Probabilistic Graphical Models for Image Analysis - Lecture 1

Alexey Gronskiy, Stefan Bauer

21 September 2018

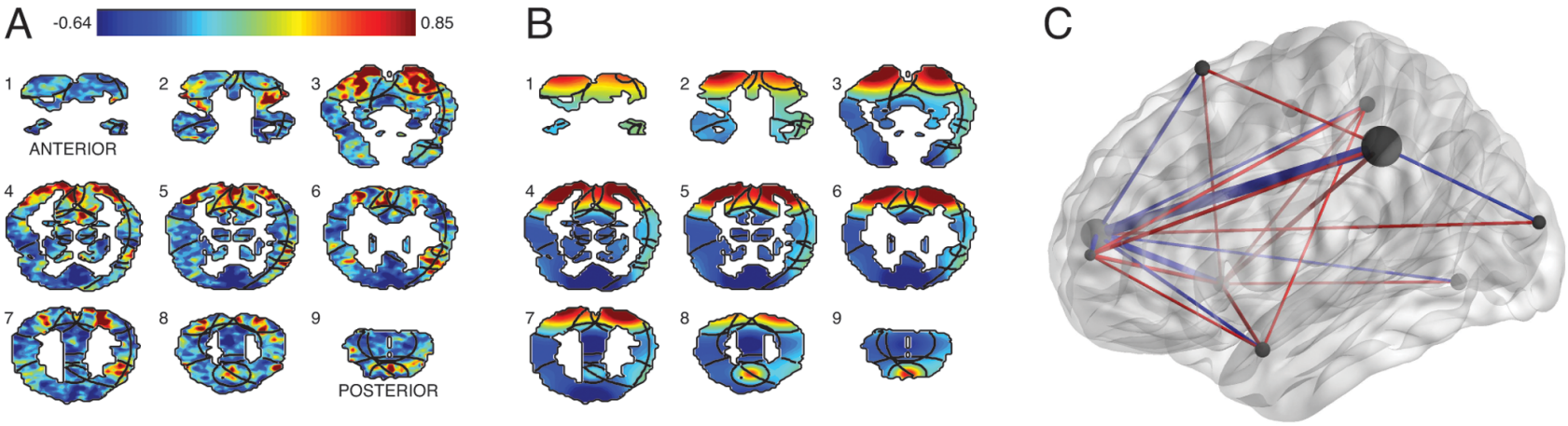
Max Planck ETH Center for Learning Systems

