• Binary label
  • $Y \sim \text{Bernoulli}(\pi)$
  • $\hat{\pi} = \frac{N_{Y=1}}{N}$
    • $N = \# \text{ of data points}$
    • $N_{Y=1} = \# \text{ of data points with label 1}$

• Binary features
  • $X_d | Y = y \sim \text{Bernoulli}(\theta_{d,y})$
  • $\hat{\theta}_{d,y} = \frac{N_{Y=y, X_d=1}}{N_{Y=y}}$
    • $N_{Y=y} = \# \text{ of data points with label } y$
    • $N_{Y=y, X_d=1} = \# \text{ of data points with label } y \text{ and feature } X_d = 1$
What if some Word-Label pair never appears in our training data?

Predictions

• Given a test data point \( x' = [x'_1, \ldots, x'_D]^T \)

\[
P(Y=1|x') \propto P(x'|Y=1)P(Y=1) = \left( \prod_{d=1}^{D} P(x'_d|Y=1) \right) P(Y=1)
\]

\[
\hat{x} = \left( \prod_{d=1}^{D} \hat{\theta}_{d,1} x'_d (1-\hat{\theta}_{d,1})^{1-x'_d} \right)^\hat{\pi} := \hat{p}_1
\]

\[
P(Y=0|x') \propto \left( \prod_{d=1}^{D} \hat{\theta}_{d,0} x'_d (1-\hat{\theta}_{d,0})^{1-x'_d} \right) (1-\hat{\pi}) := \hat{p}_0
\]

\[
\hat{y} = \begin{cases} 1 & \text{if } \hat{p}_1 \geq \hat{p}_0 \\
0 & \text{otherwise} \end{cases}
\]
What if some Word-Label pair never appears in our training data?

<table>
<thead>
<tr>
<th></th>
<th>$x_1$ (&quot;hat&quot;)</th>
<th>$x_2$ (&quot;cat&quot;)</th>
<th>$x_3$ (&quot;dog&quot;)</th>
<th>$x_4$ (&quot;fish&quot;)</th>
<th>$x_5$ (&quot;mom&quot;)</th>
<th>$x_6$ (&quot;dad&quot;)</th>
<th>$y$ (Dr. Seuss)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

The Cat in the Hat gets a Dog (by ???)

- If some $\hat{\theta}_{d,y} = 0$ and that word appears in our test data $x'$, then $P(Y = y|x') = 0$ even if all the other features in $x'$ point to the label being $y$!
- The model has been overfit to the training data
- We can address this with a prior over the parameters!
Setting the Parameters via MAP

- **Binary label**
  - \( Y \sim \text{Bernoulli}(\pi) \)
  - \( \hat{\pi} = \frac{N_{Y=1}}{N} \)
    - \( N = \# \text{ of data points} \)
    - \( N_{Y=1} = \# \text{ of data points with label 1} \)

- **Binary features**
  - \( X_d | Y = y \sim \text{Bernoulli}(\theta_{d,y}) \) and \( \theta_{d,y} \sim \text{Beta}(\alpha, \beta) \)
    - \( \hat{\theta}_{d,y} = \frac{N_{Y=y, X_d=1} + (\alpha - 1)}{N_{Y=y} + (\alpha - 1) + (\beta - 1)} \)
    - \( N_{Y=y} = \# \text{ of data points with label } y \)
    - \( N_{Y=y, X_d=1} = \# \text{ of data points with label } y \text{ and feature } X_d = 1 \)
  - Common choice: \( \alpha = 2, \beta = 2 \)