Bayes decision rule (recap)

Class \( C = \{1, \ldots, K\} \); \( C_k \) to denote \( c_k \); input features \( X = x \)

Most probable class: (maximum posterior class)

\[
k_{\text{max}} = \arg \max_{k} P(C_k | x) = \arg \max_{k} \frac{P(x | C_k) P(C_k)}{\sum_{k=1}^{K} P(x | C_k) P(C_k)}
\]

where \( P(C_k | x) \) : posterior
\( P(x | C_k) \) : likelihood
\( P(C_k) \) : prior

\( \Rightarrow \) Minimum error (misclassification) rate classification

(PRML C. M. Bishop (2006) Section 1.5)

**Avoiding the Curse of Dimensionality**

Apply the chain rule?

\[
P(x | C_k) = P(x_1, x_2, \ldots, x_D | C_k) = P(x_1 | C_k) P(x_2 | x_1, C_k) \cdots
\]

... reduction by PCA (Principal Component Analysis) / KL-transform

**More features!**?

\[
P(x | C_k) \approx \frac{n_{C_k}(x_1, \ldots, x_D)}{N_{C_k}}
\]

1D histogram

2D histogram

3D cube of numbers

... 100 binary variables, \( 2^{100} \) settings (the universe is \( 2^{100} \) picoseconds old)

In high dimensions almost all \( n_{C_k}(x_1, \ldots, x_D) \) are zero

\( \Rightarrow \) Bellman’s “curse of dimensionality”
Naive Bayes

Example - game played depending on the weather

<table>
<thead>
<tr>
<th>Outlook</th>
<th>Temperature</th>
<th>Humidity</th>
<th>Windy</th>
<th>Play</th>
</tr>
</thead>
<tbody>
<tr>
<td>sunny</td>
<td>hot</td>
<td>high</td>
<td>false</td>
<td>NO</td>
</tr>
<tr>
<td>sunny</td>
<td>hot</td>
<td>true</td>
<td>true</td>
<td>YES</td>
</tr>
<tr>
<td>rainy</td>
<td>mild</td>
<td>high</td>
<td>false</td>
<td>YES</td>
</tr>
<tr>
<td>rainy</td>
<td>cool</td>
<td>normal</td>
<td>true</td>
<td>NO</td>
</tr>
<tr>
<td>overcast</td>
<td>cool</td>
<td>normal</td>
<td>true</td>
<td>YES</td>
</tr>
<tr>
<td>sunny</td>
<td>mild</td>
<td>high</td>
<td>false</td>
<td>NO</td>
</tr>
<tr>
<td>sunny</td>
<td>cool</td>
<td>normal</td>
<td>true</td>
<td>NO</td>
</tr>
<tr>
<td>rainy</td>
<td>mild</td>
<td>normal</td>
<td>false</td>
<td>YES</td>
</tr>
<tr>
<td>sunny</td>
<td>mild</td>
<td>normal</td>
<td>true</td>
<td>YES</td>
</tr>
<tr>
<td>overcast</td>
<td>mild</td>
<td>high</td>
<td>true</td>
<td>YES</td>
</tr>
<tr>
<td>overcast</td>
<td>hot</td>
<td>normal</td>
<td>false</td>
<td>YES</td>
</tr>
<tr>
<td>rainy</td>
<td>mild</td>
<td>high</td>
<td>true</td>
<td>YES</td>
</tr>
</tbody>
</table>

\[ P(\text{Play} | O, T, H, W) = \frac{P(O, T, H, W | \text{Play}) P(\text{Play})}{P(O, T, H, W)} \]

Weather data - how to calculate probabilities?

\[ P(\text{Play} | O, T, H, W) = P(O, T, H, W | \text{Play}) P(\text{Play}) \]

If we use histograms for this 4D data:

\[ P(x | \text{Play} = Y) \times P(x | \text{Play} = N) \]

\# of bins in the histogram = 3 × 2 × 2 × 2 = 36

\# of samples available = 14. (9 for play:yes, 4 for play:no)

