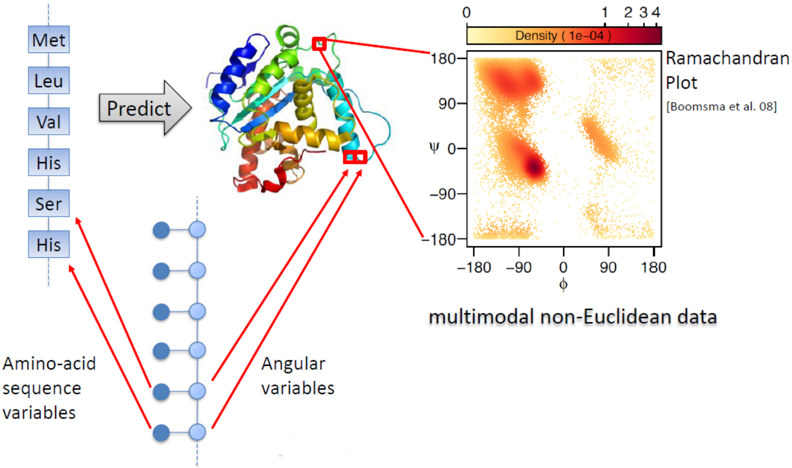
Probabilistic Learning and Graphical Models

Davide Bacciu

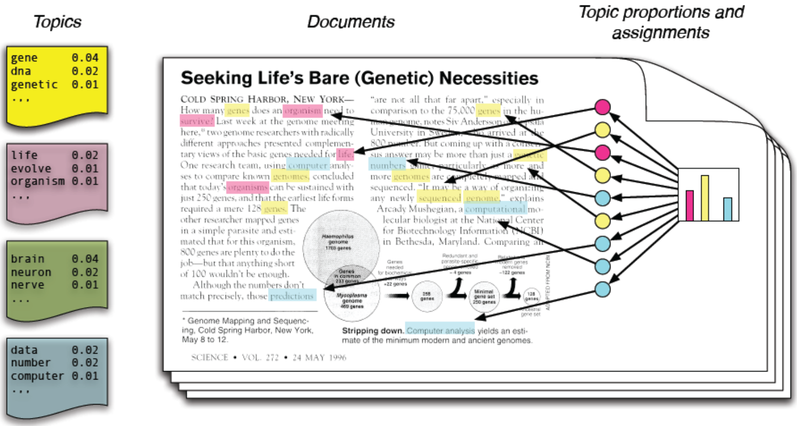
Dipartimento di Informatica
Università di Pisa
bacciu@di.unipi.it

Machine Learning: Neural Networks and Advanced Models (AA2)

Protein Secondary Structure Prediction

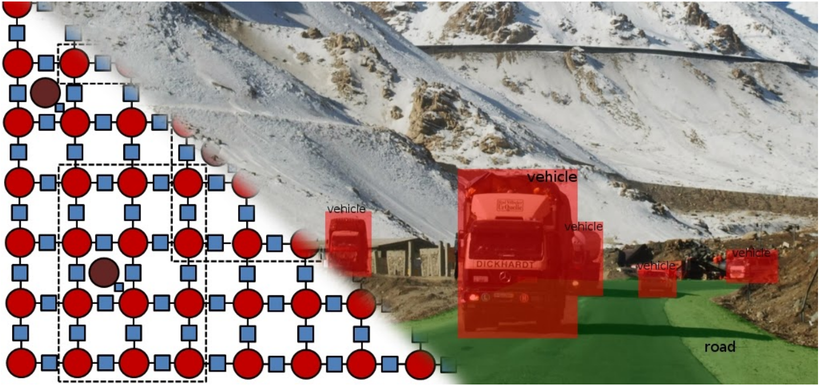


Document Understanding



D. M. Blei, Probabilistic Topic Models, Communications of the ACM, Vol. 55 No. 4, Pages 77-84

Image Understanding



Probabilistic Learning Models

- Learning models that represent knowledge inferred from data under the form of probabilities
  - Supervised, unsupervised, weakly supervised learning tasks
  - Describes how data is generated (interpretation)
  - Allow to incorporate prior knowledge on the data and on the task
- The majority of the modern task comprises large numbers of variables
  - Modeling the joint distribution of all variables can become impractical
  - Exponential size of the parameter space
  - Computationally impractical to train and predict

