# 732A96 Advanced Machine Learning

## Graphical Models and Hidden Markov Models

Jose M. Peña
IDA, Linköping University, Sweden

Lecture 1: Causal Models, Bayesian Networks and Markov Networks

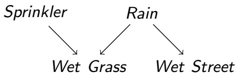
# Contents

# 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.

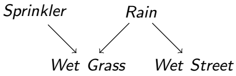
# Causal Models

- Assume that we want to represent the causal relations between a set of random variables, e.g. the latter may represent the state of the components of a system.
- A natural and intuitive representation consists of a **graph** where the random variables are the nodes, and the causal relations are the edges. We call the graph a **causal model**.




- What happens if some random variables are not modeled ?
  - Nothing if they have at most one modeled effect, e.g. mediators.
  - Problems otherwise, i.e. $WG \rightarrow WS$ and $WG \leftarrow WS$ are both wrong.
- **Exercise**. Produce a causal model for the domain *Temperature*, *Ice cream sales* and *Soda sales*.
- **Exercise**. Produce a causal model for Boyle's law, which relates the pressure and volume of a gas as $Pressure \cdot Volume = constant$ if the temperature and amount of gas remain unchanged within a closed system.

# Bayesian Networks: Definition

| DAG | Parameter values for the conditional probability distributions |
|---|---|
|  | $q(S) = (0.3, 0.7)$<br>$q(R) = (0.5, 0.5)$<br>$q(WG \mid r_0, s_0) = (0.1, 0.9)$<br>$q(WG \mid r_0, s_1) = (0.7, 0.3)$<br>$q(WG \mid r_1, s_0) = (0.8, 0.2)$<br>$q(WG \mid r_1, s_1) = (0.9, 0.1)$<br>$q(WS \mid r_0) = (0.1, 0.9)$<br>$q(WS \mid r_1) = (0.7, 0.3)$<br><br>$p(S, R, WG, WS) = q(S)q(R)q(WG \mid S, R)q(WS \mid R)$ |

- A **Bayesian network (BN)** over a finite set of **discrete** random variables $X = X_{1:n} = \{X_1, \dots, X_n\}$ consists of
  - a DAG $G$ whose nodes are the elements in $X$, and
  - parameter values $\theta_G$ specifying conditional probability distributions $q(X_i \mid pa_G(X_i))$.
- The BN represents a causal model of the system.
- The BN also represents a probabilistic model of the system, namely $p(X) = \prod_i q(X_i \mid pa_G(X_i))$.

# Bayesian Networks: Definition

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

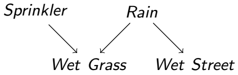
$$\sum_x \prod_i q(x_i|pa_G(X_i)) = \sum_{x_1}[q(x_1)\ldots \sum_{x_{n-1}}[q(x_{n-1}|pa_G(X_{n-1}))\sum_{x_n} q(x_n|pa_G(X_n))]\ldots] = 1$$

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

$$\begin{aligned} p(X_j|pa_G(X_j)) &= \frac{p(X_j, pa_G(X_j))}{p(pa_G(X_j))} = \frac{\sum_{X\setminus\{X_j,pa_G(X_j)\}} \prod_i q(X_i|pa_G(X_i))}{\sum_{X\setminus pa_G(X_j)} \prod_i q(X_i|pa_G(X_i))} \\ &= \frac{\sum_{X_{1:j}\setminus\{X_j,pa_G(X_j)\}} \prod_{i\leq j} q(X_i|pa_G(X_i))}{\sum_{X_{1:j}\setminus pa_G(X_j)} \prod_{i\leq j} q(X_i|pa_G(X_i))} = q(X_j|pa_G(X_j)) \end{aligned}$$

# Bayesian Networks: Separation

- We now show that many of the independencies in $p$ can be read off $G$ without numerical calculations.
- A path in $G$ is a sequence of 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 $A \to \ldots \to B$.
- Let $\rho$ be a path between the nodes $\alpha$ and $\beta$ in $G$.
- A node $B$ in $\rho$ is a **collider** when $A \to B \leftarrow C$ is a subpath of $\rho$.
- Moreover, $\rho$ is $Z$-open with $Z \subseteq X \setminus \{\alpha, \beta\}$ when
  - no non-collider in $\rho$ is in $Z$, and
  - every collider in $\rho$ is in $Z$ or has a descendant 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 $U \perp_p V|Z$.
- For instance, $S \perp_p R$, $S \not\perp_p R|WG$, $S \not\perp_p WS|WG$, $S \perp_p WS|WG, R$.

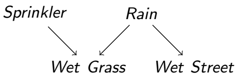
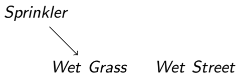



- Note that we read independencies from $G$, never dependencies.

# 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 $p(X) = \prod_i q(X_i | pa_G(X_i))$ 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.

