Lecture 06 Probabilistic Graphical Models II: From Bayesian Networks to Graph Bandits

da.holzinger@tugraz.at
Tutor: markus.plass@student.tugraz.at
http://hci-kdd.org/biomedical-informatics-big-data
ML needs a concerted effort fostering integrated research

http://hci-kdd.org/international-expert-network

Interactive

Data Mining

Knowledge Discovery

1. Data Mapping
2. Learning Algorithms
3. Graph-based Data Mining
4. Topological Data Mining
5. Entropy-based Data Mining
6. Data Visualization
7. Privacy, Data Protection, Safety and Security


© a.holzinger@hci-kdd.org
Agenda for today

- 00 Reflection – follow-up from last lecture
- 01 Graphical Models and Decision Making
- 02 Bayesian Networks
- 03 Machine Learning on Graphs
- 04 Little Excursus: What is similarity?
- 05 Probabilistic Topic Models
- 06 Graph Bandits (a very hot topic!)
00 Reflection
Quiz

1. Diagram of a network with labels a, b, c, d, e, and f.

2. Diagram of a network with nodes and edges labeled T_e, T_w, and T_u.

3. Diagram of a network with nodes A, B, C, D, and E, connected by arrows.

4. Diagram of a network with labels h1, h2, and h3, and associated text:
   - h1 = The identity of ORGANISM-1 is streptococcus
   - h2 = PATIENT-1 is febrile
   - h3 = The name of PATIENT-1 is John Jones
   - CF[h1,E] = .8: There is strongly suggestive evidence (.8) that the identity of ORGANISM-1 is streptococcus
   - CF[h2,E] = - .3: There is weakly suggestive evidence (-.3) that PATIENT-1 is not febrile
   - CF[h3,E] = + 1: It is definite (1) that the name of PATIENT-1 is John Jones

5. Diagram of a network with nodes labeled with proteins A, B, C, D, E, F, G, and H.

6. Diagram of a network with a highlighted node and connections to other nodes.
01 Graphical Models and Decision Making

\[ \mathcal{D} \equiv \{ X_1^{(i)}, X_2^{(i)}, \ldots, X_m^{(i)} \}_{i=1}^N \]
Decision Making: Learn good policy for selecting actions

Goal: Learn an **optimal policy** for selecting best actions within a given **context**

For $t = 1, \ldots, T$

1) The world produces an uncertain “context” $x_t \in X$

2) The learner selects an action $a_t \in \{1, \ldots, K\}$

3) The world reacts with a reward $r_t(a_t) \in [0,1]$
Medicine is an extremely complex application domain – dealing most of the time with uncertainties -> **probable information**!

When we have big data but little knowledge automatic ML can help to gain insight:

**Structure learning and prediction in large-scale biomedical networks with probabilistic graphical models**

If we have little data and deal with NP-hard problems we still need the **human-in-the-loop**!
Example: Why is prediction in proteins so important for us?


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)

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

**Factored:** useful for inference/learning

\[
p(x) = \prod_{s} f_s(x_s)
\]
What is the advantage of factor graphs?

<table>
<thead>
<tr>
<th></th>
<th>Dependency</th>
<th>Efficient Inference</th>
<th>Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bayesian Networks</td>
<td>Yes</td>
<td>Somewhat</td>
<td>Ancestral Generative Process</td>
</tr>
<tr>
<td>Markov Networks</td>
<td>Yes</td>
<td>No</td>
<td>Local Couplings and Potentials</td>
</tr>
<tr>
<td>Factor Graphs</td>
<td>No</td>
<td>Yes</td>
<td>Efficient, distributed inference</td>
</tr>
</tbody>
</table>

Table credit to Ralf Herbrich, Amazon
Hypothesis: most biological functions involve the interactions between many proteins, and the complexity of living systems arises as a result of such interactions.

In this context, the problem of inferring a global protein network for a given organism, - using all (genomic) data of the organism, is one of the main challenges in computational biology


- **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
Protein network inference from multiple genomic data: a supervised approach

Y. Yamanishi¹,* J.-P. Vert² and M. Kanehisa¹

¹Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan and ²Computational Biology group, Ecole des Mines de Paris, 35 rue Saint-Honoré, 77305 Fontainebleau cedex, France

\[ K_{\text{exp}} \] (Expression)
\[ K_{\text{ppi}} \] (Protein interaction)
\[ K_{\text{loc}} \] (Localization)
\[ K_{\text{phy}} \] (Phylogenetic profile)
\[ K_{\text{exp}} + K_{\text{ppi}} + K_{\text{loc}} + K_{\text{phy}} \] (Integration)

True positive

False positive
Example: Data fusion and Protein Annotation

A statistical framework for genomic data fusion

Gert R. G. Lanckriet¹, Tijl De Bie², Nello Cristianini³, Michael I. Jordan² and William Stafford Noble⁵.*

