Probabilistic Methods in Artifical Intelligence

Hadar Tal

Hebrew University of Jerusalem, Israel

May 24, 2024

1 Probability Review

Definition 1.1 (Probability Space)
A probability space is a triple $(\Omega, \mathcal{F}, P)$ *where:*

1. Ω *is the sample space*
2. $\mathcal{F}$ *is a* σ*-algebra of subsets of* Ω
3. P *is a probability measure on* $\mathcal{F}$ *such that* $P(\Omega) = 1$

Definition 1.2 (Joint Probability)
The joint probability of two events A *and* B *is:*

$$P(A, B) := P(A \cap B)$$

Definition 1.3 (Random Variable)
A random variable X *is a function* $X : \Omega \to \mathbb{R}$.
$Val(X) = Image(X) = \{x \in \mathbb{R} : \exists \omega \in \Omega \text{ s.t. } X(\omega) = x\}$

Definition 1.4 (Probability Mass Function (PMF))
The probability mass function of a random variable X *is:*

$$P(X = x) := P(\{\omega \in \Omega : X(\omega) = x\})$$

Definition 1.5 (Joint Distribution)
A joint distribution over a set of RVs $\mathcal{X} = \{X_1, X_2, \ldots, X_n\}$ *is a probability distribution* $P_{\mathcal{X}} : Val(X_1) \times Val(X_2) \times \ldots \times Val(X_n) \to [0, 1]$ *defined by:*

$$\forall x_1, \ldots, x_n : x_i \in Val(X_i) \quad P_{\mathcal{X}}(x_1, x_2, \ldots, x_n) := P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$$

Proposition 1.1 (Law of Total Probability)
For X, Y *random variables, we can write:*

$$P(X) = \sum_{y \in Val(Y)} P(X, Y = y)$$

Definition 1.6 (Conditional distribution)
For X, Y *RVs, and for any* $y \in Val(Y)$ *where* $P(Y = y) > 0$ *the conditional distribution of* X *given* $Y{=}y$ *is:*

$$P(X|y) := \frac{P_{X,Y}(X = x, Y = y)}{P_Y(Y = y)}$$

Proposition 1.2 (Chain Rule)
For any set of random variables $X_1, X_2, \ldots, X_n$*:*

$$P(X_1, X_2, \ldots, X_n) = P(X_1)P(X_2|X_1)P(X_3|X_1, X_2)\ldots P(X_n|X_1, X_2, \ldots, X_{n-1})$$

Proposition 1.3 (Bayes' Rule)
For any two random variables H, E*:*

$$P(H = h|E = e) = \frac{P(E = e|H = h)P(H = h)}{P(E = e)}$$

where we often call:

- $P(H = h)$ *the* ***prior*** *probability*
- $P(H = h|E = e)$ *the* ***posterior*** *probability in light of evidence* $E = e$
- $P(E = e|H = h)$ *the* ***likelihood*** *of the evidence* $E = e$ *given the hypothesis* $H = h$

Definition 1.7 (Marginal Independence)
Let P *be a probability distribution over a set of random variables* $\mathcal{X}$ *and let* $X, Y \in \mathcal{X}$*. We say that* X *is independent of* Y*, denoted* $P \models X \perp Y$*, if*

$$P(X|Y) = P(X)$$

Definition 1.8 (Conditional Independence)
Let P *be a probability distribution over a set of random variables* $\mathcal{X}$ *and let* $X, Y, Z \in \mathcal{X}$*. We say that* X *is independent of* Y *given* Z*, denoted* $P \models X \perp Y|Z$*, if*

$$P(X|Y, Z) = P(X|Z)$$

Lemma 1.1 (Equivalent Definitions of Conditional Independence)
Let P *be a probability distribution over a set of random variables* $\mathcal{X}$ *and let* $X, Y, Z \in \mathcal{X}$*. The following are equivalent:*

1. $P \models X \perp Y|Z$
2. $P(X, Y|Z) = P(X|Z)P(Y|Z)$
3. $P(X, Y, Z) = P(X|Z)P(Y, Z)$
4. $\exists f, g : P(X, Y, Z) = f(X, Z)g(Y, Z)$

Theorem 1.1 (Properties of Conditional Independence)
Let P *be a probability distribution over a set of random variables* $\mathcal{X}$ *and let* $X, Y, Z, W \in \mathcal{X}$*. The following hold:*

