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
  - $\text{Pa}^G_{X_i}$ 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

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

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

  $$P(X_i, \ldots, X_n) = \prod_{i \in V} P(X_i | \text{Pa}^G_{X_i})$$

- A **Bayesian Network** is a pair $B = (P, G)$ for which
  - P factorizes over G
  - P is a set of conditional probability distributions $P(X_i | \text{Pa}^G_{X_i})$
- 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 $\text{Pa}^G_{X_i}$. From $I_l(G)$ we have that $\{X_i \perp X_{v_i} \mid \text{Pa}^G_{X_i}\}$. Since $I(G) \subseteq I(P)$,

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

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:

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

  - $\sum_{k=1}^{K} \psi_k(D_k)$ 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

$$P(X_1, \ldots, X_n) \propto \exp\Bigg(- \sum_{i=1}^{M} w_i f_i(D_i)\Bigg)$$

  - 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

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

  - 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 undirected 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
  - Introduce one potential for each CPD $\phi_i(X_i, \mathrm{Pa}_{X_i}^G) = P(X_i|\mathrm{Pa}_{X_i}^G)$
  - $\mathcal{M}(G)$ is a minimal I-map for G. If G is moral, then $\mathcal{M}(G)$ is a perfect I-map for G

- Converting a Markov network H to a Bayesian network G is harder, involves adding many edges
  - An undirected graph is **chordal** if every cycle of length 3 has a shortcut between non-consecutive nodes
  - If G is a minimal I-map for H, then G must be chordal
  - Generating a BN for a Markov network involves triangulating the graph by adding edges to make the graph chordal
  - Triangulation results in a loss of independence relations present in H

Lecture 5: Conditional Random Fields

- A **generative** model requires representing the joint distribuition $P(X,Y) = P(X|Y)P(Y)$, since we can generate X given label Y

- A **discriminative** model only requires a representation of the conditional distribuition $P(Y|X)$, so we can discriminate between different Y without estimating $P(X)$

- Ex. Naive Bayes ($X_i \perp X_{-i}|Y$) vs. Logistic Regression: $P(Y=1|x;w) = \frac{1}{1+e^{-z}}$
  - Every conditional distribution that can be represented via Naive Bayes can also be represented using the logistic model
  - Ignoring dependencies might double-count evidence, i.e. spam classification and two words that always appear together (but are assumed independent)

- Tradeoffs between Generative vs. Discriminative Models:
  - Missing data: Generative allow marginalization over unseen variables, e.g. $X = \{X_O, X_U\}$,
    $P(Y|X_O) = \frac{\left(\sum_{X_U} P(X_O,X_U|Y)\right)P(Y)}{\sum_Y \sum_{X_U} P(X_O,X_U|Y)P(Y)}$ Discriminative models typically require all X be observed
  - Unlabeled data: Relativley easy with generative models, but difficult with discriminative models.
  - Adding new classes: Generative models train class-conditioned distributions separatley. Discriminative models have interactions between parameters
  - Calibrated Probabilities: Discriminative models typically yield more accurate probablities. Generative models can be overconfident due to independence assumptions
  - Data-dependent models: Discriminative models allow us to vary the model according to the data. Generative models employ data-independent parameterizations

- MLE of generative models is more efficient than training discriminative models, but may have higher asymptotic error.

Conditional Random Fields:

- Undirected graph with nodes for Y and X (alt. partially directed, with X the parent of Y)

- Parametrized by a set of factors $\phi_1(D_1), \ldots, \phi_m(D_m)$

- Represent conditional distribution $P(Y|X)$ rather than the joint

- Avoid representing dist. over X, so no potientials involving only X

$$P(Y|X) = \frac{1}{Z(X)} \prod_{i=1}^{m} \phi_i(D_i)$$
$$Z(X) = \sum_{Y} \prod_{i=1}^{m} \phi_i(D_i)$$

- Just like Markov network, except the partition function depends on the observed variables X
- The conditional distribution factiorizes as:

$$P(Y|X) = \frac{1}{Z(X)} \prod_{i=1}^{k-1} \phi_i(Y_i, Y_{i+1}) \prod_{i=1}^{k} \phi(Y_i, X_i)$$
$$Z(X) = \sum_{Y} \prod_{i=1}^{k-1} \phi_i(Y_i, Y_{i+1}) \prod_{i=1}^{k} \phi(Y_i, X_i)$$

- Hidden Markov Model: widely used to model sequential random variables

$$P(X, Y) = \prod_{t=1}^{T} P(Y_t|Y_{t-1}) P(X_t|Y_t)$$

Which requires specifying the generative model. Instead, construct discriminative version by reversing direction of arrows, ex. Maximum Entropy Markov Model:

$$P(Y|X) = \prod_{t=1}^{T} P(Y_t|Y_{t-1}, X_1, X_2, \ldots X_T, X_g)$$

Where X_g are global features. Suffers from label bias problem: observations at time t do not influence states prior to t per DAG structure. The Chain-Structured CRF version, uses undirected edges, which yields the model:

$$P(Y|X) = \frac{1}{Z(X)} \prod_{t=1}^{T} \psi(Y_t|X) \prod_{t-1}^{T-1} \phi(Y_t, Y_{t+1}|X)$$

Requires the entire observation

- Naive Markov Model: Assume X and Y are related by the following factors: $\phi_0(Y) = \exp\{w_0 \mathbf{1}[Y = 1]\}$ and $\phi_i(X_i, Y) = \exp\{w_i \mathbf{1}[X_i = 1, Y = 1]\}$. So the conditional distribution becomes:

$$P(Y = 1|x_1, \ldots, x_k) = \sigma\left(w_0 + \sum_{i=1}^{k} w_i x_i\right)$$

- CRF Parametrization: Factors may depend on a large number of variables. Typically, parameterize factors using log-linear representation:

$$\phi_c(X_c, Y_c) = \exp(w_c^T f_c(X_c, Y_c))$$