Naive Bayes
Supervised Learning Models

- **Discriminative Model**
  - Learns $P(Y \mid X)$ directly
    - Solves an optimization problem to minimize error / maximize accuracy
  - Example: Logistic Regression

- **Generative Model**
  - Takes a more circuitous route
    - $P(Y \mid X) = P(X, Y) / P(X) = P(X \mid Y) P(Y) / P(X)$
    - Depends on prior knowledge: $P(Y)$ is called the **prior**
    - Requires an intermediate step: Calculate the **likelihood**, $Pr[X \mid Y]$
    - Finally, apply Bayes’ rule to calculate $Pr[Y \mid X]$
  - Example: Naive Bayes
MNIST Database
MNIST Engineered Generative Model

- 0 is a loop
- 8 is two loops
- 1 is a line
- Etc.

(M. Kamvysselis, “Digit recognition in curvature space”, 1999)
MNIST Learned Generative Model

Original Images

Generated Images

Image Source
Maximum *a Posteriori* (MAP) Principle

- Let $X$ be a random variable ranging over the set of possible observations
- Let $C$ be a random variable whose range is the set of classes
- $P[C \mid X]$ is the probability of class $C$, given data $X$
- To classify, using MAP principle: Choose a class $c$ that maximizes $P[C \mid X]$
Maximum *a Posteriori* (MAP) Principle

- Let $X$ be a random variable ranging over the set of possible observations
- Let $C$ be a random variable whose range is the set of classes
- $P[C \mid X]$ is the probability of class $C$, given data $X$
- To classify, using MAP principle: Choose a class $c$ that maximizes $P[C \mid X]$

$$c_{MAP} \in \arg\max_{c \in C} P(C \mid X)$$
$$= \arg\max_{c \in C} \frac{P(X \mid C) P(C)}{P(X)}$$
$$\propto \arg\max_{c \in C} P(X \mid C) P(C)$$
MAP Principle

- By Bayes’ rule, \( P[Y \mid X] = P[X \mid Y] \cdot P[Y] / P[X] \)
- So, for all classes \( y \), calculate \( P[X \mid Y] \cdot P[Y = y] \)
- Choose a class \( y \) s.t. \( P[X \mid Y] \cdot P[Y = y] \) 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 \( P[X \mid Y] \cdot P[Y = y] \)
Naive Bayes Assumption

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

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

Revised Goal:

- For all classes \( y \), calculate \( \prod_i P[X_i \mid Y] \ P[Y = y] \)
Conditional Independence

\[ \Pr(R \cap B \mid Y) = \Pr(R \mid Y) \Pr(B \mid Y) \]
## Training data:

