Graphical Models

Lecture 3:
Local Conditional Probability Distributions

Andrew McCallum
mccallum@cs.umass.edu

Thanks to Noah Smith and Carlos Guestrin for some slide materials.
Conditional Probability Distributions

• Proper CPD:

\[ \sum_{x \in \text{Val}(X)} P(X = x \mid \text{Parents}(X)) = 1 \]

(for continuous case, integral)

• Everything breaks down into distributions over one variable given some others.
Conditional Probability Distributions

• So far, we’ve seen **table** representations.
  – How many parameters?
  – Where will they come from?
  – Alternatives?

• If we know more about the form of a CPD, we may be able to infer more properties of $P$.
  – Independence assertions
  – Efficient inference (later)
Deterministic CPDs

- Sometimes a variable’s value is a deterministic function of its parents’ values.

\[
P(T | G_1, G_2) = \begin{cases} 
1 & G_1 = a \text{ and } G_2 = b \\
0 & G_1 = a \text{ and } G_2 \in \{a, o\} \\
0 & G_1 = b \text{ and } G_2 \in \{b, o\} \\
0 & G_1 = G_2 = o \\
0 & G_1 = G_2 \in \{a, b\} 
\end{cases}
\]

-Allele from Mom
-Allele from Dad

Phenotypical blood type
Deterministic CPDs Affect $I(G)$

- $X \perp Y \mid T$
  (local Markov assumption)
- $X \perp Y \mid \{G_1, G_2\}$

- Can we derive such independence assertions?
Review: D-Separation

• Three sets of nodes: $X$, $Y$, and observed nodes $Z$
• $X$ and $Y$ are **d-separated** given $Z$ if there is no active trail from any $X \in X$ to any $Y \in Y$ given $Z$. 
D-Separation with Deterministic CPDs

• Query: $X \perp Y \mid Z$?
• Deterministic variables: $D$
• Algorithm:
  – Let $Z' = Z$
  – While there is an $X_i \in D$ such that $\text{Parents}(X_i) \subseteq Z'$, add $X_i$ to $Z'$
  – Calculate d-separation between $X$ and $Y$ given $Z'$
• This is sound and complete.
  – But if we know still more about the deterministic functions involved, we may be able to go farther!
Another Example

- Suppose $C = A \text{ XOR } B$
- In the case of XOR, B and C fully determine A
- $D \perp E | \{B, C\}$
Context-Specific Independence

• Suppose C = A OR B
• We are given A = 1.
  – This implies that C = 1.
• P(X | B, A=1) = P(X | A=1)
  – Independence!
• This does not work if A = 0.

Previously we defined X ⊥ Y | Z to represent the assumption P(X|Y, Z) = P(X|Z) for all values of X, Y, Z. Deterministic functions can imply a type of independence that holds only for particular values of some variables.
Context-Specific Independence

Definition

• Let $X$, $Y$, and $Z$ be disjoint sets of variables; $C$ is a set of variables that could overlap.

• Let $c \in \text{Val}(C)$.

• $X$ and $Y$ are conditionally independent given $Z$ and the context $c$:
  
  $- \quad X \perp_c Y \mid Z$

  $- \quad P(X \mid Y, Z, C = c) = P(X \mid Z, C = c)$

  whenever $P(Y, Z, C = c) > 0$
Tree CPDs

• Use a tree to represent $P(X \mid \text{Parents}(X))$
  – Each leaf is a distribution over $X$
  – Each path defines a context
Student Example

apply

A

intelligence

D

difficulty of the course

standardized test score

S

job offered

J

grade

G

letter

L

P(J | A, S, L)

| P(J | A, S, L) | 000 | 001 | 010 | 011 | 100 | 101 | 110 | 111 |
|--------------|-----|-----|-----|-----|-----|-----|-----|-----|
| yes          |     |     |     |     |     |     |     |     |
| no           |     |     |     |     |     |     |     |     |
If you don’t apply for the job, the outcome does not depend on your score or letter.

If you apply, the job offered depends on the intelligence, the difficulty of the course, the standardized test score, and the grade.

| P(J | A, S, L) | 0?? | 100 | 101 | 110 | 111 |
|------------|-----|-----|-----|-----|-----|
| yes        | yes |     |     |     |     |
| no         |     |     |     |     |     |
If your test scores are high enough, the recruiter doesn't even look at the letter.

| P(J | A, S, L) | 0??   | 100 | 101 | 11* |
|------------|-------|-----|-----|-----|
| yes        |       |     |     |     |
| no         |       |     |     |     |
Student Example

\[
P(J \mid A, S, L)
\]

| P(J | A, S, L) | 0?? | 100 | 101 | 11* |
|-------------|-----|-----|-----|-----|
| yes         | yes |     |     |     |
| no          |     |     |     |     |
Other Structured CPDs

