Lecture Notes for CS 726 - Spring 2021

Eeshaan Jain

January 11, 2022

These are my lecture notes taken during the Advanced Machine Learning (CS 726) course at IIT Bombay during the Spring 2021 session.

Contents

Probabilistic Modeling

Probability Theory

We will briefly review probability in a rigorous sense.

We define events considering we have a space of possible outcomes denoted by Ω. $\mathcal{S}$ is a set of measurable events, to which we assign probabilities, and each event $\alpha \in \mathcal{S}$ is a subset of Ω.

In a single coin toss, we have

$$\Omega = \{H, T\}$$

The event space necessarily satisfies three properties -

1. It contains the empty event $\emptyset$ and the trivial event Ω.
2. It is closed under union, i.e if $\alpha, \beta \in \mathcal{S}$, so is $\alpha \cup \beta$.
3. It is closed under complementation, i.e if $\alpha \in \mathcal{S}$, so is $\Omega - \alpha$.

Definition 1 (Probability distribution)**.** A probability distribution P over $(\Omega, \mathcal{S})$ is a mapping of events in $\mathcal{S}$ to real values satisfying

- $P(\alpha) \geq 0$ for all $\alpha \in \mathcal{S}$
- $P(\Omega) = 1$
- If $\alpha, \beta \in \mathcal{S}$ and $\alpha \cap \beta = \emptyset$, then $P(\alpha \cup \beta) = P(\alpha) + P(\beta)$

A direct consequence of the properties is:

$$P(\emptyset) = 0$$
$$P(\alpha \cup \beta) = P(\alpha) + P(\beta) - P(\alpha \cap \beta)$$

Conditional probability answers the question - after learning that event α is true, how does our belief about β change? Formally, we define

$$P(\beta|\alpha) = \frac{P(\alpha \cap \beta)}{P(\alpha)} \tag{1}$$

It can be checked that this satisfies Definition 1 and is a probability distribution. Noting that $P(\alpha \cap \beta) = P(\alpha)P(\beta|\alpha)$, we define the chain rule of conditional probabilities

$$P(\alpha_1 \cap \cdots \cap \alpha_k) = P(\alpha_1)P(\alpha_2|\alpha_1) \cdots P(\alpha_k|\alpha_1 \cap \cdots \cap \alpha_{k-1}) \tag{2}$$

We further define the Bayes' rule

$$P(\alpha|\beta) = \frac{P(\beta|\alpha)P(\alpha)}{P(\beta)} \tag{3}$$

Here, $P(\alpha|\beta)$ is called the *posterior*, $P(\beta|\alpha)$ is the *likelihood*, $P(\alpha)$ is the *prior* and $P(\beta)$ is the *marginal probability* of the structure in context. We can generalize Equation 3 as

$$P(\alpha|\beta \cap \gamma) = \frac{P(\beta|\alpha \cap \gamma)P(\alpha|\gamma)}{P(\beta|\gamma)} \tag{4}$$

Now, we formally define the notion of random variables, which intuitively can be considered to be attribute reporters.

Definition 2 (Random Variable)**.** A random variable X is a *measurable* function $X : \Omega \to \mathcal{S}$. The probability that X takes values in a set $s \in \mathcal{S}$ is written as

$$\Pr(X \in s) = \Pr(\{\omega \in \Omega | X(\omega) \in s\}) \tag{5}$$

The marginal distribution over a random variable X is the distribution over events that can be described using X, and is denoted by $P(X)$. More generally, if we want to describe a distribution over a set of random variables $\mathcal{X} = \{x_1, \cdots, x_n\}$ called the *joint distribution* denoted as $P(x_1, \cdots, x_n)$. The full assignment to the variables is denoted as $\xi \in \mathrm{Val}(\mathcal{X})$. The space corresponding to the joint assignment in $\mathcal{X}$ is called the *canonical outcome space*.

Now, we glance at independencies, a core component of Probabilistic Graphical Models.

Definition 3 (Independence)**.** An event α is independent of an event β denoted by $P \models (\alpha \perp\!\!\!\perp \beta)$, if $P(\alpha|\beta) = P(\alpha)$ or $P(\beta) = 0$.

Proposition 4. *A distribution satisfies* $(\alpha \perp\!\!\!\perp \beta)$ *if and only if*

$$P(\alpha \cap \beta) = P(\alpha)P(\beta) \tag{6}$$

Proof. Skipped (hint: Use the definition of conditional probability). □

Definition 5 (Conditional Indpendence)**.** An event α is conditionally independent of event β given γ in P, denoted by $P \models (\alpha \perp\!\!\!\perp \beta | \gamma)$ if $P(\alpha|\beta \cap \gamma) = P(\alpha|\gamma)$ or if $P(\beta \cap \gamma) = 0$.

