732A96 Advanced Machine Learning

Graphical Models and Hidden Markov Models

Jose M. Peña
IDA, Linköping University, Sweden

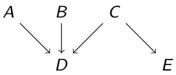
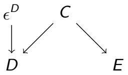
Lecture 1: Causal Models, Bayesian Networks and Markov Networks

Contents

Literature

- Main source
  - Bishop, C. M. *Pattern Recognition and Machine Learning*. Springer, 2006.
  - Pearl, J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.
  - Pearl, J. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.
- Other sources
  - Koller, D. and Friedman, N. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.
  - Koski, T. J. T. and Noble, J. M. A Review of Bayesian Networks and Structure Learning. *Mathematica Applicanda* 40, 51-103, 2012.
- R resources
  - Højsgaard, S. Graphical Independence Networks with the gRain Package for R. *Journal of Statistical Software* 46, 2012.
  - Højsgaard, S., Edwards, D. and Lauritzen, S. *Graphical Models with R*. Springer, 2012.
  - Nagarajan, R., Scutari, M. and Lébre, S. *Bayesian Networks in R*. Springer, 2013.
  - Scutari, S. Learning Bayesian Networks with the bnlearn R Package. *Journal of Statistical Software* 35, 2010.

Causal Models: Qualitative

Microscopic DAG	Macroscopic DAG	Macroscopic Violates assumption
		

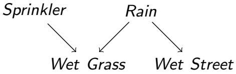
- At the microscopic level, every system is a causal system, **assuming** Reichenbach's principle of common cause: "No correlation without causation".
- Then, at the microscopic level, every system can be represented by a causal model.
- The structure of the causal model can be represented as a **directed and acyclic graph (DAG)**.
- At the macroscopic level, every system can be represented by a causal model **if no unmodeled variable** is cause of two or more modeled variables. The unmodeled causes of a variable X are aggregated into an error variable ϵ^X.

Causal Models: Quantitative

Microscopic DAG Functional	Macroscopic DAG Functional	Macroscopic DAG Probabilistic
$\begin{tikzcd} A \arrow[dr] & B \arrow[d] & C \arrow[dl] \arrow[dr] & \\ & D = f(A,B,C) & & E = f(C) \end{tikzcd}$	$\begin{tikzcd} \epsilon^D \arrow[d] & C \arrow[dl] \arrow[dr] & \\ D = f(C,\epsilon^D) & & E = f(C) \end{tikzcd}$	$\begin{tikzcd} & C \arrow[dl] \arrow[dr] & \\ D \sim q(D\|C) & & E \sim q(E\|C) \end{tikzcd}$

- **Assuming** Laplace's demon, every variable is a deterministic function of its causes at the microscopic level.
- Then, every variable is a probabilistic function of its modeled causes at the macroscopic level.
- Both Reichenbach's principle of common cause and Laplace's demon have been disproven. Human reasoning seem to comply with both of them, though.
- Then, probabilistic DAGs may not be ontological but epistemological models. They also help to identify key questions about reasoning.

Bayesian Networks: Definition

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

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

Bayesian Networks: Definition

- We now show that $p(X) = \prod_i q(X_i | pa_G(X_i))$ is a probability distribution.
- Clearly, $0 \le \prod_i q(X_i | pa_G(X_i)) \le 1$.
- 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, $p(X_i | pa_G(X_i)) = q(X_i | pa_G(X_i))$. To see it, note that

$$p(X_i | pa_G(X_i)) = \frac{p(X_i, pa_G(X_i))}{p(pa_G(X_i))} = \frac{\sum_{X \setminus \{X_i, pa_G(X_i)\}} \prod_i q(X_i | pa_G(X_i))}{\sum_{X \setminus pa_G(X_i)} \prod_i q(X_i | pa_G(X_i))} = q(X_i | pa_G(X_i))$$

Bayesian Networks: Separation

- As expected, X_i is independent in p of its non-descendants given its parents in G, i.e. $X_i \perp_p nde_G(X_i) \setminus pa_G(X_i) | pa_G(X_i)$. These independencies are called the **causal list** of G.
- Actually, p has many other independencies.
- Since p is a probability distribution, it satisfies the semi-graphoid properties:
  - Symmetry $U \perp_p V | Z \Rightarrow V \perp_p U | Z$
  - Decomposition $U \perp_p V \cup W | Z \Rightarrow U \perp_p V | Z$
  - Weak union $U \perp_p V \cup W | Z \Rightarrow U \perp_p V | Z \cup W$
  - Contraction $U \perp_p V | Z \cup W \wedge U \perp_p W | Z \Rightarrow U \perp_p V \cup W | Z$
- Then, p has all the independencies in the semi-graphoid closure of the causal list of G.
- If p is strictly positive, then it satisfies the graphoid properties:
  - Semi-graphoid properties
  - Intersection $U \perp_p V | Z \cup W \wedge U \perp_p W | Z \cup V \Rightarrow U \perp_p V \cup W | Z$
- Then, p has all the independencies in the graphoid closure of the causal list of G.
- **Exercise**. Prove the symmetry and decomposition properties.

