ECE 6504: Advanced Topics in Machine Learning
Probabilistic Graphical Models and Large-Scale Learning

Topics
- Bayes Nets
  - (Finish) Parameter Learning
  - Structure Learning

Readings: KF 18.1, 18.3; Barber 9.5, 10.4

Dhruv Batra
Virginia Tech
Administrativia

- HW1
  - Out
  - Due in 2 weeks: Feb 17, Feb 19, 11:59pm
  - Please please please please start early
  - Implementation: TAN, structure + parameter learning
  - Please post questions on Scholar Forum.
Recap of Last Time
Learning Bayes nets

<table>
<thead>
<tr>
<th>Known structure</th>
<th>Unknown structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fully observable data</td>
<td>Very easy</td>
</tr>
<tr>
<td>Missing data</td>
<td>Somewhat easy (EM)</td>
</tr>
<tr>
<td></td>
<td>Hard</td>
</tr>
<tr>
<td></td>
<td>Very very very hard</td>
</tr>
</tbody>
</table>

Data: $x^{(1)}$, ..., $x^{(m)}$

CPTs: $P(X_i | Pa_{Xi})$

Structure

Slide Credit: Carlos Guestrin
Learning the CPTs

For each discrete variable $X_i$

\[
\hat{P}_{MLE}(X_i = a \mid Pa_{X_i} = b) = \frac{\text{Count}(X_i = a, Pa_{X_i} = b)}{\text{Count}(Pa_{X_i} = b)}
\]
Plan for today

• (Finish) BN Parameter Learning
  – Parameter Sharing
  – Plate notation

• (Start) BN Structure Learning
  – Log-likelihood score
  – Decomposability
  – Information never hurts
Meta BN

• Explicitly showing parameters as variables

• Example on board
  – One variable \(X\); parameter \(\theta_X\)
  – Two variables \(X, Y\); parameters \(\theta_X, \theta_{Y|X}\)
Global parameter independence

- **Global parameter independence:**
  - All CPT parameters are independent
  - Prior over parameters is product of prior over CPTs

- **Proposition:** For fully observable data $D$, if prior satisfies global parameter independence, then

$$P(\theta \mid D) = \prod_i P(\theta_{X_i} \mid Pa_{X_i}, D)$$
Parameter Sharing

• What if $X_1, \ldots, X_n$ are $n$ random variables for coin tosses of the same coin?
Naïve Bayes vs Bag-of-Words

• What’s the difference?

• Parameter sharing!
Text classification

• Classify e-mails
  – \( Y = \{ \text{Spam}, \text{NotSpam} \} \)

• What about the features \( X \)?
  – \( X_i \) represents \( i^{\text{th}} \) word in document; \( i = 1 \) to doc-length
  – \( X_i \) takes values in vocabulary, 10,000 words, etc.

\[
h_{NB}(x) = \arg \max_y P(y) \prod_{i=1}^{\text{LengthDoc}} P(x_i|y)
\]
Bag of Words

- **Position in document doesn’t matter:**
  \[ P(X_i = x_i \mid Y = y) = P(X_k = x_i \mid Y = y) \]
  - Order of words on the page ignored
  - Parameter sharing

\[
P(y) \prod_{i=1}^{\text{LengthDoc}} P(x_i \mid y)
\]

When the lecture is over, remember to wake up the person sitting next to you in the lecture room.
Bag of Words

- **Position in document doesn’t matter:**
  \[ P(X_i=x_i|Y=y) = P(X_k=x_i|Y=y) \]
  - Order of words on the page ignored
  - Parameter sharing

\[
P(y) \prod_{i=1}^{LengthDoc} P(x_i|y)
\]
HMMs semantics: Details

Just 3 distributions:

\[ P(X_1) \]
\[ P(X_i \mid X_{i-1}) \]
\[ P(O_i \mid X_i) \]
N-grams

- Learnt from Darwin’s *On the Origin of Species*

<table>
<thead>
<tr>
<th></th>
<th>Unigrams</th>
<th>Bigrams</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.16098</td>
<td>a</td>
</tr>
<tr>
<td>2</td>
<td>0.06687</td>
<td>b</td>
</tr>
<tr>
<td>3</td>
<td>0.01414</td>
<td>c</td>
</tr>
<tr>
<td>4</td>
<td>0.02938</td>
<td>d</td>
</tr>
<tr>
<td>5</td>
<td>0.03107</td>
<td>e</td>
</tr>
<tr>
<td>6</td>
<td>0.11055</td>
<td>f</td>
</tr>
<tr>
<td>7</td>
<td>0.02325</td>
<td>g</td>
</tr>
<tr>
<td>8</td>
<td>0.01530</td>
<td>h</td>
</tr>
<tr>
<td>9</td>
<td>0.04174</td>
<td>i</td>
</tr>
<tr>
<td>10</td>
<td>0.06233</td>
<td>j</td>
</tr>
<tr>
<td>11</td>
<td>0.00060</td>
<td>k</td>
</tr>
<tr>
<td>12</td>
<td>0.00309</td>
<td>l</td>
</tr>
<tr>
<td>13</td>
<td>0.03515</td>
<td>m</td>
</tr>
<tr>
<td>14</td>
<td>0.02107</td>
<td>n</td>
</tr>
<tr>
<td>15</td>
<td>0.06007</td>
<td>o</td>
</tr>
<tr>
<td>16</td>
<td>0.06066</td>
<td>p</td>
</tr>
<tr>
<td>17</td>
<td>0.01594</td>
<td>q</td>
</tr>
<tr>
<td>18</td>
<td>0.00077</td>
<td>r</td>
</tr>
<tr>
<td>19</td>
<td>0.05265</td>
<td>s</td>
</tr>
<tr>
<td>20</td>
<td>0.05761</td>
<td>t</td>
</tr>
<tr>
<td>21</td>
<td>0.07566</td>
<td>u</td>
</tr>
<tr>
<td>22</td>
<td>0.02149</td>
<td>v</td>
</tr>
<tr>
<td>23</td>
<td>0.00993</td>
<td>w</td>
</tr>
<tr>
<td>24</td>
<td>0.01341</td>
<td>x</td>
</tr>
<tr>
<td>25</td>
<td>0.00208</td>
<td>y</td>
</tr>
<tr>
<td>26</td>
<td>0.01381</td>
<td>z</td>
</tr>
<tr>
<td>27</td>
<td>0.00039</td>
<td>z</td>
</tr>
</tbody>
</table>

