Andreas Holzinger
VO 709.049 Medical Informatics
01.02.2017 11:15-12:45

Lecture 12 Course Summary
and Future Outlook
(Reflection Lecture)

a.holzinger@tugraz.at
Tutor: markus.plass@student.tugraz.at
http://hci-kdd.org/biomedical-informatics-big-data
This whole lecture is a reflection course ... so please
... fasten your seat belts and be prepared for questions!
Lecture 1: Computer Science meets Life Sciences
Repetition of Bayes - on the work of Laplace

What is the simplest mathematical operation for us?

\[ p(x) = \sum_x (p(x, y)) \]  \hspace{1cm} (1)

How do we call repeated adding?

\[ p(x, y) = p(y|x) \times p(y) \]  \hspace{1cm} (2)

Laplace (1773) showed that we can write:

\[ p(x, y) \times p(y) = p(y|x) \times p(x) \]  \hspace{1cm} (3)

Now we introduce a third, more complicated operation:

\[ \frac{p(x, y) \times p(y)}{p(y)} = \frac{p(y|x) \times p(x)}{p(y)} \]  \hspace{1cm} (4)

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

\[ p(x, y) = \frac{p(y|x) \times p(x)}{p(y)} \quad p(h|d) = \frac{p(d|h)p(h)}{p(d)} \]  \hspace{1cm} (5)
$d \ldots \text{data}$
$h \ldots \text{hypotheses}$

$H \ldots \{H_1, H_2, \ldots, H_n\}$

$\forall h, d \ldots$

$$p(h|d) = \frac{p(d|h) \cdot p(h)}{\sum_{h' \in H} p(d|h') \cdot p(h')}$$

Posterior Probability

Likelihood

Prior Probability

Evidence = marginal likelihood = Normalization

Feature parameter $\theta$
Your MD has bad news and good news for you.
Bad news first: You are tested positive for a serious disease D, and this test T is 99% accurate
Good news: It is a rare disease, striking only 1 in 10,000 (D)

How worried would you now be – or: what is the posterior?

\[
\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}
\]

\[
p(T = 1|D = 1) = p(d|h) = 0,99 \quad \text{and} \quad p(D = 1) = p(h) = 0,0001
\]

and \( p(T = 0|D = 0) = 0,99 \)

\[
p(D = 1|T = 1) = \frac{(0,99)(0,0001)}{(1-0,99)(1-0,0001)+0,99*0,0001} = 0,0098 = 0,9%\]
Where is the highest certainty in the signal below?

\[ p(f(x) \mid D) \propto p(D \mid f(x)) \times p(f(x)) \]

Lecture 2:
Data, Information, Knowledge; Entropy and Kullback-Leibler Divergence
Warm-up Quiz

\[ H(X) = -\sum_{i=1}^{n} P(x_i) \log_b P(x_i) \]

\[ D_{KL}(P||Q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} \, dx \]

Solutions in the Appendix
What data do we have in biomedical informatics?

- Electronic health record data
- Physiological data
- Laboratory results
- Imaging data
  - X-Ray, ultrasound, MR, CT, PET, cams, observation (e.g. sleep laboratory), gait (walking)
- Genomics
- Proteomics
- Protein-Protein Interactions
- Metabolomics
- Chemical processes
- Cellular reactions
- Enzymatic reactions
  - Transcriptomics
  - RNA, mRNA, rRNA, tRNA
- Epigenetics
- Epigenetic modifications
- Foodomics, Lipidomics
- Nutrition data (Nutrigenomics)
- Diet data (allergenics)
- Exposome
- Environmental data
- Air pollution Exposure (toxicants)

Collective data

Social data

Fitness, Wellness data

Ambient Assisted Living data
(Non-medical) personal data

Collective

Individual

Tissue

Cell

Bacteria

Virus

Molecule

Atom
Question: Why is it so important to know such structures?