¹Department of Electrical Engineering and Computer Science, ²Division of Computer Science, Department of Statistics, University of California, Berkeley 94720, USA, ³Department of Electrical Engineering, ESAT-SCD, Katholieke Universiteit Leuven 3001, Belgium, ⁴Department of Statistics, University of California, Davis 95618, USA and ⁵Department of Genome Sciences, University of Washington, Seattle 98195, USA

02 Bayesian Networks
“Bayes’ Nets”
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: Directed Bayesian Network with 7 nodes

\[
p(X_1, \ldots, X_7) = \\
p(X_1)p(X_2)p(X_3)p(X_4|X_1, X_2, X_3) \cdot \\
p(X_5|X_1, X_3)p(X_6|X_4)p(X_7|X_4, X_5)
\]
Important in Clinical practice -> prognosis!

- = the prediction of the future course of a disease conditional on the patient’s history and a projected treatment strategy
- Danger: probable Information!
- Therefore valid prognostic models can be of great benefit for clinical decision making and of great value to the patient, e.g., for notification and quality-of-life decisions

Example: Breast cancer - Probability Table

<table>
<thead>
<tr>
<th>Category</th>
<th>Node description</th>
<th>State description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diagnosis</td>
<td>Breast cancer</td>
<td>Present, absent.</td>
</tr>
<tr>
<td>Clinical history</td>
<td>Habit of drinking alcoholic beverages and smoking</td>
<td>Yes, no.</td>
</tr>
<tr>
<td></td>
<td>Taking female hormones</td>
<td>Yes, no.</td>
</tr>
<tr>
<td></td>
<td>Have gone through menopause</td>
<td>Yes, no.</td>
</tr>
<tr>
<td></td>
<td>Have ever been pregnant</td>
<td>Yes, no.</td>
</tr>
<tr>
<td></td>
<td>Family member has breast cancer</td>
<td>Yes, no.</td>
</tr>
<tr>
<td>Physical findings</td>
<td>Nipple discharge</td>
<td>Yes, no.</td>
</tr>
<tr>
<td></td>
<td>Skin thickening</td>
<td>Yes, no.</td>
</tr>
<tr>
<td></td>
<td>Breast pain</td>
<td>Yes, no.</td>
</tr>
<tr>
<td></td>
<td>Have a lump(s)</td>
<td>Yes, no.</td>
</tr>
<tr>
<td>Mammographic findings</td>
<td>Architectural distortion</td>
<td>Present, absent.</td>
</tr>
<tr>
<td></td>
<td>Mass</td>
<td>Score from one to three, score from four to five, absent</td>
</tr>
<tr>
<td></td>
<td>Microcalcification cluster</td>
<td>Score from one to three, score from four to five, absent</td>
</tr>
<tr>
<td></td>
<td>Asymmetry</td>
<td>Present, absent.</td>
</tr>
</tbody>
</table>

Breast cancer – big picture – state of 1999

10 years later: Integration of microarray data

- Integrating microarray data from multiple studies to increase sample size;
- = approach to the development of more robust prognostic tests

Example: Bayes Net with four binary variables

First the structure is learned using a search strategy.

Since the number of possible structures increases super exponentially with the number of variables,

the well-known greedy search algorithm K2 can be used in combination with the Bayesian Dirichlet (BD) scoring metric:

$$p(S|D) \propto p(S) \prod_{i=1}^{n} \prod_{j=1}^{q_i} \left[ \frac{\Gamma(N'_{ij})}{\Gamma(N'_{ij} + N_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(N'_{ijk} + N_{ijk})}{\Gamma(N'_{ijk})} \right]$$

$N_{ijk}$ ... number of cases in the data set $D$

having variable $i$ in state $k$ associated with the $j$-th instantiation of its parents in current structure $S$.

$n$ is the total number of variables.
Next, $N_{ij}$ is calculated by summing over all states of a variable:

$$N_{ij} = \sum_{k=1}^{r_i} N_{ijk} \cdot N'_{ijk}$$

and $N'_{ij}$ have similar meanings but refer to prior knowledge for the parameters.

When no knowledge is available they are estimated using $N_{ijk} = N / (r_i q_i)$

with $N$ the equivalent sample size,

$r_i$ the number of states of variable $i$ and

$q_i$ the number of instantiations of the parents of variable $i$.

$\Gamma(\cdot)$ corresponds to the gamma distribution.

Finally $p(S)$ is the prior probability of the structure.

$p(S)$ is calculated by:

$$p(S) = \prod_{i=1}^{n} \prod_{l_i=1}^{p_i} p(l_i \rightarrow x_i) \prod_{m_i=1}^{o_i} p(m_i x_i)$$

with $p_i$ the number of parents of variable $x_i$ and $o_i$ all the variables that are not a parent of $x_i$.

