Probabilistic Graphical Models Class Notes
Calvin Walker

Lecture 2: Bayesian Networks
Structure

- $G = (V, E)$ is a directed acyclic graph such that:
  - One node $i \in V$ for each random variable X_i
  - $\text{Pa}_{X_i}^G$ denotes the parents of X_i
  - $\text{NonDescendants}_{X_i}$ are variables that are not descendents of X_i
- G encodes the following local independencies

$$I_l(G) = (X_i \perp \text{NonDescendants}_{X_i} | \text{Pa}_{X_i}^G)\ \forall X_i$$

  i.e. X_i is conditionally independent of $\text{NonDescendants}_{X_i}$ given $\text{Pa}_{X_i}^G$
- A distribuition P factorizes according to G if and only if

$$P(X_i, \dots, X_n) = \prod_{i \in V} P(X_i | \text{Pa}_{X_i}^G)$$

- A **Bayesian Network** is a pair $B = (P, G)$ for which
  - P factorizes over G
  - P is a set of conditional probability distributions $P(X_i | \text{Pa}_{X_i}^G)$
- So G provides a compact way to represent conditional independencies that hold under P

Independence Maps

- Let $I(P) = \{(X \perp Y \mid Z)\}$ be the set of independence assertions that hold in P
- A BN structure G is an I-map for a set of independencies I if $I(G) \subseteq I$
- A BN structure G is an I-map fpr P is G is an I-map for $I(P)$, i.e. $I(G) \subseteq I(P)$
  - Any independence asserted by G must hold in P, but the converse is not necessarily true. P may have additional independencies not reflected in G
  - So while any conditional independency expressed by G holds, the conditional dependencies expressed by G hold for some distributions that factorize over G

Representation Theorem: Given a BN structure G and joint distribution P, P factorizes G if and only if G is an I-map for P
Proof ($P \leftarrow Q$): Let T be a topological ordering on the nodes in G, and v_i be the set of nodes appearing before i in T, excluding $\text{Pa}_{X_i}^G$. From $I_l(G)$ we have that $\{X_i \perp X_{v_i} \mid \text{Pa}_{X_i}^G\}$. Since $I(G) \subseteq I(P)$,

$$P(X_1, \dots, X_n) = \prod_{i \in T} P(X_i | X_{v_i}, \text{Pa}_{X_i}^G) = \prod_{i \in T} P(X_i | \text{Pa}_{X_i}^G)$$

Active Trial: Let G be a BN structure $X_1 \leftrightarrow \cdots \leftrightarrow X_n$ be a trail in G, and Z be a subset of observed variables. The trail is active, i.e. dependency/information flow given Z if

- For every v-structure, X_i or one of its descendents is in Z
- No other node along the trail is in Z

D-seperation: let X, Y, Z be three sets of nodes in G

- X and Y are d-separated given Z if there is no active trail between any node in X to any node in Y given Z
- I.e if $\text{d-sep}_G(X, Y \mid Z)$, then $(X \perp Y \mid Z)$

For a BN structure G, we define the global Markov independencies as the set of independencies that correspond to d-separation:

$$I(G) = \{(X \perp X \mid Z : \text{d-sep}_G(X, Y|Z))\}$$

Lecture 4: Factor Graphs, Gaussian Networks

- The Markov network H does not make the structure of the distribuition explicit, i.e. maximum cliques vs. other complete graph subsets.
- A **factor graph** is a bipartite undirected graph with variable nodes (oval) and factor nodes (square). Edges exist only between variable nodes and factor nodes
- Each factor node is associated with a single potential, the scope of which is the variables that are the factor's neighbors
- Boltzmann Distribution:
  - We can rewrite a factor $\phi(D)$ as $\phi(D) = \exp(-\psi(D))$ where $\psi(D) = -\log \phi(D)$
  - The factorized distribution then becomes:

$$P(X_1, \ldots, X_n) = \frac{1}{Z} \exp\left(-\sum_{k=1}^{K} \psi_k(D_k)\right)$$

  - $\sum_{k=1}^{K} \psi_k(D_k)$ is referred to as the "free energy"
  - Can do inference as energy minimization
- Log-Linear Markov Networks with Features:
  - A feature is a function $f : \text{Val}(D_i) \mapsto \mathbb{R}$
  - A set of features $F = \{f_1(D_1) \ldots f_K(D_M)\}$ where D_i is a complete subgraph in H
  - A set of weights $\{w_1, \ldots, w_M\}$ such that

$$\propto \exp\left(-\sum_{i=1}^{M} w_i f_i(D_i)\right)$$

  - Features and weights can be reused for different factors
  - Clasically, features we hand-designed and weights learned from data
- Gaussian Markov Random Fields:
  - Consider a multivariate Gaussian density p over $x = [x_1, \ldots, x_n]^T$
  - The density function is defined as:

$$p(x) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(\frac{-1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$$

    Where the term in the exponential can be expressed as:

$$\begin{aligned}\frac{-1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu) &= \frac{-1}{2}(x-\mu)^T \Lambda(x-\mu) \\ &= \frac{-1}{2}(x^T \Lambda x - 2x^T \Lambda \mu + \mu^T \Lambda \mu)\end{aligned}$$

    This is referred to as the cononical form where $\Lambda = \Sigma^{-1}$ is the information matrix and $\eta = \Lambda\mu$ is the information vector
  - The information for parametrization $x \sim \mathcal{N}^{-1}(\eta, \Lambda)$ can also be expressed as

$$\begin{aligned}p(x) &\propto \exp\Big(-\frac{1}{2}\sum_i \Lambda_{ii} x_i^2 + 2\eta_i x_i\Big) \exp\Big(-\sum_{i,j:i\neq j} \Lambda_{ij} x_i x_j\Big) \\ &\vdots \\ &= \prod_i \phi_i(x_i) \cdot \prod_{i,j:i\neq j} \phi_{ij}(x_i, x_j)\end{aligned}$$

- Any Gaussian distribuition can be represented by a pairwise Markov network with quadradic node and edge potentials
- Two nodes x_i and x_j have an edge in the GMRF only if $\Lambda_{ij} \neq 0$
- The structure of the information matrix Λ directly encodes the Markov network graph structure

- Converting Bayesian Networks to Markov Networks
  - Moralization coverts a BN to a Markov network
  - The moral graph $\mathcal{M}(G)$ of a BN structure is an untirected graph over V that contains an edge between X_i and X_j if:
    * there is a direct edge between them
    * they are parents of the same node