**Today’s Schedule**

- Bayes decision rule review
- The curse of dimensionality
- Naive Bayes
- Text classification using Naive Bayes (introduction)

**Fish classification**

Bayesian class estimation:

\[ P(C_k | x) = \frac{P(x | C_k) P(C_k)}{P(x)} \times P(x | C_k) P(C_k) \]

Estimating the terms: (Non-Bayesian)

- Priors: \( P(C = M) \approx \frac{N_M}{N_M + N_F} \)
- Likelihoods: \( P(x | C = M) \approx \frac{n(x)}{N_M} \)

NB: These approximations work well only if we have enough data

**Avoiding the Curse of Dimensionality**

Apply the chain rule?

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

\[ = P(x_1 | C_k) P(x_2 | x_1, C_k) P(x_3 | x_1, x_2, C_k) \ldots \]

\[ \cdots \]

\[ \cdots P(x_{D-1} | x_2, \ldots, x_D, C_k) P(x_D | x_0, \ldots, x_{D-1}, C_k) \]

Solution: assume structure in \( P(x | C_k) \)

For example,

- Assume \( x_{D+1} \) depends on \( x_0 \) only
  \[ P(x | C_k) \approx P(x_1 | C_k) P(x_2 | x_1, C_k) P(x_3 | x_1, x_2, C_k) \cdots P(x_D | x_0, \ldots, x_{D-1}, C_k) \]

- Assume \( x \in R^d \) distributes in a low dimensional vector space
  * Dimensionality reduction by PCA (Principal Component Analysis) / KL-transform

**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)

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

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

⇒ Minimum error (misclassification) rate classification
(PRML C. M. Bishop (2006) Section 1.5)
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>NO</td>
</tr>
<tr>
<td>overcast</td>
<td>hot</td>
<td>high</td>
<td>false</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>rainy</td>
<td>cool</td>
<td>normal</td>
<td>true</td>
<td>YES</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>false</td>
<td>YES</td>
</tr>
<tr>
<td>sunny</td>
<td>mild</td>
<td>normal</td>
<td>false</td>
<td>YES</td>
</tr>
<tr>
<td>sunny</td>
<td>hot</td>
<td>normal</td>
<td>true</td>
<td>YES</td>
</tr>
<tr>
<td>sunny</td>
<td>mild</td>
<td>high</td>
<td>true</td>
<td>YES</td>
</tr>
<tr>
<td>sunny</td>
<td>cool</td>
<td>normal</td>
<td>true</td>
<td>YES</td>
</tr>
</tbody>
</table>

Posterior prob. of "play" given \( x = (\text{sunny}, \text{cool}, \text{humid}, \text{windy}) \)

\[
P(\text{play} = \text{Y} \mid x) \propto P(x \mid \text{play}) P(\text{play})
\]

\[
P(\text{play} = \text{Y} \mid x) \propto P(O = \text{Y}) P(T = \text{c}) P(H = \text{h}) P(W = \text{t})
\]

\[
= \frac{2 \times 3 \times 3 \times 3}{9 \times 9 \times 9 \times 9} \approx 0.0053
\]

\[
P(x \mid \text{play}) = P(O = \text{Y}) P(T = \text{c}) P(H = \text{h}) P(W = \text{t})
\]

\[
= \frac{2 \times 3 \times 3 \times 3}{9 \times 9 \times 9 \times 9} \approx 0.0053
\]

\[
P(\text{play} = \text{Y} \mid x) \propto P(O = \text{Y}) P(T = \text{c}) P(H = \text{h}) P(W = \text{t})
\]

\[
= \frac{2 \times 3 \times 3 \times 3}{9 \times 9 \times 9 \times 9} \approx 0.0053
\]

\[
P(\text{play} = \text{Y} \mid x) \propto P(O = \text{Y}) P(T = \text{c}) P(H = \text{h}) P(W = \text{t})
\]

\[
= \frac{2 \times 3 \times 3 \times 3}{9 \times 9 \times 9 \times 9} \approx 0.0053
\]

Naive Bayes properties

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

\[
P(C_i \mid x) \propto \prod_{k=1}^{D} P(x_k \mid C_i)
\]

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

Often overconfident:
Treats dependent evidence as independent

Exercise: find the odds of play, \( P(\text{play} = \text{Y} \mid x) / P(\text{play} = \text{N} \mid x) \) (answer in notes)
Another approach for the weather example

- What about applying k-NN?
- Data representation (by quantification)

\[
X = \begin{pmatrix}
3 & 3 & 2 & 0 & 0 \\
2 & 3 & 2 & 0 & 1 \\
1 & 1 & 1 & 1 & 1 \\
1 & 2 & 1 & 0 & 1 \\
3 & 2 & 2 & 0 & 0 \\
3 & 1 & 1 & 0 & 1 \\
2 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 \\
1 & 2 & 1 & 0 & 1 \\
3 & 2 & 1 & 1 & 1 \\
3 & 1 & 1 & 0 & 1 \\
1 & 2 & 2 & 1 & 0 \\
1 & 2 & 2 & 1 & 1 \\
\end{pmatrix}
\]

Rank dist. idc label
1 1.41 (7) Y
2 1.41 (8) N
3 1.41 (9) Y
4 1.41 (10) Y
5 1.41 (11) Y
6 2.00 (2) N
7 2.24 (1) N
8 2.24 (6) N
9 2.24 (14) N
10 2.45 (3) Y
11 2.45 (4) Y
12 2.45 (5) Y
13 2.65 (10) Y
14 2.65 (13) Y

Another approach for the weather example (cont.)

- k-NN
- Correlation matrix for \((O, T, H, W, P)\)

\[
\begin{pmatrix}
1 & 0.49 & 0.03 & -0.35 & -0.49 \\
0.49 & 1 & 0.49 & -0.03 & -0.14 \\
0.03 & 0.49 & 1 & -0.49 & -0.49 \\
-0.35 & -0.03 & -0.49 & 1 & 0.14 \\
-0.49 & -0.14 & -0.49 & 0.14 & 1
\end{pmatrix}
\]

Dimensionality reduction by PCA

Exercise (past exam question)

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: \(P(y = c|data)/P(y = a|data)\) for \(i = 1, 2\). State the MAP classification of each user.

Identifying Spam

- Spam?

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

- Spam?

\[\text{Congratulations to you as we bring to your notice, the results of the First Category draws of THE HOLLAND CASINO LOTTO PROMO INT. We are happy to inform you that you have emerged a winner under the First Category, which is part of our promotional draws. So that we may expedite the publication process} \]

Question

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,

\[P(x_1, \ldots, x_L | \text{Spam}) = P(x_1 | \text{Spam}) \cdots P(x_L | \text{Spam}) \]
Summary

- The curse of dimensionality
- Approximation by the Naive Bayes rule
- Example: classifying multidimensional data using Naive Bayes
- Next lecture: Text classification using Naive Bayes