732A96/TDDE15 Advanced Machine Learning

Graphical Models

Jose M. Peña
IDA, Linköping University, Sweden

Lecture 1: Bayesian and Markov Networks

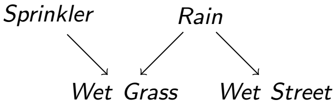
Contents

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.

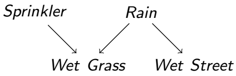
Causal Structures

- 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**.



- **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 $Pressure \cdot Volume = constant$ if the temperature and amount of gas remain unchanged within a closed system.

Bayesian Networks: Definition

<table>
<tr><th>DAG</th><th>Parameter values for the conditional probability distributions</th></tr>
<tr><td>


</td><td>

$$
\begin{aligned}
&q(s) = (0.3, 0.7) = (\theta_{s_0}, \theta_{s_1})\\
&q(r) = (0.5, 0.5) = (\theta_{r_0}, \theta_{r_1})\\
&q(wg|r_0, s_0) = (0.1, 0.9) = (\theta_{wg_0|r_0,s_0}, \theta_{wg_1|r_0,s_0})\\
&q(wg|r_0, s_1) = (0.7, 0.3) = (\theta_{wg_0|r_0,s_1}, \theta_{wg_1|r_0,s_1})\\
&q(wg|r_1, s_0) = (0.8, 0.2) = (\theta_{wg_0|r_1,s_0}, \theta_{wg_1|r_1,s_0})\\
&q(wg|r_1, s_1) = (0.9, 0.1) = (\theta_{wg_0|r_1,s_1}, \theta_{wg_1|r_1,s_1})\\
&q(ws|r_0) = (0.1, 0.9) = (\theta_{ws_0|r_0}, \theta_{ws_1|r_0})\\
&q(ws|r_1) = (0.7, 0.3) = (\theta_{ws_0|r_1}, \theta_{ws_1|r_1})
\end{aligned}
$$

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

</td></tr>
</table>

- A **Bayesian network (BN)** over a finite set of **discrete** random variables $X = X_{1:n} = \{X_1, \dots, X_n\}$ 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 $p(x) = \prod_i q(x_i|pa_i)$.

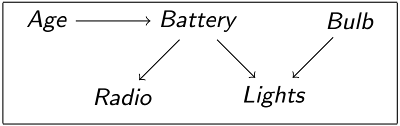
Bayesian Networks: Definition

- 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

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

Bayesian Networks: Separation

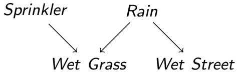
- We now show that many of the independencies in p can be read off G without numerical calculations.



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

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 $Z \subseteq X \setminus \{\alpha, \beta\}$ 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, $S \perp_p R$, $S \not\perp_p R | WG$, $S \not\perp_p WS | WG$, $S \perp_p WS | WG, R$.

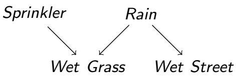


- Note that we read independencies from G, never dependencies.

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 $A \rightarrow C \rightarrow B$, $A \leftarrow C \rightarrow B$ 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 $A \rightarrow C \leftarrow B$, 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 $do(r_1)$
	
$q(s) = (0.3, 0.7)$ $q(r) = (0.5, 0.5)$ $q(wg \mid r_0, s_0) = (0.1, 0.9)$ $q(wg \mid r_0, s_1) = (0.7, 0.3)$ $q(wg \mid r_1, s_0) = (0.8, 0.2)$ $q(wg \mid r_1, s_1) = (0.9, 0.1)$ $q(ws \mid r_0) = (0.1, 0.9)$ $q(ws \mid r_1) = (0.7, 0.3)$	$q(s) = (0.3, 0.7)$ $q(wg \mid s_0) = (0.8, 0.2)$ $q(wg \mid s_1) = (0.9, 0.1)$ $q(ws) = (0.7, 0.3)$
$p(s, r, wg, ws) = q(s)q(r)q(wg \mid s, r)q(ws \mid r)$	$p(s, wg, ws) = q(s)q(wg \mid s)q(ws)$

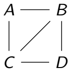
- What would be the state of the system if a random variable X_j is forced to take the state x_j, i.e. $p(x \setminus x_j | do(x_j))$?
  - Remove X_j and all the edges from and to X_j from G.
  - Remove $q(x_j | pa_j)$.
  - If $X_j \in Pa_i$, then replace $q(x_i | pa_i)$ with $q(x_i | pa_i \setminus x_j, x_j)$
  - Set $p(x \setminus x_j | do(x_j)) = \prod_i q(x_i | pa_i)$.
- 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_k 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
	$\varphi(a,b,c) = (0,0,0,0,1,1,1,1)$ $\varphi(b,c,d) = (1,2,3,4,5,6,7,8)$ $p(a,b,c,d) = \varphi(a,b,c)\varphi(b,c,d)/Z$ 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 $Z \subseteq X \setminus \{\alpha, \beta\}$ 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 ? How many if the variables have three states ?

Markov Networks: Probabilistic Inference

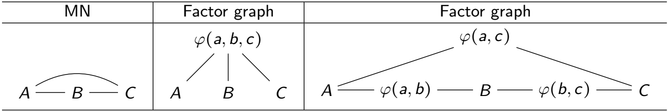
- What is the state of a random variable A if a random variable B is observed to be in the state b ?

