You want to build a classifier for new customers by learning \( P(Y \mid X) \)

<table>
<thead>
<tr>
<th>O: Is older than 35 years</th>
<th>I: Has Personal Income</th>
<th>S: Is a Student</th>
<th>J: Birthday before July 1st</th>
<th>Y: Buys computer</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

New customer i:

\[ 0, 1, 1, 1, ? \]

Want to find \( P(Y_i = 0 \mid O = 0, I = 0, S = 1, J = 1) \)

---

Announcements

- HW2 was released, due on February 15th
- Projects
  - Teams + project due February 11th (let us know if you have problems!)
  - Midway Report due March 6th
  - Poster Session on April 17th, 9am-2pm
- Final Report due May 1st

---

How many parameters must we estimate?

Suppose \( X = (X_1, X_2) \)

where \( X_i \) and \( Y \) are boolean RV's

To estimate \( P(Y \mid X_1, X_2) \)

<table>
<thead>
<tr>
<th>( X_1 )</th>
<th>( X_2 )</th>
<th>( P(Y = 1 \mid X_1, X_2) )</th>
<th>( P(Y = 0 \mid X_1, X_2) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0.1</td>
<td>0.9</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0.24</td>
<td>0.76</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0.54</td>
<td>0.46</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0.23</td>
<td>0.77</td>
</tr>
</tbody>
</table>
How many parameters must we estimate?
Suppose $X = (X_1, X_2)$
where $X_i$ and $Y$ are boolean RV's
To estimate $P(Y | X_1, X_2)$

| $X_1$ | $X_2$ | $P(Y = 1 | X_1, X_2)$ | $P(Y = 0 | X_1, X_2)$ |
|-------|-------|-----------------------|-----------------------|
| 0     | 0     | 0.1                   | 0.9                   |
| 1     | 0     | 0.24                  | 0.76                  |
| 0     | 1     | 0.54                  | 0.46                  |
| 1     | 1     | 0.23                  | 0.77                  |

4 parameters: $P(Y=0|X) = 1 - P(Y=1|X)$

If we have 30 boolean $X_i$'s: $P(Y | X_1, X_2, \ldots X_{30}) \sim 1$ Billion
Bayes Rule

\[ P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \]

Which is shorthand for:

\[ (\forall i, j) P(Y = y_i|X = x_j) = \frac{P(X = x_j|Y = y_i)P(Y = y_i)}{P(X = x_j)} \]

Equivalently:

\[ (\forall i, j) P(Y = y_i|X = x_j) = \frac{P(X = x_j|Y = y_i)P(Y = y_i)}{\sum_k P(X = x_j|Y = y_k)P(Y = y_k)} \]

Can we reduce params using Bayes Rule?

Suppose \( X = (X_1, \ldots, X_n) \)

where \( X_i \) and \( Y \) are boolean RV's

\[ P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \]

How many parameters to define \( P(X_1, \ldots, X_n | Y) \)?

| \( Y \) | \( X_1 \) | \( X_2 \) | \( \ldots \) | \( X_n \) | \( P(X|Y) \) |
|-------|-------|-------|----------|-------|--------|
| 0     | 0     | 0     | \ldots   | 0     | 0.1    |
| 0     | 1     | 0     | \ldots   | 0     | 0.02   |
| \ldots| \ldots| \ldots| \ldots   | \ldots| \ldots |
| \ldots| \ldots| \ldots| \ldots   | \ldots| \ldots |
| 0     | 1     | 1     | 1        | 1     | 0.16   |

\( 2^n \) rows!

-1 because all the probabilities sum to 1

Can we reduce params using Bayes Rule?

Suppose \( X = (X_1, \ldots, X_n) \)

where \( X_i \) and \( Y \) are boolean RV's

\[ P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \]

How many parameters to define \( P(X_1, \ldots, X_n | Y) \)?

| \( Y \) | \( X_1 \) | \( X_2 \) | \( \ldots \) | \( X_n \) | \( P(X|Y) \) |
|-------|-------|-------|----------|-------|--------|
| 0     | 0     | 0     | \ldots   | 0     | 0.1    |
| 0     | 1     | 0     | \ldots   | 0     | 0.02   |
| \ldots| \ldots| \ldots| \ldots   | \ldots| \ldots |
| \ldots| \ldots| \ldots| \ldots   | \ldots| \ldots |
| 0     | 1     | 1     | 1        | 1     | 0.16   |

\( 2^n - 1 \)

Should I compute these as well?
Can we reduce params using Bayes Rule?