What is/When do we need Shannon-Entropy?

\[ H(X) = - \sum_{i=1}^{n} P(x_i) \log_b P(x_i) \]

- Measuring uncertainty, complexity, randomness, surprise, ..., = **information!**
Mutual Information - Conditional Entropy

- In ML we need often to measure the **difference** between two probability distributions

\[
I(X; Y) = H(X) - H(X|Y)
\]

\[
I(X; Y) = \sum_{y \in Y} \sum_{x \in X} p(x, y) \log \left( \frac{p(x, y)}{p(x) p(y)} \right)
\]
Kullback-Leibler Divergence - discriminative information

For discrete distributions

\[ D_{KL}(P \| Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)} \]

For continuous distributions

\[ D_{KL}(P \| Q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} \, dx \]

KL-divergence can also be used to measure the distance between two distributions

Lecture 3: Knowledge Representation, Ontologies & Classifications
What is this, and why is it important?

http://www.kurzweilai.net/images/cell-model.png  
(Credit: UC San Diego School of Medicine)
Slide 3-7: Standardized workflow of ECG data processing

### What is a knowledge representation?

<table>
<thead>
<tr>
<th>Mathematical Logic</th>
<th>Psychology</th>
<th>Biology</th>
<th>Statistics</th>
<th>Economics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aristotle</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Descartes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Boole</td>
<td>James</td>
<td></td>
<td>Laplace</td>
<td>Bentham, Pareto</td>
</tr>
<tr>
<td>Frege, Peano</td>
<td>Hebb, Bruner</td>
<td>Lashley, Rosenblatt</td>
<td>Bernoullii</td>
<td>Friedman</td>
</tr>
<tr>
<td>Goedel, Post, Church, Turing, Davis, Putnam, Robinson</td>
<td>Miller, Newell, Simon</td>
<td>Lashley, Letwin, McCulloch, Pitts, Heubel, Weisel</td>
<td>Bayes, Tversky, Kahneman</td>
<td>Von Neumann, Simon, Raiffa</td>
</tr>
</tbody>
</table>

Why is this image important?

- Clinical repositories
- SNOMED
- Other subdomains
- OMIM
- Genetic knowledge bases
- MeSH
- Biomedical literature
- GO
- UWDA
- Anatomy
- Model organisms
- Genome annotations

TU Graz
Holzinger Group
Lecture 4: Decision, Cognition, Uncertainty, Bayesian Statistics, Probabilistic Modelling
When is the human *) better?

*) human intelligence/natural intelligence/human mind/human brain/learning

- **Natural Language Translation/Curation**
  Computers cannot understand the context of sentences [3]

- **Unstructured problem solving**
  Without a pre-set of rules, a machine has trouble solving the problem, because it lacks the creativity required for it [1]

- **NP-hard Problems**
  Processing times are often exponential and makes it almost impossible to use machines for it, but human make heuristic decisions which are often not perfect but sufficiently good [4]


When is the computer **) better?

**) Computational intelligence, Artificial Intelligence/soft computing/ML

- **High-dimensional data processing**
  Humans are very good at dimensions less or equal than 3, but computers can process data in arbitrarily high dimensions

- **Rule-Based environments**
  Difficulties for humans in rule-based environments often come from not recognizing the correct goal in order to select the correct procedure or set of rules [2]

- **Image optimization**
  Machine can look at each pixel and apply changes without human personal biases, and with more speed [1]
What does the Signal Detection Theory describe?

Two doctors, with equally good training, looking at the same CT scan, will have the same information ... but they may have a different bias/criteria!
For a single decision variable an agent can select $D = d$ for any $d \in \text{dom}(D)$.

The expected utility of decision $D = d$ is

$$E(U \mid d) = \sum_{x_1, \ldots, x_n} P(x_1, \ldots, x_n \mid d) U(x_1, \ldots, x_n, d)$$

An optimal single decision is the decision $D = d_{\text{max}}$ whose expected utility is maximal:

$$d_{\text{max}} = \arg \max_{d \in \text{dom}(D)} E(U \mid d)$$

Lecture 5: Probabilistic Graphical Models I: From Knowledge Representation to Graph Learning
This protein graph model was worth the Nobel Prize 2013...

http://sbcb.bioch.ox.ac.uk/users/oliver/software/
Remember: Decision trees are coming from Clinical Practice


Decision node

- No further surgery
  - Dies
  - Microinvasive cancer of the cervix
    - Radical hysterectomy (p=99.5%)
    - Survives (p=2%)
  - Spread (p=2%)

Death from cancer
- Probability 2%
- Utility 5%

Fertile survival
- Probability 98%
- Utility 100%

Surgical death
- Probability 0.5%
- Utility 0%

Infertile survival
- Probability 5%
- Utility 95%

Infertile survival
- Probability 5%
- Utility 5%

Physician treating a patient approx. 480 B.C.
Beazley (1963), Attic Red-figured Vase-Painters, 813, 96.
Department of Greek, Etruscan and Roman Antiquities, Sully, 1st floor, Campana Gallery, room 43 Louvre, Paris
What are Probabilistic Graphical Models?