The Graphical Models Framework

Representation

- Graphical models are a compact way to represent exponentially large probability distributions
- Encode conditional independence assumptions
- Different classes of graph structures imply different assumptions/capabilities

Inference

- How to query (predict with) a graphical model?
- Probability of unknown X given observations $\mathbf{d}$, $P(X|\mathbf{d})$
- Most likely hypothesis

Learning

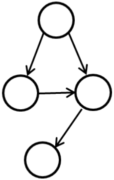
- Find the right model parameter (Parameter Learning)
- Find the right model structure (Structure Learning)
- An inference problem after all

Graphical Model Representation

A graph whose nodes (vertices) are random variables whose edges (links) represent probabilistic relationships between the variables

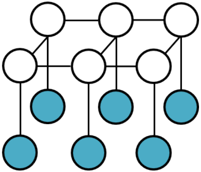
Different classes of graphs

Directed Models



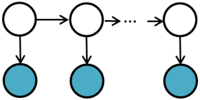
Directed edges express causal relationships

Undirected Models



Undirected edges express soft constraints

Dynamic Models



Structure changes to reflet dynamic processes

Graphical Models in this Module

Representation

- Directed graphical models: Bayesian Networks
- Undirected graphical models: Markov random fields
- Dynamic graphical models: Hidden Markov Models

Inference

- Exact inference: message passing, junction tree algorithms
- Approximate Inference: loopy belief propagation, sampling, variational inference

Learning

- Parameter learning: Expectation-Maximization algorithm
- Structure learning: PC algorithm, search-and-score

Generative Models Module

Plan of the Lectures

- Lesson 1 Introduction to Probabilistic and Graphical Models
- Lesson 2 Directed and Undirected Graphical Models
- Lesson 3 Inference in Graphical Models
- Lesson 4 Dynamic Approaches: The Hidden Markov Model
- Lesson 5 Graphical models for Structured Data
- Lesson 6 Exact and Approximate Learning: Latent Variable Models
- Lesson 7 Bridging Probabilistic and Neural: Deep Learning

Lecture Outline

- Introduction
- A probabilistic refresher
  - Probability theory
  - Conditional independence
  - Learning in probabilistic models
- Directed graphical models
  - Bayesian Networks
  - Representation
  - Conditional independence
  - Inference and Learning
- Applications and conclusions

Random Variables

- A Random Variable (RV) is a function describing the outcome of a random process by assigning unique values to all possible outcomes of the experiment
- A RV models an attribute of our data (e.g. age, speech sample,...)
- Use uppercase to denote a RV, e.g. X, and lowercase to denote a value (observation), e.g. x
- A discrete (categorical) RV is defined on a finite or countable list of values Ω
- A continuous RV can take infinitely many values

Probability Functions

- Discrete Random Variables
  - A probability function $P(X = x) \in [0, 1]$ measures the probability of a RV X attaining the value x
  - Subject to sum-rule $\sum_{x \in \Omega} P(X = x) = 1$
- Continuous Random Variables
  - A density function $p(t)$ describes the relative likelihood of a RV to take on a value t
  - Subject to sum-rule $\int_{\Omega} p(t)dt = 1$
  - Defines a probability distribution, e.g.

    $$P(X \leq x) = \int_{-\infty}^{x} p(t)dt$$
- Shorthand $P(x)$ for $P(X = x)$ or $P(X \leq x)$

Joint and Conditional Probabilities

If a discrete random process is described by a set of RVs $X_1, \dots, X_N$, then the joint probability writes

$$P(X_1 = x_1, \dots, X_N = x_n) = P(x_1 \wedge \cdots \wedge x_n)$$

The joint conditional probability of $x_1, \dots, x_n$ given y

$$P(x_1, \dots, x_n | y)$$

measures the effect of the realization of an event y on the occurrence of $x_1, \dots, x_n$

