Naive Bayes
Statistical Machine Learning

● Discriminative
  ○ $P(Y \mid X)$
  ○ Predict $Y$ given $X$ directly based on observed evidence
  ○ Linear Regression, Logistic Regression

● Generative
  ○ $P(Y \mid X) = \frac{P(X \mid Y) P(Y)}{P(X)}$
    ■ $P(X \mid Y)$ is called the likelihood function.
    ■ $P(Y)$ is called the prior.
    ■ $P(Y \mid X)$ is called the posterior.
  ○ Naive Bayes!
Naive Bayes

- By Bayes’ rule, \( P(Y \mid X) = \frac{P(X \mid Y) \, P(Y)}{P(X)} \)
- So, for all classes \( y \), calculate \( P(X \mid Y) \, P(Y = y) \)
- Choose a class \( y \) s.t. \( P(X \mid Y) \, P(Y = y) \) is maximal
- N.B. There is no need to calculate \( P(X) \) since it is constant for all classes \( y \): i.e., it does not depend on \( y \)

Present Goal:

- For all classes \( y \), calculate \( P(X \mid Y) \, P(Y = y) \)
Naive Bayes Assumption

\[ P(X \mid Y) \]
\[ = P(X_1, \ldots, X_n \mid Y) \]
\[ = \prod_i P(X_i \mid Y) \]

\[ P(X \mid Y) \]
\[ = P(\text{Outlook} = \text{Sunny}, \ldots, \text{Wind} = \text{Strong} \mid Y = \text{Yes}) \]
\[ = P(\text{Outlook} = \text{Sunny} \mid Y = \text{Yes}) \ldots P(\text{Wind} = \text{Strong} \mid Y = \text{Yes}) \]

Revised Goal:

- For all classes \( y \), calculate \( \prod_i P(X_i \mid Y) P(Y = y) \)
### Training data:

<table>
<thead>
<tr>
<th>Day</th>
<th>Outlook</th>
<th>Temp</th>
<th>Humidity</th>
<th>Wind</th>
<th>PlayTennis</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1</td>
<td>Sunny</td>
<td>Hot</td>
<td>High</td>
<td>Weak</td>
<td>No</td>
</tr>
<tr>
<td>D2</td>
<td>Sunny</td>
<td>Hot</td>
<td>High</td>
<td>Strong</td>
<td>No</td>
</tr>
<tr>
<td>D3</td>
<td>Overcast</td>
<td>Hot</td>
<td>High</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D4</td>
<td>Rain</td>
<td>Mild</td>
<td>High</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D5</td>
<td>Rain</td>
<td>Cool</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D6</td>
<td>Rain</td>
<td>Cool</td>
<td>Normal</td>
<td>Strong</td>
<td>No</td>
</tr>
<tr>
<td>D7</td>
<td>Overcast</td>
<td>Cool</td>
<td>Normal</td>
<td>Strong</td>
<td>Yes</td>
</tr>
<tr>
<td>D8</td>
<td>Sunny</td>
<td>Mild</td>
<td>High</td>
<td>Weak</td>
<td>No</td>
</tr>
<tr>
<td>D9</td>
<td>Sunny</td>
<td>Cool</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D10</td>
<td>Rain</td>
<td>Mild</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D11</td>
<td>Sunny</td>
<td>Mild</td>
<td>Normal</td>
<td>Strong</td>
<td>Yes</td>
</tr>
<tr>
<td>D12</td>
<td>Overcast</td>
<td>Mild</td>
<td>High</td>
<td>Strong</td>
<td>Yes</td>
</tr>
<tr>
<td>D13</td>
<td>Overcast</td>
<td>Hot</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D14</td>
<td>Rain</td>
<td>Mild</td>
<td>High</td>
<td>Strong</td>
<td>No</td>
</tr>
</tbody>
</table>

### Test data:

<table>
<thead>
<tr>
<th>D15</th>
<th>Sunny</th>
<th>Cool</th>
<th>High</th>
<th>Strong</th>
<th>?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Event</td>
<td>Yes</td>
<td>No</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------------</td>
<td>------</td>
<td>------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P(Yes)</td>
<td>0.64</td>
<td>0.36</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P(No)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P(Weak</td>
<td>Yes)</td>
<td>0.67</td>
<td>0.40</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P(Strong</td>
<td>Yes)</td>
<td>0.33</td>
<td>0.60</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P(High</td>
<td>Yes)</td>
<td>0.33</td>
<td>0.80</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P(Normal</td>
<td>Yes)</td>
<td>0.67</td>
<td>0.20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P(Hot</td>
<td>Yes)</td>
<td>0.22</td>
<td>0.20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P(Mild</td>
<td>Yes)</td>
<td>0.44</td>
<td>0.40</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P(Cool</td>
<td>Yes)</td>
<td>0.33</td>
<td>0.40</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P(Sunny</td>
<td>Yes)</td>
<td>0.22</td>
<td>0.60</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P(Overcast</td>
<td>Yes)</td>
<td>0.44</td>
<td>0.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P(Rain</td>
<td>Yes)</td>
<td>0.33</td>
<td>0.40</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
\[ X = (\text{Sunny, Cool, High, Strong}) \]