Suppose \( X = (X_1, \ldots, X_n) \)
where \( X_i \) and \( Y \) are boolean RV's

How many parameters to define \( P(X_1, \ldots, X_n \mid Y) \)?

\[
P(Y \mid X) = \frac{P(X \mid Y)P(Y)}{P(X)}
\]

\[
\begin{array}{cccccc}
Y & X_1 & X_2 & \ldots & X_n & P(X \mid Y) \\
0 & 0 & 0 & \ldots & 0 & 0.1 \\
0 & 1 & 0 & \ldots & \ldots & 0.02 \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
0 & 1 & 1 & 1 & 1 & 0.16 \\
\end{array}
\]

\( 2^n - 1 \)

How many parameters to define \( P(Y) \)?

\[
2^{(2^n - 1)}
\]

If \( n = 30 \), 2 billion

Naïve Bayes

Naïve Bayes assumes

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

i.e., that \( X_i \) and \( X_j \) are conditionally independent given \( Y \), for all \( i \neq j \)
**Conditional Independence**

Definition: X is conditionally independent of Y given Z, if the probability distribution governing X is independent of the value of Y, given the value of Z.

\[(\forall i, j, k) P(X = x_i | Y = y_j, Z = z_k) = P(X = x_i | Z = z_k)\]

Which we often write

\[P(X | Y, Z) = P(X | Z)\]

E.g.,

\[P(\text{thunder}|\text{raining,lightning}) = P(\text{thunder}|\text{lightning})\]

Naïve Bayes uses assumption that the \(X_i\) are conditionally independent, given Y. E.g., \(P(X_1|X_2, Y) = P(X_1|Y)\)

Given this assumption, then:

\[P(X_1, X_2|Y) = P(X_1|X_2, Y)P(X_2|Y)\]

Naïve Bayes uses assumption that the \(X_i\) are conditionally independent, given Y. E.g., \(P(X_i|X_2, Y) = P(X_i|Y)\)

Given this assumption, then:

\[P(X_1, X_2|Y) = P(X_1|X_2, Y)P(X_2|Y)\]

\[= P(X_1|Y)P(X_2|Y)\]

Chain rule

Naïve Bayes uses assumption that the \(X_i\) are conditionally independent, given Y. E.g., \(P(X_1|X_2, Y) = P(X_1|Y)\)

Given this assumption, then:

\[P(X_1, X_2|Y) = P(X_1|X_2, Y)P(X_2|Y)\]

\[= P(X_1|Y)P(X_2|Y)\]

Conditional independence
Naïve Bayes uses assumption that the $X_i$ are conditionally independent, given $Y$. E.g., $P(X_1 | X_2, Y) = P(X_1 | Y)$

Given this assumption, then:

$$P(X_1, X_2 | Y) = P(X_1 | X_2, Y)P(X_2 | Y)$$

$$= P(X_1 | Y)P(X_2 | Y)$$

in general: $P(X_1...X_n | Y) = \prod_i P(X_i | Y)$

Naïve Bayes uses assumption that the $X_i$ are conditionally independent, given $Y$. E.g., $P(X_1 | X_2, Y) = P(X_1 | Y)$

Given this assumption, then:

$$P(X_1, X_2 | Y) = P(X_1 | X_2, Y)P(X_2 | Y)$$

$$= P(X_1 | Y)P(X_2 | Y)$$

in general: $P(X_1...X_n | Y) = \prod_i P(X_i | Y)$

How many parameters to describe $P(X_1...X_n | Y)$? $P(Y)$?
- Without conditional indep assumption? $2^{(2^n - 1)}$ and 1
- With conditional indep assumption? $2n$ and 1

Naïve Bayes in a Nutshell

Bayes rule:

$$P(Y = y_k | X_1...X_n) = \frac{P(Y = y_k)P(X_1...X_n | Y = y_k)}{\sum_j P(Y = y_j)P(X_1...X_n | Y = y_j)}$$

Assuming conditional independence among $X_i$'s:

$$P(Y = y_k | X_1...X_n) = \frac{P(Y = y_k)\prod_i P(X_i | Y = y_k)}{\sum_j P(Y = y_j)\prod_i P(X_i | Y = y_j)}$$

(estimate in training)