A conditional distribution $P(x|y)$ is actually a family of distributions

- For each y, there is a distribution $P(x|y)$

Chain Rule

Definition (Product Rule a.k.a. Chain Rule)

$$P(x_1, \ldots, x_i, \ldots, x_n | y) = \prod_{i=1}^{N} P(x_i | x_1, \ldots, x_{i-1}, y)$$

Definition (Marginalization)

Using the sum and product rules together yield to the complete probability

$$P(X_1 = x_1) = \sum_{x_2} P(X_1 = x_1 | X_2 = x_2) P(X_2 = x_2)$$

Bayes Rule

Given hypothesis $h_i \in H$ and observations $\mathbf{d}$

$$P(h_i|\mathbf{d}) = \frac{P(\mathbf{d}|h_i)P(h_i)}{P(\mathbf{d})} = \frac{P(\mathbf{d}|h_i)P(h_i)}{\sum_j P(\mathbf{d}|h_j)P(h_j)}$$

- $P(h_i)$ is the prior probability of h_i
- $P(\mathbf{d}|h_i)$ is the conditional probability of observing $\mathbf{d}$ given that hypothesis h_i is true (likelihood).
- $P(\mathbf{d})$ is the marginal probability of $\mathbf{d}$
- $P(h_i|\mathbf{d})$ is the posterior probability that hypothesis is true given the data and the previous belief about the hypothesis.

Independence and Conditional Independence

- Two RV X and Y are independent if knowledge about X does not change the uncertainty about Y and vice versa

$$I(X,Y) \Leftrightarrow P(X,Y) = P(X|Y)P(Y) \\ = P(Y|X)P(X) = P(X)P(Y)$$

- Two RV X and Y are conditionally independent given Z if the realization of X and Y is an independent event of their conditional probability distribution given Z

$$I(X,Y|Z) \Leftrightarrow P(X,Y|Z) = P(X|Y,Z)P(Y|Z) \\ = P(Y|X,Z)P(X|Z) = P(X|Z)P(Y|Z)$$

- Shorthand $X \perp Y$ for $I(X,Y)$ and $X \perp Y|Z$ for $I(X,Y|Z)$

Inference and Learning in Probabilistic Models

Inference - How can one determine the distribution of the values of one/several RV, given the observed values of others?

$$P(graduate|exam_1, \ldots, exam_n)$$

Machine Learning view - Given a set of observations (data) **d** and a set of hypotheses $\{h_i\}_{i=1}^K$, how can I use them to predict the distribution of a RV X?

Learning - A very specific inference problem!

- Given a set of observations **d** and a probabilistic model of a given structure, how do I find the parameters θ of its distribution?
- Amounts to determining the best hypothesis h_θ regulated by a (set of) parameters θ

3 Approaches to Inference

Bayesian Consider all hypotheses weighted by their probabilities

$$P(X|\mathbf{d}) = \sum_i P(X|h_i)P(h_i|\mathbf{d})$$

MAP Infer X from $P(X|h_{MAP})$ where h_{MAP} is the Maximum a-Posteriori hypothesis given **d**

$$h_{MAP} = \arg\max_{h \in H} P(h|\mathbf{d}) = \arg\max_{h \in H} P(\mathbf{d}|h)P(h)$$

ML Assuming uniform priors $P(h_i) = P(h_j)$, yields the Maximum Likelihood (ML) estimate $P(X|h_{ML})$

$$h_{ML} = \arg\max_{h \in H} P(\mathbf{d}|h)$$

Considerations About Bayesian Inference

- The Bayesian approach is optimal but poses computational and analytical tractability issues

$$P(X|\mathbf{d}) = \int_H P(X|h)P(h|\mathbf{d})dh$$

- ML and MAP are point estimates of the Bayesian since they infer based only on one most likely hypothesis
- MAP and Bayesian predictions become closer as more data gets available
- MAP is a regularization of the ML estimation
  - Hypothesis prior $P(h)$ embodies trade-off between complexity and degree of fit
  - Well-suited to working with small datasets and/or large parameter spaces

Maximum-Likelihood (ML) Learning

Find the model θ that is most likely to have generated the data $\mathbf{d}$

$$\theta_{ML} = \arg\max_{\theta \in \Theta} P(\mathbf{d}|\theta)$$

from a family of parameterized distributions $P(x|\theta)$.

