Generative and Graphical Models

Davide Bacciu

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

Intelligent Systems for Pattern Recognition (ISPR)

Generative Learning

- ML models that represent knowledge inferred from data under the form of probabilities
  - Probabilities can be sampled: new data can be generated
  - Supervised, unsupervised, weakly supervised learning tasks
  - Incorporate prior knowledge on data and tasks
  - Interpretable knowledge (how data is generated)
- 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

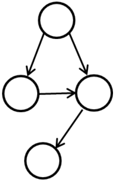
- Find the right model parameter
- An inference problem after all

Graphical Model Representation

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

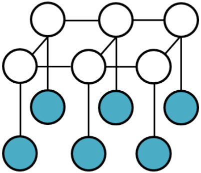
Different classes of graphs

Directed Models



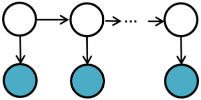
Directed edges express causal relationships

Undirected Models



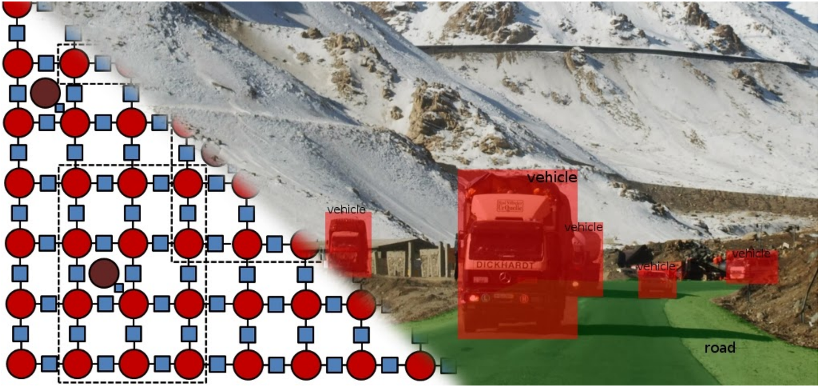
Undirected edges express soft constraints

Dynamic Models

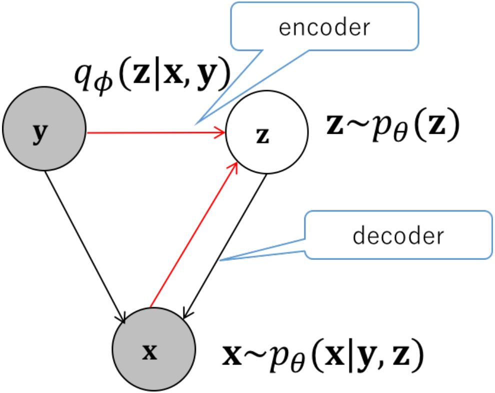


Structure changes to reflect dynamic processes

Generative Models in Machine Vision

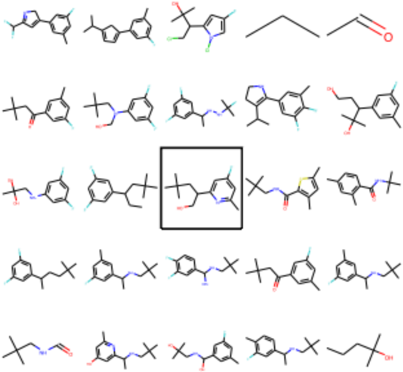


Generative Models in Deep Learning


encoder
$q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y})$
y
z
$\mathbf{z}\sim p_\theta(\mathbf{z})$
decoder
x
$\mathbf{x}\sim p_\theta(\mathbf{x}|\mathbf{y},\mathbf{z})$

Bayesian learning necessary to understand Variational Deep Learning

Generate New Knowledge



Complex data can be generated if your model is powerful enough to capture its distribution

Generative and Graphical Models Module

Lesson 1	Introduction: Directed and Undirected Graphical Models
Lesson 2-3	Dynamic GM: Hidden Markov Model
Lesson 4	Undirected GM: Markov Random Fields
Lesson 5	Bridging Neural and Generative: Boltzmann Machines
Lesson 6	Bayesian Learning and Approximated Inference: Latent Variable Models
Lesson 7	Sampling Methods

Lecture Outline

- Introduction
- A probabilistic refresher (from ML)
  - Probability theory
  - Conditional independence
  - Inference and learning in generative models
- Graphical Models
  - Directed and Undirected Representation
  - Independence assumptions, inference and learning
- Conclusions

Module content is fully covered by David Barber’s book

Probability and Learning Refresher

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

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

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)