So, to pick most probable $Y$ for $X_{new} = (X_1, ..., X_n)$

$$Y^{new} \leftarrow \arg \max_{y_k} P(Y = y_k)\prod_i P(X_i^{new} | Y = y_k)$$

(testing)
**Naïve Bayes Algorithm – discrete \( X_i \)**

- Train Naïve Bayes (examples)
  - for each value \( y_k \)
    - estimate \( \pi_k \equiv P(Y = y_k) \)
  - for each value \( x_{ij} \) of each attribute \( X_i \)
    - estimate \( \theta_{ijk} \equiv P(X_i = x_{ij} | Y = y_k) \)
- Classify \( (X_{new}) \)

\[
Y_{new}^\prime = \arg \max_{y_k} \quad P(Y = y_k) \prod_i P(X_i^{\text{new}} | Y = y_k)
\]

\[
Y_{new} = \arg \max_{y_k} \quad \pi_k \prod_i \theta_{ijk}
\]

* probabilities must sum to 1, so need estimate only \( v-1 \) of these, where \( v \) is the number of values, which is 2 in the binary case

---

**Let’s train! Buy computer? \( P(Y|O,S,J) \)**

- \( O= \) Is older than 35
- \( S= \) Is a student
- \( J= \) Birthday before July 1
- \( Y= \) buys computer

What probability parameters must we estimate?

<table>
<thead>
<tr>
<th>( O )</th>
<th>( S )</th>
<th>( J )</th>
<th>( Y )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( 0 )</td>
<td>( 0 )</td>
<td>( 0 )</td>
<td>( 0 )</td>
</tr>
<tr>
<td>( 0 )</td>
<td>( 0 )</td>
<td>( 1 )</td>
<td>( 0 )</td>
</tr>
<tr>
<td>( 1 )</td>
<td>( 0 )</td>
<td>( 1 )</td>
<td>( 1 )</td>
</tr>
<tr>
<td>( 1 )</td>
<td>( 1 )</td>
<td>( 0 )</td>
<td>( 1 )</td>
</tr>
<tr>
<td>( 1 )</td>
<td>( 1 )</td>
<td>( 1 )</td>
<td>( 0 )</td>
</tr>
<tr>
<td>( 0 )</td>
<td>( 1 )</td>
<td>( 0 )</td>
<td>( 0 )</td>
</tr>
<tr>
<td>( 0 )</td>
<td>( 0 )</td>
<td>( 0 )</td>
<td>( 0 )</td>
</tr>
<tr>
<td>( 0 )</td>
<td>( 1 )</td>
<td>( 1 )</td>
<td>( 1 )</td>
</tr>
<tr>
<td>( 0 )</td>
<td>( 1 )</td>
<td>( 0 )</td>
<td>( 0 )</td>
</tr>
<tr>
<td>( 1 )</td>
<td>( 0 )</td>
<td>( 1 )</td>
<td>( 1 )</td>
</tr>
<tr>
<td>( 1 )</td>
<td>( 0 )</td>
<td>( 0 )</td>
<td>( 0 )</td>
</tr>
</tbody>
</table>

---

**Let’s train! Buy computer? \( P(Y|O,S,J) \)**

<table>
<thead>
<tr>
<th>( O )</th>
<th>( S )</th>
<th>( J )</th>
<th>( Y )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( 0 )</td>
<td>( 0 )</td>
<td>( 0 )</td>
<td>( 0 )</td>
</tr>
<tr>
<td>( 0 )</td>
<td>( 0 )</td>
<td>( 1 )</td>
<td>( 0 )</td>
</tr>
<tr>
<td>( 1 )</td>
<td>( 0 )</td>
<td>( 1 )</td>
<td>( 1 )</td>
</tr>
<tr>
<td>( 1 )</td>
<td>( 1 )</td>
<td>( 0 )</td>
<td>( 1 )</td>
</tr>
<tr>
<td>( 1 )</td>
<td>( 1 )</td>
<td>( 1 )</td>
<td>( 0 )</td>
</tr>
<tr>
<td>( 0 )</td>
<td>( 1 )</td>
<td>( 0 )</td>
<td>( 0 )</td>
</tr>
<tr>
<td>( 0 )</td>
<td>( 0 )</td>
<td>( 0 )</td>
<td>( 0 )</td>
</tr>
<tr>
<td>( 0 )</td>
<td>( 1 )</td>
<td>( 1 )</td>
<td>( 1 )</td>
</tr>
<tr>
<td>( 0 )</td>
<td>( 1 )</td>
<td>( 0 )</td>
<td>( 0 )</td>
</tr>
<tr>
<td>( 1 )</td>
<td>( 0 )</td>
<td>( 1 )</td>
<td>( 1 )</td>
</tr>
<tr>
<td>( 1 )</td>
<td>( 0 )</td>
<td>( 0 )</td>
<td>( 0 )</td>
</tr>
</tbody>
</table>
Let's train! Buy computer? $P(Y|O,S,J)$