Optimization problem that considers the Likelihood function

$$\mathcal{L}(\theta|x) = P(x|\theta)$$

to be a function of θ.

Can be addressed by solving

$$\frac{\partial \mathcal{L}(\theta|x)}{\partial \theta} = 0$$

ML Learning with Hidden Variables

What if my probabilistic models contains both

- Observed random variables $\mathbf{X}$ (i.e. for which we have training data)
- Unobserved (hidden/latent) variables $\mathbf{Z}$ (e.g. data clusters)

ML learning can still be used to estimate model parameters

- The Expectation-Maximization algorithm which optimizes the complete likelihood

$$\mathcal{L}_c(\theta|\mathbf{X},\mathbf{Z}) = P(\mathbf{X},\mathbf{Z}|\theta) = P(\mathbf{Z}|\mathbf{X},\theta)P(\mathbf{X}|\theta)$$

- A 2-step iterative process

$$\theta^{(k+1)} = \arg\max_{\theta} \sum_{\mathbf{z}} P(\mathbf{Z}=\mathbf{z}|\mathbf{X},\theta^{(k)}) \log \mathcal{L}_c(\theta|\mathbf{X},\mathbf{Z}=\mathbf{z})$$

Joint Probabilities and Exponential Complexity

Discrete Joint Probability Distribution as a Table

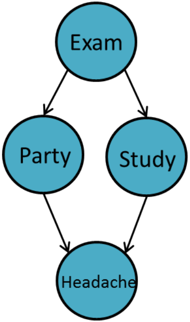
X_1	$\dots$	X_i	$\dots$	X_n	$P(X_1, \dots, X_n)$
$x_1^{'}$	$\dots$	$x_i^{'}$	$\dots$	$x_n^{'}$	$P(x_1^{'}, \dots, x_n^{'})$
x_1^{l}	$\dots$	x_i^{l}	$\dots$	x_n^{l}	$P(x_1^{l}, \dots, x_n^{l})$

- Describes $P(X_1, \dots, X_n)$ for all the RV instantiations
- For n binary RV X_i the table has 2^n entries!

Any probability can be obtained from the Joint Probability Distribution $P(X_1, \dots, X_n)$ by marginalization but again at an exponential cost (e.g. 2^{n-1} for a marginal distribution from binary RV).

Directed Graphical Models

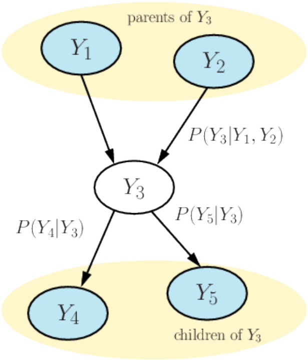
- Compact graphical representation for exponentially large joint distributions
- Simplifies marginalization and inference algorithms
- Allow to incorporate prior knowledge concerning causal relationships between RV



- Directed Graphical Models a.k.a. Bayesian Networks
- Describe conditional independence between subsets of RV by a graphical model

$$P(H|P,S,E) = P(H|P,S)$$

Bayesian Network



- Directed Acyclic Graph (DAG) $\mathcal{G} = (\mathcal{V}, \mathcal{E})$
- Nodes $v \in \mathcal{V}$ represent random variables
  - Shaded $\Rightarrow$ observed
  - Empty $\Rightarrow$ un-observed
- Edges $e \in \mathcal{E}$ describe the conditional independence relationships

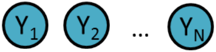
Conditional Probability Tables (CPT) local to each node describe the probability distribution given its parents

$$P(Y_1, \ldots, Y_N) = \prod_{i=1}^{N} P(Y_i|pa(Y_i))$$

A Simple Example

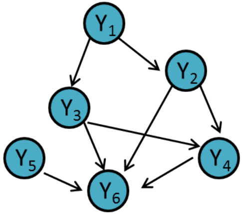
- Assume N discrete RV Y_i who can take k distinct values
- How many parameters in the joint probability distribution?
  $k^N - 1$ independent parameters

How many independent parameters if all N variables are independent? $N * (k - 1)$



$$P(Y_1, \ldots, Y_N) = \prod_{i=1}^{N} P(Y_I)$$

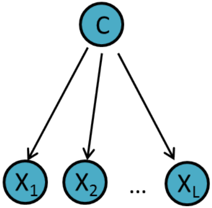
What if only part of the variables are (conditionally) independent?



