732A96/TDDE15 Advanced Machine Learning
Graphical Models

Jose M. Peña
IDA, Linköping University, Sweden

Lecture 1: Bayesian and Markov Networks

Contents

- Causal Structures
- Bayesian Networks
    - Definition
    - Causal Inference
    - Probabilistic Inference
- Markov Networks
    - Definition
    - Probabilistic Inference

Literature

- Main source
    - Bishop, C. M. *Pattern Recognition and Machine Learning*. Springer, 2006. Chapter 8.
- Additional source
    - Koski, T. J. T. and Noble, J. M. A Review of Bayesian Networks and Structure Learning. *Mathematica Applicanda* 40, 51-103, 2012.
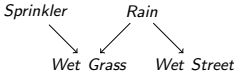
Causal Structures

- Assume that we want to represent the causal relations between a set of random variables, e.g. the variables may represent the state of the components of a system.

- A natural and intuitive representation consists of a **graph** where the nodes are the random variables, and the edges are the causal relations between the variables. We call such a graph a **causal structure**.

$$Sprinkler \qquad Rain$$
$$\searrow \qquad \swarrow \quad \searrow$$
$$Wet\ Grass \qquad Wet\ Street$$
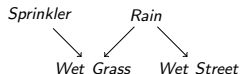
- What happens if some random variables are not modeled ?
  - Nothing if they have at most one modeled effect, e.g. mediators.
  - Problems otherwise, e.g. $WG \rightarrow WS$ and $WG \leftarrow WS$ are both wrong.

- **Exercise**. Produce a causal structure for the domain *Temperature*, *Ice cream sales* and *Soda sales*.

- **Exercise**. Produce a causal structure for Boyle's law, which relates the pressure and volume of a gas as *Pressure · Volume = constant* if the temperature and amount of gas remain unchanged within a closed system.

Bayesian Networks: Definition

DAG	Parameter values for the conditional probability distributions
Sprinkler *Rain* *Wet Grass* *Wet Street*	$q(s) = (0.3, 0.7) = (\theta_{s_0}, \theta_{s_1})$ $q(r) = (0.5, 0.5) = (\theta_{r_0}, \theta_{r_1})$ $q(wg|r_0, s_0) = (0.1, 0.9) = (\theta_{wg_0|r_0, s_0}, \theta_{wg_1|r_0, s_0})$ $q(wg|r_0, s_1) = (0.7, 0.3) = (\theta_{wg_0|r_0, s_1}, \theta_{wg_1|r_0, s_1})$ $q(wg|r_1, s_0) = (0.8, 0.2) = (\theta_{wg_0|r_1, s_0}, \theta_{wg_1|r_1, s_0})$ $q(wg|r_1, s_1) = (0.9, 0.1) = (\theta_{wg_0|r_1, s_1}, \theta_{wg_1|r_1, s_1})$ $q(ws|r_0) = (0.1, 0.9) = (\theta_{ws_0|r_0}, \theta_{ws_1|r_0})$ $q(ws|r_1) = (0.7, 0.3) = (\theta_{ws_0|r_1}, \theta_{ws_1|r_1})$ $p(s, r, wg, ws) = q(s)q(r)q(wg|s, r)q(ws|r)$

- A **Bayesian network (BN)** over a finite set of **discrete** random variables $X = X_{1:n} = \{X_1, \ldots, X_n\}$ consists of
  - a directed acyclic graph (DAG) G whose nodes are the elements in X, and
  - parameter values θ specifying probability distributions $q(x_i|pa_i)$, where Pa_i are the parents of X_i in G, i.e. the nodes with an edge into X_i.
- The BN represents a causal model of the system.
- And also a probabilistic model of the system as $p(x) = \prod_i q(x_i|pa_i)$.

Bayesian Networks: Definition

- We now show that $p(x) = \prod_i q(x_i|pa_i)$ is a probability distribution.
- Clearly, $0 \leq \prod_i q(x_i|pa_i)) \leq 1$.
- Assume without loss of generality that $Pa_i \subseteq X_{1:i-1}$ for all i. Then

$$\sum_x \prod_i q(x_i|pa_i) = \sum_{x_1}[q(x_1) \ldots \sum_{x_{n-1}}[q(x_{n-1}|pa_{n-1})\sum_{x_n} q(x_n|pa_n)]\ldots] = 1$$

- Moreover, $p(x_j|pa_j) = q(x_j|pa_j)$. To see it, note that

$$
\begin{aligned}
p(x_j|pa_j) = \frac{p(x_j, pa_j)}{p(pa_j)} &= \frac{\sum_{x \smallsetminus \{x_j, pa_j\}} \prod_i q(x_i|pa_i)}{\sum_{x \smallsetminus pa_j} \prod_i q(x_i|pa_i)} \\
&= \frac{\sum_{x_{1:j} \smallsetminus \{x_j, pa_j\}} \prod_{i \leq j} q(x_i|pa_i)}{\sum_{x_{1:j} \smallsetminus pa_j} \prod_{i \leq j} q(x_i|pa_i)} = q(x_j|pa_j)
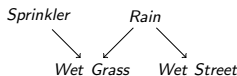\end{aligned}
$$

