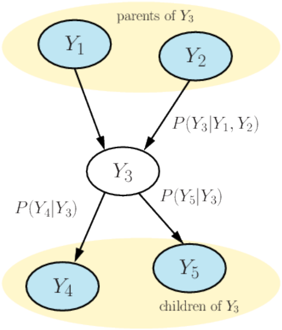
Directed and Undirected Graphical Models

Davide Bacciu

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

Machine Learning: Neural Networks and Advanced Models (AA2)

Directed Graphical Models (Bayesian Networks)



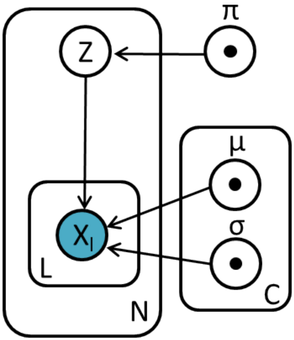
- Directed Acyclic Graph (DAG) $\mathcal{G} = (\mathcal{V}, \mathcal{E})$
- Nodes $v \in \mathcal{V}$ represent random variables
  - Shaded $\Rightarrow$ observed
  - Empty $\Rightarrow$ un-observed
- Edges $e \in \mathcal{E}$ describe the conditional independence relationships

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

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

Plate Notation

A compact representation of replication in graphical models



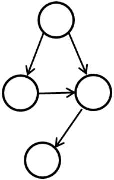
- Boxes denote replication for a number of times denoted by the letter in the corner
- Shaded nodes are observed variables
- Empty nodes denote un-observed latent variables
- Black seeds (optional) identify model parameters

Graphical Models

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

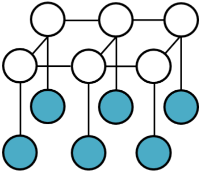
Different classes of graphs

Directed Models



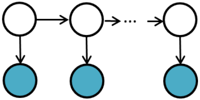
Directed edges express causal relationships

Undirected Models



Undirected edges express soft constraints

Dynamic Models



Structure changes to reflet dynamic processes

Directed Models - Local Markov Property

A variable Y_v is independent of its non-descendants given its parents and only its parents: i.e. $Y_v \perp Y_{V \backslash ch(v)} | Y_{pa(v)}$

Party and Study are marginally independent

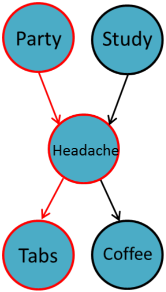
- $Party \perp Study$

However, local Markov property does not support

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

But Party and Tabs are independent given Headache

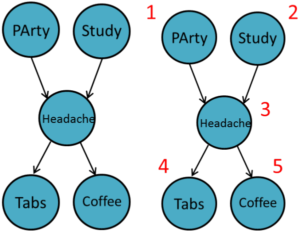
- $Tabs \perp Party | Headache$



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

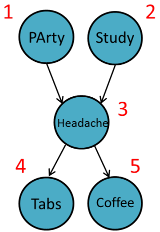


$$P(PA, S, H, T, C) =$$
$$P(PA) \cdot P(S|PA) \cdot P(H|S, PA) \cdot P(T|H, S, PA) \cdot P(C|T, H, S, PA)$$
$$= P(PA) \cdot P(S) \cdot P(H|S, PA) \cdot P(T|H) \cdot P(C|H)$$

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



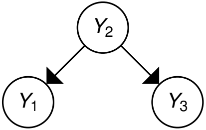
Generate i-th sample for each variable PA, S, H, T, C

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. $c_i \sim P(C|H = h_i)$

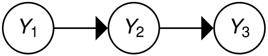
Basic Structures of a Bayesian Network

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

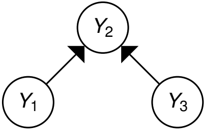
- Tail to tail (Common Cause)



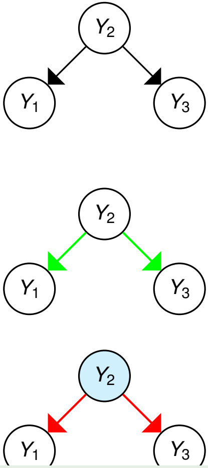
- Head to tail (Causal Effect)



- 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

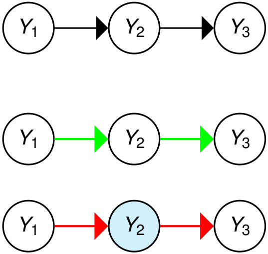
$$Y_1 \not\perp Y_3$$

- 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

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

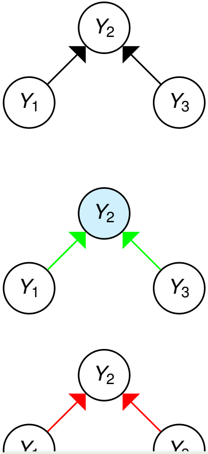
- If Y_2 is unobserved then Y_1 and Y_3 are marginally dependent

$$Y_1 \not\perp Y_3$$

- 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

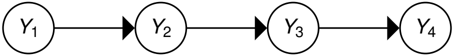
$$Y_1 \perp Y_3$$

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 $r = Y_1 \longleftrightarrow \ldots \longleftrightarrow Y_2$ 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)