Next, $p(a \rightarrow b)$ is the probability that there is an edge from $a$ to $b$ while $p(ab)$ is the inverse, i.e. the probability that there is no edge from $a$ to $b$.
Parameter learning -> second step

- Estimating the parameters of the local probability models corresponding with the dependency structure.
- CPTs are used to model these local probability models.
- For each variable and instantiation of its parents there exists a CPT that consists of a set of parameters.
- Each set of parameters was given a uniform Dirichlet prior:

$$p(\theta_{ij}|S) = Dir(\theta_{ij}|N'_{ij1}, ..., N'_{ijk}, ..., N'_{ijr_i})$$

Note: With $\theta_{ij}$ a parameter set where $i$ refers to the variable and $j$ to the $j$-th instantiation of the parents in the current structure. $\theta_{ij}$ contains a probability for every value of the variable $x_i$ given the current instantiation of the parents. $Dir$ corresponds to the Dirichlet distribution with $(N'_{ij1}, ..., N'_{ijr_i})$ as parameters of this Dirichlet distribution. Parameter learning then consists of updating these Dirichlet priors with data. This is straightforward because the multinomial distribution that is used to model the data, and the Dirichlet distribution that models the prior, are conjugate distributions. This results in a Dirichlet posterior over the parameter set:

$$p(\theta_{ij}|D, S) = Dir(\theta_{ij}|N'_{ij1} + N_{ij1}, ..., N'_{ijk} + N_{ijk}, ..., N'_{ijr_i} + N_{ijr_i})$$

with $N_{ijk}$ defined as before.
Inference in Bayes Nets is intractable (NP-complete!)

- For certain cases it is tractable if:
  - Just one variable is unobserved
  - We have singly connected graphs (no undirected loops -> belief propagation)
  - Assigning probability to fully observed set of variables
- Possibility: Monte Carlo Methods (generate many samples according to the Bayes Net distribution and then count the results)
- Otherwise: approximate solutions, NOTE:
  Sometimes it is better to have an approximate solution to a complex problem – than a perfect solution to a simplified problem
3) Machine Learning on Graphs
The two main forms of lymphoma are Hodgkin lymphoma and non-Hodgkin lymphoma (NHL). Lymphoma occurs when cells of the immune system called lymphocytes, a type of white blood cell, grow and multiply uncontrollably. Cancerous lymphocytes can travel to many parts of the body, including the lymph nodes, spleen, bone marrow, blood, or other organs, and form a mass called a tumor. The body has two main types of lymphocytes that can develop into lymphomas: B-lymphocytes (B-cells) and T-lymphocytes (T-cells).
ML tasks on graphs

- **Discover** unexplored interactions in PPI-networks and gene regulatory networks
- **Learn** the structure
- **Reconstruct** the structure

From structure to function

A Protocol for Computer-Based Protein Structure and Function Prediction

Guidelines for computer-based structural and functional characterization of protein using the iTASSER pipeline is described. Starting from query protein sequence, 3D models are generated using multiple template alignments and iterative structural assembly refinement. Functional annotations are then added.

Interesting: Hubs tend to link to small degree nodes

Nodes: proteins
Links: physical interactions (binding)

Puzzling pattern:
Hubs tend to link to small degree nodes.
Why is this puzzling?
In a random network, the probability that a node with degree \( k \) links to a node with degree \( k' \) is:

\[
p_{kk'} = \frac{kk'}{2L}
\]

\( k=50, k'=13, N=1,458, L=1746 \)

\[
p_{50,13} = 0.15 \quad p_{2,1} = 0.0004
\]

Why do we want to apply ML to graphs

- A) Discovery of unexplored interactions
- B) Learning and Predicting the structure
- C) Reconstructing the structure

Which joint probability distributions does a graphical model represent?

How can we learn the parameters and structure of a graphical model?

The chemical space

- $10^{60}$ possible small organic molecules
- $10^{22}$ stars in the observable universe

1. Find a target
2. Identify hits
3. Hit-to-lead: characterize hits
4. Lead optimization and synthesis
5. Assay

$500,000,000$ to $2,000,000,000$
Example Question: Predicting Function from Structure

How similar are two graphs? How similar is their structure? How similar are their node and edge labels?