# Bayesian Networks: Causal Reasoning

| Original | After $do(r_1)$ |
|---|---|
|  |  |
| $q(S) = (0.3, 0.7)$<br>$q(R) = (0.5, 0.5)$<br>$q(WG \mid r_0, s_0) = (0.1, 0.9)$<br>$q(WG \mid r_0, s_1) = (0.7, 0.3)$<br>$q(WG \mid r_1, s_0) = (0.8, 0.2)$<br>$q(WG \mid r_1, s_1) = (0.9, 0.1)$<br>$q(WS \mid r_0) = (0.1, 0.9)$<br>$q(WS \mid r_1) = (0.7, 0.3)$<br><br>$p(S, R, WG, WS) = q(S)q(R)q(WG \mid S, R)q(WS \mid R)$ | $q(S) = (0.3, 0.7)$<br>$q(WG \mid s_0) = (0.8, 0.2)$<br>$q(WG \mid s_1) = (0.9, 0.1)$<br>$q(WS) = (0.7, 0.3)$<br><br>$p(S, WG, WS) = q(S)q(WG \mid S)q(WS)$ |

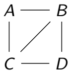
- What would be the state of the system if a random variable $X_j$ is forced to take the state $x_j$, i.e. $p(X \setminus X_j \mid do(x_j))$ ?
  - Remove $X_j$ and all the edges from and to $X_j$ from $G$.
  - Remove $q(X_j \mid pa_G(X_j))$.
  - If $X_j \in pa_G(X_i)$, then replace $q(X_i \mid pa_G(X_i))$ with $q(X_i \mid pa_G(X_i) \setminus X_j, x_j)$
  - Set $p(X \setminus X_j \mid do(x_j)) = \prod_i q(X_i \mid pa_G(X_i))$.
- So, the result of $do(x)$ on a BN **is a BN**.

# Bayesian Networks: Probabilistic Reasoning

- What is the state of the system if a random variable $X_i$ is observed to be in the state $x_i$, i.e. $p(X \setminus X_i | x_i)$ ?
  - $p(X \setminus X_i | x_i) = \frac{p(X \setminus X_i, x_i)}{p(x_i)} = \frac{p(X \setminus X_i, x_i)}{\sum_{X \setminus X_i} p(X \setminus X_i, x_i)}$
  - $p(R, WG, WS | s) = \frac{q(s) q(R) q(WG|s,R) q(WS|R)}{\sum_{r,wg,ws} q(s) q(r) q(wg|s,r) q(ws|r)}$
    $= \frac{q(s) q(R) q(WG|s,R) q(WS|R)}{q(s) \sum_r [q(r) \sum_{wg} [q(wg|s,r) \sum_{ws} q(ws|r)]]}$
- What is the state of a random variable $Y$ if a random variable $X_i$ is observed to be in the state $x_i$, i.e. $p(Y|x_i)$ ?
  - $p(Y|x_i) = \frac{p(Y, x_i)}{p(x_i)} = \frac{\sum_{X \setminus \{X_i, Y\}} p(X \setminus X_i, x_i)}{\sum_{X \setminus X_i} p(X \setminus X_i, x_i)}$
  - $p(WS|s) = \frac{\sum_{r,wg} q(s) q(r) q(wg|s,r) q(WS|r)}{\sum_{r,wg,ws} q(s) q(r) q(wg|s,r) q(ws|r)}$
    $= \frac{q(s) \sum_r [q(r) q(WS|r) \sum_{wg} q(wg|s,r)]}{q(s) \sum_r [q(r) \sum_{wg} [q(wg|s,r) \sum_{ws} q(ws|r)]]}$
- What is the state of a random variable $Y$ 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. $p(Y|x_i, do(x_j))$ ?
- Answering questions like the one above can be computationally hard.
- A BN is an efficient 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 |
|---|---|
|  | $\varphi(A,B,C) = (1,1,1,1,2,2,2,2)$<br>$\varphi(B,C,D) = (1,2,3,4,5,6,7,8)$<br><br>$p(A,B,C,D) = \varphi(A,B,C)\varphi(B,C,D)/Z$ 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 $G$.
- A clique is a maximal complete set of nodes. 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 $\rho$ between two nodes $\alpha$ and $\beta$ in $G$ is $Z$-open with $Z \subseteq X \setminus \{\alpha, \beta\}$ when no node in $\rho$ 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.

# Markov Networks: Probabilistic Reasoning

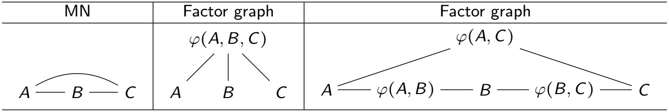
- What is the state of a random variable $A$ if a random variable $B$ is observed to be in the state $b$ ?

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

- Answering questions like the one above can be computationally hard.
- A MN is an efficient formalism to answer such questions.

# Markov Networks: Factor Graphs

- What if $\varphi(A, B, C) = \varphi(A, B)\varphi(B, C)\varphi(A, C)$ ? I.e. $\varphi(C_i) = \prod_j \varphi(C_i^j)$ with $C_i^j \subset C_i$.
- 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.




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

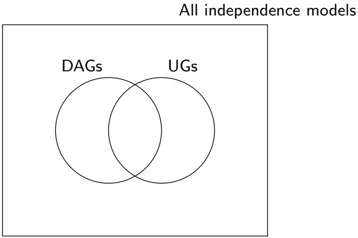
$$p(X) = \frac{1}{Z} \prod_K \varphi(K)$$

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




- An **unshielded collider** in a DAG is a subgraph of the form $A \to C \leftarrow B$ 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 $\perp_G \equiv \perp_H$ if and only if $G$ has no unshielded colliders.
- Given an UG $G$, there is an DAG $H$ such that $\perp_G \equiv \perp_H$ if and only if $G$ is triangulated.

# Contents

- Causal Models
- Bayesian Networks
  - Definition
  - Causal Reasoning
  - Probabilistic Reasoning
- Markov Networks
  - Definition
  - Probabilistic Reasoning

Thank you