Probabilistic Graphical Models Notes

Chet Corcos

June 5, 2014

Contents

1 Probability Review

$\mathbf{X}$ is a random variable and x is a possible varile of the random variable. $\mathrm{p}(\mathbf{X})$ is therefor a distribution and $\mathrm{p}(\mathbf{X} = x) = \mathrm{p}(x) \in [0, 1]$ is a probability.

- joint distribution:

$$\mathrm{p}(x, y)$$

- marginal distribution:

$$\mathrm{p}(x) = \sum_{y \in \mathbf{Y}} \mathrm{p}(x, y)$$

- conditional distribution:

$$\mathrm{p}(x|y) = \frac{\mathrm{p}(x, y)}{\mathrm{p}(y)}$$

- Bayes' Theorem:

$$\begin{aligned} \mathrm{p}(x, y) &= \mathrm{p}(x|y)\mathrm{p}(y) \\ &= \mathrm{p}(y|x)\mathrm{p}(x) \\ \mathrm{p}(x|y) &= \frac{\mathrm{p}(y|x)\mathrm{p}(x)}{\mathrm{p}(y)} \end{aligned}$$

- independence: $x \perp y$

$$\begin{aligned} \mathrm{p}(x|y) &= \mathrm{p}(x) \\ \mathrm{p}(x, y) &= \mathrm{p}(x)\mathrm{p}(y) \end{aligned}$$

- conditional independence: $x \perp y|z$

$$\begin{aligned} \mathrm{p}(x|y, z) &= \mathrm{p}(x|z) \\ \mathrm{p}(x, y|z) &= \mathrm{p}(x|z)\mathrm{p}(y|z) \end{aligned}$$

2 Bayesian Network (BN)

A Bayesian Network is a graph defined by verticies and directed edges with no cycles, also known as a directed acyclic graph (DAG).

$$\mathcal{G} = (\mathcal{V}, \mathcal{E})$$

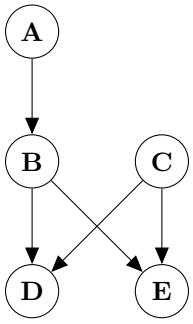
$\mathcal{V}$: a vertex reprsents a random varaible.
$\mathcal{E}$: edge represents a dependence, correlation, causality, evidence between random variables.

The structure of a BN encodes the independencies and conditional independencies.

In general, for any BN, the joint distribution can be decomposed as:

$$\mathrm{p}(\mathbf{X}_1, \mathbf{X}_2, ...\mathbf{X}_N) = \prod_{i=1}^{N} \mathrm{p}(\mathbf{X}_i|\mathrm{parents}(\mathbf{X}_i))$$

For example:



$$\mathrm{p}(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}, \mathbf{E}) = \mathrm{p}(\mathbf{A})\mathrm{p}(\mathbf{B}|\mathbf{A})\mathrm{p}(\mathbf{C})\mathrm{p}(\mathbf{D}|\mathbf{B}, \mathbf{C})\mathrm{p}(\mathbf{E}|\mathbf{B}, \mathbf{C})$$

- $\mathbf{A}$ and $\mathbf{C}$ are marginally independent (independent of all other variables).
- $\mathbf{B}$ is dependent on $\mathbf{A}$
- $\mathbf{B} \perp \mathbf{C}$
- $\mathbf{D} \perp \mathbf{E}|\mathbf{B}, \mathbf{C}$ but $\mathbf{D} \not\perp \mathbf{E}$

For discrete sets of random variables, the edges are represented as conditional probability tables. For example, if there are 2 outcomes for each random variable $\mathbf{A} = \{a_1, a_2\}$ and $\mathbf{B} = \{b_1, b_2\}$, then the relationship for $\mathrm{p}(\mathbf{B}|\mathbf{A})$ is encoded as:

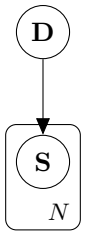
Table 1: Simple CPT Example

	a_1	a_2
b_1	0.5	0.5
b_2	0.25	0.75

Notice that the rows sum to 1 satisfying $\sum_{\mathbf{A}} \mathrm{p}(\mathbf{B}|\mathbf{A}) = 1$