**Similar Property Principle:** Molecules having similar structures should have similar activities.

**Structure-based representations:** Compare molecules by comparing substructures, e.g.
- Sets as vectors: Measure similarity by the cosine distance
- Sets as sets: Measure similarity by the Jaccard distance
- Sets as points: Measure similarity by Euclidean distance

**Problems:** Dimensionality, Non-Euclidean cases
4) Little Excursus: What is similarity?
What is Similar?

Image credit to Eamonn Keogh (2008)
Similarity and Correspondence


http://www.inf.usi.ch/bronstein/

Correspondence quality = structure similarity
(distortion)

Minimum possible correspondence distortion
Invariant Similarity

$$d(\tau X, \sigma Y) = d(X, Y)$$
Gromov-Hausdorff dist: finding the opt. correspondence


\[ d_{GH}(X, Y) = \frac{1}{2} \min_{C} \max_{(x_i, y_i) \in C} \left| \delta_X(x_i, x_j) - \delta_Y(y_i, y_j) \right| \]

\[ \forall x_i \exists y_i \text{ s.t. } (x_i, y_i) \in C \]
\[ \forall y_i \exists x_i \text{ s.t. } (x_i, y_i) \in C \]

Discrete optimization over correspondences is NP hard!
5) Probabilistic Topic Models
Topic modelling – small topic but hot topic in ML

- Probabilistic Modeling
- Statistics
- Machine Learning
- Data Science
Given the parameters $\alpha$ and $\beta$, the joint distribution of a topic mixture $\theta$, a set of $N$ topics $z$, and a set of $N$ words $w$ is given by:

$$p(\theta, z, w | \alpha, \beta) = p(\theta | \alpha) \prod_{n=1}^{N} p(z_n | \theta) p(w_n | z_n, \beta)$$

Motivation: to get insight into unknown document sets

http://agoldst.github.io/dfr-browser/demo/#/model/scaled
We only observe the docs – the other structure is hidden; then we compute the posterior $p(t,p,a|\text{docs})$. 

We find topics through a Generative Probabilistic Model

Goal: to get insight in unknown document collections
See a nice demo http://agoldst.github.io/dfr-browser/demo/#/model/grid

Each doc is a random mix of corpus-wide topics and each word is drawn from one of these topics
Output Example: 4 learned topics (details in Blei, 2008)

- human
- genome
- dna
- genetic
- genes
- sequence
- gene
- molecular
- sequencing
- map
- information
- genetics
- mapping
- project
- sequences
- evolution
- evolutionary
- species
- organisms
- life
- origin
- biology
- groups
- phylogenetic
- living
- diversity
- group
- new
- two
- common
- disease
- host
- bacteria
- diseases
- resistance
- bacterial
- new
- strains
- control
- infectious
- malaria
- parasite
- parasites
- united
- tuberculosis
- computer
- models
- information
- data
- computers
- system
- network
- systems
- model
- parallel
- methods
- networks
- software
- new
- simulations
LDA is an example for a probabilistic graphical model

- Encodes assumptions on data with a factorization of the joint
- Connects assumptions to algorithms for computing with data
- Defines the posterior (through the joint)
We can’t compute the denominator, the marginal $p\ (w)$, therefore we use approximate inference; However, this do not scale well ...
For “big data” stochastic variational inference

1. Sample a document
2. Estimate the local variational parameters using the current topics
3. Form intermediate topics from those local parameters
4. Update topics as a weighted average of intermediate and current topics
Stochastic variational inference

1: Initialize $\lambda^{(0)}$ randomly.
2: Set the step-size schedule $\rho_t$ appropriately.
3: repeat
4: Sample a document $w_d$ uniformly from the data set.
5: Initialize $\gamma_{dk} = 1$, for $k \in \{1, \ldots, K\}$.
6: repeat
7: For $n \in \{1, \ldots, N\}$ set
   \[
   \phi_{dn}^k \propto \exp \{ \mathbb{E}[\log \theta_{dk}] + \mathbb{E}[\log \beta_{k,w_{dn}}] \}, \quad k \in \{1, \ldots, K\}.
   \]
8: Set $\gamma_d = \alpha + \sum_n \phi_{dn}$.
9: until local parameters $\phi_{dn}$ and $\gamma_d$ converge.
10: For $k \in \{1, \ldots, K\}$ set intermediate topics
    \[
    \hat{\lambda}_k = \eta + D \sum_{n=1}^N \phi_{dn}^k w_{dn}.
    \]
11: Set $\lambda^{(t)} = (1 - \rho_t)\lambda^{(t-1)} + \rho_t \hat{\lambda}$.
12: until forever

Approximate inference can be difficult to achieve.

- Especially true for models that are not conditionally conjugate (Discrete choice models, Bayesian generalized linear models, ...)
- Holds us back from trying many models.
- Easily use variational inference with *any model*
- No exponential family requirements
- No mathematical work beyond specifying the model
Conclusion: What are future challenges

- **Flexible** and expressive components for building models are of utmost importance
- **Scalable** and generic inference algorithms (multi-task and transfer learning)
- **Usability** gets a totally new importance: Easy to use algorithms for the non-expert user to stretch probabilistic modeling into new areas
- Topic models are **one** approach towards detection of topics in document collections
- Example: Identifying re-occurring patterns in such data collections (gaining new knowledge)
6) Graph Bandits
The complexities of optimization: Sébastien Bubeck

I’m a bandit
Random topics on optimization, probability, and statistics. By Sébastien Bubeck

ORF523: The complexities of optimization
This page collects together the posts for the graduate course on optimization I taught at Princeton in the Spring 2013. This material has been reorganized (some parts have been cut, some have been extended) into a monograph which got recently published “Foundations and Trends in Machine Learning: Vol. 8: No. 3-4, pp 231-357. 2015” (see here for the free version):

https://blogs.princeton.edu/imabandit/


What is a bandit?

