Recitation: Bayes Nets and Friends

Lyle Ungar
Heavily adapted from slides by Mitch Marcus
With contributions from Tony Liu

What's your favorite thing to do to unwind?
Recitation Plan

- Naïve Bayes Exercise
- LDA Example
- Bayes Net Exercises
- HMM Example
Recall: Naïve Bayes

- What’s the model?
- How do you estimate the parameters?
- How is NLP Naïve Bayes different?
Consider binary classification where $x$ has 2 binary features. How many parameters are there in a Naïve Bayes classifier?
Naïve Bayes Models

- **Different models**
  - $p(y|x) \sim p(x_1|y) \times p(x_2|y) \times \cdots \times p(x_p|y) \times p(y)$
  - $p(x_j|y)$ can be Bernoulli or Gaussian or …
Naïve Bayes: Parameter Estimation

- **MAP – why not MLE?**
  - \( P(\text{"apple"}|\text{class}) = \frac{(\#\text{docs in class with "apple"})}{(\#\text{docs in class})} \)

- **Uninformed prior (Laplace smoothing)**
  - Add a document with each word to each class
  - \( (\#\text{docs in class with "apple" } + 1 ) / (\#\text{docs in class } + v) \)

- **Informed prior (Empirical Bayes)**
  - Add prior counts of each word proportionally to their frequency
  - \( (\#\text{docs in class with "apple" } + m \cdot p("apple")) / (\#\text{docs in class } + m) \)
Naïve Bayes for NLP

- What additional assumption is made in Naïve Bayes for NLP?
 Seeking Life’s Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here, two genome researchers with radically different approaches presented complementary views of the basic genes needed for life.

One research team, using computer analyses to compare known genomes, concluded that today’s organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn’t be enough.

Although the numbers don’t match precisely, these predictions are not all that far apart, especially in comparison to the 75,000 genes in the human genome, notes Siv Anderson of the Mayo Clinic, who also attended the meeting. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. It may be a way of organizing any newly sequenced genome,” explains Aravind Subramaniam, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing any newly sequenced genome with a representative set of 100 human genomes...
Recall: The LDA Model

For each document,
- Choose the topic distribution $\theta \sim \text{Dirichlet}(\alpha)$
- For each of the N words $w_n$:
  - Choose a topic $z \sim \text{Multinomial}(\theta)$
  - Then choose a word $w_n \sim \text{Multinomial}(\beta_z)$
    - Where each topic has a different parameter vector $\beta$ for the words
LDA Parameter Estimation

◆ Given a corpus of documents, find the parameters $\alpha$ and $\beta$ which maximize the likelihood of the observed data (words in documents), marginalizing over the hidden variables $\theta$, $z$

◆ **E-step**:  
  • Compute $p(\theta, z|w, \alpha, \beta)$, the posterior of the hidden variables $(\theta, z)$ given each document $w$, and parameters $\alpha$ and $\beta$.

◆ **M-step**  
  • Estimate parameters $\alpha$ and $\beta$ given the current hidden variable distribution estimates

You don’t need to know the details;  
Only what is hidden and what is observed;  
And that EM works here.
LDA: True or False?

In LDA, the words in each document are assumed to be drawn from a Dirichlet distribution. These distributions can vary across documents.
Recall: Bayes Nets

- Local Markov Assumption
- Active Trails
- D Separation
Active Trails

A trail \( \{X_1, X_2, \ldots, X_k\} \) in the graph (no cycles) is an active trail if for each consecutive triplet in the trail:

\[ X_{i-1} \rightarrow X_i \rightarrow X_{i+1}, \text{ and } X_i \text{ is not observed} \]
\[ X_{i-1} \leftarrow X_i \leftarrow X_{i+1}, \text{ and } X_i \text{ is not observed} \]
\[ X_{i-1} \leftarrow X_i \rightarrow X_{i+1}, \text{ and } X_i \text{ is not observed} \]
\[ X_{i-1} \rightarrow X_i \leftarrow X_{i+1}, \text{ and } X_i \text{ is observed or one of its descendants is observed} \]

Variables connected by active trails are not conditionally independent.
D-separation

- Variables $X_i$ and $X_j$ are independent if there is no \textit{active trail} between $X_i$ and $X_j$.
  - given a set of observed variables $O \subseteq \{X_1, \ldots, X_m\}$
  - $O$ sometimes called a “Markov Blanket”
Bayes Net Exercises

I D-separates E and L ?
Bayes Net Exercises

C ⊥ D | F?
Bayes Net Exercises

\[ D \perp I \mid E, F, K ? \]
What is the minimum number of parameters needed to represent the full joint probability $P(A, B, C, D, E, F, G, H, I, J, K, L)$ in the above network if all the variables are binary?
How are most Bayes Nets built?

- Sequentially add nodes
- Stochastic gradient descent
- Interview experts for structure
Recall: Hidden Markov Models

- Markov assumption
- Model form and parameters
- Unrolling the model
Parameters of an HMM

- **States**: A set of states $S = s_1, \ldots, s_k$
- **Markov transition probabilities**: $A = a_{1,1}, a_{1,2}, \ldots, a_{k,k}$ Each $a_{i,j} = p(s_j \mid s_i)$ represents the probability of transitioning from state $s_i$ to $s_j$.
- **Emission probabilities**: A set $B$ of functions of the form $b_i(o_t) = p(o \mid s_i)$ giving the probability of observation $o_t$ being emitted by $s_i$
- **Initial state distribution**: the probability $\pi_i$ that $s_i$ is a start state
Markov Model Example

\[ S_1 = [0, 1] \]

<table>
<thead>
<tr>
<th>Today's Weather</th>
<th>Sunny</th>
<th>Rainy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sunny</td>
<td>0.8</td>
<td>0.2</td>
</tr>
<tr>
<td>Rainy</td>
<td>0.6</td>
<td>0.4</td>
</tr>
</tbody>
</table>

Markov Transition Matrix A

What is the expected value of \( s_1 \)?

What is the expected value of \( s_{1,000,000} \)?

Steady state at [0.75, 0.25]
(first eigenvector, with eigenvalue of 1)
Hidden Markov Model Example

\[ S_1 = [0.5, 0.5] \]

We observe: (umbrella, no umbrella)

We can ask questions like:
- What is the joint probability of the states (rain, sun) and our observations?

### Markov Transition Matrix A

<table>
<thead>
<tr>
<th>Today's Weather</th>
<th>Sunny</th>
<th>Rainy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sunny</td>
<td>0.8</td>
<td>0.2</td>
</tr>
<tr>
<td>Rainy</td>
<td>0.6</td>
<td>0.4</td>
</tr>
</tbody>
</table>

### Emission Probabilities B

<table>
<thead>
<tr>
<th>Weather</th>
<th>Sunny</th>
<th>Rainy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Umbrella</td>
<td>0.1</td>
<td>0.8</td>
</tr>
<tr>
<td>No Umbrella</td>
<td>0.9</td>
<td>0.2</td>
</tr>
</tbody>
</table>
HMM Exercise

True or False? The following statement about hidden Markov models holds for all $1 \leq t \leq T$ and $k$

\[ P(O_{t+1} = o_{t+1}, \ldots, O_T = o_T | O_1 = o_1, \ldots, O_t = o_t, S_t = k) \]
\[ = P(O_{t+1} = o_{t+1}, \ldots, O_T = o_T | S_t = k) \]