- $O$: Is older than 35
- $S$: Is a student
- $J$: Birthday before July 1
- $Y$: buys computer

What probability parameters must we estimate?

- $P(Y=1)$
- $P(O=1 | Y=1)$
- $P(O=1 | Y=0)$
- $P(S=1 | Y=1)$
- $P(S=1 | Y=0)$
- $P(J=1 | Y=1)$
- $P(J=1 | Y=0)$
- $P(Y=0)$
- $P(O=0 | Y=1)$
- $P(O=0 | Y=0)$
- $P(S=0 | Y=1)$
- $P(S=0 | Y=0)$
- $P(J=0 | Y=1)$
- $P(J=0 | Y=0)$

<table>
<thead>
<tr>
<th>O</th>
<th>S</th>
<th>J</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>O</th>
<th>S</th>
<th>J</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Let's train! Buy computer? $P(Y|O,S,J)$

- $O$: Is older than 35
- $S$: Is a student
- $J$: Birthday before July 1
- $Y$: buys computer

What probability parameters must we estimate?

- $P(Y=1)$
- $P(O=1 | Y=1)$
- $P(O=1 | Y=0)$
- $P(S=1 | Y=1)$
- $P(S=1 | Y=0)$
- $P(J=1 | Y=1)$
- $P(J=1 | Y=0)$
- $P(Y=0)$
- $P(O=0 | Y=1)$
- $P(O=0 | Y=0)$
- $P(S=0 | Y=1)$
- $P(S=0 | Y=0)$
- $P(J=0 | Y=1)$
- $P(J=0 | Y=0)$

<table>
<thead>
<tr>
<th>O</th>
<th>S</th>
<th>J</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Let’s train! Buy computer? \( P(Y,O,S,J) \)

- \( O \): Is older than 35
- \( S \): Is a student
- \( J \): Birthday before July 1
- \( Y \): buys computer

What probability parameters must we estimate?

\[
\begin{align*}
P(Y=1) & : \frac{9}{14} & P(O=1 | Y=1) & : \frac{4}{9} & P(S=1 | Y=1) & : \frac{1}{5} & P(J=1 | Y=1) & : \frac{5}{9} \\
P(O=1 | Y=0) & : \frac{5}{14} & P(O=0 | Y=1) & : \frac{1}{5} & P(S=1 | Y=0) & : \frac{1}{5} & P(J=0 | Y=1) & : \frac{4}{9} \\
P(S=1 | Y=1) & : \frac{4}{9} & P(S=0 | Y=1) & : \frac{3}{9} & P(J=1 | Y=0) & : \frac{5}{9} & P(O=0, S=1, J=1 | Y=1) & : \frac{1}{5} \times \frac{4}{5} \times \frac{5}{9} \times \frac{9}{14} = 0.0057
\end{align*}
\]

Let’s test! Buy computer? \( P(Y,O,S,J) \)

- \( O \): Is older than 35
- \( S \): Is a student
- \( J \): Birthday before July 1
- \( Y \): buys computer

What probability parameters must we estimate?

\[
\begin{align*}
P(Y=1) & : \frac{5}{14} & P(O=1 | Y=0) & : \frac{1}{5} & P(S=1, J=1 | Y=1) & : \frac{1}{5} \times \frac{1}{5} \times \frac{2}{5} \times \frac{5}{14} = 0.0057
\end{align*}
\]
Let's test! Buy computer? P(Y|O,S,J)

- O= Is older than 35
- S= Is a student
- J = Birthday before July 1
- Y= buys computer

What probability parameters must we estimate?