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

- 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.



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

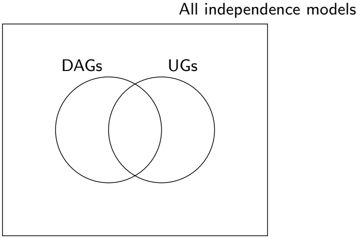
$$p(x) = \frac{1}{Z} \prod_K \varphi(k)$$

where Z is a normalization constant, i.e.

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

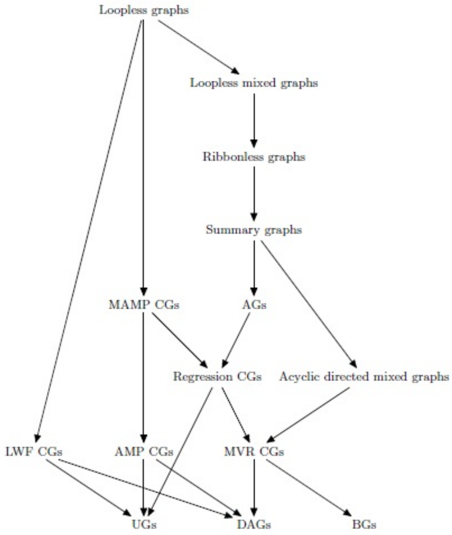
- Factor graphs: Finer-grained parameterization of MNs.

Intersection of Bayesian and Markov Networks

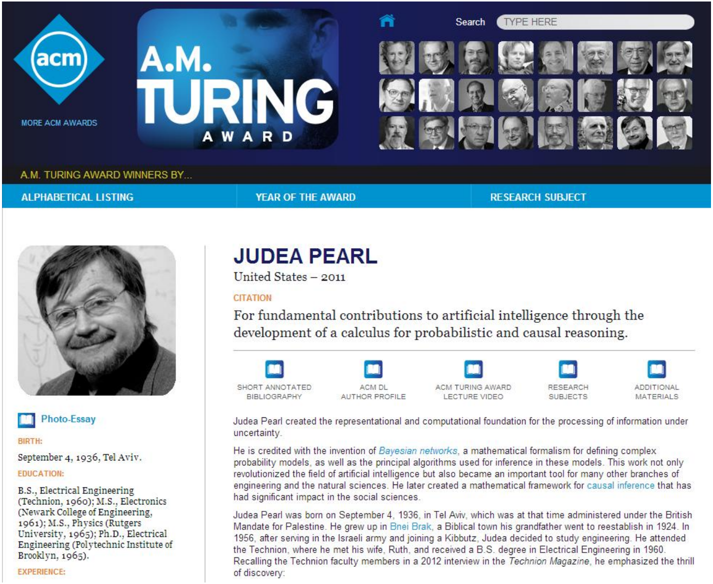


- An **unshielded collider** in a DAG is a subgraph of the form $A \to C \leftarrow B$ such that A and B are not adjacent in the DAG.
- An UG is **triangulated** if every cycle in it contains a chord, i.e. an edge between two non-consecutive nodes in the cycle.
- Given a DAG G, there is an UG H such that 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



Relevance of Graphical Models



acm

MORE ACM AWARDS

A.M. TURING AWARD

Search TYPE HERE

A.M. TURING AWARD WINNERS BY...

ALPHABETICAL LISTING | YEAR OF THE AWARD | RESEARCH SUBJECT

Photo-Essay

BIRTH:

September 4, 1936, Tel Aviv.

EDUCATION:

B.S., Electrical Engineering (Technion, 1960); M.S., Electronics (Newark College of Engineering, 1961); M.S., Physics (Rutgers University, 1965); Ph.D., Electrical Engineering (Polytechnic Institute of Brooklyn, 1965).

EXPERIENCE:

JUDEA PEARL

United States – 2011

CITATION

For fundamental contributions to artificial intelligence through the development of a calculus for probabilistic and causal reasoning.

SHORT ANNOTATED BIBLIOGRAPHY | ACM DL AUTHOR PROFILE | ACM TURING AWARD LECTURE VIDEO | RESEARCH SUBJECTS | ADDITIONAL MATERIALS

Judea Pearl created the representational and computational foundation for the processing of information under uncertainty.

He is credited with the invention of *Bayesian networks*, a mathematical formalism for defining complex probability models, as well as the principal algorithms used for inference in these models. This work not only revolutionized the field of artificial intelligence but also became an important tool for many other branches of engineering and the natural sciences. He later created a mathematical framework for causal inference that has had significant impact in the social sciences.

Judea Pearl was born on September 4, 1936, in Tel Aviv, which was at that time administered under the British Mandate for Palestine. He grew up in Bnei Brak, a Biblical town his grandfather went to reestablish in 1924. In 1956, after serving in the Israeli army and joining a Kibbutz, Judea decided to study engineering. He attended the Technion, where he met his wife, Ruth, and received a B.S. degree in Electrical Engineering in 1960. Recalling the Technion faculty members in a 2012 interview in the *Technion Magazine*, he emphasized the thrill of discovery:

Contents

Thank you