10-708: Probabilistic Graphical Models 10-708, Spring 2018

1 : Directed GMs: Bayesian Networks

Lecturer: Kayhan Batmanghelich *Scribes: Sumedha Singla*

1 Types of Graphical Models

A graphical model, is a way of expressing dependences between random variables. The nodes in the graph corresponds to the random variables and the edges corresponds to the relation or correspondence between the random variables. Depending on the type of edges, there are two types of graphical models

1. Directed Graphical Model (DGM) or Bayesian Network (BN)

   It's a directed acyclic graph (DAG), where the edges represents the conditional dependences (or causality) between the random variables. The model represents a factorization of the joint probability of all random variables.

2. Undirected Graphical Model (UGM) or Markov Network (MN).

   It's an un-directed graph which may have cycles, where edges represents correlation between the random variables.

2 Notations

1. Variable, value, index: Its a placeholder X, that may take a value $X = x$. x denotes a particular value that variable X may take. Index i in X_i, denotes the i-th observation for variable X.
2. Random variable: Expressed in capital letter e.g. X
3. Random vector: Express in bold letter e.g. $\boldsymbol{X}$
4. Random matrix: Express in bold capitalized letter e.g. $\boldsymbol{X} = \{X_{ij}\}$
5. Parameters: E.g. θ, α, β. They may have index.

3 The Dishonest Casino

A fair dice is the one which have equal probability of landing on any face. While in a loaded dice, the probability for each face is no longer same. Consider the casino game, where the person (player or casino person) who rolls highest number wins. We are given the sequence of the rolls by the casino player (X). These are the observations for our current model. The casino player may use a fair or loaded dice to get this sequence. The choice of the dice (Y) can be interpreted as a hidden binary random variable, which affects the roll (X).

- Evaluation

  Given the model of the game, how likely is the sequence X. Finding the $p(X)$.

- Decoding

  Finding the probability of using a fair or loaded dice given the observation sequence i.e $p(Y|X)$

- Learning

  Learning the parameters of the given model. To answer the questions like, how "loaded" is the loaded dice?, we are aiming at find the probability of each face for the loaded dice. To answer the question "How often does the casino player change from fair to loaded, and back?", we aim at learning the relation between different dice rolling events, if they are independent or have some dependency. Example, if the output of the first dice roll, decided whether the casino player chooses (loaded/fair) dice in next roll, then the dice rolling events are not independent.

4 Knowledge Engineering

Constructing a graphical model for a given problem, involves picking the random variables. Depending on the problem, some variables are observed, some are hidden, they can be either continuous or discrete, etc. Next, one have to decide, how the variables interact with each other, e.g define the causuality relation. Usually prior knowledge, domain knowledge can help establishing these relations.

4.1 Hidden Markov Model

Its an example of DGM with hidden and observed variables. In HMM, the hidden state is not directly visible, but the output, dependent on the state, is visible and observed. Each state has a probability distribution over the possible output. Therefore, the sequence of tokens generated by an HMM gives some information about the sequence of states.

For example, in parts of speech tagging, the sequence of words is given or observed and the goal is to mark each word with the most probable part-of-speech.

4.2 Understanding the graph

Given a DGM, we can define the join distribution over the random variables, by factorizing over the conditional dependences suggested by the graph. In the absence of the dependency information, we can write joint distribution using the Bayes theorem. This is the generalized form, which makes no assumption about the dependencies and hence, consider all the possible values for all the random variables. If we have n binary random variables, then the joint distribution table will have 2^n rows which is exponential in terms of n. To find marginal probability, we have to integrate out rest of the variables from the joint distribution. In the absence of dependency information, it also requires exponential time to find marginal probability for any random variable in the graph.

5 Bayesian Network

In BN, we can start generating sample from the nodes, which don't have parent and can follow the directed edges to subsequently generate sample from child given the parent. By following the chain of conditional dependence, we can generate samples from the joint distribution given by the graph.

Given a BN, we can find the join distribution over all the random variables, by using the **factorization theorem**. BN don't provide the specific parametrization of individual probability distributions i.e it doesn't specify the parameters of the probability distribution p in $p(X|Y)$.

To find the dependency information between different random variables in a given graph, we can follow a **qualitative specification** which depends on domain knowledge, expert knowledge (e,g gene A causes symptom B) or **quantitative specification** in which we evaluate new dependency information from the given condition probability tables (CPT). In CPT, either rows or columns sums to 1, while in joint probability table, the entire table sums to 1.

Graphical models have three fundamental local structures that composes bigger graph structures.

- Common parent

  Its a diverging connection where two random variables (A, C) have a common parent B. Knowing the parent B, decouples the two random variables (A, C). It holds the conditional independence $A \perp C|B$.

- Cascade

  Its a series connection from one random variable (A) to another (C) via a third variable (B). Knowing B blocks the flow of information/evidence between A and C, hence knowing B decouples the two random variables (A, C). It holds the conditional independence $A \perp C|B$.