2.1 Naive Bayes

Take an example of diagnosing diseases $\mathbf{D}$ from a set of N symptoms $\{\mathbf{S}_1, \mathbf{S}_2, ...\mathbf{S}_N\}$. To clarify, $\mathbf{D}$ could have multiple outcomes which are different diseases $\mathbf{D} = \{d_1, d_2, ...\}$. The symptoms could be binary $\mathbf{S}_i = \{1, 0\}$ or a probability $\mathbf{S}_i = [0, 1]$.



$$\mathrm{p}(\mathbf{S}_1, \mathbf{S}_2, ...\mathbf{S}_N, \mathbf{D}) = \mathrm{p}(\mathbf{D}) \prod_{i=1}^{N} \mathrm{p}(\mathbf{S}_i|\mathbf{D})$$

We can determine the probability of a disease:

$$\begin{aligned} \mathrm{p}(\mathbf{D}|\mathbf{S}_1, \mathbf{S}_2, ...\mathbf{S}_N) &= \frac{\mathrm{p}(\mathbf{S}_1, \mathbf{S}_2, ...\mathbf{S}_N, \mathbf{D})}{\mathrm{p}(\mathbf{S}_1, \mathbf{S}_2, ...\mathbf{S}_N)} \\ &= \frac{\mathrm{p}(\mathbf{D}) \prod_{i=1}^{N} \mathrm{p}(\mathbf{S}_i|\mathbf{D})}{\sum_{\mathbf{D}} \mathrm{p}(\mathbf{D}) \prod_{i=1}^{N} \mathrm{p}(\mathbf{S}_i|\mathbf{D})} \end{aligned}$$

To determine the most likely disease:

$$\mathbf{D}^* = \underset{\mathbf{D}}{\operatorname{argmax}}\, \mathrm{p}(\mathbf{D}) \prod_{i=1}^{N} \mathrm{p}(\mathbf{S}_i|\mathbf{C})$$

3 BN Semantics

A BN $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ can be defined by the set of all independencies of the graph, $\mathcal{I}(\mathcal{G})$. The set of independencies can by derived from looking at each node locally or looking at every path globally.

3.1 Local Factorization

The set of all local independencies an be determined by the set of all local independencies:

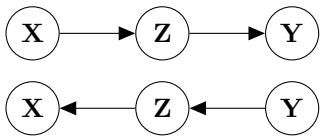
$$\mathcal{I}(\mathcal{G}) = \{\mathcal{V}_i \perp \text{non-decendants}(\mathcal{V}_i)|\text{parents}(\mathcal{V}_i)\}$$

non-decendants($\mathcal{V}_i$) are any nodes that can by reached from following directed edged from $\mathcal{V}_i$.

3.2 D-Separation

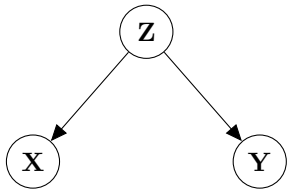
First we need to define active trail. Active trail means that there is a flow of information. Consider whether the trail between $\mathbf{X} \to \mathbf{Y}$ is active or inactive the following cases for local structures:

- Case 1



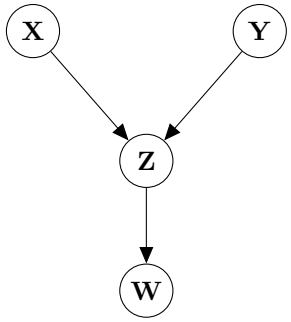
  - active if $\mathbf{Z}$ is unobserved: $\mathbf{X} \not\perp \mathbf{Y}$
  - inactive if $\mathbf{Z}$ is observed: $\mathbf{X} \perp \mathbf{Y} | \mathbf{Z}$

- Case 2



  - active if $\mathbf{Z}$ is unobserved: $\mathbf{X} \not\perp \mathbf{Y}$
  - inactive if $\mathbf{Z}$ is observed: $\mathbf{X} \perp \mathbf{Y} | \mathbf{Z}$

- Case 3



  - active if $\mathbf{Z}$ or any decentant($\mathbf{Z}$) such as $\mathbf{W}$ is observed: $\mathbf{X} \not\perp \mathbf{Y} | \mathbf{Z}$ and $\mathbf{X} \not\perp \mathbf{Y} | \mathbf{W}$
  - inactive if $\mathbf{Z}$ and all decentant($\mathbf{Z}$) such as $\mathbf{W}$ is unobserved: $\mathbf{X} \perp \mathbf{Y}$