- PGM can be seen as a combination between
- **Graph Theory + Probability Theory + Machine Learning**
- One of the most exciting AI advances in the last decades
- Compact representation for exponentially-large probability distributions
- Example Question: “Is there a path connecting two proteins?”
- $Path \ (X, Y) = edge \ (X, Y)$
- $Path \ (X, Y) = edge \ (X, Y), path \ (Z, Y)$
- This can NOT be expressed in first-order logic
- Need a Turing-complete fully-fledged language

- Take patient information, e.g., observations, symptoms, test results, -omics data, etc. etc.
- Reach conclusions, and predict into the future, e.g. how likely will the patient be re-admissioned
- **Prior** = belief before making a particular observation
- **Posterior** = belief after making the observation and is the prior for the next observation – intrinsically incremental

\[
p(x_i|y_j) = \frac{p(y_j|x_i)p(x_i)}{\sum p(x_i, y_j)p(x_i)}
\]
Identifying Networks in Disease Research

Network Basics on the Example of Bioinformatics

\[ G(V, E) \]
\[ V \text{ ... vertex} \]
\[ E \text{ ... edge \{a, b\}} \]
\[ a, b \in V; a \neq b \]

Lecture 6: Probabilistic Graphical Models II: From Bayesian Networks to Graph Bandits
Problem: Is Graph Isomorphism NP-complete?


- Important for health informatics: Discovering relationships between biological components
- Unsolved problem in computer science:
  - Can the graph isomorphism problem be solved in polynomial time?
    - So far, no polynomial time algorithm is known.
    - It is also not known if it is NP-complete
    - We know that subgraph-isomorphism is NP-complete
• is a **probabilistic model**, consisting of two parts:
  1) a dependency structure and
  2) local probability models.

\[ p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i \mid Pa(x_i)) \]

Where \( Pa(x_i) \) are the parents of \( x_i \).

BN inherently model the **uncertainty in the data**. They are a successful marriage between probability theory and graph theory; allow to model a multidimensional probability distribution in a sparse way by searching independency relations in the data. Furthermore this model allows different strategies to integrate two data sources.

Example: Bayes Net with four binary variables

Three types of Probabilistic Graphical Models

Undirected: Markov random fields, useful e.g. for computer vision (Details: Murphy 19)

\[ P(X) = \frac{1}{Z} \exp \left( \sum_{ij} W_{ij} x_i x_j + \sum_i x_i b_i \right) \]

Directed: Bayes Nets, useful for designing models (Details: Murphy 10)

\[ p(x) = \prod_{k=1}^{K} p(x_k | pa_k) \]

Factored: useful for inference/learning

\[ p(x) = \prod_{s} f_{s}(x_{s}) \]

Tutorial on Factor Graphs http://deepdive.stanford.edu/inference
Lecture 7: Dimensionality Reduction and Subspace Clustering with the Doctor-in-the-Loop
Discovery vs. Prediction

Typical questions include:

- Is this protein functioning as an enzyme?
- Does this gene sequence contain a splice site?
- Is this melanoma malignant?

Given object $x$ – predict the class label $y$

- If $y \in \{0, 1\}$ → binary classification problem
- If $y \in \{1, \ldots, n\}$ and is $n \in \mathbb{N}$ → multiclass problem
- If $y \in \mathbb{R}$ → regression problem
Why do we need Clustering?

- Group similar objects into clusters together, e.g.
  - For image segmentation
  - Grouping genes similarly affected by a disease
  - Clustering patients with similar diseases
  - Cluster biological samples for category discovery
  - Finding subtypes of diseases
  - Visualizing protein families

- Inference: given $x_i$, predict $y_i$ by learning $f$
- No training data set – learn model and apply it
Why do we need Subspace Clustering in medical data?

