
  \hat{\theta} = \arg \max_{\theta} P(X|\theta) = \arg \max_{\theta} P(X|\theta) P(\theta)
  \]
  - “Best” means being consistent with our “prior” knowledge and explaining data well
  - Problem: how to define prior?

Illustration of Bayesian Estimation

\[
\text{posterior } p(x) = \frac{\text{likelihood } \times \text{ prior } p(x)}{\text{evidence}}
\]

05 Information Theory & Entropy
- Information is the reduction of uncertainty
- If something is 100% certain its uncertainty = 0
- Uncertainty is a max. if all choices are equally probable
- Uncertainty (as information) sums up for independent sources

An overview on the History of Entropy

Bernoulli (1713)
Principle of Insufficient Reason

Maxwell (1859), Boltzmann (1871),
Gibbs (1902) Statistical Modeling of problems in physics

Bayes (1763), Laplace (1770)
How to calculate the state of a system with a limited number of expectation values

Jeffreys, Cox (1939-1948)
Statistical Inference

Shannon (1948)
Information Theory

Pearson (1900)
Goodness of Fit measure

Fisher (1922)
Maximum Likelihood

Generalized Entropy

Entropy Methods

Bayesian Statistics

Generalized Entropy

Towards a Taxonomy of Entropic Methods

Jaynes (1957)
Maximum Entropy (MaxEn)

Adler et al. (1965)
Topology Entropy (TopEn)

Moshkovitz (1969)
Graph Entropy (MinEn)

Posner (1975)
Minimum Entropy (MinEn)

Tsallis (1980)
Tsallis-Entropy

Pincus (1991)
Approximate Entropy (ApEn)

Richman (2000)
Sample Entropy (SampEn)

Rubinstein (1997)
Cross Entropy (CE)

Generalized Entropy

Renyl (1961)
Renyl-Entropy


http://www.scottaraomson.com

Example of the usefulness of ApEn (1/3)


Example of the usefulness of ApEn (2/3)

Let: \( \langle x_n \rangle = \{x_1, x_2, \ldots, x_N\} \)

\[ \tilde{x}_i = (x_i, x_{i+1}, \ldots, x_{i+m-1}) \]

\[ \| \tilde{x}_i, \tilde{x}_j \| = \max_{k=1,2,\ldots,m} \left( |x_{i+k-1} - x_{j+k-1}| \right) \]

\[ \tilde{H}(m, r) = \lim_{N \to \infty} \left[ \phi^m(r) - \phi^{m+1}(r) \right] \]

\[ C_r^m(i) = \frac{N_m(i)}{N-m+1} \quad \phi^m(r) = \frac{1}{N-m+1} \sum_{i=1}^{N-m+1} \ln C_r^m(i) \]


Example: ApEn (2)

ApEn

Heart Rate Variability (HRV) can be used as a marker of cardiovascular health status.

Entropy measures represent a family of new methods to quantify the variability of the heart rate.

A promising approach, due to its ability to discover certain patterns and shifts in the "apparent ensemble amount of randomness" of stochastic processes,

- measure randomness and predictability of processes.

- Entropy:
  - Measure for the uncertainty of random variables
- Kullback-Leibler divergence:
  - comparing two distributions
- Mutual Information:
  - measuring the correlation of two random variables

\[ H[X] = - \sum_x p(x) \log_2 p(x) \]


Important quantity in:
- coding theory
- statistical physics
- machine learning
**Conditional Entropy**

\[ H[y|x] = -\int \int p(y, x) \ln p(y|x) \, dy \, dx \]

\[ H[x, y] = H[y|x] + H[x] \]

**The Kullback-Leibler Divergence**

\[ KL(p||q) = - \int p(x) \ln q(x) \, dx - \left( - \int p(x) \ln p(x) \, dx \right) \]

\[ = - \int p(x) \ln \left( \frac{q(x)}{p(x)} \right) \, dx \]

\[ KL(p||q) \simeq \frac{1}{N} \sum_{n=1}^{N} \{- \ln q(x_n|\theta) + \ln p(x_n)\} \]

\[ KL(p||q) \geq 0 \]

KL-divergence is often used to measure the distance between two distributions.

**Note: KL is not symmetric!**

\[ q^* = \arg\min_q D_{KL}(p||q) \]

\[ q^* = \arg\min_q D_{KL}(q||p) \]

\[ KL(p||q) \neq KL(q||p) \]

**Entropy measures generally ...**

- ... are robust against noise;
- ... can be applied to complex time series with good replication;
- ... is finite for stochastic, noisy, composite processes;
- ... the values correspond directly to irregularities – good for detecting anomalies.