We can no define two nodes, $\mathbf{X}$ and $\mathbf{Y}$, as being d-separated if given $\mathbf{Z}$, there is no active undirected path between them: d-sep($\mathbf{X}, \mathbf{Y} | \mathbf{Z}$). And thus, the set of all independencies of a graph can be defined as:

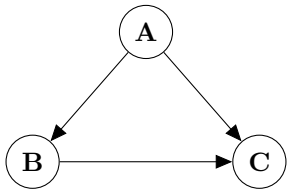
$$\mathcal{I}(\mathcal{G}) = \{\mathbf{X} \perp \mathbf{Y} | \mathbf{Z} : \text{d-sep}(\mathbf{X}, \mathbf{Y} | \mathbf{Z})\}$$

3.3 I-map

Given a joint distribution P with the set of independancies $\mathcal{I}(P)$, we say that a BN graph, $\mathcal{G}$, is I-map of P if $\mathcal{I}(\mathcal{G}) \subseteq \mathcal{I}(P)$.

A graph $\tilde{\mathcal{G}}$ formed by adding edges to $\mathcal{G}$ implies $\mathcal{I}(\tilde{\mathcal{G}}) \subseteq \mathcal{I}(\mathcal{G})$. Thus is $\mathcal{G}$ is I-map of P, also $\tilde{\mathcal{G}}$ is I-map of P.

Note that $\mathcal{I}(\mathcal{G}) = \emptyset$ is a valid I-map, as is the case for a fully connected graph:



Thus $\emptyset$ is I-map of any distribution P. Thus, we define a *minimal I-map* as an I-map such that removing any edge from a graph will result in a graph that is not I-map of the underlying distribution.

A minimal I-map *always* exists. However it is not necessarily unique.

Theorem 3.1: (PGM Book by Koller) states that if $\mathcal{G}$ is and I-map of P, the P is factorizable to the graph.

Theorem 3.2: (PGM Book by Koller) states if P factorizes according to a BN graph $\mathcal{G}$, then $\mathcal{G}$ is an I-map of P.

These two theorems basically just say that it is valid to represent a distribution as a graph or a graph as a distribution. Any distribution can be represented by a minimal I-map graph, however, it is possible that a graph cannot encode all of the independencies of a distribution.

Theorem 3.3 (Soundness): (PGM Book by Koller) states that if P factorized according to $\mathcal{G}$, then $\mathcal{I}(\mathcal{G}) \subseteq \mathcal{I}(P)$.

Theorem 3.4 (Completeness): (PGM Book by Koller) states that for almost all P that factorizes with respect to $\mathcal{G}$, we have $\mathcal{I}(\mathcal{G}) = \mathcal{I}(P)$.

How these two sets of theorems differ? I'm not entirely sure. But it's important to note that not any set of statistical independencies can be represented by a BN. This can be proven with the gossip example.

3.4 Gossip Example

Suppose A gossips with B, B gossips with C, C gossips with D, and D gossips with A. A and C don't like each other, nor to B and D. Thus the set of independencies we would like to capture are $A \perp C|B, D$ and $B \perp D|A, C$. In fact, no BN can represent these two independencies at the same time. Thus we must introduce undirected graphical models, also known as Markov Random Fields (MRF).

4 Markov Random Fields (MRF)

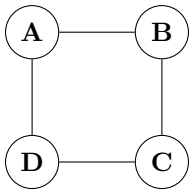
Markov Random Fields are represented by an undirected graph $\mathcal{H} = (\mathcal{V}, \Phi)$ as a set of verticies and factors. A factor represents a compatibility score between two or more verticies. A factor must be positive and can be thought of as an unnormalized probability.

$$\phi(D) \in \mathbb{R}^+ : D \subseteq \mathcal{V}$$

The joint distribution (called a Gibbs distribution) is the product of all factors normalized by a partition function, Z.

$$\begin{aligned} \mathrm{p}(\mathcal{V}) &= \frac{1}{Z}\phi_1(D_1)\phi_2(D_2)...\phi_k(D_k) \\ Z &= \sum_{\mathcal{V}} \phi_1(D_1)\phi_2(D_2)...\phi_k(D_k) \end{aligned}$$