Proposition 6. *P satisfies* $(\alpha \perp\!\!\!\perp \beta|\gamma)$ *if and only if*

$$P(\alpha \cap \beta|\gamma) = P(\alpha|\gamma)P(\beta|\gamma) \tag{7}$$

Definition 7. Let $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ be sets of random variables. $\mathbf{X}$ is conditionally independent of $\mathbf{Y}$ given $\mathbf{Z}$ in a distribution P if P satisfies $(\mathbf{X} = \mathbf{x} \perp\!\!\!\perp \mathbf{Y} = \mathbf{y} | \mathbf{Z} = \mathbf{z})$ for all values of $\mathbf{x} \in \mathrm{Val}(\mathbf{X}), \mathbf{y} \in \mathrm{Val}(\mathbf{Y})$ and $\mathbf{z} \in \mathrm{Val}(\mathbf{Z})$. We say that the variables in $\mathbf{Z}$ are *observed*. If $\mathbf{Z}$ is empty, then we say that $\mathbf{X}$ and $\mathbf{Y}$ are marginally independent.

Proposition 8. *The distribution P satisfies* $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z})$ *if and only if*

$$P(\mathbf{X}, \mathbf{Y}|\mathbf{Z}) = P(\mathbf{X}|\mathbf{Z})P(\mathbf{Y}|\mathbf{Z}) \tag{8}$$

The following properties hold for conditional independencies:

1. *Symmetry:* $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z}) \implies (\mathbf{Y} \perp\!\!\!\perp \mathbf{X}|\mathbf{Z})$

2. *Decomposition:* $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y}, \mathbf{W}|\mathbf{Z}) \implies (\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z})$

3. *Weak Union:* $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y}, \mathbf{W}|\mathbf{Z}) \implies (\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z}, \mathbf{W})$

4. *Contraction:* $(\mathbf{X} \perp\!\!\!\perp \mathbf{W} | \mathbf{Z}, \mathbf{Y})$ & $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} | \mathbf{Z}) \implies (\mathbf{X} \perp\!\!\!\perp \mathbf{Y}, \mathbf{W} | \mathbf{Z})$

If our distribution is positive (i.e, for all non-empty $\alpha \in \mathcal{S}, P(\alpha) > 0$), we have another property

- ⋄ *Intersection:* $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} | \mathbf{Z}, \mathbf{Z})$ & $(\mathbf{X} \perp\!\!\!\perp \mathbf{W} | \mathbf{Z}, \mathbf{Y}) \implies (\mathbf{X} \perp\!\!\!\perp \mathbf{Y}, \mathbf{W} | \mathbf{Z})$

Probabilistic Graphical Models

Given a set of n random variables $\mathcal{X} = \{x_1, x_2, \cdots x_n\}$ where n is large, we want to build a joint probability distribution P over this set. Explicitly representing the joint distribution is computationally expensive, since just having binary values variables requires the joint distribution to specify $2^n - 1$ numbers, and for more practical variables, the count is too large.

We want to efficiently represent, estimate and answer inference queries on the distribution.

An example of a query can be - `Estimate the fraction of people with a bachelor's degree`.

Alternatives to explicit joint distributions

▷ Can we assume all columns are independent? **NO** - this is obviously a very bad assumption.
▷ Can we use data to detect highly correlated column pairs, and estimate their pairwise frequencies? **MAYBE** - but there might be too many correlated pairs, and the method is ad hoc.
To solve the above two not so good ways, we explore conditional independencies. It may be possible that `income` $\not\perp\!\!\!\perp$ `age` but `income` $\perp\!\!\!\perp$ `age`$|$`experience`.

Note that we write that a set X is conditionally independent of Y given Z, i.e $X \perp\!\!\!\perp Y | Z$ if

$$\Pr(X|Y, Z) = \Pr(X|Z)$$

Probabilistic graphical models use a graph-based representation as the basis for compactly encoding a complex distribution over a high-dimensional space.

It is convenient to represent the independence assumption using a graph. The so called graphical model has nodes as the variables (continuous or discrete), and the edges represent direct interaction. If we consider directed edges, we talk about Bayesian Networks, and if we consider undirected edges, we talk about Markov Random Fields.

Essentially the graphical model is a combination of the graph and potentials.

