Naïve Bayes
Naïve Bayes derivation

- Predict $\text{argmax}_t P(t \mid x)$
  
  $= \text{argmax}_t P(x \mid t) P(t) / P(x)$
  $= \text{argmax}_t P(x \mid t) P(t)$

- Naïve independence assumption
  
  $P(x \mid t) = \prod_j P(x_j \mid t)$

- Predict the label $t$ maximizing
  
  $P(t) \prod_j P(x_j \mid t)$

- Uses generative model: pick $t$ then generate $x$ using $t$
Naïve Bayes example using max likelihood estimates (empirical counts)

• Data: \((\text{boolean})\)

\[
\begin{array}{ccc}
\text{x} & \text{t} & \\
T,T & +1 & \\
T,F & +1 & \\
F,T & +1 & \\
F,T & +1 & \\
F,F & -1 & \\
T,F & -1 & \\
F,T & -1 & \\
\end{array}
\]

• Predict on \(x=(T,F)\) using max likelihood estimates from data

\[
P(t=+1) = \frac{4}{7}; \quad P(t=-1) = \frac{3}{7}
\]

\[
P(x_1=T \mid t=+1) = \frac{1}{2}
\]

\[
P(x_2=F \mid t=+1) = \frac{1}{4}
\]

\[
P(x_1=T \mid t=-1) = \frac{1}{3}
\]

\[
P(x_2=F \mid t=-1) = \frac{2}{3}
\]

For “+1”: \((4/7)(1/2)(1/4) = 1/14\)

For “-1”: \((3/7)(1/3)(2/3) = 2/21\)

Predict “-1”
Naïve Bayes example using max likelihood estimates

- Data: (boolean)

<table>
<thead>
<tr>
<th>x</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>T,T</td>
<td>+1</td>
</tr>
<tr>
<td>T,F</td>
<td>+1</td>
</tr>
<tr>
<td>F,T</td>
<td>+1</td>
</tr>
<tr>
<td>F,T</td>
<td>+1</td>
</tr>
<tr>
<td>F,F</td>
<td>-1</td>
</tr>
<tr>
<td>T,T</td>
<td>-1</td>
</tr>
<tr>
<td>F,F</td>
<td>-1</td>
</tr>
</tbody>
</table>

- Predict on \( x=(T,F) \) using max likelihood estimates from data

\[
P(t = +1) = \frac{4}{7}; \quad P(t = -1) = \frac{3}{7}
\]

\[
P(x_1=T \mid t=+1) = \frac{1}{2}
\]

\[
P(x_2=F \mid t=+1) = \frac{1}{4}
\]

\[
P(x_1=T \mid t=-1) = \frac{1}{3}
\]

\[
P(x_2=F \mid t=-1) = \frac{2}{3}
\]

For “+1”: \( (4/7)(1/2)(1/4) = 1/14 \)

For “-1”: \( (3/7)(1/3)(2/3) = 2/21 \)

Predict “-1”, even on +1 example!
Naïve Bayes discussion

• Straight from data, no searching
  – But need to estimate class conditional prob’s – the probabilities of words given the class

• Successful applications:
  – Diagnosis,
  – Classifying text (Joachims, 1996) 89% accuracy for identifying source from 20 newsgroups (1000 documents each group, 2/3 train 1/3 test)
  – Newsweeder (Lang, 1995) interesting articles up from 16% to 59% after filtering
Newsgroup classification accuracy vs training size (Mitchell)
Naïve Bayes Issues

1. Conditional independence optimistic, but…
   Don’t have to get probabilities right, just the predictions
2. What if an attributeValue-label pair not in training set?
   • Use Laplace estimation.
3. Numeric Features: use Gaussian or other density (Poisson, exponential)
4. Attributes for text classification?
   • Bag of words model
Naïve Bayes for Text
(see Mitchell’s book)

• Let $V$ be the vocabulary (all words/symbols in all training documents)
• For each class $t$, let $\text{Docs}_t$ be the concatenation of all docs labeled $t$
• For each word $w$ in $V$, let $\#w(\text{Docs}_t)$ be # of times $w$ occurs in $\text{Docs}_t$
• Set $P(w \mid t)$ to:
  $$(\#w(\text{docs}_t) + 1) / (|V| + \sum_w \#w(\text{docs}_y))$$
Naïve bayes for text (2)

• Predict on new document $x$ with class $t$ maximizing

$$P(t) \prod_{w \in x} P(w \mid t)$$

Note: repeated words multiplied in multiple times
boundary (boolean features)

• For \( \mathbf{x} \), the \( y \) maximizing: \( P(y) \prod_j P(x_j \mid y) \)

Also maximizes: \( \log(P(y)) + \sum_j \log(P(x_j \mid y)) \)

• Let \( a_j = \log(P(x_j = 1 \mid y)); b_j = \log(P(x_j = 0 \mid y)) \)

\[
\sum_j \log(P(x_j \mid y)) = \sum_j (a_j x_j + b_j(1-x_j))
= \sum_j x_j(a_j-b_j) + \sum_j b_j
= \mathbf{w}_y \cdot \mathbf{x} + c_y
\]

• So predict with the class maximizing a set of linear functions - a LTU for two class with boolean features.
Exercise:

• Repeat slide 3 example using Laplacian probability estimates. Calculate the “vote” for each of the two classes for the new instance \( x = (T,F) \).

• Use Naïve Bayes in Weka for iris2.arff (iris.arff)

• Data: (boolean)
  
  T,T  +1
  T,F  +1
  F,T  +1
  F,F  -1
  T,F  -1
  F,T  -1