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

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

$$\begin{aligned} I(X,Y) \Leftrightarrow P(X,Y) &= P(X|Y)P(Y) \\ &= P(Y|X)P(X) = P(X)P(Y) \end{aligned}$$

- 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

$$\begin{aligned} I(X,Y|Z) \Leftrightarrow P(X,Y|Z) &= P(X|Y,Z)P(Y|Z) \\ &= P(Y|X,Z)P(X|Z) = P(X|Z)P(Y|Z) \end{aligned}$$

- 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 **X** (i.e. for which we have training data)
- Unobserved (hidden/latent) variables **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

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

Graphical Models

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}	$P(x_1^{l}, \dots, x_n^{l})$

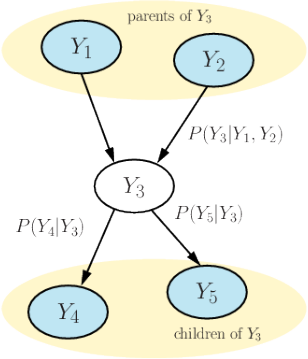
- Describes $P(X_1, \dots, X_n)$ for all the RV instantiations
- For n binary RV X_i the table has 2^n entries!

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).

Graphical Models

- Compact graphical representation for exponentially large joint distributions
- Simplifies marginalization and inference algorithms
- Allow to incorporate prior knowledge concerning causal relationships and associations between RV
    - Directed Graphical Models a.k.a. Bayesian Networks
    - Undirected Graphical Models a.k.a. Markov Random Fields

Bayesian Network



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

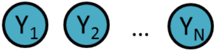
Conditional Probability Tables (CPT) local to each node describe the probability distribution given its parents

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

A Simple Example

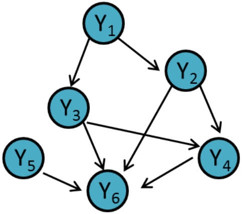
- Assume N discrete RV Y_i who can take k distinct values
- How many parameters in the joint probability distribution?
  $k^N - 1$ independent parameters

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



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

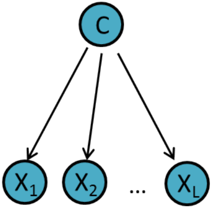
What if only part of the variables are (conditionally) independent?



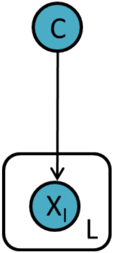
If the N nodes have a maximum of L children ⇒ $(k - 1)^L \times N$ independent parameters

A Compact Representation of Replication

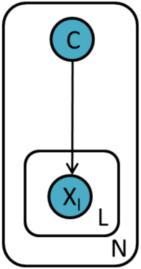
If the same causal relationship is replicated for a number of variables, we can compactly represent it by plate notation



The Naive Bayes Classifier

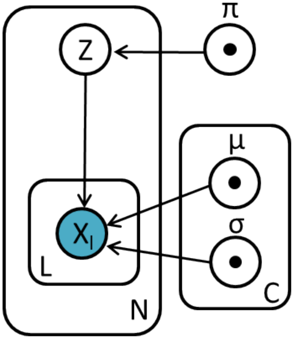


Replication for *L* attributes



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

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

Party and Study are marginally independent

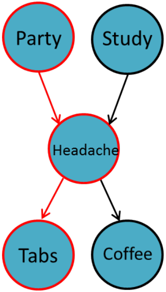
- $Party \perp Study$

However, local Markov property does not support

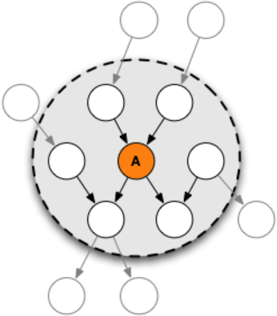
- $Party \perp Study | Headache$
- $Tabs \perp Party$

But Party and Tabs are independent given Headache

- $Tabs \perp Party | Headache$



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

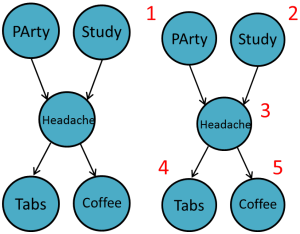
$$P(A|Mb(A), Z) = P(A|Mb(A)) \ \forall Z \notin Mb(A)$$