Mutual Information and Point Wise MI

Example: Disease-Disease Relationship

Let two words, $w_i$ and $w_j$, have probabilities $P(w_i)$ and $P(w_j)$. Then their mutual information $PMI(w_i, w_j)$ is defined as:

$$PMI(w_i, w_j) = \log \left(\frac{P(w_i, w_j)}{P(w_i) P(w_j)}\right)$$

For $w_i$ denoting rheumatoid arthritis and $w_j$ representing diffuse scleritis the following simple calculation yields:

$$P(w_i) = \frac{94,030}{1,000,000} \quad P(w_j) = \frac{24}{1,000,000}$$

$$P(w_i, w_j) = \frac{12}{1,000,000} \quad PMI(w_i, w_j) = 7.7.$$
Open Questions, future outlook, hot topics, challenges

- 1) Challenges include -omics data analysis, where KL divergence and related concepts could provide important measures for discovering biomarkers.
- 2) Hot topics are new entropy measures suitable for computations in the context of complex/uncertain data for ML algorithms.
- Inspiring is the abstract geometrical setting underlying ML main problems, e.g. Kernel functions can be completely understood in this perspective. Future work may include entropic concepts and geometrical settings.

Limitations and Open Problems

- The case of higher order statistical structure in the data – nonlinear and hierarchical?
- Outliers in the data – noise models?
- There are $\frac{D(D+1)}{2}$ parameters in a multi-variate Gaussian model – what happens if $D \gg$ ?
  dimensionality reduction

Questions

Thank you!
Appendix
ApEn
Given a signal \( x(n) \), ApEn can be calculated as follows (1):

1. Form \( m \)-vectors, \( X(1) \) to \( X(N-m+1) \) defined by:
   \[ X(i) = [x(i), x(i+1), ..., x(i+m-1)] \quad i = 1, N-m+1 \]

2. Define the distance \( d_p(X(i), X(j)) \) between vectors \( X(i) \) and \( X(j) \) as the maximum absolute difference between their respective scalar components:
   \[ d_p(X(i), X(j)) = \max_{k=0,1,...,m-1} |x(i+k) - x(j+k)| \]

3. Define for each \( i \), for \( i = 1, N-m+1 \), let
   \[ C^p_m(i) = \frac{1}{N-m+1} \sum_{j=1}^{N-m+1} \delta_p^m(X(i), X(j)) \]

4. Take the natural logarithm of each \( C^p_m(i) \), and average it over \( i \) as defined in step 5:
   \[ \delta_p^m(i) = \frac{1}{N-m+1} \sum_{i=1}^{N-m+1} \ln(C^p_m(i)) \]

5. Increase the dimension to \( m+1 \) and repeat steps 1) to 4).

6. Calculate ApEn value for a finite data length of \( N \):
   \[ \text{ApEn}(m, r, N) = \delta^m_N - \delta^{m+1}(r) \]


---

**Backup Slide: Comparison ApEn - SampEn**

SampEn
Given a signal \( x(n) \), SampEn can be calculated as follows (5):

1. Form \( m \)-vectors, \( X(1) \) to \( X(N-m+1) \) defined by:
   \[ X(i) = [x(i), x(i+1), ..., x(i+m-1)] \quad i = 1, N-m+1 \]

2. Define the distance \( d_p(X(i), X(j)) \) between vectors \( X(i) \) and \( X(j) \) as the maximum absolute difference between their respective scalar components:
   \[ d_p(X(i), X(j)) = \max_{k=0,1,...,m-1} |x(i+k) - x(j+k)| \]

3. Define for each \( i \), for \( i = 1, N-m+1 \), let
   \[ C^p_m(i) = \frac{1}{N-m+1} \sum_{j=1}^{N-m+1} \delta_p^m(X(i), X(j)) \]

4. Take the natural logarithm of each \( C^p_m(i) \), and average it over \( i \) as defined in step 5:
   \[ \delta^m_N(i) = \frac{1}{N-m+1} \sum_{i=1}^{N-m+1} \ln(C^p_m(i)) \]

5. Increase the dimension to \( m+1 \) and repeat steps 1) to 4).

6. Calculate SampEn value for a finite data length of \( N \):
   \[ \text{SampEn}(m, r, N) = -\ln \left( \frac{A^m_N}{A^{m+1}(r)} \right) \]


---

**Backup Slide: Graph Entropy Measures**

- The most important question: Which kind of structural information does the entropy measure detect?
- The topological complexity of a molecular graph is characterized by its number of vertices and edges, branching, cyclicity etc.
### Backup: English/German Subject Codes OEFOS 2012

