Bayesian Networks

Goal: represent a joint distribution $P$ over random variables $X = \{X_1, \ldots, X_n\}$

<table>
<thead>
<tr>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$P(x_1, x_2)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>false</td>
<td>false</td>
<td>0.1</td>
</tr>
<tr>
<td>false</td>
<td>true</td>
<td>0.2</td>
</tr>
<tr>
<td>true</td>
<td>false</td>
<td>0.3</td>
</tr>
<tr>
<td>true</td>
<td>true</td>
<td>0.4</td>
</tr>
</tbody>
</table>

Bayesian Networks

• If variables are binary, the joint distribution has $2^n - 1$ parameters
  – Expensive space usage
  – Human expert has hard time determining these numbers
  – Need large amounts of data to estimate these numbers accurately

• How do we represent a joint probability distribution compactly?
  – Solution: Exploit independence properties

Bayesian Networks

• Suppose we toss $n$ coins and let $X_i$ be the outcome of coin toss $i$

• The joint distribution $P(X_1, \ldots, X_n)$ has $2^n - 1$ parameters
Bayesian Networks

- Now assume the coin tosses are marginally independent i.e. \( X_i \perp X_j \) for any \( i, j \)
- The joint distribution \( P(X_1, ..., X_n) = P(X_1)P(X_2) ... P(X_n) \)

For each \( i \), we have the following table:

<table>
<thead>
<tr>
<th>( X_i )</th>
<th>( P(X_i) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>false</td>
<td>( 1 - \theta_i )</td>
</tr>
<tr>
<td>true</td>
<td>( \theta_i )</td>
</tr>
</tbody>
</table>

There are only \( n \) parameters \( (\theta_1, ..., \theta_n) \) to specify!

The Conditional Parameterization

- Define 2 random variables: Intelligence (I) and SAT score (S)
- We could represent the joint distribution as follows:

<table>
<thead>
<tr>
<th>I</th>
<th>S</th>
<th>( P(I, S) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>low</td>
<td>low</td>
<td>0.665</td>
</tr>
<tr>
<td>low</td>
<td>high</td>
<td>0.035</td>
</tr>
<tr>
<td>high</td>
<td>low</td>
<td>0.06</td>
</tr>
<tr>
<td>high</td>
<td>high</td>
<td>0.24</td>
</tr>
</tbody>
</table>

The Conditional Parameterization

- An alternative representation: \( P(I, S) = P(I)P(S|I) \)
- Note: represents the causal process i.e. intelligence affects SAT score

Prior distribution over I

<table>
<thead>
<tr>
<th>I</th>
<th>( P(I) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>low</td>
<td>0.7</td>
</tr>
<tr>
<td>high</td>
<td>0.3</td>
</tr>
</tbody>
</table>

Conditional probability distribution of \( S \) given \( I \)

| I   | \( P(S|I) \) |
|-----|--------------|
| low | low: 0.95, high: 0.05 |
| high| high: 0.8 |

| I   | \( P(S|I) \) |
|-----|--------------|
| low | low: 0.95, high: 0.05 |
| high| high: 0.8 |

- There are 3 binomial distributions here: \( P(I), P(S|I = low), P(S|I = high) \)
- Only 3 independent parameters are needed:

\( \theta_{I=high}, \theta_{S=high|I=low}, \theta_{S=high|I=high} \)
The Conditional Parameterization

The joint distribution (conditional parameterization version) drawn as a Bayesian network looks like:

\[ P(I, S) = P(I)P(S|I) \]

Naïve Bayes

- Now assume we have 3 random variables:
  - Intelligence: low, high
  - SAT score: low, high
  - Grade: A, B, C
- No independencies that hold:
  - Intelligence correlated with SAT score and grade
  - SAT score and grade not independent

Naïve Bayes

- But conditional independencies hold!
- If a student has high intelligence, a high SAT score no longer gives us information about the student’s grade
- Formally: \( S \perp G | I \)

Note: This is only true if intelligence is the only reason by his grade and SAT score might be correlated.
Naive Bayes

- This leads to the following factored representation:

\[ P(I, S, G) = P(S, G | I) P(I) \]

\[ = P(S | I) P(G | I) P(I) \]

(As before)

[Conditional independence:]

\[ P(S, G | I) = P(S | I) P(G | I) \]

There are 3 binomial distributions:

- \( P(I) \) with parameter: \( \theta_I = \text{high} \)
- \( P(S | I = \text{low}) \) with parameter: \( \theta_S = \text{high} | I = \text{low} \)
- \( P(S | I = \text{high}) \) with parameter: \( \theta_S = \text{high} | I = \text{high} \)

And 2 three-valued multinomial distributions:

- \( P(G | I = \text{low}) \) with parameters: \( \theta_G = A | I = \text{low}, \theta_G = B | I = \text{low} \)
- \( P(G | I = \text{high}) \) with parameters: \( \theta_G = A | I = \text{high}, \theta_G = B | I = \text{high} \)

7 vs 11 independent parameters for a full joint distribution

---

**Intelligence**

**Grade**

**SAT**

Based on these assumptions, the joint distribution factorizes as:

\[ P(C, X_1, \ldots, X_m) = P(C) \prod_{i=1}^{m} P(X_i | C) \]

If all the variables are binary, there are a total of \((2m + 1)\) independent parameters needed to specify the naive Bayes model.
A Bayesian network is composed of:
• The DAG structure
• The conditional probability distributions in each node

- A Bayesian network is represented as a Directed Acyclic Graph (DAG) $G$
  - Nodes are random variables
  - Edges correspond to the direct influence of one random variable on another
- $G$ can be viewed in two different ways:
  - The skeleton for a compact, factored representation of a joint distribution
  - A compact representation for a set of conditional independence assumptions about a distribution
- Both are equivalent

DAG Structure: intuitively, each variable depends directly only on its parents
Bayesian Networks

Each node has a local probability model
- Captures the conditional probability distribution of the node given its parents $P(X|Parents(X))$
- Specifies a distribution over each value of $X$ given each possible joint assignment of values to its parents
- A node with no parents eg. $P(I)$ is conditioned on the empty set of variables and is a marginal distribution

Bayesian Networks

With a Bayesian network, you can compute the value of any state of the joint probability distribution

$$P(I = \text{high}, D = \text{low}, G = B, S = \text{high}, L = \text{weak})$$

$$= P(I = \text{high})P(D = \text{low})P(G = B | I = \text{high}, D = \text{low}) \ast$$

$$P(S = \text{high} | I = \text{high})P(L = \text{weak} | G = B)$$

$$= 0.3 \ast 0.6 \ast 0.08 \ast 0.8 \ast 0.4 = 0.004608$$

This uses the chain rule for Bayesian networks (more on this later)