P(Y=1) : 5/14
P(O=1 | Y=1) : 4/9
P(O=1 | Y=0) : 5/9
P(S=1 | Y=1) : 6/9
P(S=1 | Y=0) : 1/5
P(J=1 | Y=1) : 5/9
P(J=1 | Y=0) : 2/5

P(Y=0) :
P(O=0 | Y=1) :
P(O=0 | Y=0) :
P(S=0 | Y=1) :
P(S=0 | Y=0) :
P(J=0 | Y=1) :
P(J=0 | Y=0) :

P(O=0,S=1,J=1|Y=0) * P(Y=0) = 1/5 * 1/5 * 2/5 * 5/14 = 0.0057
P(O=0,S=1,J=1|Y=1) * P(Y=1) = 5/9 * 6/9 * 5/9 * 9/14 = 0.13

P(Y=0|O=0,S=1,J=1) = 0.04
P(Y=1|O=0,S=1,J=1) = 0.96  Choice is Label 1

Let's test! Buy computer? P(Y|O,S,J)

- O= Is older than 35
- S= Is a student
- J = Birthday before July 1
- Y= buys computer

What probability parameters must we estimate?

P(Y=1) : 9/14
P(O=1 | Y=1) : 4/9
P(O=1 | Y=0) : 5/9
P(S=1 | Y=1) : 6/9
P(S=1 | Y=0) : 1/5
P(J=1 | Y=1) : 5/9
P(J=1 | Y=0) : 2/5

P(Y=0|O=0,S=1,J=1) = 0.04
P(Y=1|O=0,S=1,J=1) = 0.96  Choice is Label 1

Assume we have a lot of data

- We estimate these probabilities
- We obtain 68% test accuracy.
- What does this mean?

Our estimated P(Y=1) was 64.5%, let's say this is actually the truth.
- Is 68% impressive?
• What is chance performance?

• What happens if I flip an unbiased coin?

• If P(Y=1)>0.5, then we can just predict 1 all the time!
  • What will be the accuracy?

• What happens if you predict Y=1 with probability 0.645 ==
  this is called probability matching in cognitive science

**Naïve Bayes: Point #1**

Often the \(X_i\) are not really conditionally independent

• We use Naïve Bayes in many cases anyway, and it often works pretty well
  – often the right classification, even when not the right probability (see [Domingos&Pazzani, 1996])

• What is effect on estimated \(P(Y|X)\)?
  – Extreme case: what if we add two copies: \(X_i = X_k\)

**Extreme case: what if we add two copies: \(X_i = X_k\)**

\[
P(Y=1) \cdot P(O=1|Y=1) \cdot P(S=0|Y=1) \cdot P(J=0|Y=1) \\
--------------------------------------- \\
P(Y=1) \cdot P(O=1|Y=1) \cdot P(S=0|Y=1) \cdot P(J=0|Y=1) + P(Y=0) \cdot P(O=1|Y=0) \cdot P(S=0|Y=0) \cdot P(J=0|Y=0)
\]

**Naïve Bayes: Point #2**

Irrelevant variables, do they affect you?

\[
P(Y=1) \cdot P(O=1|Y=1) \cdot P(S=0|Y=1) \cdot P(J=0|Y=1) \\
--------------------------------------- \\
P(Y=1) \cdot P(O=1|Y=1) \cdot P(S=0|Y=1) + P(Y=0) \cdot P(O=1|Y=0) \cdot P(S=0|Y=0) \cdot P(J=0|Y=0)
\]
### Naïve Bayes: Point #2

Irrelevant variables, do they affect you?

\[
P(Y=1) P(O=1|Y=1) P(S=0|Y=1) P(J=0|Y=1) + P(Y=0) P(O=1|Y=0) P(S=0|Y=0) P(J=0|Y=0)
\]

\[
P(Y=1) P(O=1|Y=1) P(S=0|Y=1) P(J=0|Y=1)
\]

If J is not relevant then \(P(J|Y=0) = P(J|Y=1) = P(J)\)

Does it hurt classification?

---

### Naïve Bayes: Point #3

Another way to view Naïve Bayes (Boolean \(Y, X_i\)'s):

Decision rule: is this quantity greater or less than 1?

\[
\frac{P(Y = 1|X_1...X_n)}{P(Y = 0|X_1...X_n)} = \frac{P(Y = 1)}{P(Y = 0)} \prod_i P(X_i|Y = 1) / \prod_i P(X_i|Y = 0)
\]

(is \(q_1\) larger than \(q_0\)?)