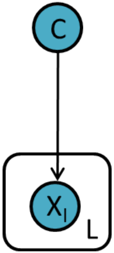
If the N nodes have a maximum of L children $\Rightarrow (k-1)^L \times N$ independent parameters

A Compact Representation of Replication

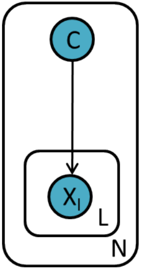
If the same causal relationship is replicated for a number of variables, we can compactly represent it by plate notation



The Naive Bayes Classifier

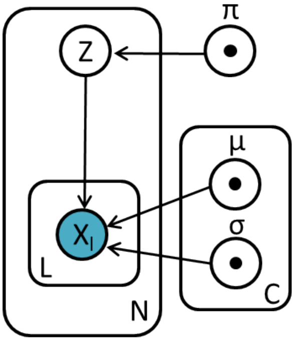


Replication for *L* attributes



Replication for *N* data samples

Full Plate Notation



Gaussian Mixture Model

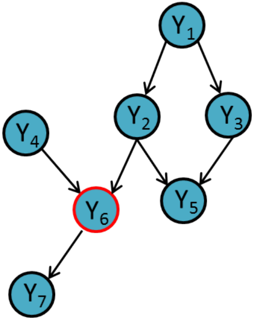
- Boxes denote replication for a number of times denoted by the letter in the corner
- Shaded nodes are observed variables
- Empty nodes denote un-observed latent variables
- Black seeds (optional) identify model parameters
  - $\pi \rightarrow$ multinomial prior distribution
  - $\mu \rightarrow$ means of the C Gaussians
  - $\sigma \rightarrow$ std of the C Gaussians

Local Markov Property

Definition (Local Markov property)

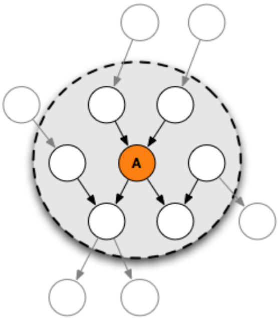
Each node / random variable is conditionally independent of all its non-descendants given a joint state of its parents

$$Y_v \perp Y_{V \setminus ch(v)} | Y_{pa(v)} \text{ for all } v \in \mathcal{V}$$



$$P(Y_1, Y_3, Y_5, Y_6 | Y_2, Y_4) = P(Y_6 | Y_2, Y_4) \times P(Y_1, Y_3, Y_5 | Y_2, Y_4)$$

Markov Blanket



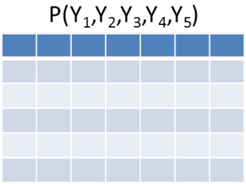
- The Markov Blanket $Mb(A)$ of a node A is the minimal set of vertices that shield the node from the rest of Bayesian Network
- The behavior of a node can be completely determined and predicted from the knowledge of its Markov blanket

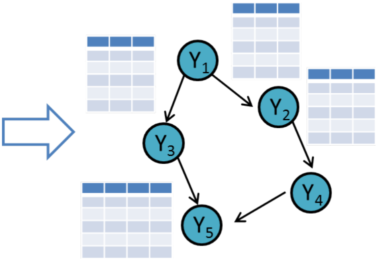
$$P(A|Mb(A), Z) = P(A|Mb(A)) \ \forall Z \notin Mb(A)$$

- The Markov blanket of A contains
  - Its parents $pa(A)$
  - Its children $ch(A)$
  - Its children's parents $pa(ch(A))$

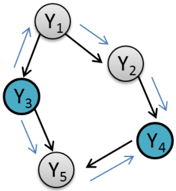
Why Using Bayesian Network?