Lecture 8: Decision Making under Uncertainty: Decision Support Systems
Two types of decisions (Diagnosis vs. Therapy)

- **Type 1 Decisions**: related to the diagnosis, i.e. computers are used to assist in diagnosing a disease on the basis of the individual patient data. Questions include:
  - What is the probability that this patient has a myocardial infarction on the basis of given data (patient history, ECG, ...)?
  - What is the probability that this patient has acute appendices, given the signs and symptoms concerning abdominal pain?

- **Type 2 Decisions**: related to therapy, i.e. computers are used to select the best therapy on the basis of clinical evidence, e.g.:
  - What is the best therapy for patients of age x and risks y, if an obstruction of more than z % is seen in the left coronary artery?
  - What amount of insulin should be prescribed for a patient during the next 5 days, given the blood sugar levels and the amount of insulin taken during the recent weeks?

Helps to make rational decisions (risks vs. success)

Expected Value of Surgery

Surgery

- **Death**: 0.05
- **Survival**: 0.95

No Surgery

- **Death**: 0
- **Wheelchair**: 3
- **Full mobility**: 10
- **Poor mobility**: 6

Values:
- **7.7**
- **0.05**
- **0.95**
- **0.05**
- **0.95**
- **0.6**
- **0.4**
Original Example from MYCIN

\[ h_1 = \text{The identity of ORGANISM-1 is streptococcus} \]
\[ h_2 = \text{PATIENT-1 is febrile} \]
\[ h_3 = \text{The name of PATIENT-1 is John Jones} \]

\[ CF[h_1,E] = .8 \quad : \quad \text{There is strongly suggestive evidence (.8) that the identity of ORGANISM-1 is streptococcus} \]
\[ CF[h_2,E] = - .3 \quad : \quad \text{There is weakly suggestive evidence (.3) that PATIENT-1 is not febrile} \]
\[ CF[h_3,E] = + 1 \quad : \quad \text{It is definite (1) that the name of PATIENT-1 is John Jones} \]

State-of-the-art architecture of DSS

Why is this model so interesting for us?

Corchado et al. (2009)
Lecture 9: Interactive Visualization and Visual Analytics
**Scatterplot** = oldest, point-based technique, projects data from n-dim space to an arbitrary k-dim display space;

**Parallel coordinates** = (PCP), originally for the study of high-dimensional geometry, data point plotted as polyline;

**RadViz** = Radial Coordinate visualization, is a “force-driven” point layout technique, based on Hooke’s law for equilibrium;
Radar chart (star plot, spider web, polar graph, polygon plot) = radial axis technique;

Heatmap = a tabular display technique using color instead of figures for the entries;

Glyph = a visual representation of the entity, where its attributes are controlled by data attributes;

Chernoff face = a face glyph which displays multivariate data in the shape of a human face
1) What facets of the target information should be visualized?

2) What data source should each facet be linked to and what relationships these facets have?

3) What layout algorithm should be used to visualize each facet?

4) What interactive techniques should be used for each facet and for which infovis tasks?

---

Lecture 10: Biomedical Information Systems and Knowledge Management
A workflow is defined as a process that contains tasks $T$, and the respective rules on how those tasks are executed:

- **Workflow $W$**: $(T, P, C, A, S_0)$ where
  - $T = \{T_1, T_2, \ldots, T_m\}$ A set of tasks, $m \geq 1$
  - $P = (p_{ij})_{m \times m}$ **Precedence matrix of the task set**
  - $C = (c_{ij})_{m \times m}$ **Conflict matrix of the task set**
  - $A = (A(T_1), A(T_2), \ldots, A(T_m))$ Pre-Condition set for each task
  - $S_0 \in \{0, 1, 2, 3\}_m$ is the initial state

---

\[ T = \{T_1, T_2, \ldots, T_8\}, \]

\[ A(T_1) = \emptyset, A(T_2) = \{\{T_1\}, \{T_6\}\}, A(T_3) = \{\{T_1\}\}, \]

\[ A(T_4) = \{\{T_2\}\}, A(T_5) = \{\{T_4\}\}, \]

\[ A(T_6) = A(T_7) = \{\{T_5\}\}, A(T_8) = \{\{T_3, T_7\}\}. \]

\[ S_0 = (1, 0, 0, 0, 0, 0, 0, 0). \]

\[
\begin{bmatrix}
0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}
\]

\[
\begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}
\]

Lecture 11: Privacy, Data Protection, Safety, Security & Privacy Aware Machine Learning
What are the Requirements of an electronic patient record

Anonymization: Personal data cannot be re-identified (e.g. k-Anonymization)
Pseudonymization: The personal data is replaced by a “pseudonym”, which allows later tracking back to the source data (re-identification)
Which Definitions do we need to know?