- Slot-machine (bandit - robs your money)
- One-armed bandit
- Very simple model for sequential decision making under uncertainty
- Main challenge: exploration versus exploitation
- Many application domains: A/B-Testing, Crowdsourcing, optimization, search, ...
Multi-armed bandit:= a gambler strategically operating multiple machines in order to draw the highest possible profits

- There are $n$ slot-machines (“einarmige Banditen”)
- Each machine $i$ returns a reward $y \approx P(y; \Theta_i)$
- Challenge: The machine parameter $\Theta_i$ is unknown
- Which arm of a slot machine should a gambler pull to maximize his cumulative reward over a sequence of trials? (stochastic setting or adversarial setting)
Each arm $a$ either
wins (reward=1) with fixed (unknown) probability $\mu_a$, or
loses (reward=0) with fixed (unknown) probability $1 - \mu_a$

- All draws are independent given $\mu_1 \ldots \mu_k$
- Problem:
  How to pull arms to maximize the total reward?
Underlying Principle of the k-Armed Bandits problem

- Let \( a_t \in \{1, \ldots, n\} \) be the choice of a machine at time \( t \)
- Let \( y_t \in \mathbb{R} \) be the outcome with a mean of \( \langle y_{at} \rangle \)
- Now, the given policy maps all history to a new choice:

\[
\pi : [(a_1, y_1), (a_2, y_2), \ldots, (a_{t-1}, y_{t-1})] \mapsto a_t
\]

- The problem: Find a policy \( \pi \) that \( \max \langle y_T \rangle \)
- Now, two effects appear when choosing such machine:
  - You collect more data about the machine (=knowledge)
  - You collect reward

- Exploration and Exploitation
  - **Exploration**: Choose the next action \( a_t \) to \( \min \langle H(b_t) \rangle \)
  - **Exploitation**: Choose the next action \( a_t \) to \( \max \langle y_t \rangle \)

models an agent that simultaneously attempts to acquire new knowledge (called "exploration") and optimize his or her decisions based on existing knowledge (called "exploitation"). The agent attempts to balance these competing tasks in order to maximize total value over the period of time considered.

**MAP-Principle: “Optimism in the face of uncertainty”**

\[ a_t = \max_{a \in A} \left( \hat{r}_t(a) + \sqrt{\frac{\log(1/\delta)}{T_t(a)}} \right) \]

**Exploitation**

*the higher the (estimated) reward the higher the chance to select the action*

**Exploration**

*the higher the (theoretical) uncertainty the higher the chance to select the action*

A bandit in a graph is still a bandit 😊

- Let $G$ a known graph with $K$ nodes $\{1, 2, \ldots, K\}$
- Let $f$ be a unknown function defined on the set of nodes
- For $t = 1$ to $n$,
  - Select a node $l_t$
  - Observe reward $r_t = f(l_t) + \epsilon_t$
- Goal: maximize sum of expected rewards
- Equivalently minimize regret:
  \[
  R_n = \sum_{t=1}^{n} (f^* - f(l_t)),
  \]
  where $f^* = \max_{1 \leq i \leq K} f(i)$.
- We care about the case when $K > n$
Smooth Graph Function
Knowledge can be represented in two ways:

1) as full history
\[ h_t = [(a_1, y_1), (a_2, y_2), ..., (a_{t-1}, y_{t-1})] \]

or

2) as belief
\[ b_t(\theta) = P(\theta|h_t) \]

where \( \Theta \) are the unknown parameters of all machines.

The process can be modelled as belief MDP:

\[
P(b'|y, a, b) = \begin{cases} 
1 & \text{if } b' = b'_{[b,a,y]} \\
0 & \text{otherwise} 
\end{cases}, \quad P(y|a, b) = \int_{\theta_a} b(\theta_a) P(y|\theta_a)
\]
The optimal policies can be modelled as belief MDP

\[
P(b'|s', s, a, b) = \begin{cases} 
1 & \text{if } b' = b[s', s, a] \\
0 & \text{otherwise}
\end{cases}, \quad P(s'|s, a, b) = \int_{\theta} b(\theta) P(s'|s, a, \theta)
\]

\[
V(b, s) = \max_{a} \left[ \mathbb{E}(r|s, a, b) + \sum_{s'} P(s'|a, s, b) V(s', b') \right]
\]

Clinical trials: potential treatments for a disease to select from new patients or patient category at each round, see:

Games: Different moves at each round, e.g. GO

Adaptive routing: finding alternative paths, also finding alternative roads for driving from A to B

Advertisement placements: selection of an ad to display at the Webpage out of a finite set which can vary over time, for each new Web page visitor
Randomized clinical trials have changed little in 70 years, and it’s time to revamp the approach by merging clinical research with clinical practice.

