Machine Learning
4771
Instructor: Tony Jebara
Topic 7

- Review: Discrete, Bernoulli & Naïve Bayes
- Multinomial
- Text: Multinomial Counts
- Naïve Bayes Independence
- Preview: Graphical Models
- Preview: Examples of Graphical Models
- Continuous Probability Models
- Gaussian Distribution
- Structured Gaussian
Review: Discrete PDF Models

• **Bernoulli:** recall binary (coin flip) probability, just 1x2 table

\[ p(x) = \alpha^x (1 - \alpha)^{1-x} \quad \alpha \in [0,1] \quad x \in \{0,1\} \]

<table>
<thead>
<tr>
<th>( x = 0 )</th>
<th>( x = 1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.73</td>
<td>0.27</td>
</tr>
</tbody>
</table>

• **Multidimensional Probability Table:** multiple binary events

<table>
<thead>
<tr>
<th>( x_1 = 0 )</th>
<th>( x_1 = 1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_2 = 0 )</td>
<td>0.4</td>
</tr>
<tr>
<td>( x_2 = 1 )</td>
<td>0.1</td>
</tr>
</tbody>
</table>


• Why do we write these as an equations instead of tables?

• To do things like... maximum likelihood...
• Fill in the table so that it matches real data...

• Example: coin flips H,H,T,T,T,H,T,H,H ???

• Why is this correct?
Review: Text & Naïve Bayes

- Text classification: simplest model
- There are about 50,000 words in English
- Each document is 50,000 dimensional binary vector $x^i$
- Each dimension is a word, set to 1 if word in the document

$\begin{align*}
\text{Dim1: } & \text{“the” } = 1 \\
\text{Dim2: } & \text{“hello” } = 0 \\
\text{Dim3: } & \text{“and” } = 1 \\
\text{Dim4: } & \text{“happy” } = 1 \\
\end{align*}$

- Naïve Bayes: assumes each word is independent

$$p(x_1, x_2, \ldots, x_{50000}) = \prod_{d=1}^{50000} p(x_d) = \prod_{d=1}^{50000} \alpha_d^{x_d} (1 - \alpha_d)^{1-x_d}$$

- Each 1 dimension is a Bernoulli
- The whole vector is multivariate Bernoulli

$\in \begin{cases} 
\text{religion} \\
\text{politics} 
\end{cases}$
Review: Text & Naïve Bayes

• Maximum likelihood: assume we have several IID vectors
• Have i=1..N documents, each 50,000 dimensional binary
• Each dimension is a word, set to 1 if word in the document

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Word</th>
<th>x^1</th>
<th>x^2</th>
<th>x^3</th>
<th>x^4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dim1: “the”</td>
<td></td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Dim2: “hello”</td>
<td></td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Dim3: “and”</td>
<td></td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Dim4: “happy”</td>
<td></td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

• Likelihood: \( L(\theta) = \prod_{i=1}^{N} p(x^i) = \prod_{i=1}^{N} \prod_{d=1}^{50000} \alpha_d^{x^i_d} (1 - \alpha_d)^{(1-x^i_d)} \)

• Maximum likelihood solution: for each word count number of documents it appears in divided by total N documents

\[ \alpha_d = \frac{N_d}{N} \]

• Classification (later): build two models & compare \( p(x|\alpha): N \) religion docs > ? < \( p(x|\alpha'): N \) politics docs
Multinomial Probability Models

- **Multinomial**: beyond binary multi-category event (dice)

\[
p(x) = \prod_{m=1}^{M} \alpha_m^{x_m} \quad \sum_m \alpha_m = 1
\]

\[x \in \mathbb{B}^M \quad ; \quad \sum_m x_m = 1\]

- **Maximum Likelihood (IID)**:

\[
\sum_{i=1}^{N} \log p(x^i) = \sum_{i=1}^{N} \log \prod_{m=1}^{M} \alpha_m^{x_{m}^i} = \sum_{i=1}^{N} \sum_{m=1}^{M} x_{m}^i \log (\alpha_m)
\]

- Can’t just take gradient, constraint:

- Need to use Lagrange multipliers:

\[
\frac{\partial}{\partial \alpha_q} \left( \sum_{i=1}^{N} \sum_{m=1}^{M} x_{m}^i \log (\alpha_m) - \lambda \left( \sum_{m=1}^{M} \alpha_m - 1 \right) \right) = 0
\]

\[
\sum_{i=1}^{N} x_{m}^i \frac{1}{\alpha_q} - \lambda = 0
\]

\[
\alpha_q = \frac{\sum_{i=1}^{N} x_{m}^i}{\lambda}
\]
Multinomial Probability (ML)

• Had the gradient:

$$\alpha_q = \frac{\sum_{i=1}^{N} x_{q}^i}{\lambda}$$

Holds for each q...

