COMPSCI 240: Reasoning Under Uncertainty

Andrew Lan and Nic Herndon

University of Massachusetts at Amherst

Spring 2019

Lecture 28: Bayesian Networks

Review

- Chain Rule for n random variables

$$P(X_n, \cdots, X_1) = P(X_n | X_{n-1}, \cdots, X_1) P(X_{n-1}, \cdots, X_1)$$

- Marginal probabilities for multiple discrete random variables $X_1, \cdots X_n$ with joint PMF, denoted as $P(X_1, \cdots X_n)$, could be computed as

$$P(X_1 = x_1) = \sum_{x_2} \cdots \sum_{x_n} P(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n)$$

- In practice, it is much common to encounter real-world problems that involve measuring multiple random variables $X_1, ..., X_n$ for each repetition of the experiment.
- These random variables $X_1, ..., X_n$ may have complex relationships among themselves.

The Curse of Dimensionality

- Suppose we have an experiment where we obtain the values of d random variables $X_1, ..., X_d$, where each variable has binary outcomes (for simplicity).
- **Question:** How many **numbers** does it take to write down a joint distribution for them?
- **Answer:** We need to define a probability for each d-bit sequence:

$$
\begin{gathered}
P(X_1 = 0, X_2 = 0, ..., X_d = 0) \\
P(X_1 = 1, X_2 = 0, ..., X_d = 0) \\
\vdots \\
P(X_1 = 1, X_2 = 1, ..., X_d = 1)
\end{gathered}
$$

- The number of d-bit sequences is 2^d. Because we know that the probabilities have to add up to 1, we need to write down $2^d - 1$ numbers to specify the full joint PMF on d binary variables.

How Fast is Exponential Growth?

- $2^d - 1$ grows **exponentially** as d increases **linearly**:

d	$2^d - 1$
1	1
10	1023
100	1,267,650,600,228,229,401,496,703,205,375
$\vdots$	$\vdots$

- Storing the full joint PMF for 100 binary variables would take about 10^{30} real numbers or about 10^{18} **terabytes** of storage!
- Joint PMFs grow in size so rapidly, we have no hope whatsoever of storing them explicitly for problems with more than about 30 (binary) random variables.

Factorizing Joint Distributions

- To address this, we start by *factorizing the joint distribution*, i.e., re-writing the joint distribution as a product of conditional PMFs over single variables (called factors).
- If we know some conditional independency between the variables, we can save some space.

Conditional Independence: Simplification 2

- Suppose we instead only assume that:
  - $P(X_2 = a_2 | X_1 = a_1, X_3 = a_3) = P(X_2 = a_2 | X_1 = a_1)$ for all a_1, a_2, a_3.
- This gives the "conditional independence model": X_2 is conditionally independent of X_3 given X_1

$$
\begin{aligned}
&P(X_1 = a_1, X_2 = a_2, X_3 = a_3) \\
&\quad = P(X_1 = a_1) P(X_3 = a_3 | X_1 = a_1) P(X_2 = a_2 | X_1 = a_1, X_3 = a_3) \\
&\quad = P(X_1 = a_1) P(X_3 = a_3 | X_1 = a_1) P(X_2 = a_2 | X_1 = a_1)
\end{aligned}
$$

- How many numbers do we need to store for three binary random variables in this case?
  $1 + 2 + 2 = 5$ (as opposed to $2^3 - 1 = 7$ if we encoded the full joint)

Example

- Toothache: boolean variable indicating whether the patient has a toothache caused by a cavity
- Cavity: boolean variable indicating whether the patient has a cavity
- Catch/Find: whether the dentist finds the cavity
- If the patient has a cavity, the probability that the dentist finds the cavity doesn't depend on whether he/she has a toothache

$$P(Find|Toothache, Cavity) = P(Find|Cavity)$$

  Therefore, Find is conditionally independent of Toothache given Cavity

- Likewise, Toothache is conditionally independent of Find given Cavity

$$P(Toothache|Find, Cavity) = P(Toothache|Cavity)$$

- What is the space requirement to represent the Joint Distribution $P(Toothache, Find, Cavity)$?

Bayesian Networks

- Keeping track of all the conditional independence assumptions gets tedious when there are a lot of variables.
  - Consider the following situation: You live in a quiet neighborhood in the suburbs of LA. There are two reasons the alarm system in your house will go off: your house is broken into or there is an earthquake. If your alarm is on, you might get a call from the police department. You might also get a call from your neighbor.
- To get around this problem, we use "Bayesian Networks" to express the conditional independence structure of these models.

Bayesian Networks

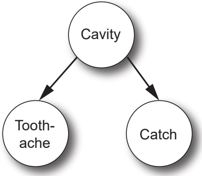
- A Bayesian network uses conditional independence assumptions to more compactly represent a joint PMF of many random variables.
- We use a Directed Acyclic Graph (DAG) to encode conditional independence assumptions.
  - Nodes X_i in the graph G represent random variables.
  - A directed edge $X_j \rightarrow X_i$ means X_i directly depends on X_j (not causation!).
  - We also define that X_j is a "parent" of X_i.
  - The set of variables that are parents of X_i is denoted Pa_i.
  - X_i is independent of all its nondescendants given Pa_i.
  - The factor associated with variable X_i is $P(X_i|Pa_i)$.

Example: Bayesian Network

- Toothache: boolean variable indicating whether the patient has a toothache
- Cavity: boolean variable indicating whether the patient has a cavity
- Catch/Find: whether the dentist's probe catches in the cavity
- We had

$$P(Find|Toothache, Cavity) = P(Find|Cavity)$$
$$P(Toothache|Find, Cavity) = P(Toothache|Cavity)$$

- This can be graphically represented as

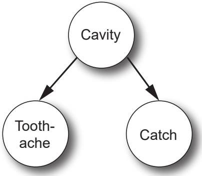


Bayesian Networks vs. Markov Chains

- Do not confuse the *Bayesian Networks* and the *Transition Probability Graphs of Markov Chains*.
- These two graphs look similar (both have circles with arrows) but represent two vastly different entities.
- In Transition Probability Graphs, **nodes** represent all possible **states**, and **arrows** represents the **probability of transition** from one state to another (with numbers written on it).
- In Bayesian Networks, **nodes** represent all possible **random variables**, and **arrows** represents **dependencies** between the random variables (no numbers associated with it).

Example: Bayesian Network

- Given a BayesNet (DAG),



- Thus,

$$\begin{aligned} P(C, T, F) &= P(C|T, F)P(T|F)P(F) \\ &= P(C|F)P(T|F)P(F) \end{aligned}$$

$$\begin{aligned} P(C, T, F) &= P(T|C, F)P(C|F)P(F) \\ &= P(T|F)P(C|F)P(F) \end{aligned}$$