Definition 9 (Potentials)**.** Potentials $\psi_c(\mathbf{x}_c)$ are scores for assignment of values to subsets c of directly interacting variables. We factorize the probability as a product of these potentials, i.e

$$\Pr(\mathbf{x} = x_1, \cdots, x_n) \propto \prod \psi_s(\mathbf{x}_s) \tag{9}$$

Bayesian Networks

Bayesian Networks, also referred to as *directed graphical models* are a family of probability distributions that has a compact parameterization representable using a directed graph.

It is known that

$$\Pr(x_1, x_2, \cdots, x_n) = \Pr(x_1)\Pr(x_2|x_1)\Pr(x_3|x_2, x_1)\cdots\Pr(x_n|x_{n-1}, \cdots, x_1) \tag{10}$$

A compact Bayesian Network is a distribution in which each factor in the above equation depends on the *parent* variables represented by $\text{Pa}(x_i)$ for variable x_i. Thus, we have

$$\Pr(x_i|x_{i-1}, x_{i-2}, \cdots, x_1) = \Pr(x_i|\text{Pa}(x_i)) \tag{11}$$

and the corresponding potentials at each node in terms of its parents are

$$\psi_i(x_i, \text{Pa}(x_i)) = \Pr(x_i, \text{Pa}(x_i)) \tag{12}$$

Thus,

$$\Pr(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} \Pr(x_i|\text{Pa}(x_i)) \tag{13}$$

Consider the situation when each variable can take d values. The naive approach gives us $\mathcal{O}(d^n)$ parameters. If we think of the potentials as probability tables (with the rows corresponding to $\text{Pa}(x_i)$) and columns corresponding to the values of x_i, with entries as $\psi_i(x_i, \text{Pa}(x_i))$, we can notice that if $|\text{Pa}(x_i)| \leq k$, then the number of parameters are $\mathcal{O}(d^{k+1})$, and for n variables, we have $\mathcal{O}(nd^{k+1})$, which provides us the compact representation.

Definition

Now we formally define these -

Definition 11 (Bayesian Network)**.** A Bayesian Network is a directed graph $G = (V, E)$ together with

- ◇ a random variable x_i for each node $i \in V$
- ◇ a potential $\psi_i(x_i, \text{Pa}(x_i))$ for each node $i \in V$

For a variable x_i in our Bayesian Network $\mathcal{G}$, denote $\text{ND}(x_i)$ as the non-descendents of x_i. The following local conditional independencies hold in $\mathcal{G}$ -

$$x_i \perp\!\!\!\perp \text{ND}(x_i)|\text{Pa}(x_i) \tag{14}$$

Example 10 shows the independencies in a simple Bayesian Network.

Example 10.

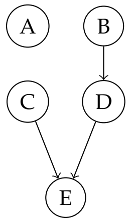

Figure 1: Sample BN

Consider the BN above. We will consider each variable at a time.

- ◇ A has no parent, and has no descendent. Thus,

$$A \perp\!\!\!\perp B, C, D, E$$

- ◇ B has no parent, but has D as a descendent. Thus,

$$B \perp\!\!\!\perp A, C$$

- ◇ C has no parent, but has E as a descendent. Thus,

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

- ◇ D has B as a parent, and has E as the descendent. Thus,

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

- ◇ E has C and D as parents, but has no descenent. Thus,

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

Definition 12 (Factorization)**.** Let $\mathcal{G}$ be a Bayesian Network graph over the variables $\{X_i\}_{i=1}^n$. We say that a distribution P over the same space factorizes according to $\mathcal{G}$ if P can be expressed as a product described in Equation 13. Such factorization is also known as the chain rule for Bayesian Networks, and is denoted as Factorize$(P, \mathcal{G})$.

Definition 13. Let P be a distribution over $\mathcal{X}$. We define $\mathcal{I}(P)$ to be the set of independent assertions of the form $\mathbf{X} \perp\!\!\!\perp \mathbf{Y} | \mathbf{Z}$ that hold in P.

We can now write "P satisfies the local independencies associated with $\mathcal{G}$" as $\mathcal{I}_\ell(\mathcal{G}) \subseteq \mathcal{I}(P)$.

Definition 14 (Independency-Map)**.** Let $\mathcal{K}$ be any graph object associated with a set of independencies $\mathcal{I}(\mathcal{K})$. We call $\mathcal{K}$ an I-map for a set of independencies $\mathcal{I}$ if $\mathcal{I}(\mathcal{K}) \subseteq \mathcal{I}$.

Thus for $\mathcal{G}$ to be an I-map for P, any independence that asserts in $\mathcal{G}$ must also assert in P, but P can have additional independencies not reflected in $\mathcal{G}$.