1. ***Symmetry*** - $(X \perp Y|Z) \implies (Y \perp X|Z)$
2. ***Decomposition*** - $(X \perp Y, W|Z) \implies (X \perp Y|Z) \wedge (X \perp W|Z)$
3. ***Weak Union*** - $(X \perp Y, W|Z) \implies (X \perp Y|W, Z)$
4. ***Contraction*** - $(X \perp Y|Z) \wedge (X \perp W|Y, Z) \implies (X \perp Y, W|Z)$
5. ***Intersection*** - *For strictly positive distributions,*

$$(X \perp Y|W, Z) \wedge (X \perp W|Y, Z) \implies (X \perp Y, W|Z)$$

2 Bayesian Networks

2.1 Bayesian Networks Basics

Definition 2.1 (Probabilistic Graphical Model (PGM))
A probabilistic graphical model is a pair $(\mathcal{G}, P)$ where:

1. *$\mathcal{G}$ is a graph*
2. *P is a probability distribution*

Definition 2.2 (Bayesian Network)
A Bayesian Network $\mathcal{B}$ is:

1. ***Bayesian Network Structure*** *- A directed acyclic graph (DAG) $\mathcal{G} = (\mathcal{X}, E)$ $(|\mathcal{X}| = n)$*
2. ***Set of CPDs*** *- $\{P_i(X_i|Pa(X_i))\}_{i=1}^n$*

the network defines a probability distribution:

$$P_{\mathcal{B}}(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P_i(X_i|Pa(X_i))$$

A Bayesian Network is the tuple $\mathcal{B} = (\mathcal{G}, P_{\mathcal{B}})$.

Theorem 2.1 (Bayesian Network defines a probability distribution)
For any Bayesian Network B, $P_B(X_1, X_2, \ldots, X_n)$ is a joint probability distribution over the variables $X_1, X_2, \ldots, X_n$.

Definition 2.3 (Descendants of a node)
Let $G = (V, E)$ be a directed graph and let $X_i \in V$. The descendants of X_i are:

$$D(X_i) = \{X_j \in \mathcal{X} : \exists \text{ directed path } X_i \to \cdots \to X_j\}$$

Definition 2.4 (Naive Bayes Model)
A Naive Bayes Model is a Bayesian Network where all the features are non adjacent children of the class node.

Definition 2.5 (Naive Bayes Classifier)
A Naive Bayes Classifier is a classifier that uses the Naive Bayes Model to classify instances.

$$\hat{c} = \underset{c \in C}{\operatorname{argmax}}\, P(c|x_1, x_2, \ldots, x_n) = \underset{c \in C}{\operatorname{argmax}}\, P(c, x_1, x_2, \ldots, x_n) = \underset{c \in C}{\operatorname{argmax}}\, P(c) \prod_{i=1}^{n} P(x_i|c)$$

2.2 Independencies and Factorization in Bayesian Networks

Definition 2.6 ($I_{LM}(\mathcal{G})$)
*The **Local Markov Independencies Set** of a Bayesian Network B is the set of all independencies that hold in the network:*

$$I_{LM}(\mathcal{G}) = \{(X_i \perp ND(X_i)|Pa(X_i))\}_{i=1}^{|\mathcal{X}|}$$

Definition 2.7 (I(P))
The set of independencies that hold in a distribution P over $\mathcal{X}$ is:

$$I(P) = \{(X \perp Y|Z) : (X, Y, Z) \subseteq \mathcal{X}, \quad P \models (X \perp Y|Z)\}$$

Definition 2.8 (I-map)
A DAG $\mathcal{G}$ is an I-map of a distribution P if all independencies assumptions of $\mathcal{G}$ hold in P:

$$I_{LM}(\mathcal{G}) \subseteq I(P)$$

Theorem 2.2 (Factorization)
If G is an I-map of P, then we can write:

$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | Pa(X_i))$$

Definition 2.9 (Factorization)
We say that P factorizes over $\mathcal{G}$ if there exist CPDs $\{P_i\}_{i=1}^n$ such that:

$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P_i(X_i | Pa(X_i))$$

Corollary 2.1 (Independencies implies Factorization)
If $\mathcal{G}$ is an I-map of P ($P \models I_{LM}(\mathcal{G})$), then P factorizes over $\mathcal{G}$.

Corollary 2.2 (Independencies implies Factorization (2))
If $\mathcal{G}$ is an I-map of P ($P \models I_{LM}(\mathcal{G})$), then $(\mathcal{G}, P)$ is a Bayesian Network.

Theorem 2.3 (Independencies in P_B)
For $P_{\mathcal{B}}$ it holds for all i that

1. $X_i \perp ND(X_i) | Pa(X_i) \quad (I_{LM}(\mathcal{G}))$
2. $P_{\mathcal{B}}(X_i | ND(X_i)) = P_i(X_i | Pa(X_i))$

Corollary 2.3 (Factorization implies Independencies)
If P factorizes over $\mathcal{G}$, then $\mathcal{G}$ is an I-map of P ($P \models I_{LM}(\mathcal{G})$).