• Rules
  − CPD trees can always be represented compactly as rules, but the converse does not hold

• Decision diagrams

• Any kind of partition structure of $\text{Val}(X) \times \text{Val}(\text{Parents}(X))$

• Context-specific CPDs can make some edges spurious (given the context)!
Independent Causes

• A different kind of CPD structure
• Consider random variable Y and its parents, $\text{Parents}(Y) = X$
• Two examples
  – Noisy OR
  – Generalized linear models
Another Professor (Perfect World)

- Letter quality depends on whether you participated by asking good questions (Q), and on whether you wrote a good final paper (F)

| P(L | Q, F) | Q = 1 and F = 1 | Q = 1 and F = 0 | Q = 0 and F = 1 | Q = 0 and F = 0 |
|-----------|-----------------|-----------------|-----------------|-----------------|
| high | 1 | 1 | 1 | 0 |
| low | 0 | 0 | 0 | 1 |
Another Professor (Real World)

• Letter quality depends on whether you participated by asking good questions (Q), and on whether you wrote a good final paper (F)

| P(L | Q, F)       | Q = 1 and F = 1 | Q = 1 and F = 0 | Q = 0 and F = 1 | Q = 0 and F = 0 |
|-----------------|-----------------|-----------------|-----------------|-----------------|
| high            | 0.98            | 0.8             | 0.9             | 0               |
| low             | 0.02            | 0.2             | 0.1             | 1               |
Another Professor (Real World)

\[
P(Q' \mid Q) \begin{array}{c|cc}
Q = 1 & Q = 0 \\
\hline
\text{high} & 0.8 & 0 \\
\text{low} & 0.2 & 1 \\
\end{array}
\]

\[
L = Q' \lor F'
\]

\[
P(F' \mid F) \begin{array}{c|cc}
F = 1 & F = 0 \\
\hline
\text{high} & 0.9 & 0 \\
\text{low} & 0.1 & 1 \\
\end{array}
\]
Another Professor

\[
P(Q' \mid Q) \begin{array}{c|c|c}
Q = 1 & Q = 0 \\
\hline
\text{high} & 0.8 & 0 \\
\text{low} & 0.2 & 1 \\
\end{array}
\]

\[L = Q' \lor F'
\]

\[
P(F' \mid F) \begin{array}{c|c|c}
F = 1 & F = 0 \\
\hline
\text{high} & 0.9 & 0 \\
\text{low} & 0.1 & 1 \\
\end{array}
\]

noise parameter, \( \lambda_i \)
Another Professor

<table>
<thead>
<tr>
<th>$P(Q' \mid Q)$</th>
<th>$Q = 1$</th>
<th>$Q = 0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>high</td>
<td>0.8</td>
<td>0</td>
</tr>
<tr>
<td>low</td>
<td>0.2</td>
<td>1</td>
</tr>
</tbody>
</table>

$L = Q' \lor F' \lor \text{Leak}$

<table>
<thead>
<tr>
<th>leak probability, $\lambda_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$1$</td>
</tr>
<tr>
<td>$0$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$P(F' \mid F)$</th>
<th>$F = 1$</th>
<th>$F = 0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>high</td>
<td>0.9</td>
<td>0</td>
</tr>
<tr>
<td>low</td>
<td>0.1</td>
<td>1</td>
</tr>
</tbody>
</table>
Noisy OR Model

\( \lambda_i \) is the noise parameter for \( X_i \).

\[
P(Y = 0 \mid X) = (1 - \lambda_0) \prod_{i : X_i = 1} (1 - \lambda_i)
\]

\[
P(Y = 1 \mid X) = 1 - \left[ (1 - \lambda_0) \prod_{i : X_i = 1} (1 - \lambda_i) \right]
\]

Or equivalently written (where \( x^1 = 1 \) and \( x^0 = 0 \))

\[
P(y^0 \mid x_1, \ldots, x_k) = (1 - \lambda_0) \prod_{i=1}^k (1 - \lambda_i)^{x_i}
\]
Noisy OR as a Conditional Bayesian Network
(What is a Conditional Bayesian Network?)

• Conditional Bayesian Network is a BN with three types of variables:
  – Inputs, always observed, no parents: \( X \)
  – Outputs: \( Y \)
  – Encapsulated: \( Z \)

\[
P(Y, Z \mid X) = \prod_{W \in Y \cup Z} P(W \mid \text{Parents}(W))
\]
Independent Causes

- Many “additive” effects combine to score $X$
- $P(Y = 1)$ is defined as a function of $X$

\[
\text{sigmoid}(z) = \frac{e^z}{1 + e^z}
\]
Generalized Linear Model

• Score is defined as a *linear* function of $\mathbf{X}$:

$$f(\mathbf{X}) = w_0 + \sum_i w_i X_i$$

$Z = f(\mathbf{X})$ is a random variable!