Image Credit: Kevin Murphy
Plate Notation

• $X_1, \ldots, X_n$ are $n$ random variables for coin tosses of the same coin

• Plate denotes replication
Plate Notation

Plates denote replication of random variables
Hierarchical Bayesian Models

- Why stop with a single prior?

Figure 1: Graphical model representation of LDA. The boxes are “plates” representing replicates. The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document.
BN: Parameter Learning: What you need to know

• Parameter Learning
  – MLE
    • Decomposes; results in counting procedure
    • Will shatter dataset if too many parents
  – Bayesian Estimation
    • Conjugate priors
    • Priors = regularization (also viewed as smoothing)
    • Hierarchical priors
  – Plate notation
  – Shared parameters
Learning Bayes nets

<table>
<thead>
<tr>
<th>Data structure</th>
<th>Known structure</th>
<th>Unknown structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fully observable data</td>
<td>Very easy</td>
<td>Hard</td>
</tr>
<tr>
<td>Missing data</td>
<td>Somewhat easy (EM)</td>
<td>Very very hard</td>
</tr>
</tbody>
</table>

CPTs – $P(X_i | Pa_{Xi})$

(C) Dhruv Batra

Slide Credit: Carlos Guestrin
Goals of Structure Learning

• Prediction
  – Care about a good structure because presumably it will lead to good predictions

• Discovery
  – I want to understand some system

Data

\[ x^{(1)} \quad \ldots \quad x^{(m)} \]

\[ \text{structure} \]

\[ \text{CPTs} - P(X_i | \text{Pa}_{X_i}) \]

\[ \text{parameters} \]
Types of Errors

• Truth:

• Recovered:
Learning the structure of a BN

- **Constraint-based approach**
  - Test conditional independencies in data
  - Find an I-map

- **Score-based approach**
  - Finding a structure and parameters is a density estimation task
  - Evaluate model as we evaluated parameters
    - Maximum likelihood
    - Bayesian
    - etc.

Data

\[
\begin{align*}
\langle x_1^{(1)}, \ldots, x_n^{(1)} \rangle \\
\vdots \\
\langle x_1^{(m)}, \ldots, x_n^{(m)} \rangle
\end{align*}
\]
Score-based approach

Possible structures

Learn parameters

Score structure -52

Learn parameters

Score structure -60

Learn parameters

Score structure -500
How many graphs?

- N vertices.

- How many (undirected) graphs?

- How many (undirected) trees?
What’s a good score?

- Score(G) = log-likelihood(G : D, \( \theta_{\text{MLE}} \))
Information-theoretic interpretation of Maximum Likelihood Score

• Consider two node graph
  – Derived on board
Information-theoretic interpretation of Maximum Likelihood Score

- For a general graph $G$

$$ \log \hat{P}(\mathcal{D} \mid \theta, G) = m \sum_i \sum \hat{P}(x_i, \text{Pa}_{x_i}, G) \log \hat{P}(x_i \mid \text{Pa}_{x_i}, G) $$

$$ \log \hat{P}(\mathcal{D} \mid \theta, G) = m \sum_i \hat{I}(X_i, \text{Pa}_{X_i}) - m \sum_i \hat{H}(X_i) $$
Information-theoretic interpretation of Maximum Likelihood Score

\[ \log \hat{P}(D \mid \theta, G) = m \sum_i \hat{I}(X_i, Pa_{X_i}) - m \sum_i \hat{H}(X_i) \]

- Implications:
  - Intuitive: higher mutual info \( \rightarrow \) higher score
  - Decomposes over families in BN (node and it’s parents)
  - Same score for I-equivalent structures!
  - Information never hurts!
Chow-Liu tree learning algorithm 1

- For each pair of variables $X_i, X_j$
  - Compute empirical distribution:
    \[
    \hat{P}(x_i, x_j) = \frac{\text{Count}(x_i, x_j)}{m}
    \]
  - Compute mutual information:
    \[
    \hat{I}(X_i, X_j) = \sum_{x_i, x_j} \hat{P}(x_i, x_j) \log \frac{\hat{P}(x_i, x_j)}{\hat{P}(x_i)\hat{P}(x_j)}
    \]

- Define a graph
  - Nodes $X_1, \ldots, X_n$
  - Edge $(i, j)$ gets weight $\hat{I}(X_i, X_j)$
Chow-Liu tree learning algorithm 2

• Optimal tree BN
  – Compute maximum weight spanning tree
  – Directions in BN: pick any node as root, and direct edges away from root
    • breadth-first-search defines directions
Can we extend Chow-Liu?

- Tree augmented naïve Bayes (TAN) [Friedman et al. ’97]
  - Naïve Bayes model overcounts, because correlation between features not considered

  - Same as Chow-Liu, but score edges with:

\[
\hat{I}(X_i, X_j \mid C) = \sum_{c, x_i, x_j} \hat{P}(c, x_i, x_j) \log \frac{\hat{P}(x_i, x_j \mid c)}{\hat{P}(x_i \mid c) \hat{P}(x_j \mid c)}
\]