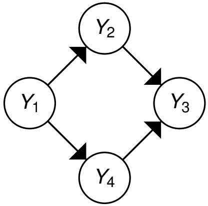
The Markov blanket $Mb(Y)$ is the minimal set of nodes which d-separates a node Y from all other nodes (i.e. it makes Y conditionally independent of all other nodes in the BN)

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

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

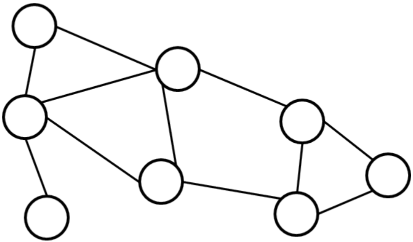
Directed models cannot represent some (bidirectional) dependencies in the distributions



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

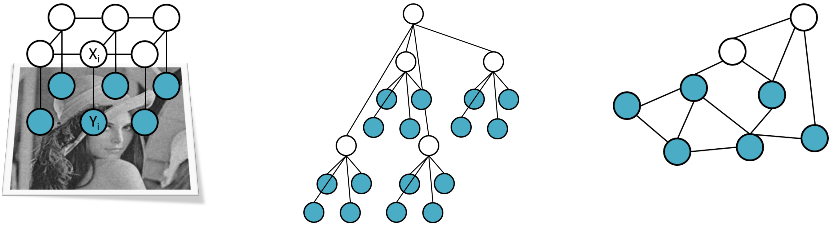
Markov Random Fields



- 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

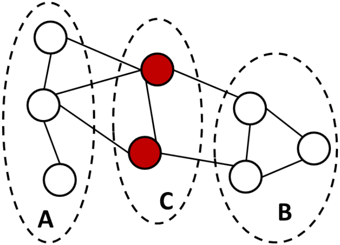
Image Processing



- 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?

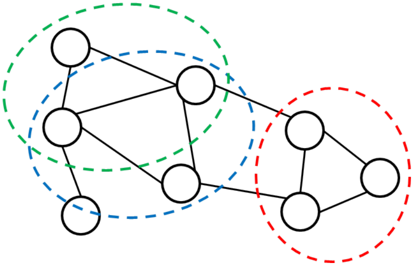
Cliques

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 $\mathbf{X} = X_1, \ldots, X_N$ 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)

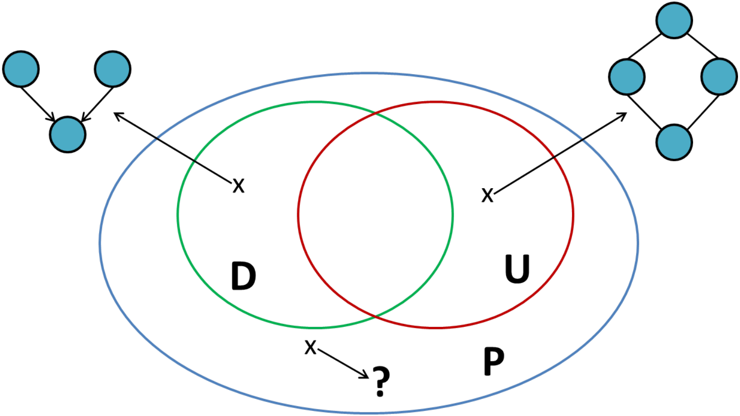
A convenient and widely used strictly positive representation of the potential functions is