Limitations of drug design for rare diseases due to:

- Lack of understanding of the underlying principles of the rare disease
  - Motivation: Research advances
- Unbalanced economic motivation (cost/benefit)
  - Motivation: Orphan Drug Act and other regulations
- Unavailability of # patients for standard trials
  - This is the true bottleneck!

The goal of Standard Randomized Controlled Trials (RCT) are a controlled learning setting:
- Control for Type I and Type II errors, dependent of trial size $n_{RCT}$
- In the case if the patient population $N$ is smaller than the trial size $n_{RCT}$: underpowered trial – problem!

If we change the goal to

“learning sufficient - to treat $N$ as effectively as possible”,

then bandit strategies – optimal policy for max. the expected reward - are perfectly suited!

Learning vs. Earning dilemma

- Learning → experimenting with all treatments
- Earning → selecting one treatment only, based on experimentation results

**Question 1:** How much learning is best – for an optimal treatment of \( N \) patients?

- Suppose \( N \) patients with a rare disease:
  - Experimental Group E and control group C
  - e.g. control = response rate pc and little information about experimental group

**Question 2:** How many allocations of treatment to E are necessary (= how much experimentation?)
DYNAMIC PROGRAMMING AND LAGRANGE MULTIPLIERS
BY RICHARD BELLMAN
RAND CORPORATION, SANTA MONICA, CALIFORNIA
Communicated by Einar Hille, August 13, 1956

1. Introduction.—The purpose of this note is to indicate how a suitable combination of the classical method of the Lagrange multiplier and the functional-equation method of the theory of dynamic programming can be used to solve numerically, and treat analytically, a variety of variational problems that cannot readily be treated by either method alone.

A series of applications of the method presented here will appear in further publications.

2. Functional Equation Approach.—Consider the problem of maximizing the function

$$F(x_1, x_2, \ldots, x_N) = \sum_{i=1}^{N} q_i(x_i),$$

subject to the constraints

$$(a) \quad \sum_{j=1}^{N} a_{ij}(x_j) \leq c_i, \quad i = 1, 2, \ldots, M,$$

$$(b) \quad x_i \geq 0,$$

Some Statistics on rare diseases (orphan diseases)

- 7,000 + different types - more being discovered every day
- >10% of the world population is suffering (if all of the people with rare diseases lived in one country, it would be the world’s 3rd most populous country)
- 80% of rare diseases are genetic, so are present throughout a person’s lifetime, even if symptoms do not immediately appear
- >50% of the people affected by rare diseases are children
- Are responsible for 35% of deaths in the first year of life
- The prevalence distribution is skewed – 80% of all rare disease patients are affected by 350 rare diseases
- >50% of rare diseases do not have a disease specific foundation supporting or researching their rare disease

https://www.hon.ch/HONselect/RareDiseases/
Cerebral Autosomal Dominant Arteriopathy with Subcortical Infarcts and Leukoencephalopathy is a hereditary disease affecting all the small cerebral arteries. It causes subcortical infarcts and damages the white matter (leukoencephalopathy) and it is due to various mutations of the Notch3 gene situated on chromosome 19:

Conclusion and Future Challenges
Bandit strategy: Is experimentation worth it for a small number N?
- Reconcile clinical trials and clinical practice
- Extensions should deal with randomization, delayed responses and uncertainty around N
- Bayesian bandits need Online-ML
- Bandits are a great source of inspirations and building blocks for solving many problems
- Future work: convex optimization, contextual, combinatorial, ...

Thank you!
Questions
Sample Questions (1/3)

- What kind of graphical models are used in medical informatics?
- Which type of graph is particularly useful for inference and learning?
- What is the key challenge in the application of graphical models for health informatics?
- What was Judea Pearl (1988) discussing in his paper, for which he received the Turing award?
- What main difficulties arise during breast cancer prognosis?
- What can be done to increase the robustness of prognostic cancer tests?
- Inference in Bayes Nets is NP-complete, but there are certain cases where it is tractable, which ones?
Why do we want to apply ML to graphs?

Describe typical ML tasks on the example of blood cancer cells!

If you have a set of points – which similarity measures are useful?

What is the advantage of factor graphs?

Why is the Gromov-Hausdorff distance useful?

What is the central goal of a generative probabilistic model?