4.1 Gossip Example, continued



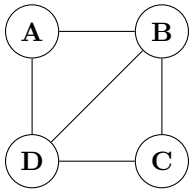
The factors are defined by $D_1 = (A, B), D_2 = (B, C), D_3 = (C, D), D_4 = (D, A)$ such that:

$$\begin{aligned} \mathrm{p}(A, B, C, D) &= \frac{1}{Z}\phi_1(A, B)\phi_2(B, C)\phi_2(C, D)\phi_2(D, A) \\ Z &= \sum_{A,B,C,D} \phi_1(A, B)\phi_2(B, C)\phi_2(C, D)\phi_2(D, A) \end{aligned}$$

4.2 Graph Factorization

A distribution P_Φ with $\Phi = \{\phi_1(D_1)\phi_2(D_2)...\phi_k(D_k)\}$ factorizes over an undirected graph $\mathcal{H}$ if $\forall\ k$, D_k is a complete subgraph (commonly referred to as a clique) of $\mathcal{H}$.

For example, the following distributions all factorize over the graph:



- $P_\Phi = \frac{1}{Z}\phi_1(A)\phi_2(B,C,D)$
- $P_\Phi = \frac{1}{Z}\phi_1(A,D,B)\phi_2(B,C,D)$
- $P_\Phi = \frac{1}{Z}\phi_1(A,B)\phi_2(B,D)\phi_3(C,D)\phi_4(D,A)\phi_5(B,C)$

4.3 Reduced MRF

Suppose we have observed some of the verticies. Given the distribution $P_\Phi(\mathcal{V})$, if we observe $U = u \in \mathcal{V}$ the reduced Gibbs Distribution $P_{\Phi[u]}(\mathcal{V})$ is given by:

$$P_\Phi = P_\Phi(\mathcal{V}) = \frac{1}{Z}\prod_{k=1}^{K}\phi_k(D_k)$$

$$P_{\Phi[u]} = P_\Phi(\mathcal{V}\backslash U|U=u) = \frac{1}{Z}\prod_{k=1}^{K}\phi_k[u](D_k\backslash U)$$

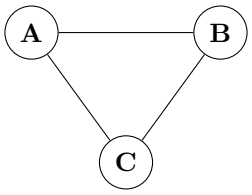
(note that the partition function Z changed).

The reduced graph, $\mathcal{H}[u]$ will have fewer edges than the unreduced graph $\mathcal{H}$ – no edges are added.

The reduced graph theorem states that $P_\Phi(\mathcal{V}\backslash U|U=u)$ factorizes over $\mathcal{H}[u]$

4.4 Factor Graphs

Consider the following undirected graph:

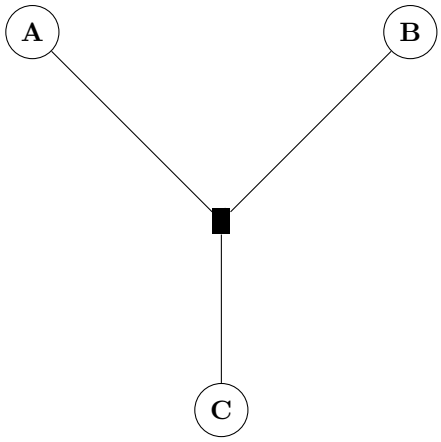


This can be represented by two different Gibbs distributions:

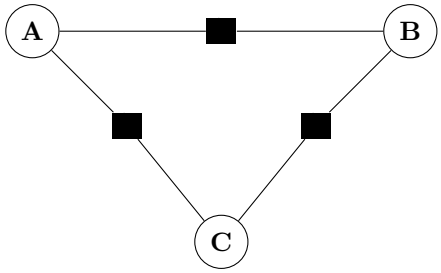
$$P^1_\Phi(A,B,C) = \frac{1}{Z}\phi(A,B,C)$$
$$P^2_\Phi(A,B,C) = \frac{1}{Z}\phi_1(A,B)\phi_2(B,C)\phi_3(C,A)$$

However, these represent two different factor graphs:

- $P^1_\Phi(A,B,C) = \frac{1}{Z}\phi(A,B,C)$:

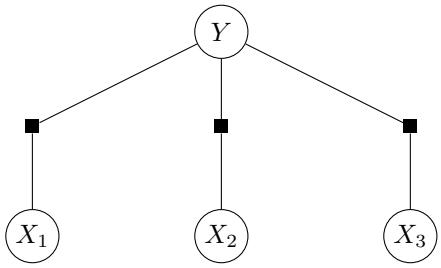


- $P^2_\Phi(A, B, C) = \frac{1}{Z}\phi_1(A, B)\phi_2(B, C)\phi_3(C, A)$:

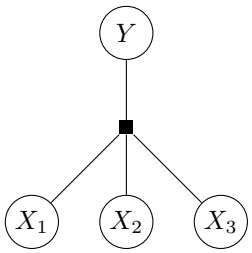


These actually represent different distributions. A more concrete example:

- X_1, X_2, X_3 are independent conditioned on Y.