Theorem 2.4 (Fundmental Theorem of Bayesian Networks)
Let $\mathcal{G}$ be a BN structure over $\mathcal{X} = X_1, X_2, \ldots, X_n$ and let P be a joint distribution over $\mathcal{X}$. Then $\mathcal{G}$ is an I-map of $P \Leftrightarrow P$ factorizes over $\mathcal{G}$.

Definition 2.10 (Minimal I-map)
A DAG $\mathcal{G}$ is a minimal I-map of a distribution P if

1. *$\mathcal{G}$ is an I-map of P*
2. *If $\mathcal{G}' \subset \mathcal{G}$ then $\mathcal{G}'$ is not an I-map of P*

2.3 Reasoning Patterns in Bayesian Networks

Definition 2.11 (Reasoning Patterns in Bayesian Networks)
There are 4 main reasoning patterns in Bayesian Networks:

- ***Downstream (causal) reasoning*** - $X \rightarrow Z \rightarrow Y$
- ***Upstream (evidential) reasoning*** - $X \leftarrow Z \leftarrow Y$
- ***Common Causal reasoning*** - $X \leftarrow Z \rightarrow Y$
- ***Common Effect reasoning*** - $X \rightarrow Z \leftarrow Y$

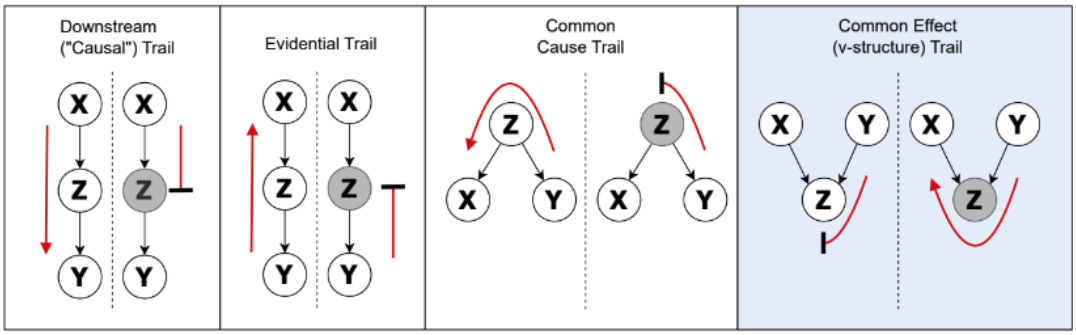

Figure 1: Reasoning Patterns in Bayesian Networks

2.4 D-separation

Question 2.1
If P factorizes over $\mathcal{G}$, then $\mathcal{G}$ is an I-map of P ($P \models I_{LM}(\mathcal{G})$).
Can p satisfy more independencies than those implied by $\mathcal{G}$? Yes.
Given $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z} \in \mathcal{X}$, we would like to characterize when does $P \models I_{LM}(\mathcal{G}) \implies P \models X \perp Y|Z$.
Or characterize the complement - Can we find P that factorizes over $\mathcal{G}$ but $P \not\models X \perp Y|Z$?

Definition 2.12 (Active Trail)
A trail $X = X_1 - X_2 - \cdots - X_n$ between X and Y in a BN is active given a set of observed RVs Z, if whenever there is a v-structure along the trail $X_{i-1} \rightarrow X_i \leftarrow X_{i+1}$, then X_i or one of its descendants is in Z, and all other nodes along the trail are not in Z.

Definition 2.13 (d-separation)
The sets $\boldsymbol{X}$ and $\boldsymbol{Y}$ are d-separated given $\boldsymbol{Z}$ in $\mathcal{G}$, denoted $d\text{-}sep_{\mathcal{G}}(\boldsymbol{X}; \boldsymbol{Y}|\boldsymbol{Z})$, if there is no active trail between any node in $\boldsymbol{X}$ and any node in $\boldsymbol{Y}$ given $\boldsymbol{Z}$.

Definition 2.14 (Global Markov Independencies)
The set of global Markov Independencies of a BN structure $\mathcal{G}$ is the set of all independencies that correspond to d-separation:

$$I(\mathcal{G}) := I_{GM}(\mathcal{G}) := \{(X \perp Y|Z) : d\text{-}sep_{\mathcal{G}}(X; Y|Z)\}$$

d-separation characterizes precisely the full set of independencies that a BN structure encodes.

Theorem 2.5 (Soundness)
If a distribution P factorizes over a BN structure $\mathcal{G}$, then $I(\mathcal{G}) \subseteq I(P)$.

Note - the other direction is not true. If a distribution P factorizes over $\mathcal{G}$, then it is not necessarily true that $I(P) \subseteq I(\mathcal{G})$.

Theorem 2.6 (completness)
If $(X \perp Y|Z) \notin I(\mathcal{G})$, then there exists a distribution P that factorizes over $\mathcal{G}$ in which $P \not\models (X \perp Y|Z)$.

3 Bayesian Networks