<table>
<thead>
<tr>
<th>Day</th>
<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>1</td>
<td>Sunny</td>
<td>Hot</td>
<td>High</td>
<td>Weak</td>
<td>No</td>
</tr>
<tr>
<td>D2</td>
<td>2</td>
<td>Sunny</td>
<td>Hot</td>
<td>High</td>
<td>Strong</td>
<td>No</td>
</tr>
<tr>
<td>D3</td>
<td>3</td>
<td>Overcast</td>
<td>Hot</td>
<td>High</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D4</td>
<td>4</td>
<td>Rain</td>
<td>Mild</td>
<td>High</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D5</td>
<td>5</td>
<td>Rain</td>
<td>Cool</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D6</td>
<td>6</td>
<td>Rain</td>
<td>Cool</td>
<td>Normal</td>
<td>Strong</td>
<td>No</td>
</tr>
<tr>
<td>D7</td>
<td>7</td>
<td>Overcast</td>
<td>Cool</td>
<td>Normal</td>
<td>Strong</td>
<td>Yes</td>
</tr>
<tr>
<td>D8</td>
<td>8</td>
<td>Sunny</td>
<td>Mild</td>
<td>High</td>
<td>Weak</td>
<td>No</td>
</tr>
<tr>
<td>D9</td>
<td>9</td>
<td>Sunny</td>
<td>Cool</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D10</td>
<td>10</td>
<td>Rain</td>
<td>Mild</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D11</td>
<td>11</td>
<td>Sunny</td>
<td>Mild</td>
<td>Normal</td>
<td>Strong</td>
<td>Yes</td>
</tr>
<tr>
<td>D12</td>
<td>12</td>
<td>Overcast</td>
<td>Mild</td>
<td>High</td>
<td>Strong</td>
<td>Yes</td>
</tr>
<tr>
<td>D13</td>
<td>13</td>
<td>Overcast</td>
<td>Hot</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D14</td>
<td>14</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>Day</th>
<th>Day</th>
<th>Outlook</th>
<th>Temp</th>
<th>Humidity</th>
<th>Wind</th>
</tr>
</thead>
<tbody>
<tr>
<td>D15</td>
<td>15</td>
<td>Sunny</td>
<td>Cool</td>
<td>High</td>
<td>Strong</td>
</tr>
<tr>
<td>Event</td>
<td>Probability if Yes</td>
<td>Probability if 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>Yes</td>
<td>0.64</td>
<td>0.36</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weak</td>
<td>0.67</td>
<td>0.40</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Strong</td>
<td>0.33</td>
<td>0.60</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>0.33</td>
<td>0.80</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normal</td>
<td>0.67</td>
<td>0.20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hot</td>
<td>0.22</td>
<td>0.40</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mild</td>
<td>0.44</td>
<td>0.40</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cool</td>
<td>0.33</td>
<td>0.20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sunny</td>
<td>0.22</td>
<td>0.60</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overcast</td>
<td>0.44</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rain</td>
<td>0.33</td>
<td>0.40</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
\[ X = [\text{Sunny}, \text{Cool}, \text{High}, \text{Strong}] \]

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

- \( P[\text{Sunny} \mid \text{No}] \cdot P[\text{Cool} \mid \text{No}] \cdot P[\text{High} \mid \text{No}] \cdot P[\text{Strong} \mid \text{No}] \cdot P[\text{No}] \)
  
  \[ = [0.60] \cdot [0.20] \cdot [0.80] \cdot [0.60] \cdot [0.36] = 0.021 \]

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

So our NB classifier outputs \textit{No}
MLE for Naive Bayes

\[
\begin{align*}
\arg\max_{\theta} & \prod_{i=1}^{n} P(x_i, y_i \mid \theta) \\
= & \arg\max_{\theta} \prod_{i=1}^{n} P(y_i \mid \theta) P(x_i \mid y_i \mid \theta) \\
= & \arg\max_{\theta} \prod_{i=1}^{n} P(y_i \mid \theta) \prod_{j=1}^{m} P(x_{ij} \mid y_i, \theta) \\
= & \arg\max_{\theta} \log \prod_{i=1}^{n} P(y_i \mid \theta) \prod_{j=1}^{m} P(x_{ij} \mid y_i, \theta) \\
= & \arg\max_{\theta} \sum_{i=1}^{n} \log P(y_i \mid \theta) + \sum_{i=1}^{n} \sum_{j=1}^{m} \log P(x_{ij} \mid y_i, \theta)
\end{align*}
\]
MLE for Naive Bayes (cont’d)

Likelihood of class labels:

\[ P(y_i = 1) = p \quad \text{and} \quad P(y_i = 0) = 1 - p \]

\[ P(y_i | \theta) = p[y_i=1](1 - p)[y_i=0] \]
MLE for Naive Bayes (cont’d)

Likelihood of feature values, given class labels:

\[
P(x_{ij} = 1 \mid y_i = 1) = a_j \quad \text{and} \quad P(x_{ij} = 0 \mid y_i = 1) = 1 - a_j
\]

\[
P(x_{ij} = 1 \mid y_i = 0) = b_j \quad \text{and} \quad P(x_{ij} = 0 \mid y_i = 0) = 1 - b_j
\]

\[
P(x_{ij} \mid y_i, \theta) = a_j^{[y_i=1, x_{ij}=1]} (1 - a_j)^{[y_i=1, x_{ij}=0]} b_j^{[y_i=0, x_{ij}=1]} (1 - b_j)^{[y_i=0, x_{ij}=0]}
\]
MLE for Naive Bayes (cont’d)

Plug $P(y_i \mid \theta)$ and $P(x_{ij}, y_i \mid \theta)$ back into the log likelihood function.