Weather data summary

Counts

<table>
<thead>
<tr>
<th>Outlook</th>
<th>Temperature</th>
<th>Humidity</th>
<th>Windy</th>
<th>Play</th>
</tr>
</thead>
<tbody>
<tr>
<td>sunny</td>
<td>hot</td>
<td>high</td>
<td>false</td>
<td>NO</td>
</tr>
<tr>
<td>sunny</td>
<td>hot</td>
<td>true</td>
<td>true</td>
<td>YES</td>
</tr>
<tr>
<td>rainy</td>
<td>mild</td>
<td>high</td>
<td>false</td>
<td>YES</td>
</tr>
<tr>
<td>rainy</td>
<td>cool</td>
<td>normal</td>
<td>true</td>
<td>NO</td>
</tr>
<tr>
<td>overcast</td>
<td>cool</td>
<td>normal</td>
<td>true</td>
<td>YES</td>
</tr>
<tr>
<td>sunny</td>
<td>mild</td>
<td>high</td>
<td>false</td>
<td>NO</td>
</tr>
<tr>
<td>sunny</td>
<td>cool</td>
<td>normal</td>
<td>true</td>
<td>NO</td>
</tr>
<tr>
<td>rainy</td>
<td>mild</td>
<td>normal</td>
<td>false</td>
<td>YES</td>
</tr>
<tr>
<td>sunny</td>
<td>mild</td>
<td>normal</td>
<td>true</td>
<td>YES</td>
</tr>
<tr>
<td>overcast</td>
<td>mild</td>
<td>high</td>
<td>true</td>
<td>YES</td>
</tr>
<tr>
<td>overcast</td>
<td>hot</td>
<td>normal</td>
<td>false</td>
<td>YES</td>
</tr>
<tr>
<td>rainy</td>
<td>mild</td>
<td>high</td>
<td>true</td>
<td>YES</td>
</tr>
</tbody>
</table>

Relative frequencies

\[ P(x | \text{Play} = Y) \times P(x | \text{Play} = N) \]

Test example

\[ x = (\text{sunny}, \text{cool}, \text{true}, \text{true})? \]

Naive Bayes properties

Easy and cheap:
Record counts, convert to frequencies, score each class by multiplying prior and likelihood terms

\[ P(C_k | x) \propto \left( \prod_{i=1}^{n} P(x_i | C_k) \right) P(C_k) \]

Statistically viable:
Simple count-based estimates work in 1D

Often overconfident:
Treats dependent evidence as independent
Exercise (past exam question)

Identifying Spam

The table gives a small dataset. Tick marks indicate which movies 3 children (marked c) and 4 adults (marked a) have watched. The final two rows give the movies watched by two users of the system of unknown age.

Apply maximum likelihood estimation of the priors and likelihoods to this data, using the naive Bayes assumption for the likelihoods. Hence find the odds that the test user y, is child: \( P(y = c|\text{data})/P(y = a|\text{data}) \) for \( i = 1, 2 \). State the MAP classification of each user.

Identifying Spam

Spam?

Dear Dr. Steve Renals, The proof for your article, Combining Spectral Representations for Large-Vocabulary Continuous Speech Recognition, is ready for your review. Please access your proof via the user ID and password provided below. Kindly log in to the website within 48 HOURS of receiving this message so that we may expedite the publication process.

Spam?

Congratulations to you as we bring to your notice, the results of the First Category draws of THE HOLLAND CASINO LOTTO PROMO INT. We are happy to inform you that you have emerged a winner over a large number of tickets purchased. This charge is for your country's information directory during my desperate search for someone who can assist me secretly and confidentially in relocating and managing some family fortunes.

Spam?

How can we identify an email as spam automatically?

Text classification: classify email messages as spam or non-spam (ham), based on the words they contain.

With the Bayes decision rule,
\[
P(\text{Spam} | x_1, \ldots, x_l) \propto P(x_1, \ldots, x_l | \text{Spam}) P(\text{Spam})
\]

Using the naive Bayes assumption,
\[
P(x_1, \ldots, x_l | \text{Spam}) = P(x_1 | \text{Spam}) \cdots P(x_l | \text{Spam})
\]