Overview

1. Motivation - Why Graphical Models
2. Course
3. Probabilistic Modeling
4. Probabilistic Inference

Motivation - Why Graphical Models

Neuroscience

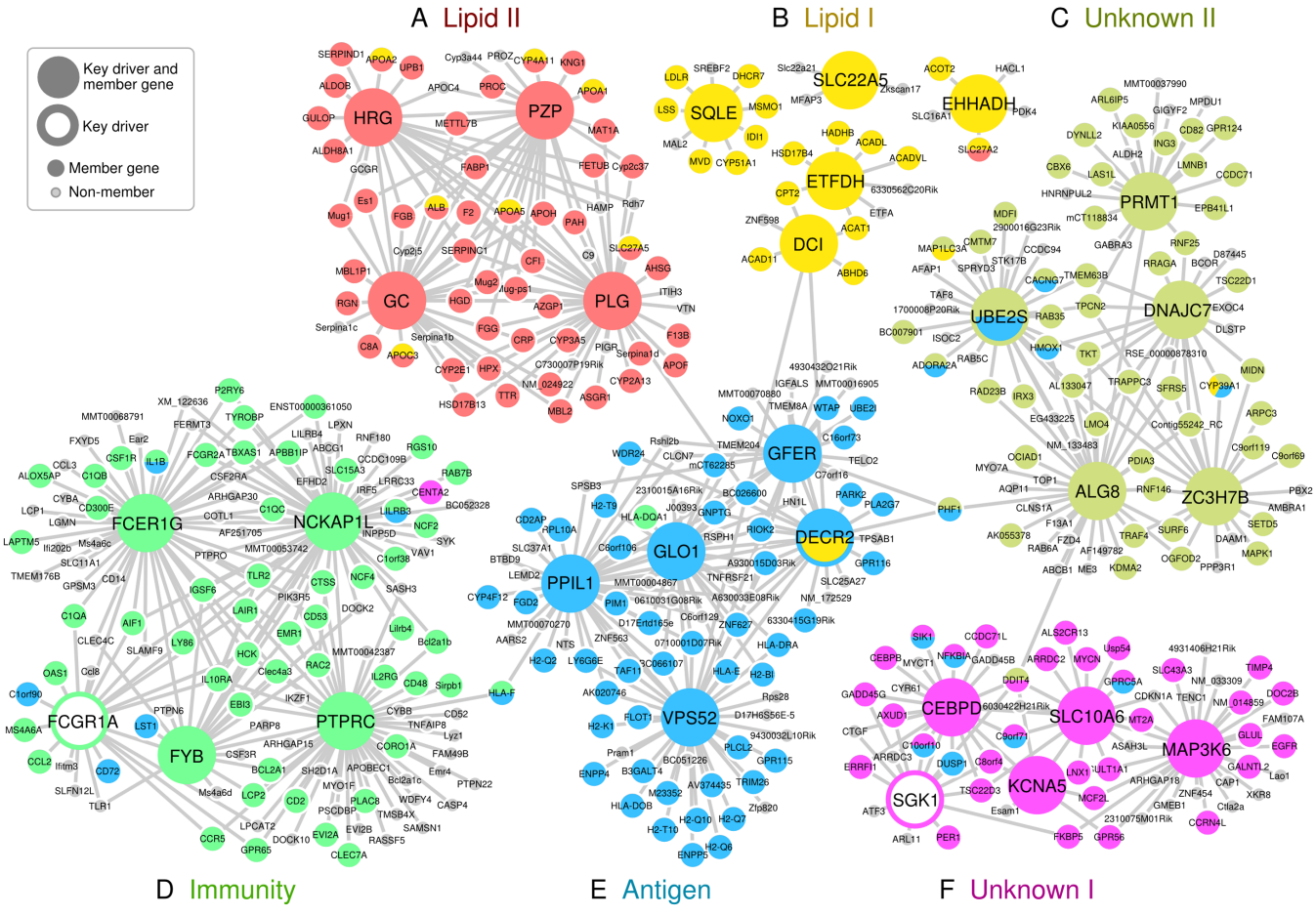


Manning et al. Topographic Factor Analysis: A Bayesian Model for Inferring Brain Networks from Neural Data, *PloS one*, 2014.

Image Generation

Karras et al. Progressive growing of gans for improved quality, stability, and variation.*ICLR* 2018.

Genomics



Mäkinen et al. Integrative Genomics Reveals Novel Molecular Pathways and Gene Networks for Coronary Artery Disease, *PloS genetics*, 2014.

Robotics

Grisetti et al. A tutorial on graph-based SLAM. *IEEE Intelligent Transportation Systems Magazine* 2010

Course

Administrative Information

- **Exam**
  20 minute oral exam in English
- **Exercises**
  Each Tuesday from 15-16:00 in CAB G56; Exercises will be incorporated into the Lectures.
- **Contact**
  Please ask questions through the open forum using `https://piazza.com/`.
- **Additional Information**
  `http://www.vvz.ethz.ch/Vorlesungsverzeichnis/lerneinheit.view?semkez=2018W&ansicht=KATALOGDATEN&lerneinheitId=126518&lang=en`

Head TA - Philippe Wenk

Exercise Administration

Time & Room

Tuesday, 15-16, CAB G56

Exercise

- One exercise sheet per week.
- Hand it in one week afterwards.
- We return it to you the week after.
- No exercise next week!

Testat

- No Testat requirement for the exam.
- If you are a PhD student you might want to talk to your Studiensekretariat.