Use calculus to solve for the optimal $\theta$: i.e., take derivatives & set equal to zero.

\[
p = \frac{\text{Count } (y_i = 1)}{\text{Count } (y_i = 1) + \text{Count } (y_i = 0)}
\]

\[
a_j = \frac{\text{Count } (y_i = 1, x_{ij} = 1)}{\text{Count } (y_i = 1)}
\]

\[
b_j = \frac{\text{Count } (y_i = 0, x_{ij} = 1)}{\text{Count } (y_i = 0)}
\]
Back to MNIST: Bernoulli Model

The parameter vector $\theta$ consists of $p$, $a_j$, and $b_j$, for all $j \in \{1, \ldots, m\}$.

There are 784 features, so there are 784 $a_j$ and 784 $b_j$ parameters, so 1569 in total.

In the full joint, there are $2^{784}$ parameters.

15% error rate on 10,000 test images
Back to MNIST: Gaussian Model

< 5% error rate on 10,000 test images
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
Naive Bayes

We estimate the requisite probabilities by counting.
Naive Bayes (cont’d)

We estimate the requisite probabilities by counting.

\[ P[\bullet] = ? \quad P[\bullet] = ? \quad P[+] = ? \quad P[-] = ? \]

\[ P[\bullet | +] = ? \quad P[\bullet | -] = ? \quad P[\bullet | +] = ? \quad P[\bullet | -] = ? \]
Naive Bayes (cont’d)

We estimate the requisite probabilities by counting.

\[
P[\cdot] = \frac{7}{10} \quad P[\cdot] = \frac{3}{10} \quad P[+] = \frac{7}{10} \quad P[-] = \frac{3}{10}
\]

\[
P[\cdot | +] = \frac{6}{7} \quad P[\cdot | -] = \frac{1}{3} \quad P[\cdot | +] = \frac{1}{7} \quad P[\cdot | -] = \frac{2}{3}
\]
Naive Bayes (cont’d)

We estimate the requisite probabilities by counting.

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

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

Great!

Now how do we classify a new \( \bullet \)?
Naive Bayes (cont’d)

We estimate the requisite probabilities by counting.

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

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

\[ P[+ | \bullet] = P[\bullet | +] P[+] = \left(\frac{6}{7}\right) \left(\frac{7}{10}\right) = \frac{6}{10}. \]
\[ P[- | \bullet] = P[\bullet | -] P[-] = \left(\frac{1}{3}\right) \left(\frac{3}{10}\right) = \frac{1}{10}. \]

So \( \bullet \) is classified as a +.
Naive Bayes (cont’d)

We estimate the requisite probabilities by counting.

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

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

\[
P[+ | \bullet] = P[\bullet | +] P[+] = (\frac{1}{7})(\frac{7}{10}) = \frac{1}{10}.
\]

\[
P[- | \bullet] = P[\bullet | -] P[-] = (\frac{2}{3})(\frac{3}{10}) = \frac{2}{10}.
\]

So \(\bullet\) is classified as a -.
Sanity Check

We estimate the requisite probabilities by counting.

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

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

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

\[ P[+ | \bullet] > P[- | \cdot], \text{ so } \bullet \text{ is classified as a } +, \text{ and } P[- | \cdot] > P[+ | \cdot] \text{ so } \cdot \text{ is classified as a } -. \]