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:*

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

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

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

2 Bayesian Networks

Theorem 2.1 (Chain Rule)
For any set of random variables $X_1, X_2, \ldots, X_n$:

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

Definition 2.1 (Conditional probability distribution)
The conditional probability distribution of a random variable X given another random variable Y is:

$$P(X_i|Y) = \frac{P(X_i, Y)}{P(Y)} = \frac{P(X_i, Y)}{\sum_i P(X_i, Y)}$$

Definition 2.2 (Bayesian Network)
A Bayesian Network B is:

1. *A directed acyclic graph (DAG) $G = (V, E)$*
2. *A set of conditional probability distributions $P_i(X_i|Pa(X_i))$ for each node X_i in the graph*

the network defines a probability distribution:

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

Theorem 2.2 (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 ($I_{LM}(G)$)
The ***Local Markov Independencies Set*** *of a Bayesian Network B is the set of all independencies that hold in the network:*

$$I_{LM}(G) = \{(X_i \perp ND(X_i)|Pa(X_i)) \quad : i \in |V|\}$$

Definition 2.4 (I-map)
A DAG G is an I-map of a distribution P if all independencies assumptions of G hold in P:

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

Theorem 2.3 (If G is an I-map of P, then P factorizes according to G)
If G is an I-map of P, then

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

Theorem 2.4 (Independencies in P_B)
For P_B it holds for all i that

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

Definition 2.5 (Minimal I-map)
A DAG G is a minimal I-map of a distribution P if

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