Related Courses

- Some overlap with the following courses:
  Advanced Machine Learning - Prof. Buhmann
  http://ml2.inf.ethz.ch/courses/ml/
  Introduction to Machine Learning - Prof. Krause
  https://las.inf.ethz.ch/teaching/introml-s18
  Probabilistic Artificial Intelligence - Prof. Krause
  https://las.inf.ethz.ch/teaching/pai-f18
  Computational Intelligence Lab - Prof. Hofmann
  http://cil.inf.ethz.ch
  Deep Learning - Prof. Hofmann http://www.da.inf.ethz.ch/teaching/2017/DeepLearning/
- Other courses with some overlap are offered by the computer vision group.

Reading

Online Resources:

- **Barber**: `http://web4.cs.ucl.ac.uk/staff/D.Barber/pmwiki/pmwiki.php?n=Brml.Online`
- **MacKay** `http://www.inference.phy.cam.ac.uk/mackay/itila/book.html`
- **Wainwright & Jordan** `http://www.eecs.berkeley.edu/~wainwrig/Papers/WaiJor08_FTML.pdf`

Classics:

- **Bishop:** Pattern recognition and machine learning
- **Koller and Friedman:** Probabilistic graphical models

Course outline

- Learning from data
- Probabilistic models
- Graphical models
- Variational Inference
- State Space Models
- Dynamical Systems
- Factor Analysis
- Autoencoders
- Current research

Probabilistic Modeling

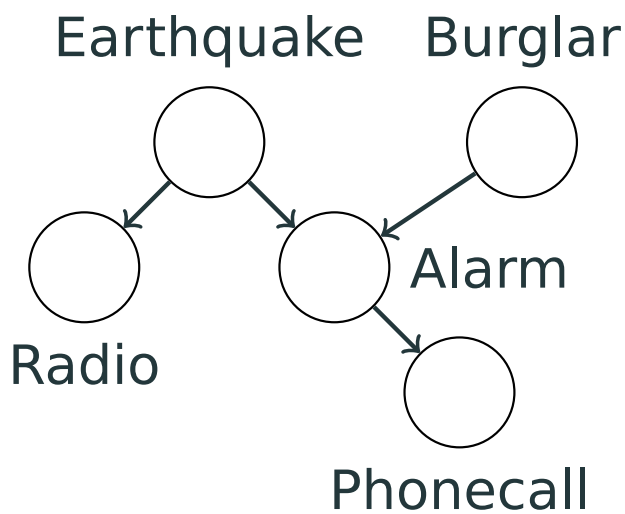
Reasoning under Uncertainty

Knowledge Representation

Use domain knowledge to design representative models

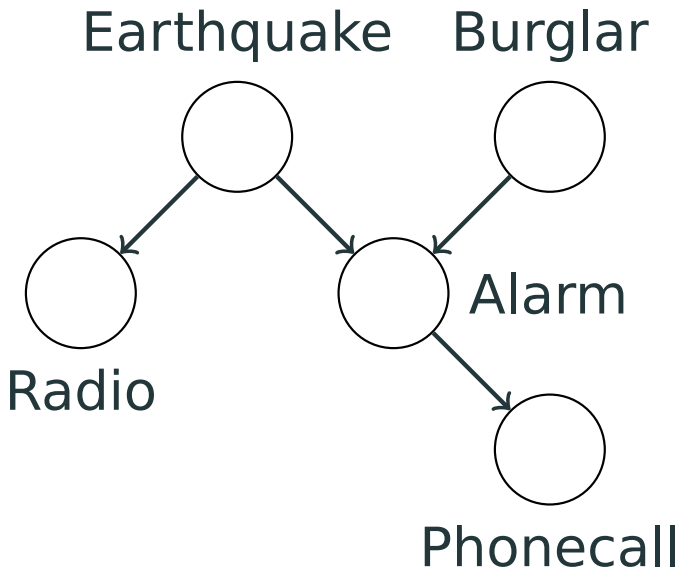
Automated Reasoning

Develop a general suite of algorithms for reasoning.



Deal with uncertainty inherent in the real world using the notion of probability.

Graphical Models



Graphical Models provide a compact representation of two equivalent perspectives:

- representation of a set of independencies
- factorization of the joint distribution

Algorithms

Representation

What do the different edges mean and how these can be used to model tasks in image analysis.

Learning

Given some empirical measurements (data), estimate the parameters of the model which best explains it.

Inference

For a given model, how to use it for making decisions and reasoning about the task at hand.