Bayesian Networks: Separation

- We now show that many of the independencies in p can be read off G without numerical calculations.

$$Age \longrightarrow Battery \qquad Bulb$$
$$Radio \qquad Lights$$
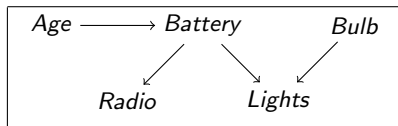
- **Chain**: $Age \to Battery \to Radio$
  - $Age \not\perp Radio | \varnothing$
  - $Age \perp Radio | Battery$
- **Fork**: $Radio \leftarrow Battery \to Lights$
  - $Radio \not\perp Lights | \varnothing$
  - $Radio \perp Lights | Battery$
- **Collider**: $Battery \to Lights \leftarrow Bulb$
  - $Battery \perp Bulb | \varnothing$
  - $Battery \not\perp Bulb | Lights$
- **Chain + collider**: $Age \to Battery \to Lights \leftarrow Bulb$
  - $Age \perp Bulb | \varnothing$
  - $Age \not\perp Bulb | Lights$
  - $Age \perp Bulb | Lights, Battery$

Bayesian Networks: Separation

- A path in G is a sequence of distinct and adjacent nodes, i.e. the direction of the edge is irrelevant. A node B is a descendant of a node A in G if there is a path $A \to \ldots \to B$.
    - E.g., $Age \to Battery \to Lights \leftarrow Bulb$ is a path.
    - E.g., $Lights$ is a descendant of Age.
- Let ρ be a path in G between the nodes α and β.
- A node B in ρ is a **collider** when $A \to B \leftarrow C$ is a subpath of ρ.
    - E.g., $Lights$ is a collider in the path $Age \to Battery \to Lights \leftarrow Bulb$.
- Moreover, ρ is Z-**open** with $Z \subseteq X \smallsetminus \{\alpha, \beta\}$ when
    - no non-collider in ρ is in Z, and
    - every collider in ρ is in Z or has a descendant in Z.
    - E.g., the path $Age \to Battery \to Lights \leftarrow Bulb$ is Z-open with $Z = \{Lights\}$.
- Let U, V and Z be three disjoint subsets of X. Then, U and V are **separated** given Z in G (i.e. $U \perp_G V | Z$) when there is no Z-open path in G between a node in U and a node in V.
    - E.g., $Age \perp_G Bulb | \varnothing$.
- The separation criterion is **sound**, i.e. if $U \perp_G V | Z$ then $U \perp_p V | Z$.
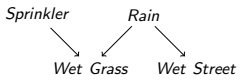- For instance, $S \perp_p R$, $S \not\perp_p R | WG$, $S \not\perp_p WS | WG$, $S \perp_p WS | WG, R$.

$$Sprinkler \qquad Rain$$
$$\searrow \swarrow \qquad \searrow$$
$$Wet\ Grass \qquad Wet\ Street$$

- Note that we read independencies from G, never dependencies.

Bayesian Networks: Separation

- Moreover, the separation criterion is also **complete**, i.e. p may be such that $U \perp_G V | Z$ if and only if $U \perp_p V | Z$.
- Moreover, p factorizes as $p(X) = \prod_i q(x_i | pa_i)$ if and only if it satisfies all the independencies identified by the separation criterion.
- **Exercise**. Prove that $A \perp_p B | C$ for the DAGs $A \to C \to B$, $A \leftarrow C \to B$ and $A \leftarrow C \leftarrow B$, i.e. prove that $p(a, b | c) = p(a|c)p(b|c)$.
- **Exercise**. Prove that $A \perp_p B | \varnothing$ for the DAG $A \to C \leftarrow B$, i.e. prove that $p(a, b) = p(a)p(b)$.
- **Exercise**. Find the minimal set of nodes that separates a given node from the rest. This set is called the Markov blanket of the given node.
- **Exercise**. How many free parameters do we have in the wet grass BN ? How many do we have if we specify the distribution without the assistance of a BN ?