• Probability distribution over binary value $Y$ is commonly* defined by:

$$P(Y = 1) = \text{sigmoid}(f(\mathbf{X}))$$

Logistic regression
Maximum entropy classifier

* not the only choice

$$\text{sigmoid}(z) = \frac{e^z}{1 + e^z}$$
aka “logit” function
Logistic CPD Model as a Conditional Bayesian Network

Compare: Naïve Bayes model
Logistic Models

• Weight $w_i$ can be positive or negative.
• The $X$ and $Y$ do not need to be binary.
  – Very useful:
    multinomial logistic to allow many values for $Y$;
    indicator variables to allow many values for $X$
• “Multinomial logit”
CPDs with Causal Independence
Captures Noisy-OR and Generalized linear models

individual noise models

deterministic and symmetrically decomposable (binary, commutative, associative operation on the $Z_i$), then
“Symmetric Independence of Causal Influence” (symmetric ICI)

Noisy OR:
$X \rightarrow Z_i = \text{simple noisy model}$
$Z_i s \rightarrow Z = \text{OR}$
$Z \rightarrow Y = \text{copy}$

GLM:
$X \rightarrow Z_i = w_i X_i$
$Z_i s \rightarrow Z = \text{sum}$
$Z \rightarrow Y = \text{sigmoid}$

All interaction among $X_i$ happens here!
Something Interesting Happened!

- We are now representing conditional probability distributions as conditional Bayesian Networks!
| P(Y | X) | 000 | 001 | 010 | 011 | 100 | 101 | 110 | 111 |
|-------|-----|-----|-----|-----|-----|-----|-----|-----|
| 1     |     |     |     |     |     |     |     |     |
| 0     |     |     |     |     |     |     |     |     |
Continuous Random Variables

First case: continuous child and parents

• Gaussian distribution is a CPD:
  \[ P(Y \mid \text{Mean} = \mu, \text{Variance} = \sigma^2) = \text{Normal} (\mu, \sigma^2) \]

• **Linear Gaussian** CPD: \[ \text{Normal} (\text{linear} (x), \sigma^2) \]
  – Y is a linear function of the variables \( X \), with Gaussian noise that has variance \( \sigma^2 \)
Example

• $X(t)$ is a vehicle’s position at time $t$
• $V(t)$ is its velocity at time $t$

\[ X(t+1) \approx X(t) + V(t) \]

• Allow for some randomness in the motion:

\[
p(Y| x_1, \ldots x_k) = \mathcal{N} \left( \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k; \sigma^2 \right)
\]
Continuous Random Variables

**Second case**: continuous child with discrete *and* continuous parents \((X_{\text{disc}} \text{ and } X_{\text{cont}})\)

- "Conditional linear Gaussian":

\[
P(Y \mid X_{\text{disc}} = x_{\text{disc}}, X_{\text{cont}} = x_{\text{cont}}) = \mathcal{N}(w_{x_{\text{disc}},0} + \sum_j w_{x_{\text{disc}},j} x_{\text{cont},j} \sigma_{x_{\text{disc}}}^2)
\]

- different weights for each \(x_{\text{disc}}\)

- Induces a **Gaussian mixture** for \(Y\)
Continuous Random Variables

Third case:
discrete child Y with continuous parent X

• Threshold model
  – Makes $P(Y \mid X)$ discontinuous in X’s value

• Multinomial logit (see slide 27 “Generalized Linear Model”)
CPDs can be Bayesian Networks!

• “Conditional Bayesian Network” is a BN with three types of variables:
  – Inputs, always observed, no parents: \( \mathbf{X} \)
  – Outputs: \( \mathbf{Y} \)
  – Encapsulated: \( \mathbf{Z} \)

\[
P(\mathbf{Y}, \mathbf{Z} | \mathbf{X}) = \prod_{W \in \mathbf{Y} \cup \mathbf{Z}} P(W | \text{Parents}(W))
\]

• A CPD is an “Encapsulated CPD” if it can be represented by a Conditional Bayesian Network.
  – Construct a complex BN, with components composed of other BN subcomponents! ...object-oriented style!
Encapsulated CPDs (K&F Figure 5.15)
Where We Are

• Structure in CPDs
• Effects on independence assertions
• Examples:
  – Determinism
  – Context-specificity
  – Independent causes
  – Continuous distributions
  – CPDs represented by Conditional BNs