• Recall the constraint:

$$\sum_{m} \alpha_m - 1 = 0$$

• Plug in $\alpha$’s solution:

$$\sum_{m} \frac{\sum_{i=1}^{N} x_{m}^i}{\lambda} - 1 = 0$$

$$\lambda = \sum_{m=1}^{M} \sum_{i=1}^{N} x_{m}^i = N$$

• Final answer:

$$\alpha_q = \frac{\sum_{i=1}^{N} x_{q}^i}{\lambda} = \frac{N_q}{N}$$

• Example: Rolling dice

1,6,2,6,3,6,4,6,5,6

\[
\begin{array}{ccccccc}
    x=1 & \times=2 & \times=3 & \times=4 & \times=5 & \times=6 \\
    0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.5 \\
\end{array}
\]
Text: Multinomial Counts

- **Multinomial**: can also *count many* multi-category events
  
  Dice: 1,3,1,4,6,1,1  Word Dice: the, dog, jumped, the

- If each (of 2000) words is IID multi-category event:
  
  \[ p(x) = p(x^1, x^2, \ldots, x^{2000}) = \prod_{w=1}^{2000} p(x^w) = \prod_{w=1}^{2000} \prod_{d=1}^{50000} \alpha_d^x \]

- Get count of each time an event occurred
  
  \[ p(x) = \prod_{w=1}^{2000} \prod_{d=1}^{50000} \alpha_d^{x^w} = \prod_{d=1}^{50000} \alpha_d^{\sum_{w=1}^{2000} x^w} = \prod_{d=1}^{50000} \alpha_d^{X_d} \]

- **BUT**: order shouldn’t matter when “counting” so multiply by # of possible orderings. Choosing \( X_1, \ldots X_M \) from \( N \)

  \[ \binom{N}{X_1, X_2, \ldots, X_M} = \frac{N!}{\prod_{m=1}^{M} X_m!} \]

- Bag-of-words model (only # of words matters, not order):

  \[ p(X) = \frac{\left( \sum_{m=1}^{M} X_m \right)!}{\prod_{m=1}^{M} X_m!} \prod_{m=1}^{M} \alpha_m^{X_m} \quad \sum_{m} \alpha_m = 1 \quad X \in \mathbb{Z}_+^{M} \]
**Text: Multinomial Counts**

- **Text classification: bag-of-words model**
- Each document is 50,000 dimensional vector
- Each dimension is a word, set to \# times word in doc

<table>
<thead>
<tr>
<th></th>
<th>$X^1$</th>
<th>$X^2$</th>
<th>$X^3$</th>
<th>$X^4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dim1: “the”</td>
<td>9</td>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Dim2: “hello”</td>
<td>0</td>
<td>5</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Dim3: “and”</td>
<td>6</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Dim4: “happy”</td>
<td>2</td>
<td>5</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

- Each document is a vector of multinomial counts

\[
p(X) = \frac{(\sum_{m=1}^{M} X_m)!}{\prod_{m=1}^{M} X_m!} \prod_{m=1}^{M} \alpha_m^{X_m} \quad \sum_{m} \alpha_m = 1 \quad X \in \mathbb{Z}^M_+
\]

- Likelihood:

\[
l(\alpha) = \sum_{i=1}^{N} \log p(X^i) = \sum_{i=1}^{N} \log \left( \frac{(\sum_{m=1}^{M} X^i_m)!}{\prod_{m=1}^{M} X^i_m!} \prod_{m=1}^{M} \alpha_m^{X^i_m} \right)
\]

\[\propto \sum_{i=1}^{N} \sum_{m=1}^{M} X^i_m \log (\alpha_m) \quad \text{same formula as Multinomial ML}\]
Text: Models Comparison

- For text modeling (McCallum & Nigam '98)
  Bernoulli better for small vocabulary
  Multinomial better for large vocabulary
Text: Newsgroup Recognition

- Model text from 12 newsgroups each with a multinomial
- Use speech recognizer to create a document of past 200 spoken words
- IBM ViaVoice
- 50% accuracy
- See which newsgroup is most similar
- 95%+ accuracy
- Project text from related newsgroup onto table
Structuring Probability Models

- Quickly explodes if we have many variables...

\[ p(x) = p(\text{flu}?, \text{headache}?, \ldots, \text{temperature}?) \]

- For D true/false “medical” variables \( \text{table size} = 2^D \)

- Exponential blow-up of the size of the probability table

- Example: 8x8 binary images of digits

- If multinomial with M choices, probabilities are how big?

- As in Naïve Bayes or Multivariate Bernoulli, if words were independent: much more efficient

\[ p(x) = p(\text{flu}?) p(\text{headache}?) \ldots p(\text{temperature}?) \]

\[
\begin{array}{cccc}
0.73 & 0.27 & 0.2 & 0.8 \\
0.54 & 0.46 & \\
\end{array}
\]

- For D true/false “medical” variables \( \text{table size} = 2 \times D \)

(really even less than that...)