---

### Naïve Bayes: Point #2

Irrelevant variables, do they affect you?

\[
P(Y=1) P(O=1|Y=1) P(S=0|Y=1) P(J=0|Y=1) + P(Y=0) P(O=1|Y=0) P(S=0|Y=0) P(J=0|Y=0)
\]

\[
P(Y=1) P(O=1|Y=1) P(S=0|Y=1) P(J=0|Y=1)
\]

If J is not relevant then \(P(J|Y=0) = P(J|Y=1) = P(J)\)

Does it hurt classification?

If we had the correct estimate, performance is not affected!
If we have noisy estimates \(\Rightarrow\) performance is affected

---

### Naïve Bayes: Point #3

Another way to view Naïve Bayes (Boolean \(Y, X_i\)'s):

Decision rule: is this quantity greater or less than 1?

\[
\frac{P(Y = 1|X_1...X_n)}{P(Y = 0|X_1...X_n)} = \frac{P(Y = 1)}{P(Y = 0)} \prod_i P(X_i|Y = 1) / \prod_i P(X_i|Y = 0) > \text{ or } < 1 ?
\]

\[
\log \frac{P(Y = 1|X_1...X_n)}{P(Y = 0|X_1...X_n)} = \log \frac{P(Y = 1)}{P(Y = 0)} + \sum_i \log \frac{P(X_i|Y = 1)}{P(X_i|Y = 0)} > \text{ or } < 0 ?
\]

\[
\theta_{ik} = \hat{P}(X_i = 1|Y = k)
\]

\[
1 - \theta_{ik} = \hat{P}(X_i = 0|Y = k)
\]
Naïve Bayes: Point #3

Another way to view Naïve Bayes (Boolean $Y$, $X_i$'s):
Decision rule: is this quantity greater or less than 1?

$$\frac{P(Y = 1|X_1 \ldots X_n)}{P(Y = 0|X_1 \ldots X_n)} = \frac{P(Y = 1) \prod_i P(X_i|Y = 1)}{P(Y = 0) \prod_i P(X_i|Y = 0)}$$

$$\log \frac{P(Y = 1|X_1 \ldots X_n)}{P(Y = 0|X_1 \ldots X_n)} = \log \frac{P(Y = 1)}{P(Y = 0)} + \sum_i \log \frac{P(X_i|Y = 1)}{P(X_i|Y = 0)}$$

$$= \log \frac{P(Y = 1)}{P(Y = 0)} + \sum_i X_i \log \frac{\theta_i}{\theta_0} + (1 - X_i) \frac{1 - \theta_i}{1 - \theta_0}$$

$$\theta_{ik} = \hat{P}(x_i = 1|Y = k)$$

$1 - \theta_{ik} = \hat{P}(x_i = 0|Y = k)$

What happens when $X$ has a large number of dimensions?

Naïve Bayes: Point #4

If unlucky, our MLE estimate for $P(X_i | Y)$ might be zero.
(for example, $X_i = \text{birthdate. } X_i = \text{Jan}_25\_1992$)

- Why worry about just one parameter out of many?
- What can be done to address this?
Estimating Parameters

- Maximum Likelihood Estimate (MLE): choose $\theta$ that maximizes probability of observed data $D$

\[
\hat{\theta} = \arg \max_{\theta} P(D \mid \theta)
\]

- Maximum a Posteriori (MAP) estimate: choose $\theta$ that is most probable given prior probability and the data

\[
\hat{\theta} = \arg \max_{\theta} P(\theta \mid D)
\]

\[
= \arg \max_{\theta} \frac{P(D \mid \theta)P(\theta)}{P(D)}
\]

Learning to classify text documents

- Classify which emails are spam?
- Classify which emails promise an attachment?
- Classify which web pages are student home pages?

How shall we represent text documents for Naïve Bayes?

Baseline: Bag of Words Approach

```
<table>
<thead>
<tr>
<th>Word</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>aardvark</td>
<td>0</td>
</tr>
<tr>
<td>about</td>
<td>2</td>
</tr>
<tr>
<td>all</td>
<td>2</td>
</tr>
<tr>
<td>Africa</td>
<td>1</td>
</tr>
<tr>
<td>apple</td>
<td>0</td>
</tr>
<tr>
<td>anxious</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
<tr>
<td>gas</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
<tr>
<td>oil</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
<tr>
<td>Zaire</td>
<td>0</td>
</tr>
</tbody>
</table>
```