Describe the LDA-model and its application for topic modelling!
Briefly describe the stochastic variational inference algorithms!

What is the principle of a bandit?

How does a multi-armed bandit (MAB) work?

In which ways can a MAB represent knowledge?

What is the main problem of a clinical trail – and maybe the main problem in clinical medicine?

Why are rare diseases both important and relevant? Describe an example disease!

What is the big problem in clinical trials for rare diseases?

What did Richard Bellman (1956) describe with dynamic programming?

Why are graph bandits a hot topic for ML research?
1. This is a factor graph of an undirected graph – we have seen this in protein networks (refer to slide Nr. 70 in lecture 5). Factor graph is bipartite and has two types of nodes: Variables, which can be either evidence variables (when we know its value) or query variables (when the value is unknown and we want to predict the value); and factors, which define the relationship between variables in the graph. Each factor can be connected to many variables and comes with a factor function to define the relationship between these variables. For example, if a factor node is connected to two variables nodes A and B, a possible factor function could be `imply(A,B)`, meaning that if the random variable A takes value 1, then so must the random variable B. Each factor function has a weight associated with it, which describes how much influence the factor has on its variables in relative terms. For more information please consult: [http://deepdive.stanford.edu/inference](http://deepdive.stanford.edu/inference).

2. This is the decomposition of a tree, rooted at nodes into subtrees.

3. Metabolic and physical processes that determine the physiological and biochemical properties of a cell. As such, these networks comprise the chemical reactions of metabolism, the metabolic pathways, as well as the regulatory interactions that guide these reactions. With the sequencing of complete genomes, it is now possible to reconstruct the network of biochemical reactions in many organisms, from bacteria to human. Several of these networks are available online: Kyoto Encyclopedia of Genes and Genomes (KEGG)[1], EcoCyc [2], BioCyc [3] and metaTIGER [4]. Metabolic networks are powerful tools for studying and modelling metabolism.

4. MYCIN – expert system that used early AI (rule-based) to identify bacteria causing severe infections, such as bacteremia and meningitis, and to recommend antibiotics, with the dosage adjusted for patient's body weight — the name derived from the antibiotics themselves, as many antibiotics have the suffix "-mycin".

5. Protein-Protein Interaction network (undirected graph here).

6. PPI with critical node, bottleneck, hub, etc.
Appendix
Key Idea: Conditional independence assumptions are very useful – however: Naïve Bayes is extreme!

- $X$ is \textit{conditionally independent} of $Y$, given $Z$, if the $P(X)$ governing $X$ is independent of value $Y$, given value of $Z$:

$$(\forall i, j, k) P(X = x_i | Y = y_j, Z = z_k) = P(X = x_i | Z = z_k)$$

can be abbr. with $P(X|Y, Z) = P(X|Z)$

- Graphical models express sets of conditional independence assumptions via graph structure
- The graph structure plus associated parameters define joint probability distribution over the set of variables

GM are amongst the most important ML developments
Where do the data come from?

http://www.ebi.ac.uk/intact/
... are libraries of life science data, collected from scientific experiments and computational analyses.

... contain (clinical, biological, ...) data from clinical work, genomics, proteomics, metabolomics, microarray gene expression, phylogenetics, etc.

Examples:
- Text: e.g. PubMed, OMIM (Online Mendelian Inheritance in Man);
- Sequence data: e.g. Entrez, GenBank (DNA), UniProt (protein).
- Protein structures: e.g. PDB, Structural Classification of Proteins (SCOP), CATH (Protein Structure Classification);
van Kampen (2012), Bioinformatics Laboratory, Academic Medical Center, NL
Ensembl is a joint project between EMBL - EBI and the Wellcome Trust Sanger Institute to develop a software system which produces and maintains automatic annotation on selected eukaryote genomes. Ensembl receives major funding from the Wellcome Trust. Our acknowledgements page includes a list of additional current and previous funding bodies.

http://www.ensembl.org/index.html
ArrayExpress - functional genomics data

ArrayExpress is a database of functional genomics experiments that can be queried and the data downloaded. It includes gene expression data from microarray and high throughput sequencing studies. Data is collected to MIAME and MIANSEQE standards. Experiments are submitted directly to ArrayExpress or are imported from the NCBI GEO database.