- \[ P(\text{Sunny} \mid \text{Yes}) \cdot P(\text{Cool} \mid \text{Yes}) \cdot P(\text{High} \mid \text{Yes}) \cdot P(\text{Strong} \mid \text{Yes}) \cdot P(\text{Yes}) \]
  \[ = (0.22) \cdot (0.33) \cdot (0.33) \cdot (0.33) \cdot (0.64) = 0.0051 \]

- \[ P(\text{Sunny} \mid \text{No}) \cdot P(\text{Cool} \mid \text{No}) \cdot P(\text{High} \mid \text{No}) \cdot P(\text{Strong} \mid \text{No}) \cdot P(\text{No}) \]
  \[ = (0.34) \cdot (0.60) \cdot (0.40) \cdot (0.80) \cdot (0.60) = 0.039 \]

\[ P(X \mid \text{No}) > P(X \mid \text{Yes}) \]

So our NB classifier outputs **No**
Naive Bayes in R
Naive Bayes in R

You must install the `e1071` package to use Naive Bayes in R

```r
> install.packages('e1071')
> library(e1071)
> classifier <- naiveBayes(training_data, labels)
> predict(classifier, testing_data)
```
Extras
Estimating Probabilities

We estimate the requisite probabilities by counting
Estimating Probabilities

\[
P(\bullet) = \frac{7}{10} \quad P(\bullet) = \frac{3}{10} \quad P(+)= \frac{7}{10} \quad P(-)= \frac{3}{10}
\]

\[
P(\bullet|+) = \frac{6}{7} \quad P(\bullet|-)= \frac{1}{3} \quad P(\bullet|+) = \frac{1}{7} \quad P(\bullet|-) = \frac{2}{3}
\]

\[
P(+|\bullet) = \frac{6}{7} \quad P(-|\bullet) = \frac{1}{7} \quad P(+|\bullet) = \frac{1}{3} \quad P(-|\bullet) = \frac{2}{3}
\]
Estimating Probabilities

\[ P(\bullet) = \frac{7}{10} \quad P(\bullet) = \frac{3}{10} \quad P(+) = \frac{7}{10} \quad P(-) = \frac{3}{10} \]

\[ P(\bullet | +) = \frac{6}{7} \quad P(\bullet | -) = \frac{1}{3} \quad P(\bullet | +) = \frac{1}{7} \quad P(\bullet | -) = \frac{2}{3} \]

\[ P(+ | \bullet) = \frac{6}{7} \quad P(- | \bullet) = \frac{1}{7} \quad P(+ | \bullet) = \frac{1}{3} \quad P(- | \bullet) = \frac{2}{3} \]

Great!

Now how do we classify a new \( \bullet \)?
Estimating Probabilities

\[
P(\bullet) = \frac{7}{10} \quad P(\bullet) = \frac{3}{10} \quad P(+) = \frac{7}{10} \quad P(-) = \frac{3}{10}
\]

\[
P(\bullet | +) = \frac{6}{7} \quad P(\bullet | -) = \frac{1}{3} \quad P(\bullet | +) = \frac{1}{7} \quad P(\bullet | -) = \frac{2}{3}
\]

\[
P(+ | \bullet) = \frac{6}{7} \quad P(- | \bullet) = \frac{1}{7} \quad P(+ | \bullet) = \frac{1}{3} \quad P(- | \bullet) = \frac{2}{3}
\]

\[
P(+ | \bullet) > P(- | \bullet).
\]
So \(\bullet\) is classified as a +.
Bayes’ Optimal Classifier

- Let A be a random variable ranging over the set of possible observations
- Let B be a random variable whose range is the classes
- \( P[A] \) is the prior probability of the observed data (\( A = a \))
- \( P[B] \) is the prior probability of each class (\( B = b \))
- \( P[B \mid A] \) is the probability of class B, given data A
- \( P[A \mid B] \) is the probability of data A, given class B
- We can learn these probabilities from data
- We can then classify using the maximum a posteriori (MAP) principle:
  - Given the data and a new observation, choose a class for which \( P[B \mid A] \) is maximized
  - I.e., by Bayes’ theorem, choose a B that maximizes \( P[A \mid B] \cdot P[B] \)