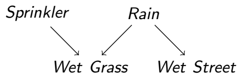
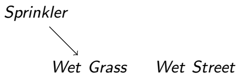
Bayesian Networks: Separation

- Let ρ be a path between two nodes α and β in a DAG G.
- A node B in ρ is a **collider** when $A \rightarrow B \leftarrow C$ is a subpath of ρ.
- 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.
- 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 identifies **all and only** the independencies in the semi-graphoid closure of the causal list of G.
- Then, the separation criterion is **sound**.
- Moreover, it is also **complete**, i.e. $I(p) = \{U \perp_p V | Z\}$ may coincide with $I(G) = \{U \perp_G V | Z\}$.
- Such so-called **faithful** probability distributions exist.

Bayesian Networks: Separation

- **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**. Prove that $A \perp_p B | C, D$ for the DAG $A \to C \to D \to B$.
- **Exercise**. Prove that $A \perp_p B | C, D$ for the DAG $A \to C \to D \leftarrow B$.
- **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**. We have seen that if p factorizes as $p(X) = \prod_i q(X_i | pa_G(X_i))$, then it satisfies all the independencies identified by the separation criterion. **The opposite is also true**. Prove it.

Bayesian Networks: Causal Reasoning

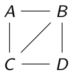
Original	After $do(r_1)$
 $q(S) = (0.3, 0.7)$ $q(R) = (0.5, 0.5)$ $q(WG \mid r_0, s_0) = (0.1, 0.9)$ $q(WG \mid r_0, s_1) = (0.7, 0.3)$ $q(WG \mid r_1, s_0) = (0.8, 0.2)$ $q(WG \mid r_1, s_1) = (0.9, 0.1)$ $q(WS \mid r_0) = (0.1, 0.9)$ $q(WS \mid r_1) = (0.7, 0.3)$ $p(S, R, WG, WS) = q(S)q(R)q(WG \mid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

- 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 the questions above is NP-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

UG	Potentials assuming binary random variables
	$\varphi(A,B,C) = (1,1,1,1,1,1,1,1)$ $\varphi(B,C,D) = (2,2,2,2,2,2,2,2)$ $p(A,B,C,D) = \varphi(A,B,C)\varphi(B,C,D)/Z$ 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.
- 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

- X_i is independent in p of its non-adjacent nodes given its adjacent nodes in G, i.e. $X_i \perp_p X \setminus ad_G(X_i) | ad_G(X_i)$. These independencies are called the **adjacency list** of G.
- Actually, p has many other independencies. Specifically, it has all the independencies in the semi-graphoid closure of the adjacency list. Moreover, if p is strictly positive then it has all the independencies in graphoid closure of the adjacency list of G.
- A path ρ between two nodes α and β in an UG G 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 identifies **all and only** the independencies in the graphoid closure of the adjacency list of G.
- Then, the separation criterion is **sound**.
- Moreover, it is also **complete**, i.e. $I(p) = \{U \perp_p V | Z\}$ may coincide with $I(G) = \{U \perp_G V | Z\}$.
- Such so-called **faithful** probability distributions exist.

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.
- We have seen that if p factorizes as $p(X) = \frac{1}{Z} \prod_{K \in Cl(G)} \varphi(K)$, then it satisfies all the independencies identified by the separation criterion. **The opposite is also true if p is strictly positive**.
- Specifically, $p(X) = \prod_{K \in Cs(G)} \psi(K)$ where
  - $Cs(G)$ are the complete sets of nodes of G
  - $\psi(K) = \exp \phi(K)$
  - $\phi(k) = \sum_{B \subseteq K} (-1)^{|K \setminus B|} H(b)$
  - $H(b) = \log p(b, \overline{b}^*)$
  - x^* is an arbitrary but fixed state and $\overline{b}^*$ denote the values of $X \setminus B$ consistent with x^*.
- This result is known as Hammersley-Clifford theorem.

Markov Networks: Factor Graphs

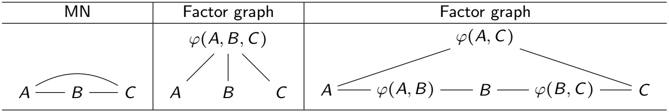
- What if $\varphi(C_i) = \prod_j \phi(C_i^j)$ with $C_i^j \subseteq C_i$?
- A MN may obscure the structure of the potentials. Solution: Factor graphs.
- A factor graph over X consists of an UG G with two types of nodes: The elements in X and a set of potentials $\varphi(K)$ over subsets of X. All the edges in G are between a potential and the elements of X that are in its 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.



Markov Networks: Probabilistic Reasoning

- In the first example above, what is the state of a random variable A if a random variable B is observed to be in the state b ?

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

- Answering questions like the one above is typically NP-hard.
- A MN is an efficient formalism to answer such questions.

Intersection of Bayesian Networks 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 $I(G) = I(H)$ if and only if G has no unshielded colliders.
- Given an UG G, there is an DAG H such that $I(G) = I(H)$ if and only if G is triangulated.