Updated today at 06:00
- 43495 experiments
- 1233850 assays
- 18.51 TB of archived data

Latest News
1 November 2013 - Need to keep your unpublished ArrayExpress microarray data private for longer?
Microarray experiment submitters, have you ever wondered if you could just change the release date of unpublished ArrayExpress data by yourself without emailing curators? Now you can! Use our new release date changing tool (more details on this help page). Submitters of high-throughput sequencing experiments, please continue to email us at miameexpress@ebi.ac.uk for release date changes so we can make sure the sequence read records at the European Nucleotide Archive are kept in sync.

http://www.ebi.ac.uk/arrayexpress/
IntAct provides a freely available, open source database system and analysis tools for molecular interaction data. All interactions are derived from literature curation or direct user submissions and are freely available. To perform a search in the IntAct database use the search box above.

http://www.ebi.ac.uk/intact/

Cardiovascular disease is by far the most prevalent disease in ageing populations. Correlated with alterations in lipid metabolism profiles, it has estimated incidence rate of 30-40% in the UK population, over the age 65. Low Density Lipoprotein Cholesterol (LDL-C) is a prominent component in lipid metabolism, stands out as a major contributory factor. Furthermore, it is apparent that neither nutritional status nor physical activity have any effect on the rising levels of LDL-C with age.

Besides its well publicized detrimental effects, cholesterol is also an important component of all cell membranes, being a hormone precursor and playing a crucial role in absorption of lipid soluble vitamins. Its absorption from the gut is documented as being inefficient, and also displays high variability between individuals (30-80%). The precise transport and enzymatic mechanisms involved, particularly pertaining to how cholesterol traverses enterocyte membranes, is not well established.

The hepatic system is central in cholesterol metabolism, with the liver able to synthesize VLDs (very low density lipoproteins), which are converted into IDLs (intermediate density lipoproteins) through the action of lipoprotein lipase (LPL). LPLs can be taken up by the liver directly, or further hydrolysed into LDLs, the main cholesterol carrier in the blood. LDLs may also be taken up through the LDL-receptor (LDLR), which is highly expressed in the liver, and expressed in peripheral tissues. The hepatic receptor is transcriptionally regulated by intracellular cholesterol levels.

It has been demonstrated that: a) There is age-associated decline in the clearance rate of LDL-C from the blood, as well as a decrease in the number of hepatic LDLRs. b) Intestinal cholesterol absorption increases with age in some species.

In this paper, the authors take a mechanistic approach to construct a model, with these observations in mind, making extensive use of published experimental measurements over the last seventy years. The model incorporates dietary cholesterol absorption in the intestine, and hepatic LDL-C clearance from the plasma [1, BIOMD0000000434]. It consists of 6 compartments (Figure 1), and is composed of a series of coupled ODEs.

http://www.ebi.ac.uk/biomodels-main/
Counts the number of “i-dimensional holes”

bi is the “i-th Betti number”

Enrico Betti
(1823-1892)

Emmy Noether
(1882-1935)

\[ b_1 = 1 \quad b_1 = 0 \quad b_1 = 2 \]

\[ b_2 = 0 \quad b_2 = 1 \quad b_2 = 1 \]

Betti numbers are computed as dimensions of Boolean vector spaces (E. Noether)

Structural Patterns are often hidden in weakly str. data

- Statement of Vin de Silva (2003), Pomona College:
  - Let $M$ be a topological or metric space, known as the *hidden parameter space*;
  - let $\mathbb{R}^d$ be a Euclidean space, the *observation space*,
  - and let $f: M \rightarrow \mathbb{R}^d$ be a continuous embedding.
  - Furthermore, let $X \subset M$ be a finite set of data points, perhaps the realization of a stochastic process, i.e., a family of random variables $\{X_i, i \in I\}$ defined on a probability space $(\Omega, \mathcal{F}, P)$, and denote $Y = f(X) \subset \mathbb{R}^d$ the images of these points under the mapping $f$.
  - We refer to $X$ as *hidden data*, and $Y$ as the *observed data*.
  - $M$, $f$ and $X$ are unknown, but $Y$ is - so can we identify $M$?

De Silva, V. 2004. GEOMETRY AND TOPOLOGY OF POINT-CLOUD DATA SETS: A STATEMENT OF MY RESEARCH INTERESTS.
https://www.pomona.edu/directory/people/vin-de-silva
Mega Problem: To date none of our known methods, algorithms and tools scale to the massive amount and dimensionalities of data we are confronted in practice;

we need much more research efforts towards making computational topology successful as a general method for data mining and knowledge discovery

Topic model toolkits

- **Particular topic models**
  - Stanford topic model toolbox
  - Topic modeling at Princeton
  - MALLET (Java) [http://mallet.cs.umass.edu](http://mallet.cs.umass.edu)
  - Network topic models: Bayes-stack
    [https://github.com/bgamari/bayes-stack](https://github.com/bgamari/bayes-stack)
  - Gensim (Python) [http://radimrehurek.com/gensim/](http://radimrehurek.com/gensim/)
  - R package for Topic models. [http://epub.wu.ac.at/3987/](http://epub.wu.ac.at/3987/)

- **Frameworks for generative models**
  - Variational inference: Infer.net
  - Gibbs sampling: OpenBUGS [http://openbugs.net/](http://openbugs.net/)