- V-structure

  Its a converging connection, where two random variables (A, B) have a common child C. Knowing the common child C, allows the flow of information between previously independent parents A and B. In this case, A and B are marginally independent i.e $A \perp B|\emptyset$ but the independence is lost given the common child C i.e. $A \not\perp B|C$

6 I-Map

A BN have two components a graph structure ($\mathcal{G}$) and a probability distribution (P). $\mathcal{G}$ encodes a set of independencies $I(P)$ like ($A \perp B|C$) which holds for every P that can be represented by the graph $\mathcal{G}$. Hence, we can say if P factorizes over $\mathcal{G}$ then $\mathcal{G}$ is an I-map for P.

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

Any independence that $\mathcal{G}$ asserts must also hold in P. Conversely, P may have additional independencies that are not reflected in $\mathcal{G}$.

Referring to the **slide: 19**, The I-map for the given graphs and probability distributions are as follow:

$$I(\mathcal{G}_{\emptyset}) = \{(X \perp Y|\emptyset), \emptyset\}$$

$$I(\mathcal{G}_{X\to Y}) = \{\emptyset\}$$

$$I(\mathcal{G}_{Y\to X}) = \{\emptyset\}$$

$$I(P_1) = \{(X \perp Y|\emptyset), \emptyset\}$$

$$I(P_2) = \{\emptyset\}$$

Hence we can say

$$I(\mathcal{G}_{\emptyset}) \subseteq I(P_1)$$

$$I(\mathcal{G}_{X\to Y}) \subseteq I(P_1), I(P_2)$$
$$I(\mathcal{G}_{Y\to X}) \subseteq I(P_1), I(P_2)$$

We can choose a parametrization for our probability distribution (P) such that it enforces additional independencies which were not enforce by the underlying graph structure ($\mathcal{G}$) as in case of P_1 for graph $\mathcal{G}_{X\to Y}$ and $\mathcal{G}_{Y\to X}$.

Given a BN, its easier to write $I(\mathcal{G})$ based on the structure of the graphical model. We can write $I(\mathcal{G})$ as a set of local conditional independencies.

7 d-connection and d-separation

The "d" in d-separation and d-connection stands for dependence. Consider $\mathcal{X}, \mathcal{Y}$ and $\mathcal{Z}$ as 3 disjoint sets of random variables in a graph $\mathcal{G}$. A path is active if it carries information, or dependence. $\mathcal{X}$ and $\mathcal{Y}$ are d-connected, if there is any active path between them that passes through $\mathcal{Z}$. A variable is a collider when it is causally influenced by two or more variables. A collider forms an active path between the influencing variables. Here, the influencing variables are in $\mathcal{X}$ and $\mathcal{Y}$ and the collider variable are in set $\mathcal{Z}$.

D-separation helps determining which variables are conditionally independent given other variables in a BN. One way to determine d-separation is through **Moralized Ancestral Graph**. It follows the below process

1. Draw the ancestral graph

   This is a reduced version of the original BN, consisting only of the variables which are parents of some other variables.

2. Moralize the ancestral graph by marrying the parents.

   For each pair of variables with a common child (V-structure), draw an undirected edge between them. (If a variable has more than two parents, draw undirected edges between every pair of parents.). Here the common child is the collider.

3. Disorient the graph by replacing the directed edges with undirected edges.

4. Delete the given variables or the variables in set $\mathcal{Z}$ and their edges.

5. If the variables are disconnected in this graph, they are guaranteed to be independent.

Another way for testing d-separation is using **Bayes Ball Algorithm**.

For the graph shown on **slide: 25** the I-map is

$$I(\mathcal{G}) = \{X_2 \perp \{X_1, X4\}, X_2 \perp X_4|X_1, X_2 \perp X_4|\{X_1, X_3\}, X_2 \perp X_1|X_4, X_3 \perp X_4|X_1, X_3 \perp X_4|\{X_1, X_2\}\}$$

8 Equivalence

It's difficult to write $\mathcal{D}_1$, the family of all distributions that satisfies $I(\mathcal{G})$. For that we have to enumerate over all the possible variations of the variables in graph $\mathcal{G}$. Its easier to represent a distribution $\mathcal{D}_2$ that factorizes over the graph $\mathcal{G}$. As per **Equivalence Theorem**, $\mathcal{D}_1 \equiv \mathcal{D}_2$.

Two graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ are I-equivalent if $I(\mathcal{G}_1) = I(\mathcal{G}_2)$. The skeleton of a BN graph $\mathcal{G}$ is an undirected graph that contains an undirected edge for every directed edge in original graph $\mathcal{G}$. If two graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ have the same skeleton and the same set of v-structures then they are I-equivalent.

9 Plate Notation

Plate notation is a method of representing variables that repeat in a graphical model. Instead of drawing each repeated variable individually, a plate or rectangle is used to group variables into a subgraph that repeat together, and a number is drawn on the plate to represent the number of repetitions of the subgraph in the plate. The variables within a plate are replicated in a conditionally independent manner.

In Hidden Markov Model (HMM), each observed variable X_i is conditionally dependent on the hidden state Y_i. We can represent this in a plate model, by putting relation $X \rightarrow Y$ in a plate with N on side, denoting the relation is repeated N times. In dynamic mixture model, parameter θ defines the proportion of observation X in kth mixture. Since θ enumerates over number of mixtures k, its represented in a plate with k on side.