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

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

Directed Vs Undirected Models

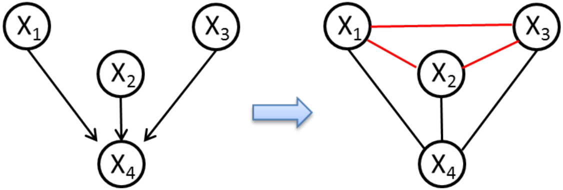
Long story short



From Directed To Undirected

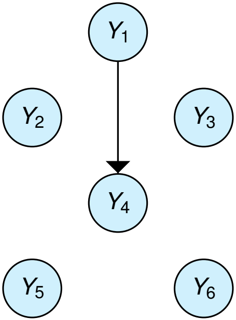
Straightforward in some cases

Requires a little bit of thinking for v-structures



Moralization a.k.a. marrying of the parents

The BN Structure Learning Problem

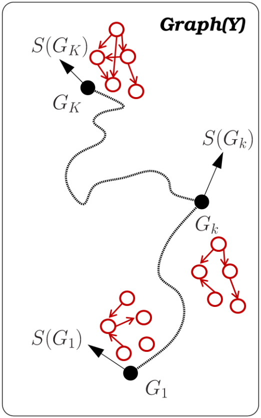


Y_1	Y_2	Y_3	Y_4	Y_5	Y_6
1	2	1	0	3	4
4	0	0	0	1	2
...	...	...	...	...	...
...	...	...	...	...	...
0	0	1	3	2	1

- Observations are given for a set of fixed random variables
- But the structure of the Bayesian Network is not specified
  - How do we determine which arcs exist in the network (causal relationships)?
- Determining causal relationships between variables entails
  - Deciding on arc presence
  - Directing edges

Structure Finding Approaches

- Search and Score
  - Model selection approach
  - Search in the space of the graphs
- Constraint Based
  - Use tests of conditional independence
  - Constrain the network
- Hybrid
  - Model selection of constrained structures



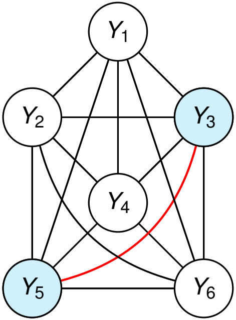
Constraint-based Models Outline

- Tests of conditional independence $I(X_i, X_j|Z)$ determine edge presence (network skeleton)
  - Estimate mutual information $MI(X_i, X_j|Z)$ and assume conditional independence if MI is below a threshold, e.g. $I(X_i, X_j|Z) = MI(X_i, X_j|Z) < \alpha_{cut}$
- Testing order is the fundamental choice for avoiding super-exponential complexity
  - Level-wise testing: tests $I(X_i, X_j|Z)$ are performed in order of increasing size of the conditioning set Z (PC algorithm by Spirtes, 1995)
  - Nodes that enter Z are chosen in the neighborhood of X_i and X_j
- Markovian dependencies determine edge orientation (DAG)
  - Deterministic rules based on the 3 basic substructures seen previously

PC Algorithm Skeleton Identification

1. Initialize a fully connected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$
2. **for each** edge $(Y_i, Y_j) \in \mathcal{V}$
   - **if** $I(Y_i, Y_j)$ **then** prune (Y_i, Y_j)
3. $K \leftarrow 1$
4. **for each** test of order $K = |Z|$
   - **for each** edge $(Y_i, Y_j) \in \mathcal{V}$
     - $Z \leftarrow$ set of conditioning sets of K-th order for Y_i, Y_j
     - **if** $I(Y_i, Y_j|z)$ **for any** $z \in Z$ **then** prune (Y_i, Y_j)
   - $K \leftarrow K + 1$
5. **return** $\mathcal{G}$

PC Algorithm
Order 0 Tests



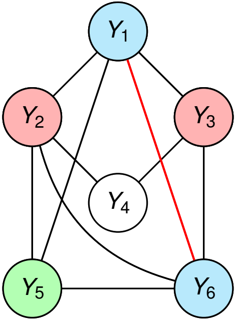
Y_1	Y_2	Y_3	Y_4	Y_5	Y_6
1	2	1	0	3	4
4	0	0	0	1	2
...	...	...	...	...	...
...	...	...	...	...	...
0	0	1	3	2	1

Step 1 Initialize

Step 2 Check unconditional independence $I(Y_i, Y_j)$

Step 3 Repeat unconditional tests for all edges

PC Algorithm
Order 1 Tests



Y_1	Y_2	Y_3	Y_4	Y_5	Y_6
1	2	1	0	3	4
4	0	0	0	1	2
...	...	...	...	...	...
...	...	...	...	...	...
0	0	1	3	2	1

Step 4 Select an edge (Y_i, Y_j)

Step 5 Add the neighbors to the conditioning set Z

Step 6 Check independence for each $z \in Z$

Step 7 Iterate until convergence

Take Home Messages

- Directed graphical models
  - Represent asymmetric (causal) relationships between variables and provide a compact representation of conditional probabilities
  - Difficult to assess conditional independence relationships (v-structures)
  - Straightforward to incorporate prior knowledge and to interpret
- Undirected graphical models
  - Represent bi-directional relationships between variables (e.g. constraints)
  - Factorization in terms of generic potential functions which, however, are typically not probabilities
  - Easy to assess conditional independence, but difficult to interpret the encoded knowledge
  - Serious computational issues associated with computation of normalization factor

Next Lecture

- Inference in Graphical Models
- Exact inference
  - Inference on a chain
  - Inference in tree-structured models
  - Sum-product algorithm
- Elements of approximate inference
  - Variational algorithms
  - Sampling-based methods