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

Joint Probability Factorization

An application of Chain rule and Local Markov Property

1. Pick a topological ordering of nodes
2. Apply chain rule following the order
3. Use the conditional independence assumptions



$$P(PA, S, H, T, C) =$$

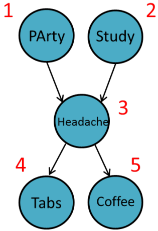
$$P(PA) \cdot P(S|PA) \cdot P(H|S, PA) \cdot P(T|H, S, PA) \cdot P(C|T, H, S, PA)$$

$$= P(PA) \cdot P(S) \cdot P(H|S, PA) \cdot P(T|H) \cdot P(C|H)$$

Sampling from a Bayesian Network

A BN describes a generative process for observations

1. Pick a topological ordering of nodes
2. Generate data by sampling from the local conditional probabilities following this order



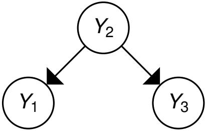
Generate i-th sample for each variable PA, S, H, T, C

1. $pa_i \sim P(PA)$
2. $s_i \sim P(S)$
3. $h_i \sim P(H|S = s_i, PA = pa_i)$
4. $t_i \sim P(T|H = h_i)$
5. $c_i \sim P(C|H = h_i)$

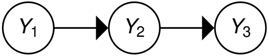
Basic Structures of a Bayesian Network

There exist 3 basic substructures that determine the conditional independence relationships in a Bayesian network

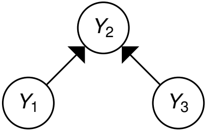
- Tail to tail (Common Cause)



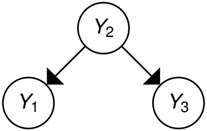
- Head to tail (Causal Effect)

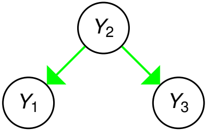


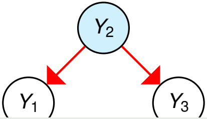
- Head to head (Common Effect)



Tail to Tail Connections







- Corresponds to

$$P(Y_1, Y_3 | Y_2) = P(Y_1 | Y_2) P(Y_3 | Y_2)$$

- If Y_2 is unobserved then Y_1 and Y_3 are marginally dependent

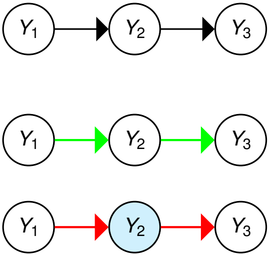
$$Y_1 \not\perp Y_3$$

- If Y_2 is observed then Y_1 and Y_3 are conditionally independent

$$Y_1 \perp Y_3 | Y_2$$

When Y_2 in observed is said to block the path from Y_1 to Y_3

Head to Tail Connections



Observed Y_2 blocks the path from Y_1 to Y_3

- Corresponds to

$$
\begin{aligned}
P(Y_1, Y_3|Y_2) &= P(Y_1)P(Y_2|Y_1)P(Y_3|Y_2) \\
&= P(Y_1|Y_2)P(Y_3|Y_2)
\end{aligned}
$$

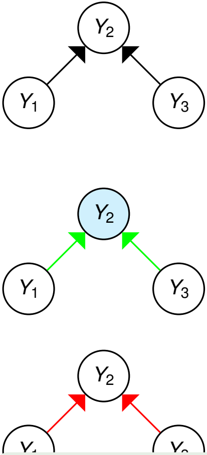
- If Y_2 is unobserved then Y_1 and Y_3 are marginally dependent

$$Y_1 \not\perp Y_3$$

- If Y_2 is observed then Y_1 and Y_3 are conditionally independent

$$Y_1 \perp Y_3|Y_2$$

Head to Head Connections



- Corresponds to

$$P(Y_1, Y_2, Y_3) = P(Y_1)P(Y_3)P(Y_2|Y_1, Y_3)$$

- If Y_2 is observed then Y_1 and Y_3 are conditionally dependent

$$Y_1 \not\perp Y_3 | Y_2$$

- If Y_2 is unobserved then Y_1 and Y_3 are marginally independent

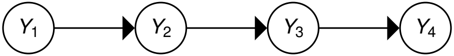
$$Y_1 \perp Y_3$$

If any Y_2 descendants is observed it unlocks the path

Derived Conditional Independence Relationships

A Bayesian Network represents the local relationships encoded by the 3 basic structures plus the derived relationships

Consider



Local Markov Relationships

$$Y_1 \perp Y_3 | Y_2$$

$$Y_4 \perp Y_1, Y_2 | Y_3$$

Derived Relationship

$$Y_1 \perp Y_4 | Y_2$$

d-Separation

Definition (d-separation)

Let $r = Y_1 \longleftrightarrow \cdots \longleftrightarrow Y_2$ be an undirected path between Y_1 and Y_2, then r is d-separated by Z if there exist at least one node $Y_c \in Z$ for which path r is blocked.