Bayesian Networks: Causal Inference

Original	After $do(r_1)$

Sprinkler   *Rain*   *Sprinkler*

Wet Grass   *Wet Street*   *Wet Grass*   *Wet Street*

$q(s) = (0.3, 0.7)$
$q(r) = (0.5, 0.5)$
$q(wg|r_0, s_0) = (0.1, 0.9)$ $q(s) = (0.3, 0.7)$
$q(wg|r_0, s_1) = (0.7, 0.3)$ $q(wg|s_0) = (0.8, 0.2)$
$q(wg|r_1, s_0) = (0.8, 0.2)$ $q(wg|s_1) = (0.9, 0.1)$
$q(wg|r_1, s_1) = (0.9, 0.1)$ $q(ws) = (0.7, 0.3)$
$q(ws|r_0) = (0.1, 0.9)$
$q(ws|r_1) = (0.7, 0.3)$ $p(s, wg, ws) = q(s)q(wg|s)q(ws)$

$p(s, r, wg, ws) = q(s)q(r)q(wg|s, r)q(ws|r)$

▸ What would be the state of the system if a random variable X_j is forced to take the state x_j, i.e. $p(x \smallsetminus x_j | do(x_j))$?
  ▸ Remove X_j and all the edges from and to X_j from G.
  ▸ Remove $q(x_j | pa_j)$.
  ▸ If $X_j \in Pa_i$, then replace $q(x_i | pa_i)$ with $q(x_i | pa_i \smallsetminus x_j, x_j)$
  ▸ Set $p(x \smallsetminus x_j | do(x_j)) = \prod_i q(x_i | pa_i)$.
▸ So, the result of $do(x)$ on a BN **is a BN**. More on causality in Lecture 5.

Bayesian Networks: Probabilistic Inference

- What is the state of a random variable X_j if a random variable X_i is observed to be in the state x_i, i.e. $p(x_k|x_i)$?

  - $p(x_k|x_i) = \frac{p(x_k, x_i)}{p(x_i)} = \frac{\sum_{x \setminus \{x_i, x_k\}} p(x)}{\sum_{x \setminus x_i} p(x)}$

  - $p(ws|s) = \frac{\sum_{r,wg} p(r, wg, ws, s)}{\sum_{r, wg, ws} p(r, wg, ws, s)}$

    $= \frac{\sum_{r,wg} q(s)q(r)q(wg|s,r)q(ws|r)}{\sum_{r, wg, ws} q(s)q(r)q(wg|s,r)q(ws|r)} = \frac{q(s)\sum_r[q(r)q(ws|r)\sum_{wg} q(wg|s,r)]}{q(s)\sum_r[q(r)\sum_{wg}[q(wg|s,r)\sum_{ws} q(ws|r)]]}$

- What is the state of a random variable X_k if a random variable X_i is observed to be in the state x_i, after forcing a random variable X_j to take the state x_j, i.e. $p(x_k|x_i, do(x_j))$?

- Answering questions like the one above can be computationally hard.

- A BN is an efficient (because it uses the independences encoded) formalism to compute a posterior probability distribution from a prior probability distribution in the light of observations, hence the name.

Markov Networks: Definition

▸ A BN represents asymmetric (causal) relations, whereas a Markov network represents **symmetric** relations, e.g. physical laws.

UG	Potentials assuming binary random variables
A —— B $\mid$ $\diagup$ $\mid$ C —— D	$\varphi(a, b, c) = (0, 0, 0, 0, 1, 1, 1, 1)$ $\varphi(b, c, d) = (1, 2, 3, 4, 5, 6, 7, 8)$ $p(a, b, c, d) = \varphi(a, b, c)\varphi(b, c, d)/Z$ with $Z = \sum_{a,b,c,d} \varphi(a, b, c)\varphi(b, c, d)$

▸ A **Markov network (MN)** over X consists of
  ▸ an undirected graph (UG) G whose nodes are the elements in X, and
  ▸ a set of non-negative functions $\varphi(k)$ over the cliques $Cl(G)$ of G, i.e. the maximal complete sets of nodes in G. The functions are called potentials. They represent **compatibility** relations between the random variables in the cliques.

▸ The MN represents a probabilistic model of the system, namely

$$p(x) = \frac{1}{Z} \prod_{K \in Cl(G)} \varphi(k)$$

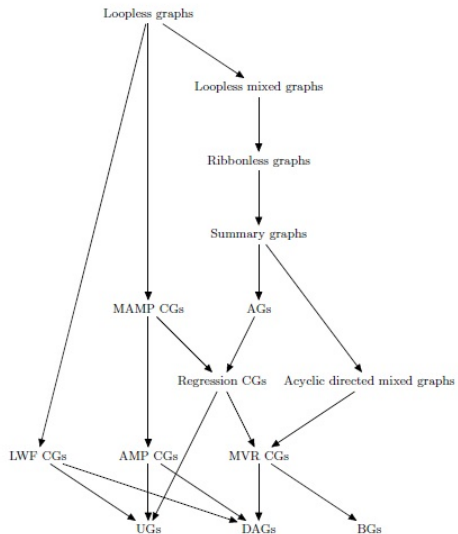where Z is a normalization constant, i.e.

$$Z = \sum_{x} \prod_{K \in Cl(G)} \varphi(k)$$

▸ Clearly, $p(x)$ is a probability distribution.