To do list for modeling

Representation

Develop techniques to represent dependencies between variables using graphical models.

Learning

Given some empirical measurements (x, y), estimate the parameters of the model $(\theta_{ML}, \theta_{MAP}, \ldots)$ which best explains them.

Inference

For a given model (θ), investigate various ways of using it to make predictions, de-noise images, etc.

Machine Learning

Model the machine learning task as the **joint probability**

$$P(x, y).$$

example	$x_i \in \mathcal{X}$ for example a set of images.
label	$y_i \in \mathcal{Y}$ for example whether the digit is 1,3,4 or 8.
Training Data	Consists of examples and associated labels, which are used to train the machine.
Testing Data	Consists of examples and associated labels. Labels are **never** used to train machine, only to evaluate its performance
Predictions	Output of the trained machine.

Supervised learning

Let $X = \{\text{set of } d \times d \text{ images}\}$
and $Y = \{0, 1, \ldots, 9\} = \{\text{set of labelings}\}$.

Problem: Given labeled training data

$$\mathcal{D} = \big\{(X_1, Y_1), \ldots, (X_n, Y_n)\big\} \subset X^n \times Y^n,$$

find a function

$$\widehat{f} : \underbrace{X}_{\text{images}} \to \underbrace{Y}_{\text{labels}}$$

that correctly classifies all images – including new ones.

Learning a function

Suppose we have a class of functions $\mathcal{F} = \{f_\theta : X \to Y | \theta \in \Theta\}$.

Empirical risk minimization Define

$$\widehat{\theta} := \operatorname*{argmin}_{\theta \in \Theta} \mathbb{E}_{\widehat{P}}\Big[\mathbb{1}[f_\theta(x_i) \neq y_i]\Big]$$

How to interpret $\mathbb{E}_{\widehat{P}}\Big[\mathbb{1}[f_\theta(x_i) \neq y_i]\Big]$?

Number of incorrect classifications made by f_θ on the training data.

Empirical risk minimization finds the function $f_{\widehat{\theta}}$ that **makes the fewest mistakes** on the training data.

Linear Regression

Consider the problem of linear regression,

$$y = \theta^\mathsf{T}\mathbf{x} + \varepsilon$$

where the loss is the mean squared error:

$$\min_\theta \sum_{i=1}^{n} (\theta^\mathsf{T}\mathbf{x}_i - y_i)^2$$

1. How are examples represented?
2. How are labels represented?
3. What does the machine estimate?
4. What does the machine output?

Unsupervised learning

Problem: Given

- class of models $P(X;\theta)$ for $\theta \in \Theta$ and
- unlabeled data $\mathcal{D} = (x_1, \ldots, x_n) \in X^n$
- sampled i.i.d. from unknown distribution $P(X)$,

find the model $\widehat{\theta}$ that "best fits" the data.

In other words, want $P(X;\widehat{\theta}) \approx \widehat{P}(X)$ or, **better**, $P(X;\widehat{\theta}) \approx P(X)$.

The likelihood

Suppose we have a class of probability distributions

$$\mathcal{M} = \{P(x|\theta)|\theta \in \Theta\} \text{ on } X.$$

The probability of the observations taken as a **function where x is fixed and θ varies**

$$L(\theta|x) := P(x|\theta),$$

is called the **likelihood**.

$L(\theta|x)$:

- "how likely is parameter θ to have generated x?"

or

- "how well does θ explain x?"

Learning a model

Maximum likelihood estimation (ML)

Given data $\mathcal{D} = (x^1, \ldots, x^n) \in X^n$, let

$$L(\theta|\mathcal{D}) := P(\mathcal{D}|\theta) := \prod_{i=1}^{N} P(x^i|\theta)$$

be the likelihood of the parameters given the data.

Note: Assume datapoints are independently, identically distributed (i.i.d.) given θ!