Estimating Parameters: $Y, X_i$ discrete-valued

Maximum likelihood estimates:

\[
\hat{\pi}_k = \hat{P}(Y = y_k) = \frac{\# D \{Y = y_k\}}{|D|}
\]

\[
\hat{\theta}_{ijk} = \hat{P}(X_i = x_j \mid Y = y_k) = \frac{\# D \{X_i = x_j \wedge Y = y_k\}}{\# D \{Y = y_k\}}
\]

MAP estimates (Beta, Dirichlet priors):

\[
\pi_k = \hat{P}(Y = y_k) = \frac{\# D \{Y = y_k\} + (\beta_k - 1)}{|D| + \sum_m (\beta_m - 1)}
\]

\[
\theta_{ijk} = \hat{P}(X_i = x_j \mid Y = y_k) = \frac{\# D \{X_i = x_j \wedge Y = y_k\} + (\beta_k - 1)}{\# D \{Y = y_k\} + \sum_m (\beta_m - 1)}
\]

Only difference: “imaginary” examples

How many counts should we add?
How can we express X?

- Y discrete valued. e.g., Spam or not
- X = (X_1, X_2, ... X_n) with n the number of words in English.

What are the problems with this representation?

- Some words always present
- Some words very infrequent
- Doesn’t count how often a word appears
- Conditional independence assumption is false...

Learning to classify document: P(Y|X)
the “Bag of Words” model

- Y discrete valued. e.g., Spam or not
- X = (X_1, X_2, ... X_n) = document

- X_i is a random variable describing the word at position i in the document
- possible values for X_i: any word w_k in English
- X_i represents the i^{th} word position in document
- X_1 = “I”, X_2 = “am”, X_3 = “pleased”
- and, let’s assume the X_i are iid (indep, identically distributed)

\[ P(X_i|Y) = P(X_j|Y) \quad (\forall i, j) \]
Learning to classify document: \( P(Y|X) \)
The “Bag of Words” model

- \( Y \) discrete valued. e.g., Spam or not
- \( X = (X_1, X_2, \ldots X_n) = \) document

- \( X_i \) is a random variable describing the word at position \( i \) in the document
- possible values for \( X_i \): any word \( w_k \) in English

- Document = bag of words: the vector of counts for all \( w_k \)'s
  - like #heads, #tails, but we have many more than 2 values
  - assume word probabilities are position independent
  (i.i.d. rolls of a 50,000-sided die)

Naïve Bayes Algorithm – discrete \( X_i \)

- Train Naïve Bayes (examples)
  for each value \( y_k \)
  estimate \( \pi_k \equiv P(Y = y_k) \)
  for each value \( x_j \) of each attribute \( X_i \)
  estimate \( \theta_{ijk} \equiv P(X_i = x_j|Y = y_k) \)

- Classify \((X_{\text{new}})\)
  \( Y^\text{new} \leftarrow \arg \max_{y_k} P(Y = y_k) \prod_i P(X_{i}^{\text{new}}|Y = y_k) \)
  \( Y^\text{new} \leftarrow \arg \max_{y_k} \pi_k \prod_i \theta_{ijk} \)

* Additional assumption: word probabilities are position independent
  \( \theta_{ijk} = \theta_{mj_k} \) for all \( i, m \)

MAP estimates for bag of words

Map estimate for multinomial

\[
\theta_{jk} = \frac{\alpha_{jk} + \beta_{jk} - 1}{\sum_m \alpha_{mk} + \sum_m \beta_{mk} - 1}
\]
MAP estimates for bag of words

Map estimate for multinomial

\[
\theta_{jk} = \frac{\alpha_{jk} + \beta_{jk} - 1}{\sum_m \alpha_{mk} + \sum_m \beta_{mk} - 1}
\]

What \( \beta \)'s should we choose?
- Probabilities over all classes?
- Constant per word?

Prior is Dirichlet
Posterior is also Dirichlet
(\( \text{Dirichlet is the conjugate prior for a multinomial likelihood function} \))

For code and data, see www.cs.cmu.edu/~tom/mlbook.html click on "Software and Data"

Learning Curve for 20 Newsgroups

19 Prior probabilities

Accuracy vs. Training set size (1/3 withheld for test)

Naive Bayes: 89% classification accuracy
For code and data, see www.cs.cmu.edu/~tom/mlbook.html click on ”Software and Data"

### Performance can be very good

- Even when taking half of the email
- Assumption doesn’t hurt the particular problem?
- Redundancy?
- Leads less examples to train?
  Converges faster to asymptotic performance? (Ng and Jordan)