Remark 15 (Notation Alert)*.* Note that we will use the following interchangeably - P satisfies the local conditional independencies satisfied by $\mathcal{G}$ and $\mathcal{G}$ is an I-map for P, i.e

$$\text{Local-CI}(P, \mathcal{G}) \equiv \mathcal{I}_\ell(\mathcal{G}) \subseteq \mathcal{I}(P) \tag{15}$$

Theorem 16. *Given a distribution $P(x_1, x_2, \cdots, x_n)$ and a directed acyclic graph (DAG) $\mathcal{G}$,*

$$\textit{Local-CI}(P, \mathcal{G}) \iff \textit{Factorize}(P, \mathcal{G}) \tag{16}$$

Proof. ($\Longrightarrow$) We essentially need to show that if $\mathcal{G}$ is an I-map for P, then P factorizes according to $\mathcal{G}$. Consider a topologically sorted order $x_1, x_2, \cdots, x_n$ in $\mathcal{G}$. Local-CI$(P, \mathcal{G})$ tells us that

$$\Pr(x_i | x_1, \cdots, x_{i-1}) = \Pr(x_i | \text{Pa}(x_i))$$

We can write

$$P(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} P(x_i | x_1, \cdots, x_{i-1})$$

Each term in the product can be simplified due to the notion of Local-CI stated above, and we reach Equation 13, proving factorization.

($\Longleftarrow$) Proof has been skipped. □

Minimal Construction

Our goal is to construct a minimal and correct BN $\mathcal{G}$ to represent P. A DAG $\mathcal{G}$ is correct if all Local-CIs that are implied in $\mathcal{G}$ hold in P, and a DAG $\mathcal{G}$ is minimal if we cannot remove any edge(s) from $\mathcal{G}$ and still get a correct BN for P.

In the setting, we define our oracle $\mathscr{O}$ to whom we can ask any query of the type "`Is X ⊥⊥ Y|Z`?" pertaining to P and get a boolean answer. We will query the oracle several times to build up our BN. The following algorithm constructs such a BN -

```
Variables: $x_1, x_2, \cdots, x_n \longleftarrow$ ordered variables in $\mathcal{X}$
Independencies: $\mathcal{I} \longleftarrow$ set of independencies
$\mathcal{G} \longleftarrow$ Empty graph over $\mathcal{X}$
for $i = 1$ to $n$ do
    $\mathbf{U} \longleftarrow \{x_1, \cdots, x_{i-1}\}$ // Set of candidate parents of $x_i$
    for $U' \subseteq \{x_1, \cdots, x_{i-1}\}$ do
        if $U' \subset U$ and $(x_i \perp\!\!\!\perp \{x_1, \cdots, x_{i-1}\} - U' | U') \in \mathcal{I}$ then
            $\mathbf{U} \longleftarrow U'$
        end
    end
    // Now we have the minimal set $\mathbf{U}$ satisfying
     $(x_i \perp\!\!\!\perp \{x_1, \cdots, x_{i-1}\} - \mathbf{U} | \mathbf{U})$
    // Now we set $\mathbf{U}$ to be the parents of $x_i$
    for $x_j \in \mathbf{U}$ do
        Add $x_j \rightarrow x_i$ in $\mathcal{G}$
    end
end
return $\mathcal{G}$
```

Algorithm 1: Minimal Bayesian Network Construction (I-Map)

We know sketch rough proofs for the claims of the algorithm.

Theorem 17. *The BN $\mathcal{G}$ constructed by algorithm 1 is minimal, i.e we cannot remove any edge from the BN while maintaining the correctness of the BN for P.*

Proof. By construction. A subset of $\text{ND}(x_i)$ were available when we chose parents of $\mathbf{U}$ minimally. □

Theorem 18. *$\mathcal{G}$ constructed by the above algorithm is correct, i.e, the local-CIs induced by $\mathcal{G}$ hold in P.*

Proof. The construction is such that $\text{Factorize}(P, \mathcal{G})$ holds everytime. Since $\text{Factorize}(P, \mathcal{G}) \implies \text{Local-CI}(P, \mathcal{G})$, the constructed BN satisfies the local-CIs of P. □

Example 19. To be added.

Remark 20 (Importance of ordering). It is possible that a different ordering in $\mathcal{X}$ gives rise to a different BN, which although may be minimal, but may not be *optimal*. A minimal BN is defined for a given ordering, while an optimal BN is defined over all orderings. Example 19 shows such a case.