Goal Find the parameter θ that is most likely to have generated the data $\mathcal{D} = (v^1, \ldots, v^N)$.
Definition: the maximum likelihood estimator (MLE) is

$$\widehat{\theta}_{ML} = \arg\max_{\theta \in \Theta} L(\theta|\mathcal{D}).$$

The negative log-likelihood

We usually minimize the **negative Log Likelihood** $-\log P(x|\theta)$ since,

- Logarithm is monotonic: $\arg\min f(x) = \arg\min \log f(x)$

  {minimize negative log-likelihood} $\leftrightarrow$ {maximize likelihood}

- Simplifies math: $\log \prod_i f(x_i) = \sum_i \log f(x_i)$

- Better numerical behavior:

Relation to empirical risk minimization (**Exercise:**) Show computing the MLE $\leftrightarrow$ minimizing empirical risk for

$$\ell(f, x) = -\log f(x).$$

Probabilistic Inference

(Conditional) Independence of RVs

Independence of Random Variables

RVs X, Y are independent iff

$$p(x, y) = p(x)p(y).$$

I.e.: In a simultaneous draw, the value taken by X does not depend on the value taken by Y and vice versa.

Conditional Independence of Random Variables

RVs X, Y are conditionally independent given Z, iff

$$p(x, y|z) = p(x|z)p(y|z).$$

This however does not imply independence of the RVs X and Y.

Marginalization and its exponential running time

Marginalization

Assume a joint distribution of RVs $X_1, \ldots, X_N$ is given

$$p(x_1, \ldots, x_N).$$

Often we're interested in *marginals* of RVs X_S, where S is a subset of indices of the variable indices. The marginal can be computed as

$$p(x_S) = \sum_{x_{\{1,\ldots,N\} \notin S}} p(x_1, \ldots, x_N)$$

Summation is expensive

Assume each X can take K different values, and $S = \{1\}$. A full summation will have K^{N-1} summands, i.e. running time exponential in the number of variables!

Expressing Joint distributions

Example: Autonomous shooting device

Assume we have a simple model for a self-shooting device with three variables: $X_{\text{burglar}}, X_{\text{alarm}}, X_{\text{shoot}}$ each taking values in $\{0, 1\}$, the joint probability can be specified by giving the probability for each configuration $P(x_{\text{bu}}, x_{\text{al}}, x_{\text{sh}})$.

Joint distribution

We can specify a joint distribution as follows:

X_{burglar}	X_{alarm}	X_{shoot}	$P(\cdot)$
0	0	0	0.576
0	0	1	0.144
0	1	0	0.024
0	1	1	0.056
1	0	0	0.032
1	0	1	0.008
1	1	0	0.048
1	1	1	0.112

Graphical Models

Say we have the following graphical model for the previous example:

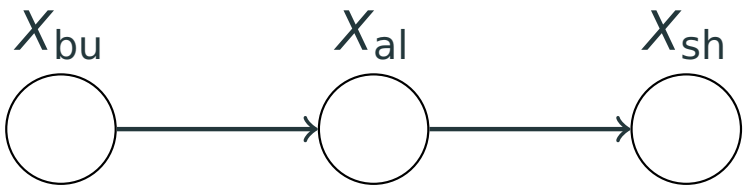


Figure 1: Illustrating the joint distribution by a graphical model.

Here the arrows show conditional dependencies. Again: you can *not* conclude that X_{sh} is independent of X_{bu} only because they are not connected.

Factorising joint distributions

Incorporating Independence

From the graphical model we can write the joint as:

$$p(x_{\text{bu}}, x_{\text{al}}, x_{\text{sh}}) = p(x_{\text{sh}}|x_{\text{al}})p(x_{\text{al}}|x_{\text{bu}})p(x_{\text{bu}})$$

Alternative specification of the joint distribution

Use conditional probability tables according to the dependencies.