Markov Networks: Separation

- We now show that many of the independencies in p can be read off G without numerical calculations.
- A path ρ in G between two nodes α and β is Z-**open** with $Z \subseteq X \smallsetminus \{\alpha, \beta\}$ when no node in ρ is in Z.
- Let U, V and Z be three disjoint subsets of X. Then, U and V are **separated** given Z in G (i.e. $U \perp_G V | Z$) when there is no Z-open path in G between a node in U and a node in V.
- The separation criterion is **sound**, i.e. if $U \perp_G V | Z$ then $U \perp_p V | Z$.
- Moreover, it is also **complete**, i.e. p may be such that $U \perp_G V | Z$ if and only if $U \perp_p V | Z$.
- Moreover, p factorizes as $p(x) = \frac{1}{Z} \prod_{K \in Cl(G)} \varphi(k)$ if and only if it satisfies all the independencies identified by the separation criterion.

Markov Networks: Separation

- **Exercise**. Prove that $A \perp_p B | C$ for the UG $A - C - B$, i.e. prove that $p(a, b | c) = f(a, c) g(b, c)$ for some functions f and g.
- **Exercise**. Find the minimal set of nodes that separates a given node from the rest. This set is called the Markov blanket of the given node.
- **Exercise**. How many free parameters do we have in the ABCD MN ? How many do we have if we specify the distribution without the assistance of a MN ?

Markov Networks: Probabilistic Inference
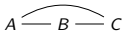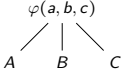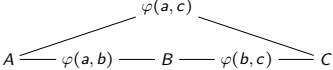
- What is the state of a random variable A if a random variable B is observed to be in the state b ?

$$p(a|b) = \frac{\sum_{c,d} \varphi(a,b,c)\varphi(b,c,d)/Z}{\sum_{a,c,d} \varphi(a,b,c)\varphi(b,c,d)/Z} = \frac{\sum_{c}[\varphi(a,b,c)\sum_{d}\varphi(b,c,d)]}{\sum_{a,c}[\varphi(a,b,c)\sum_{d}\varphi(b,c,d)]}$$

- Answering questions like the one above can be computationally hard.
- A MN is an efficient (because it uses the independences encoded) formalism to answer such questions.

Markov Networks: Factor Graphs

- ▶ What if $\varphi(a, b, c) = \varphi(a, b)\varphi(b, c)\varphi(a, c)$? That is, $\varphi(k) = \prod_j \varphi(k_j)$ with $K_j \subset K$ for all j.
- ▶ A MN may obscure the structure of the potentials. Solution: Factor graphs.
- ▶ A **factor graph** over X consists of an UG G with two types of nodes: The elements in X and a set of potentials $\varphi(k)$ over subsets of X. All the edges in G are between a potential and the elements of X that are in the potential's domain.

MN	Factor graph	Factor graph
$A \frown B \frown C$	$\varphi(a, b, c)$ $A \quad B \quad C$	$\varphi(a, c)$ $A \frown \varphi(a, b) \frown B \frown \varphi(b, c) \frown C$

- ▶ The factor graph represents a probabilistic model of the system, namely

$$p(x) = \frac{1}{Z} \prod_K \varphi(k)$$

where Z is a normalization constant, i.e.

$$Z = \sum_x \prod_K \varphi(k)$$

- ▶ Factor graphs: Finer-grained parameterization of MNs.

Intersection of Bayesian and Markov Networks

All independence models

DAGs   UGs
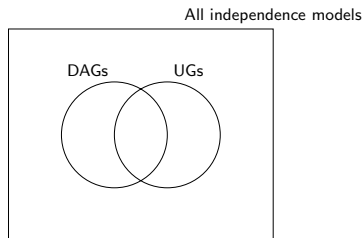
- An **unshielded collider** in a DAG is a subgraph of the form $A \rightarrow C \leftarrow B$ such that A and B are not adjacent in the DAG.
- An UG is **triangulated** if every cycle in it contains a chord, i.e. an edge between two non-consecutive nodes in the cycle.
- Given a DAG G, there is an UG H such that G and H represent the same separations if and only if G has no unshielded colliders.
- Given an UG G, there is an DAG H such that G and H represent the same separations if and only if G is triangulated.

Families of Graphical Models

Loopless graphs → Loopless mixed graphs → Ribbonless graphs → Summary graphs → MAMP CGs, AGs → Regression CGs, Acyclic directed mixed graphs → LWF CGs, AMP CGs, MVR CGs → UGs, DAGs, BGs

Relevance of Graphical Models

Contents

Thank you