- **Safety** = any protection from harm, injury, or damage;
- Data Protection = all measures to ensure availability and integrity of data
- **Privacy** = (US pron. “prai ...”; UK pron. “pri ...”; from Latin: privatus "separated from the rest", are the individual rights of people to protect their personal life and matters Confidentiality = secrecy ("ärztliche Schweigepflicht")

This poses a big privacy problem

87% of the population in the USA can be uniquely re-identified by Zip-Code, Gender and date of birth

<table>
<thead>
<tr>
<th>Hospital Patient Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Birthdate</td>
</tr>
<tr>
<td>1/21/76</td>
</tr>
<tr>
<td>4/13/68</td>
</tr>
<tr>
<td>2/28/76</td>
</tr>
<tr>
<td>1/21/76</td>
</tr>
<tr>
<td>4/13/86</td>
</tr>
<tr>
<td>2/28/76</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Voter Registration Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
</tr>
<tr>
<td>Andre</td>
</tr>
<tr>
<td>Beth</td>
</tr>
<tr>
<td>Carol</td>
</tr>
<tr>
<td>Dan</td>
</tr>
<tr>
<td>Ellen</td>
</tr>
</tbody>
</table>


**K-Anonymity** ... a release of data is said to have the *k-anonymity property* if the information for each person contained in the release cannot be distinguished from at least \( k - 1 \) individuals whose information also appear in the release.

**L-Diversity** ... extension requiring that the values of all confidential attributes within a group of \( k \) sets contain at least \( L \) clearly distinct values

**t-Closeness** ... extension requiring that the distribution of the confidential attribute within a group of \( k \) records is similar to the confidential attribute’s distribution in the whole data set (local distribution must resemble the global distribution)
The Right to Be Forgotten: Towards Machine Learning on Perturbed Knowledge Bases

Bernd Malle$^{1,2}$, Peter Kieseberg$^{1,2}$, Edgar Weippl$^2$, and Andreas Holzinger$^1$(S)

$^1$ Holzinger Group HCI-KDD, Institute for Medical Informatics, Statistics and Documentation, Medical University Graz, Graz, Austria
{b.malle, a.holzinger}@hci-kdd.org

$^2$ SBA Research gGmbH, Favoritenstrae 16, 1040 Vienna, Austria
PKieseberg@sba-research.org

Abstract. Today’s increasingly complex information infrastructures represent the basis of any data-driven industries which are rapidly becoming the 21st century’s economic backbone. The sensitivity of these infrastructures to disturbances in their knowledge bases is therefore of crucial interest for companies, organizations, customers and regulating bodies. This holds true with respect to the direct provisioning of such information in crucial applications like clinical settings or the energy industry, but also when considering additional insights, predictions and personalized services that are enabled by the automatic processing of those data. In the light of new EU Data Protection regulations applying from 2018 onwards which give customers the right to have their data deleted on request, information processing bodies will have to react to these changing jurisdictional (and therefore economic) conditions. Their choices include a re-design of their data infrastructure as well as preventive actions like anonymization of databases per default. Therefore, insights into the effects of perturbed/anonymized knowledge bases on


Keywords: Machine learning · Knowledge bases · Right to be forgotten · Perturbation · Anonymization · k-anonymity · SaNGreeA ·
Very recent: News from 28.01.2017

Europäischer Datenschutz in der Big-Data-Welt

28.01.2017 12:30 Uhr  -  Monika Ermert, dpa

(Dbild: Håkan Dahlström CC BY 2.0)


Der Beirat der Datenschutzkonvention des Europarats legte zum Internationalen Datenschuthtag „Richtlinien zum Schutz persönlicher Daten in einer Big Data-Welt“ vor.

https://www.heise.de/newsticker/meldung/Europaeischer-Datenschutz-in-der-Big-Data-Welt-3609737.html
KEEP CALM AND GOOD LUCK WITH THE EXAM!