Lecture Notes for CS 726 - Spring 2021

Eeshaan Jain

January 9, 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

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 1 (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

$$\Pr(\mathbf{x} = x_1, \cdots, x_n) \propto \prod \psi_s(\mathbf{x}_s) \tag{1}$$

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

$$\Pr(x_1, x_2, \cdots, x_n) = \Pr(x_1)\Pr(x_2|x_1)\Pr(x_3|x_2, x_1)\cdots\Pr(x_n|x_{n-1}, \cdots, x_1) \tag{2}$$

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

$$\Pr(x_i|x_{i-1}, x_{i-2}, \cdots, x_1) = \Pr(x_i|\text{Pa}(x_i)) \tag{3}$$

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

$$\psi_i(x_i, \text{Pa}(x_i)) = \Pr(x_i, \text{Pa}(x_i)) \tag{4}$$

Thus,

$$\Pr(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} \Pr(x_i|\text{Pa}(x_i)) \tag{5}$$

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.

Now we formally define these -

Definition 2 (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 $i \in V$