Compacting parameter space





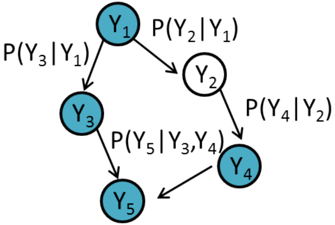
Reducing inference costs



- Variable elimination: e.g. compute $P(Y_4 | Y_3)$
- Inference algorithms exploiting sparse graph structure

Learning in Bayesian Networks

Parameter learning

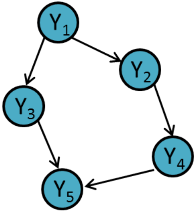


Infer the θ parameters of each conditional probability from data

- Includes hidden random variables (e.g. Y_2)

Learning network structure

Y_1	Y_2	Y_3	Y_4	Y_5
1	2	1	0	3
4	0	0	0	1
...	...	...	...	...
...	...	...	...	...
0	0	1	3	2

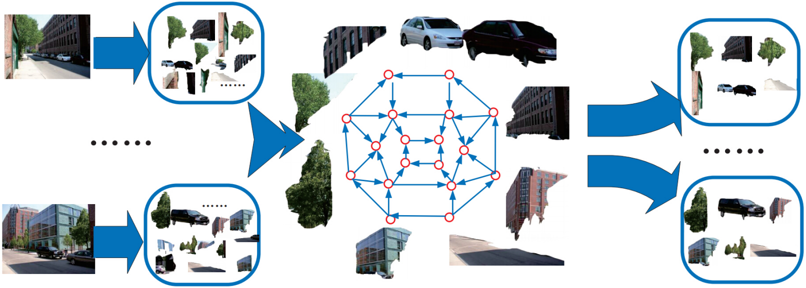


Determine edge presence and orientation from data

- Infer causal relationships

Learning to Segment Image Parts

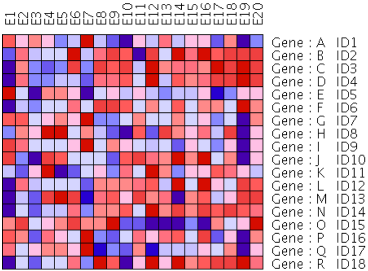
Latent Topics Network

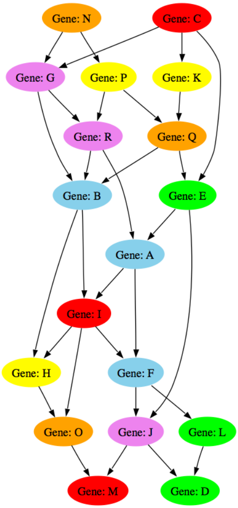


Yuan et al, A novel topic-level random walk framework for scene image co-segmentation, ECCV 2014

Discovering Gene Interaction Networks

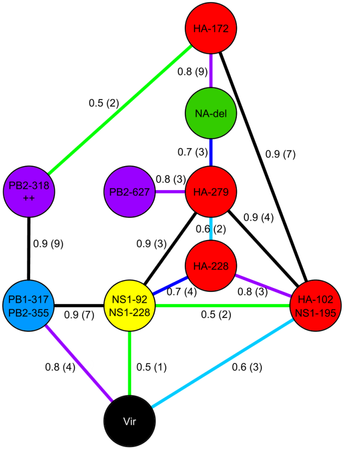
Gene Expression Data





Assessing Influenza Virulence

Inferred Gene Bayesian Graphical Model for H5N1 Virulence



http://hiv.bio.ed.ac.uk/samLycett/research.html

Take Home Messages

- Graphical models provide a compact representation for probabilistic models with a large number of variables
  - Use conditional independence to simplyfy joint probability
  - Efficient inference methods that exploit the sparse graph structure
  - Learning is a special-case of inference performed on the parameters of the probability functions in the graphical model
- Directed graphical models (Bayesian Networks)
  - Directed edges - Provide a representation of the causal relationships between variables
  - Parameter learning - Estimate the conditional probabilities associated with the nodes (visible or unobserved)
  - Structure learning - Use data to determine edge presence and orientation

Next Lecture

- Directed graphical models
  - Bayesian Networks
  - Determining conditional independence (d-separation)
  - Structure learning
- Undirected graphical models
  - Markov Random Fields
  - Joint probability factorization
  - Ising model