Naive Bayes

Recall: MLE: model \( P(x, y) \) as \( \hat{P}_\theta \) for \( \theta \in \Theta \)

\[ \hat{\theta}_{\text{MLE}} = \arg\max_{\theta \in \Theta} P(\theta | \text{Data}) \]

MAP: \( \hat{\theta}_{\text{MAP}} = \arg\max_{\theta \in \Theta} P(\theta | \text{Data}) = P(\text{Data} | \theta) \frac{1}{P(\theta)} \]

<table>
<thead>
<tr>
<th>( \Theta )</th>
<th>( Y )</th>
<th>( X = \text{Favorite dish} )</th>
<th>( X = { \text{Soup}, \text{Mac N Cheese}, \text{Tacos} } )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adult</td>
<td>Soup</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Child</td>
<td>Mac N Cheese</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Child</td>
<td>Mac N Cheese</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adult</td>
<td>Tacos</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adult</td>
<td>Soup</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Child</td>
<td>Tacos</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adult</td>
<td>Soup</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adult</td>
<td>Mac N Cheese</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Child</td>
<td>Mac N Cheese</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

What is the issue with this?

<table>
<thead>
<tr>
<th>( \Theta )</th>
<th>( Y )</th>
<th>( X_{CD1} = \text{Favorite dish} )</th>
<th>( X_{CD2} = \text{# words known} )</th>
<th>( \text{XCD3} = \text{movie} )</th>
<th>( \text{XCD4} = \text{hours of sleep} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adult</td>
<td>Soup</td>
<td>20000</td>
<td>Godfather</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>Child</td>
<td>Mac N Cheese</td>
<td>200</td>
<td>Frozen</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>Child</td>
<td>Mac N Cheese</td>
<td>400</td>
<td>Frozen</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>Adult</td>
<td>Tacos</td>
<td>17000</td>
<td>Eternals</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Adult</td>
<td>Soup</td>
<td>15000</td>
<td>Godfather</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Child</td>
<td>Tacos</td>
<td>1000</td>
<td>Eternals</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Adult</td>
<td>Soup</td>
<td>21000</td>
<td>Avengers</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Adult</td>
<td>Mac N Cheese</td>
<td>11000</td>
<td>Avengers</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>Child</td>
<td>Mac N Cheese</td>
<td>700</td>
<td>Avengers</td>
<td>11</td>
<td></td>
</tr>
</tbody>
</table>

\[ \hat{P}(Y = \text{Adult} | X = \{ \text{Soup}, 20000, \text{Avengers}, 8 \}) \]
Naive Bayes Model

Assumption: \[ P(x=x \mid y=y) = \prod_{d=1}^{D} P(x_{cd} = x_{cd} \mid y=y) \]

"Given it's a child's favorite dish, # words known, hours of sleep are independent"

Why is this useful?

\[
P(y=y \mid x=x) = \frac{P(x=x \mid y=y) P(y=y)}{P(x=x)}
\]

\[
= \prod_{d=1}^{D} \frac{P(x_{cd} = x_{cd} \mid y=y) P(y=y)}{P(x=x)}
\]

\[
= \prod_{d=1}^{D} P(x_{cd} = x_{cd} \mid y=y) P(y=y)
\]

\[
= \prod_{d=1}^{D} P(x_{cd} = x_{cd} \mid y=y) P(y=y)
\]

\[
\leq \sum_{y \in C} P(x_{cd} = x_{cd} \mid y \in C) P(y=c)
\]

\[P(y=y) \quad \text{and} \quad P(x_{cd} = x_{cd} \mid y=y)\]

are easy to estimate.

\[\text{Eg. Estimate}\]

\[\hat{P}(y = \text{adult} \mid x = \{\text{soup, 20,000, Avengers, 8}\})\]?

\[h(x) = \arg\max_{y \in C} \hat{P}(y=c \mid x=x)
\]

\[= \arg\max_{y \in C} \prod_{d=1}^{D} P(x_{cd} = x_{cd} \mid y=y) P(y=y)
\]

\[= \arg\max_{y \in C} \sum_{y \in C} P(y) P(x_{cd} = x_{cd} \mid y=y)
\]
When $X_{\alpha}^{(d)}$ are counts

$E_{\gamma} X_{\alpha}^{(d)} = \gamma$ means $\alpha$th word in the dictionary

occurs $\gamma$ times in the document $x_i$

$x$ is an $m$ word document: $X_{\alpha}^{(d)} \in \{0, 1, \ldots, m\}$

$\sum_{\alpha=1}^{A} X_{\alpha}^{(d)} = m$

Multinomial distribution:

$p(x=x_i | m, y=y) = \frac{m!}{x_{\alpha}^{(d)}! x_{\alpha+1}^{(d)}! \ldots x_{A}^{(d)}!} \prod_{\alpha=1}^{A} (\theta_{\alpha}^{(d)} x_{\alpha}^{(d)})$

MLE estimate:

$\hat{\theta}_{\alpha}^{(d)} = \frac{\sum_{i=1}^{D} I(y_i = c) x_{\alpha}^{(d)}}{\sum_{i=1}^{D} I(y_i = c) m_i}$

$m_i$: # words in document $i$, $\sum_{i=1}^{D} m_i = m$

(Map hallucinate examples per document)

$h(x) = \text{argmax}_{y \in C} p(y = y) \prod_{\alpha=1}^{A} \hat{\theta}_{\alpha}^{(d)} x_{\alpha}^{(d)}$

$X_{\alpha}^{(d)}$'s are Continuous variables Gaussian distribution conditioned on $y$

$p(x_i | y) = \frac{1}{\sqrt{2\pi \sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2 \sigma_y^2}\right)$

Parameter estimation:

$\hat{\mu}_y = \frac{\sum_{i=1}^{D} I(y_i = y) x_i}{\sum_{i=1}^{D} I(y_i = y)}$

$\hat{\sigma}_y = \sqrt{\frac{\sum_{i=1}^{D} I(y_i = y) (x_i - \hat{\mu}_y)^2}{\sum_{i=1}^{D} I(y_i = y)}}$

1. For both multinomial case and Gaussian case (with variance between class per feature fixed) classification boundary is linear.

2. For Gaussian case

$p(y = y | x) = \frac{1}{1 + \exp(-y(x^T \theta + b))}$

Logistic link function.