In other words, d-separation holds if at least one of the following holds

- r contains an head-to-tail structure $Y_i \longrightarrow Y_c \longrightarrow Y_j$ (or $Y_i \longleftarrow Y_c \longleftarrow Y_j$) and $Y_c \in Z$
- r contains a tail-to-tail structure $Y_i \longleftarrow Y_c \longrightarrow Y_j$ and $Y_c \in Z$
- r contains an head-to-head structure $Y_i \longrightarrow Y_c \longleftarrow Y_j$ and neither Y_c nor its descendants are in Z

Markov Blanket and d-Separation

Definition (Nodes d-separation)

Two nodes Y_i and Y_j in a BN $\mathcal{G}$ are said to be d-separated by $Z \subset \mathcal{V}$ (denoted by $Dsep_{\mathcal{G}}(Y_i, Y_j|Z)$) if and only if all undirected paths between Y_i and Y_j are d-separated by Z

Definition (Markov Blanket)

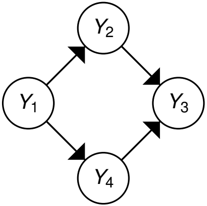
The Markov blanket $Mb(Y)$ is the minimal set of nodes which d-separates a node Y from all other nodes (i.e. it makes Y conditionally independent of all other nodes in the BN)

$$Mb(Y) = \{pa(Y), ch(Y), pa(ch(Y))\}$$

Are Directed Models Enough?

- Bayesian Networks are used to model asymmetric dependencies (e.g. causal)
- What if we want to model symmetric dependencies
  - Bidirectional effects, e.g. spatial dependencies
  - Need undirected approaches

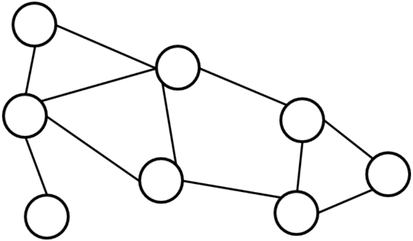
Directed models cannot represent some (bidirectional) dependencies in the distributions



What if we want to represent $Y_1 \perp Y_3 | Y_2, Y_4$?
What if we also want $Y_2 \perp Y_4 | Y_1, Y_3$?

Cannot be done in BN! Need undirected model

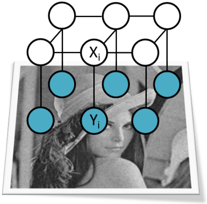
Markov Random Fields

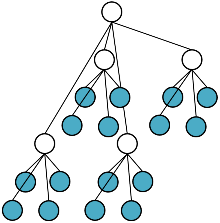


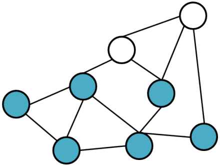
- Undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ (a.k.a. Markov Networks)
- Nodes $v \in \mathcal{V}$ represent random variables X_v
  - Shaded $\Rightarrow$ observed
  - Empty $\Rightarrow$ un-observed
- Edges $e \in \mathcal{E}$ describe bi-directional dependencies between variables (constraints)

Often arranged in a structure that is coherent with the data/constraint we want to model

Image Processing



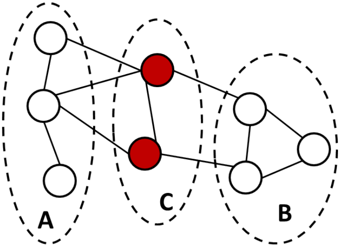




- Often used in image processing to impose spatial constraints (e.g.smoothness)
- Image de-noising example
  - Lattice Markov Network (Ising model)
  - $Y_i \rightarrow$ observed value of the noisy pixel
  - $X_i \rightarrow$ unknown (unobserved) noise-free pixel value
- Can use more expressive structures
  - Complexity of inference and learning can become relevant

Conditional Independence

What is the undirected equivalent of d-separation in directed models?



$A \perp B | C$

Again it is based on node separation, although it is way simpler!

- Node subsets $A, B \subset \mathcal{V}$ are conditionally independent given $C \subset \mathcal{V} \setminus \{A, B\}$ if all paths between nodes in A and B pass through at least one of the nodes in C
- The Markov Blanket of a node includes all and only its neighbors

Joint Probability Factorization

What is the undirected equivalent of conditional probability factorization in directed models?