X_{bu}	$P(X_{\text{bu}})$
0	0.8
1	0.2

X_{bu}	X_{al}	$P(X_{\text{al}}\|X_{\text{bu}})$
0	0	0.9
0	1	0.1
1	0	0.2
1	1	0.8

X_{al}	X_{sh}	$P(X_{\text{sh}}\|X_{\text{al}})$
0	0	0.8
0	1	0.2
1	0	0.3
1	1	0.7

Elimination Algorithm

In the previous example we might be interested in the probability of the shooting device shooting or not shooting some target, i.e. $P(x_{\text{sh}})$:

$$P(x_{\text{sh}}) = \sum_{x_{\text{al}}, x_{\text{bu}}} P(x_{\text{bu}}, x_{\text{al}}, x_{\text{sh}})$$

We can use the factorization from above:

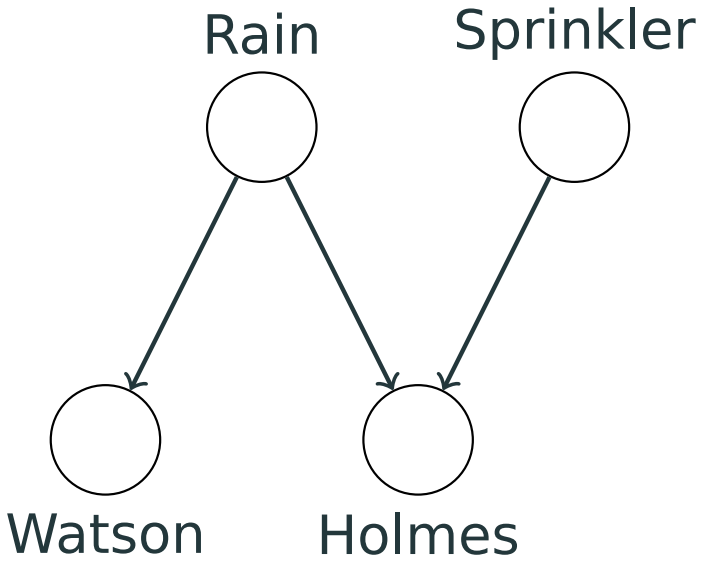
$$P(x_{\text{sh}}) = \sum_{x_{\text{al}}, x_{\text{bu}}} P(x_{\text{sh}}|x_{\text{al}})P(x_{\text{al}}|x_{\text{bu}})P(x_{\text{bu}})$$

We can push in the sum:

$$P(x_{\text{sh}}) = \sum_{x_{\text{al}}} P(x_{\text{sh}}|x_{\text{al}}) \sum_{x_{\text{bu}}} P(x_{\text{al}}|x_{\text{bu}})P(x_{\text{bu}}).$$

For a larger chain this makes a huge difference computationally: $\mathcal{O}(2^N)$ vs. $\mathcal{O}(N)$, with N the number of variables.

Inference in Bayesian Networks (Williams, 2006 and Barber, 2012)



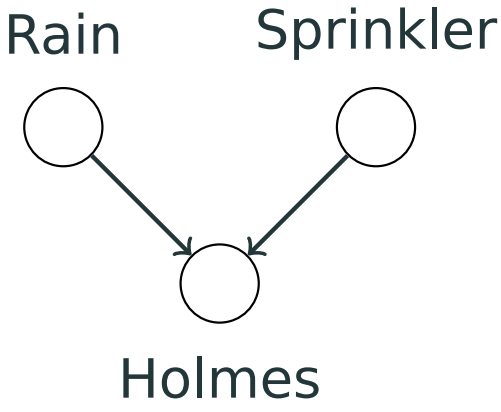
Interesting questions:

- If Holmes' lawn is wet, was it the rain or the sprinkler?
- If Watson's lawn is also wet how does this change things?

To answer these question we must do *inference*.

Lets plug some numbers in...

Why *Bayesian* Networks?



Recall: Bayes' Theorem

$$posterior \propto likelihood \times prior$$

In our example:

- prior (probability of the events occuring)

$$p(r = yes, s = yes) = p(r = yes)p(s = yes)$$

- "likelihood" (observation model) $p(h = yes|r, s)$
- posterior (conditioning on observed evidence) $p(s|h = yes),\ \ p(r|h = yes)$

Next week & Open Master’s thesis topics

Open master’s thesis topics

- Brain Connectivity Estimation a
- Online Outlier Detection
- Sleep Classification
- Model Selection for Dynamical Systems
- Simulator for Robotics Platform
- Large-Scale Parameter Inference in Dynamical Systems
- Analysis of the effects of smoking using mass-spec. data

Plan for next week

- Gaussian Mixture (EM)
- Variational Inference
- Belief Propagation

Questions?