Naive Bayes

Say we have two tests for a specific disease. Each test \((f_1, f_2)\) can come back either positive “pos” or negative “neg”, and the true underlying condition of the patient is represented by \(y\) \((y = 1\) is “healthy” and \(y = 2\) is “disease”). We observe this training data where \(n = 7\) and \(p = 2\):

\[
\begin{array}{ccc}
\mathbf{x} & f_1 & f_2 & y \\
\mathbf{x}_1 & \text{pos} & \text{neg} & 1 \\
\mathbf{x}_2 & \text{pos} & \text{pos} & 2 \\
\mathbf{x}_3 & \text{pos} & \text{neg} & 2 \\
\mathbf{x}_4 & \text{neg} & \text{neg} & 1 \\
\mathbf{x}_5 & \text{pos} & \text{neg} & 2 \\
\mathbf{x}_6 & \text{neg} & \text{neg} & 1 \\
\mathbf{x}_7 & \text{neg} & \text{pos} & 2 \\
\end{array}
\]

1. To estimate the probability \(p(y = k)\), for \(k = 1, 2, \cdots, K\), we will use the formula:

\[
\theta_k = \frac{N_k + 1}{n + K}
\]

where \(N_k\) is the count (“Number”) of data points where \(y = k\). Compute \(\theta_1\) and \(\theta_2\). What would \(\theta_1\) and \(\theta_2\) be if we in fact had no training data?

2. To estimate the probabilities \(p(x_j = v | y = k)\) for all features \(j\), values \(v\), and class label \(k\), we will use the formula:

\[
\theta_{j,v,k} = \frac{N_{j,v,k} + 1}{N_k + |f_j|}
\]

where \(N_{j,v,k}\) is the count of data points where \(x_j = v\) and \(y = k\), and \(|f_j|\) is the number of possible values that \(f_j\) (feature \(j\)) can take on. Fill in the following tables with these \(\theta\) values.
3. Say we have a new data point $\mathbf{x}_{\text{test}} = \text{[neg, pos]}$. Our goal is to predict based on the Naive Bayes posterior probability:

$$p(y = k | \mathbf{x}) \propto p(y = k) \prod_{j=1}^{p} p(x_j | y = k).$$

In practice, we will compute this probability for each class $k$, based on our estimates ($\theta_k$ and $\theta_{j,v,k}$ terms). Then we will assign this data point the class label with maximum probability:

$$\hat{y} = \arg \max_{k \in \{1, 2, \ldots, K\}} p(y = k) \prod_{j=1}^{p} p(x_j | y = k).$$

For this $\mathbf{x}_{\text{test}}$, compute $p(y = 1 | \mathbf{x})$ and $p(y = 2 | \mathbf{x})$ and then assign a prediction label $\hat{y}$.

4. **Confusion Matrix.** Say for a test dataset with $m = 5$, we have these predictions and true labels:

<table>
<thead>
<tr>
<th>$\mathbf{x}_{\text{test}}$</th>
<th>$\hat{y}$</th>
<th>$y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathbf{x}_1$</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>$\mathbf{x}_2$</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>$\mathbf{x}_3$</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>$\mathbf{x}_4$</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>$\mathbf{x}_5$</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Draw a confusion matrix for this dataset with true labels on the rows and predicted labels on the columns. Normalize so that each row sums to 1. What would an *ideal* confusion matrix look like?