- We seek a product of functions defined over a set of nodes associated with some local property of the graph
- Markov blanket tells that nodes that are not neighbors are conditionally independent given the remainder of the nodes

$$P(X_v, X_i | X_{\mathcal{V} \setminus \{v,i\}}) = P(X_v | X_{\mathcal{V} \setminus \{v,i\}}) P(X_i | X_{\mathcal{V} \setminus \{v,i\}})$$

- Factorization should be chosen in such a way that nodes X_v and X_i are not in the same factor

What is a well-known graph structure that includes only nodes that are pairwise connected?

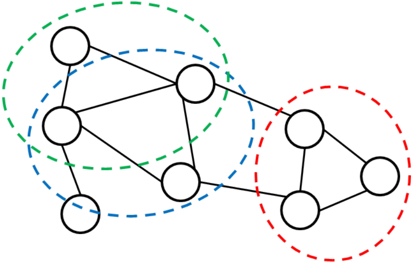
Cliques

Definition (Clique)

A subset of nodes C in graph $\mathcal{G}$ such that $\mathcal{G}$ contains an edge between all pair of nodes in C

Definition (Maximal Clique)

A clique C that cannot include any further node from the graph without ceasing to be a clique



Maximal Clique Factorization

Define $\mathbf{X} = X_1, \dots, X_N$ as the RVs associated to the N nodes in the undirected graph $\mathcal{G}$

$$P(\mathbf{X}) = \frac{1}{Z} \prod_{C} \psi(\mathbf{X}_C)$$

- $\mathbf{X}_C \rightarrow$ RV associated with nodes in the maximal clique C
- $\psi(\mathbf{X}_C) \rightarrow$ potential function over the maximal cliques C
- $Z \rightarrow$ partition function ensuring normalization

$$Z = \sum_{\mathbf{X}} \prod_{C} \psi(\mathbf{X}_C)$$

Partition function is the computational bottleneck of undirected modes: e.g. $O(K^N)$ for N discrete RV with K distinct values

Potential Functions

- Potential functions $\psi(\mathbf{X}_C)$ are not probabilities!
- Express which configurations of the local variables are preferred
- If we restrict to strictly positive potential functions, the Hammersley-Clifford theorem provides guarantees on the distribution that can be represented by the clique factorization

Definition (Boltzmann distribution)

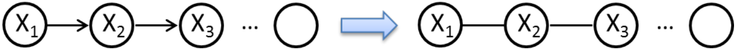
A convenient and widely used strictly positive representation of the potential functions is

$$\psi(\mathbf{X}_C) = \exp\left\{-E(\mathbf{X}_C)\right\}$$

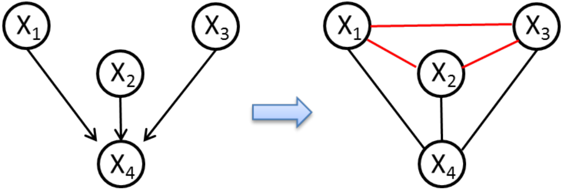
where $E(\mathbf{X}_C)$ is called energy function

From Directed To Undirected

Straightforward in some cases



Requires a little bit of thinking for v-structures



Moralization a.k.a. marrying of the parents

Take Home Messages

- Generative models as a gateway for next-gen deep learning
- Directed graphical models
  - Represent asymmetric (causal) relationships between RV and conditional probabilities in compact way
  - Difficult to assess conditional independence (v-structures)
  - Ok for prior knowledge and interpretation
- Undirected graphical models
  - Represent bi-directional relationships (e.g. constraints)
  - Factorization in terms of generic potential functions (not probabilities)
  - Easy to assess conditional independence, but difficult to interpret
  - Serious computational issues due to normalization factor

Generative Models in Code

- PyMC3 - Bayesian statistics and probabilistic ML; gradient-based Markov chain Monte Carlo variational inference (Python, Theano)
- Edward - Bayesian statistics and ML, deep learning, and probabilistic programming (Python, TensorFlow)
- Pyro - Deep probabilistic programming (Python, PyTorch)
- TensorFlow Probability - Combine probabilistic models and deep learning with GPU/TPU support (Python)
- PyStruct - Markov Random Field models in Python (some of them)
- Pgmpy - Python package for Probabilistic Graphical Models
- Stan - Probabilistic programming language for statistical inference (native C++, PyStan package)

Next Lecture

Hidden Markov Model (HMM)

- A dynamic graphical model for sequences
- Unfolding learning models on structures
- Exact inference on a chain with observed and unobserved variables
- The Expectation-Maximization algorithm for HMMs