<table>
<thead>
<tr>
<th>Code</th>
<th>English</th>
<th>German</th>
</tr>
</thead>
<tbody>
<tr>
<td>106005</td>
<td>Bioinformatics</td>
<td>Bioinformatik</td>
</tr>
<tr>
<td>106007</td>
<td>Biostatistics</td>
<td>Biostatistik</td>
</tr>
<tr>
<td>304005</td>
<td>Medical Biotechnology</td>
<td>Medizinische Biotechnologie</td>
</tr>
<tr>
<td>305901</td>
<td>Computer-aided diagnosis and therapy</td>
<td>Computerunterstützte Diagnose und Therapie</td>
</tr>
<tr>
<td>304003</td>
<td>Genetic engineering, technology</td>
<td>Gentechnik,-technologie</td>
</tr>
<tr>
<td>3906 (old)</td>
<td>Medical computer sciences</td>
<td>Medizinische Computerwissenschaften</td>
</tr>
<tr>
<td>305906</td>
<td>Medical cybernetics</td>
<td>Medizinische Kybernetik</td>
</tr>
<tr>
<td>305904</td>
<td>Medical documentation</td>
<td>Medizinische Dokumentation</td>
</tr>
<tr>
<td>305905</td>
<td>Medical informatics</td>
<td>Medizinische Informatik</td>
</tr>
<tr>
<td>305907</td>
<td>Medical statistics</td>
<td>Medizinische Statistik</td>
</tr>
</tbody>
</table>

http://www.statistik.at

---

### Advance Organizer (1/2)

- **Abduction** = cyclical process of generating possible explanations (i.e., identification of a set of hypotheses that are able to account for the clinical case on the basis of the available data) and testing those (i.e., evaluation of each generated hypothesis on the basis of its expected consequences) for the abnormal state of the patient at hand;
- **Abstraction** = data are filtered according to their relevance for the solution problem and chunked in schemas representing an abstract description of the problem (e.g., abstracting that an adult male with haemoglobin concentration less than 14g/dl is an anemic patient);
- ** Artefact/surrogate** = error or anomaly in the perception or representation of information though the involved method, equipment or process;
- **Data** = physical entities at the lowest abstraction level which are, e.g., generated by a patient (patient data) or a (biological) process; data contain no meaning;
- **Data quality** = includes quality parameter such as: Accuracy, Completeness, Update status, Relevance, Consistency, Reliability, Accessibility;
- **Data structure** = way of storing and organizing data to use it efficiently;
- **Deduction** = deriving a particular valid conclusion from a set of general premises;
- **DIK-Model** = Data-Information-Knowledge three level model;
- **DIKW-Model** = Data-Information-Knowledge-Wisdom four level model;
- **Disparity** = containing different types of information in different dimensions;
- **Heart rate variability (HRV)** = measured by the variation in the beat-to-beat interval;
- **HRV artifact** = noise through errors in the location of the instantaneous heart beat, resulting in errors in the calculation of the HRV, which is highly sensitive to artifact and errors in as low as 2% of the data will result in unwanted biases in HRV calculations;

---

### Advance Organizer (2/2)

- **Induction** = deriving a likely general conclusion from a set of particular statements;
- **Information** = derived from the data by interpretation (with feedback to the clinician);
- **Information Entropy** = a measure for uncertainty: highly structured data contain low entropy, if everything is in order there is no uncertainty; no surprise, ideally H = 0;
- **Knowledge** = obtained by inductive reasoning with previously interpreted data, collected from many similar patients or processes, which is added to the “body of knowledge” (explicit knowledge). This knowledge is used for the interpretation of other data and to gain implicit knowledge which guides the clinician in taking further action;
- **Large Data** = consist of at least hundreds of thousands of data points;
- **Multi-Dimensionality** = containing more than three dimensions and data are multi-variate;
- **Multi-Modality** = a combination of data from different sources;
- **Multivariate** = encompassing the simultaneous observation and analysis of more than one statistical variable;
- **Reasoning** = process by which clinicians reach a conclusion after thinking on all facts;
- **Spatiality** = contains at least one (non-scalar) spatial component and non-spatial data;
- **Structural Complexity** = ranging from low-structured (simple data structure, but many instances, e.g., flow data, volume data) to high-structured data (complex data structure, but only a few instances, e.g., business data);
- **Time-Dependency** = data is given at several points in time (time series data);
- **Voxel** = volumetric pixel = volumetric picture element;
### Mathematical Notation

"In mathematics you don't understand things. You just get used to them" – John von Neumann

**Data**
- \( n \) \text{ Number of samples }
- \( d \) \text{ Number of input variables }
- \( X = [x_1, \ldots, x_d] \) \text{ Matrix of input samples }
- \( y = [y_1, \ldots, y_d] \) \text{ Vector of output samples }
- \( Z = [z_1, \ldots, z_d] \) \text{ Combined input-output training data or representation of data points in a feature space }