D-Separation

Our goal is to know when we can guarantee $\mathbf{X} \perp\!\!\!\perp \mathbf{Y} | \mathbf{Z}$ holds given a BN $\mathcal{G}$. The further discussion provides some cases where we can guarantee $\mathbf{X} \not\perp\!\!\!\perp \mathbf{Y} | \mathbf{Z}$.

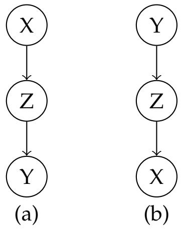

Figure 2: Causal and evidential effect

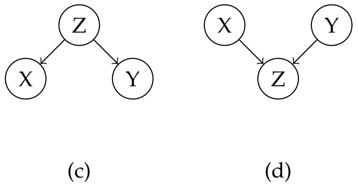

Figure 3: Common cause and common effect

1. **Direct Connection:** If there is an edge $X \to Y$, then regardless of any $\mathbf{Z}$, we can find examples where they influence each other.
2. **Indirect Connection:** This means that there is a trail between the nodes in the graph. We consider the simple case when we a 3-node graph and Z is between X and Y. Consider the 4 diagrams to the left for reference.
   (a) *Indirect causal effect:* X cannot influence Y via Z if Z is observed.
   (b) *Indirect evidential effect:* This is similar to the previous case as dependence is a symmetric notion. Thus, X can influence Y via Z, only if Z is not observed.
   (c) *Common cause:* The conclusion is similar to (a) and (b).
   (d) *Common effect:* (v-structure) This case is a bit tricky to understand, but the crux is that X can influence Y when either Z or one of Z's descendents is observed.

   If we have flow of influence from X to Y via Z, we say that the trail $X \rightleftharpoons Y \rightleftharpoons Z$ is active.

$$\begin{gathered}\left.\begin{aligned}\text{Causal trail: } & X \to Z \to Y\\ \text{Evidential trail: } & Y \to Z \to X\\ \text{Common cause: } & X \leftarrow Z \to Y\end{aligned}\right\}\text{Active if and only if } Z \text{ is observed}\\ \left.\star\ \text{Common effect: } X \to Z \leftarrow Y\right\}\\ \hookrightarrow \text{Active if and only if } Z \text{ or one of } Z\text{'s descendent is observed}\end{gathered}\tag{17}$$

Now, we can create a general notion of trails -

Definition 21. Let $\mathcal{G}$ be a BN, and $x_1 \rightleftharpoons \cdots \rightleftharpoons x_n$ be a trail in $\mathcal{G}$. Let $\mathbf{Z} \subset \{\text{observed variables}\}$. The trail is active given $\mathbf{Z}$ if

- ◇ Whenever we have a v-structure $x_{i-1} \to x_i \leftarrow x_{i+1}$, then x_i or one of its descendents are in $\mathbf{Z}$
- ◇ No other node along the trail is in $\mathbf{Z}$.

We can see that if $x_1 \in \mathbf{Z}$ or $x_n \in \mathbf{Z}$, then the trail is inactive.

Definition 22 (d-separation)**.** Let $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ be three sets of nodes in $\mathcal{G}$. We say that $\mathbf{X}$ and $\mathbf{Y}$ are d-separated given $\mathbf{Z}$, i.e d-sep$_{\mathcal{G}}(\mathbf{X}; \mathbf{Y} | \mathbf{Z})$ if there is no active trail between any node $x \in \mathbf{X}$ and $y \in \mathbf{Y}$ given $\mathbf{Z}$.

Essentially in a DAG, $\mathbf{Z}$ d-separates $\mathbf{X}$ from $\mathbf{Y}$ if all paths $\mathcal{P}$ from any $\mathbf{X}$ to $\mathbf{Y}$ is blocked by $\mathbf{Z}$.

A path $\mathfrak{P}$ is *blocked* if it is active, i.e there is flow of influence.

Definition 23 (Global Markov independencies)**.** The set

$$\mathcal{I}(\mathcal{G}) \overset{\text{def}}{=} \{(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} | \mathbf{Z}) : \text{d-sep}_{\mathcal{G}}(\mathbf{X}; \mathbf{Y} | \mathbf{Z})\} \tag{18}$$

denoting the set of independencies corresponding to d-separation is the set of global Markov independencies.

We use the same notation as $\mathcal{I}(P)$ as we can show that the independencies in $\mathcal{I}(\mathcal{G})$ are those guaranteed to hold for every distribution over $\mathcal{G}$ (Theorem 24).

Theorem 24. *The d-separation test identifies the complete set of conditional independencies that hold in all distributions that conform to a given Bayesian Network.*

Proof. Skipped. □