- X_1, X_2, X_3 are NOT independent conditioned on Y.



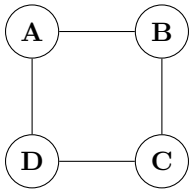
5 MRF Semantics

5.1 Local Factorization

If $X, Y \in \mathcal{V}$ but are not directly connected, then:

$$X \perp Y | \mathcal{V} \setminus \{X, Y\}$$

For example, given:



$$A \perp B | C, D$$

Applying this local factorization to all nodes, we get the Markov Blanket.

5.2 Markov Blanket

A Markov Blanket, $\text{MB}(X)$ is the set of directly connected neightbors of X:

$$X \perp \mathcal{V} \setminus \{X, \text{MB}(X)\} | \text{MB}(X)$$

5.3 Separation

We define two nodes X and Y as being separated with respect to Z so long as Z is an intermediate node for all paths $X \to Y$.

5.4 I-map

Given an undirected graph, $\mathcal{H}$:

$$\mathcal{I}(\mathcal{H}) = \{\mathbf{X} \perp \mathbf{Y} | \mathbf{Z} : \text{sep}(\mathbf{X}, \mathbf{Y} | \mathbf{Z})\}$$

Once again, we have soundness and completeness:

- **Soundness**: If P factorizes with $\mathcal{H}$, then $\mathcal{H}$ is an I-map of P, $\mathcal{I}(\mathcal{H}) \subseteq \mathcal{I}(P)$

- **Hammersley-Clifford Theorem**: If $P > 0 : \forall\, x, \mathrm{p}(x) > 0$ and $\mathcal{H}$ is an I-map of P, then P factorizes over $\mathcal{H}$.
- **Completeness**: If $\not{sep}(X, Y|Z)$ then $X \not\perp Y|Z$ for some distribution P. For almost all P, $\mathcal{I}(P) = \mathcal{I}(\mathcal{H})$

6 Review

Given some distribution, $\mathrm{p}(x_1, x_2, ..., x_N)$, how to we compute inferences, $\mathrm{p}(x_1|x_2)$?

The answer is factorization. To do aid in the process, we can create a graph such as a Bayesian Network (BN), Markov Network (MN), or a factor graph. BN are directed acyclic graphs while MN are undirected and often log-linear models (exponential family distributions) or conditional random fields (CRF).

This all leads to graph semantics about statistical independance. The holy grail of which is d-separation which says everything you need to know about the graph structure.

From here, we have exact inference techniques, approximate inference techniques, and learning.

For exact inference, we learned variable elimination with dynamic programming techniques such as belief propagation on a clique tree. We also learning about tree width which is the computational complexity of solving the graph.

For approximate inference, we learned about sampling such as monte carlo markov chain (MCMC) sampling, most notably particle filtering and gibbs sampling. We leaned about variational method and how it was derived from KL divergence. We also learned about loopy belief propagation.

For learning, we have complete data and incomplete data. Both of which involve maximum likelihood estimation (MLE). Complete data is easy and convex while incomplete data requires the expectation maximization (EM) algorithm and perhaps MCMC EM ;)

A huge topic we did not learn about in this course is structure learning which is much much harder.

There are three main methods - constraint-based, score-based and optimiation/regression-based.

The constraint based methods suck, but they are very interesting in how they consider causality. Some well known methods are PC and K2. The goal is to learn a minimum i-map.

Score-based methods are either bayesian or non-bayesian. Bayesian approiaches high highly regularized with a prior and are widely used in biology learning which genes cause which diseases. For non-bayesian techniques, its just a simple MLE but its "bullshit" because it always overfits and is an NP-HARD problem.

An interesting topic in structure learning is "compressive sensing" for sprise driven models. This basically just learns the adjacency matrix of a MN.