### What if we have continuous $X_i$ ?

Eg., image classification: $X_i$ is real-valued $i^{th}$ pixel

Naïve Bayes requires $P(X_i \mid Y=y_k)$, but $X_i$ is real (continuous)

\[
P(Y = y_k \mid X_1 \ldots X_n) = \frac{P(Y = y_k) \prod_i P(X_i \mid Y = y_k)}{\sum_j P(Y = y_j) \prod_i P(X_i \mid Y = y_j)}
\]

Common approach: assume $P(X_i \mid Y=y_k)$ follows a Normal (Gaussian) distribution
GNB Example: Classify a person’s cognitive state, based on brain image

- reading a sentence or viewing a picture?
- reading the word describing a “Tool” or “Building”?
- answering the question, or getting confused?

Stimuli for the study:
60 distinct exemplars, presented 6 times each

What if we have continuous \( X_i \)?

Gaussian Naïve Bayes (GNB): assume

\[
p(X_i = x | Y = y_k) = \frac{1}{\sqrt{2\pi\sigma_{ik}^2}} e^{-\frac{(x - \mu_{ik})^2}{2\sigma_{ik}^2}}
\]

Sometimes assume variance
- is independent of \( Y \) (i.e., \( \sigma_i \)),
- or independent of \( X_i \) (i.e., \( \sigma_k \))
- or both (i.e., \( \sigma \))

Gaussian Naïve Bayes Algorithm – continuous \( X_i \)

- Train Naive Bayes (examples)
  
  for each value \( y_k \) estimate* \( \pi_k \equiv P(Y = y_k) \)
  for each attribute \( X_i \) estimate \( P(X_i | Y = y_k) \)
  
  - class conditional mean \( \mu_{ik} \), variance \( \sigma_{ik} \)

- Classify \( (X^{new}) \)

\[
Y^{new} \leftarrow \arg\max_{y_k} P(Y = y_k) \prod_i P(X_i^{new} | Y = y_k)
\]

\[
Y^{new} \leftarrow \arg\max_{y_k} \pi_k \prod_i N(X_i^{new}; \mu_{ik}, \sigma_{ik})
\]

* probabilities must sum to 1, so need estimate only \( n-1 \) parameters...
Estimating Parameters: $Y$ discrete, $X_i$ continuous

Maximum likelihood estimates:

$$\hat{\mu}_{ik} = \frac{1}{\sum_j \delta(Y^j = y_k)} \sum_j X_i^j \delta(Y^j = y_k)$$

$$\hat{\sigma}^2_{ik} = \frac{1}{\sum_j \delta(Y^j = y_k)} \sum_j (X_i^j - \hat{\mu}_{ik})^2 \delta(Y^j = y_k)$$

Mean activations over all training examples for $Y$=“bottle”

$Y$ is the mental state (reading “house” or “bottle”) $X_i$ are the voxel activities,

this is a plot of the $\mu$’s defining $P(X_i | Y$=“bottle”)
Questions to think about:

- How can we extend Naïve Bayes if just 2 of the Xᵢ’s are dependent?
- What error will the classifier achieve if Naïve Bayes assumption is satisfied and we have infinite training data?
- What does the decision surface of a Naïve Bayes classifier look like?
- Can you use Naïve Bayes for a combination of discrete and real-valued Xᵢ?

We covered:

- Bayes classifiers to learn P(Y|X)
- MLE and MAP estimates for parameters of P
- Conditional independence
- Naïve Bayes \( \rightarrow \) make Bayesian learning practical
- Text classification
- Naïve Bayes and continuous variables Xᵢ:
  - Gaussian Naïve Bayes classifier

Next:

- Learn P(Y|X) directly
  - Logistic regression, Regularization, Gradient ascent
- Naïve Bayes or Logistic Regression?
  - Generative vs. Discriminative classifiers

Gaussian Naïve Bayes – Big Picture

\[
Y^{new} \leftarrow \arg \max_{y \in \{0,1\}} P(Y = y) \prod_i P(X_i^{new} | Y = y) \quad \text{assume } P(Y=1) = 0.5
\]

Example: Y= PlayBasketball (boolean), X1=Height, X2=MLgrade

Logistic Regression

Idea:

- Naïve Bayes allows computing P(Y|X) by learning P(Y) and P(X|Y)

- Why not learn P(Y|X) directly?