**Distribution**
- \( p \) \text{ Probability }
- \( F(x) \) \text{ Cumulative probability distribution function (cdf) }
- \( f(x) \) \text{ Probability density function (pdf) }
- \( p(x, y) \) \text{ Joint probability density function }
- \( p(y|x) \) \text{ Probability density function, which is parameterized }
- \( p(y|x) \) \text{ Conditional density }
- \( t(x) \) \text{ Target function }

---

### Glossary
- \( \text{ApEn} \) = Approximate Entropy;
- \( C_{\text{data}} \) = Data in computational space;
- \( \text{DIK} \) = Data-Information-Knowledge-3-Level Model;
- \( \text{DIKW} \) = Data-Information-Knowledge-Wisdom-4-Level Model;
- \( \text{GraphEn} \) = Graph Entropy;
- \( H \) = Entropy (General);
- \( \text{HRV} \) = Heart Rate Variability;
- \( \text{MaxEn} \) = Maximum Entropy;
- \( \text{MinEn} \) = Minimum Entropy;
- \( \text{NE} \) = Normalized entropy (measures the relative informational content of both the signal and noise);
- \( I^{\text{data}} \) = Data in perceptual space;
- \( \text{PDB} \) = Protein Data Base;
- \( \text{SampEn} \) = Sample Entropy;

---

### Scientists in data integration - selection - incomplete!

---

### Clinical view on data – information, and knowledge

Status as of 04.04.2016
Clinical View of Data, Information, Knowledge

From Patient Data to Medical Knowledge

Human Information Processing Model

Our definition of Knowledge – adaptive agent


A space as a set of points

Let us collect $n$-dimensional $i$ observations: $x_i = [x_{i1}, ..., x_{in}]$

Example Metric Space

A set $S$ with a metric function $d$ is a metric space

$\quad \quad d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}$


Point Cloud Data Sets

Point cloud in $\mathbb{R}^2$  topological space  metric space


ML-Jungle Top Level View and the focus of today ...

Always with a focus/application in health informatics

Holzinger Group, hci-kdd.org

117

MADE Health Verona 02
Grand challenges in Machine Learning for Health

- Big data with many training sets (this is good for MLI)
- Small number of data sets, rare events
- Very-high-dimensional problems
- Complex data – NP-hard problems
- Missing, dirty, wrong, noisy, ..., data

GENERALISATION

TRANSFER


Discrete versus continuous random variable

- X: S → R (“measure” of outcome)
- Events can be defined according to X
  - E(X=a) = {s_i | X(s_i)=a}
  - E(X≥a) = {s_i | X(s_i) ≥ a}
- Consequently, probabilities can be defined on X
  - P(X=a) = P(E(X=a))
  - P(a≥X) = P(E(a≥X))
- partitioning the sample space

My greatest concern was what to call it. I thought of calling it “information”, but the word was overly used, so I decided to call it “uncertainty”. When I discussed it with John von Neumann, he had a better idea. Von Neumann told me, “You should call it entropy, for two reasons. In the first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important, nobody knows what entropy really is, so in a debate you will always have the advantage.”

A measure for uncertainty (2/3)

\[ \log_2 \frac{1}{p} = -\log_2 p \]

\[ H = - \sum_{i=1}^{N} p_i \log_2 (p_i) \]


Entropy H as a measure for uncertainty (3/3)

\[ H_B = - \sum_{k=1}^{B} \log_2 p_k = -1 \cdot \log_2 (1) = 0 \]

\[ H_B = - \sum_{k=1}^{B} \frac{1}{B} \log_2 \left( \frac{1}{B} \right) = \log_2 (B) \]

\[ H = H_{\text{max}} = \log_2 N \]

Background on Information Theory

- Developed by Claude Shannon in the 1940s
- Maximizing the amount of information that can be transmitted over an imperfect communication channel
- Data compression (entropy)
- Transmission rate (channel capacity)

Vapnik – Chervonenkis

- The VC dimension is a measure of the capacity of a space of functions that can be learned by a statistical classification algorithm. It is defined as the cardinality of the largest set of points that the algorithm can shatter. It is a core concept in Vapnik–Chervonenkis theory


\[ Q \ldots P = \{ p_1, \ldots, p_n \} \]
\[ H(Q) = - \sum_{i=1}^{n} (p_i \ast \log p_i) \]

\[ Qb = \{ a_1, a_2 \} \text{ with } P = \{ p, 1-p \} \]
\[ H(Qb) = p \ast \log \frac{1}{p} + p \ast \log \frac{1}{1-p} \]


Entropic methods – what for?

- 1) Set of noisy, complex data
- 2) Extract information out of the data
- 3) to support a previous set hypothesis
- Information + Statistics + Inference
- = powerful methods for many sciences
- Application e.g. in biomedical informatics for analysis of ECG, MRI, CT, PET, sequences and proteins, DNA, topography, for modeling etc. etc.