CS 4824/ECE 4424: Naïve Bayes

Acknowledgement:
Many of these slides are derived from Tom Mitchell, Pascal Poupart, Pieter Abbeel, Eric Eaton, Carlos Guestrin, William Cohen, and Andrew Moore.
Let’s learn classifiers by learning $P(Y \mid X)$

- Suppose we want to learn the function $f: \langle G, H \rangle \rightarrow W$
- Or $P(W \mid G, H)$

<table>
<thead>
<tr>
<th>Gender</th>
<th>hours_worked</th>
<th>wealth</th>
<th>$P(\text{poor} \mid G, H)$</th>
<th>$P(\text{rich} \mid G, H)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>v0:40.5-</td>
<td>poor</td>
<td>0.253122</td>
<td>0.0245695</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rich</td>
<td>0.0421768</td>
<td>0.0116293</td>
</tr>
<tr>
<td>Male</td>
<td>v0:40.5-</td>
<td>poor</td>
<td>0.331313</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>rich</td>
<td>0.0971295</td>
<td></td>
</tr>
<tr>
<td></td>
<td>v1:40.5+</td>
<td>poor</td>
<td>0.134106</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>rich</td>
<td>0.105933</td>
<td></td>
</tr>
</tbody>
</table>

$P(\text{rich} \mid G, H) = \frac{A}{A + B}$
How many parameters must we estimate?

- Suppose $X = \langle X_1, X_2, \ldots, X_n \rangle$, where $X_i$ and $Y$ are boolean RVs.

- To estimate $P(Y | X_1, X_2, \ldots, X_n)$, how many parameters do we need? $2^n$.

- How can we design a learning algorithm that is practical?

- Can Bayes Rule help?

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

\[
\text{(equivalently)}
\]

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

\[
\text{this is a shorthand for}
\]

\[
\sum_{k} P(X = x_j | Y = y_k) P(Y = y_k)
\]

\[
\text{with}
\]

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

\[
\text{and}
\]

\[
\sum_{k} P(X = x_j | Y = y_k) P(Y = y_k)
\]
Can we reduce parameters using Bayes Rule?

- Suppose $X = \langle X_1, X_2, \ldots, X_n \rangle$, where $X_i$ and $Y$ are boolean RVs.

- Bayes Rule:
  
  $P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$

- How many parameters needed to estimate $P(X_1, X_2, \ldots, X_n | Y)$?
  - Case 1: $Y = 1$ \quad $P(X_1 \ldots X_n | Y = 1) = 2^n - 1$
  - Case 2: $Y = 0$ \quad $P(X_1 \ldots X_n | Y = 0) = 2^n - 1$

- How many parameters needed to estimate $P(Y) = 1$?

So, if we use Bayes Rule, we need $2(2^n - 1) + 1$ parameters!
Naïve Bayes

- Naïve Bayes assumes

\[
P(X_1, \ldots, X_n \mid Y) = \prod_{i} P(X_i \mid Y)
\]

- That is, \( X_i \) and \( X_j \) are conditionally independent given \( Y \) \( \forall i \neq j \)
Conditional independence

- $X$ is conditionally independent of $Y$ given $Z$, if the probability distribution governing $X$ is independent of the value of $Y$ given the value of $Z$
  
  \[
  \forall i, j, k \quad P(X = x_i | Y = y_j, Z = z_k) = P(X = x_i | Z = z_k)
  \]
  
  - Or equivalently, $P(X | Y Z) = P(X | Z)$

- $P(\text{Thunder} | \text{Rain, Lightning}) = P(\text{Thunder} | \text{Lightning})$

  - That is, Thunder and Rain are conditionally independent given Lightning.
Naïve Bayes assumes conditional independence

- Under the conditional independence assumption, then
  \[
  P(X_1, X_2 | Y) = P(X_1 | Y) P(X_2 | Y)
  \]

  \[
  P(X_1, ..., X_n | Y) = \prod_{i=1}^{n} P(X_i | Y)
  \]
Naïve Bayes assumes conditional independence

- In General,
  \[ P(X_1, \ldots, X_n | Y) = \prod_{i} P(X_i | Y) \]
- How many parameters to describe \( P(X_1, \ldots, X_n | Y) \)? \( P(Y) \)?
  \[ \frac{2n}{2n + 1} + 1 \]
- Without conditional independence:
  \[ \frac{2(2^n) - 1}{2n + 1} + 1 \]
- With conditional independence:
Naïve Bayes summary

- Bayes Rule:
  \[ P(Y = y_k | X_1, \ldots, X_n) = \frac{P(Y = y_k)P(X_1, \ldots, X_n | Y = y_k)}{\sum_j P(Y = y_j)P(X_1, \ldots, X_n | Y = y_j)} \]

- Assuming conditional independence among \( X_i \)'s
  \[ P(Y = y_k | X_1, \ldots, X_n) = \frac{P(Y = y_k) \prod_i P(X_i | Y = y_k)}{\sum_j P(Y = y_j) \prod_i P(X_i | Y = y_j)} \]

- How to pick the most probable \( Y \) for \( X^{\text{New}} = <X_1, X_2, \ldots, X_n> \)? \( y \in \{0, 1\} \)
  \[ \max \left\{ \frac{P(Y = 0) \prod_i P(X_i^{\text{New}} | Y = 0)}{P(Y = 1) \prod_i P(X_i^{\text{New}} | Y = 1)} \right\} \]
Naïve Bayes algorithm - discrete $X_i$

- Train Naïve Bayes (examples)
  - For each value $y_k$
    - Estimate $\pi_k = P(Y = y_k)$
    - For each value $x_{ij}$ of each attribute $X_i$
      - Estimate $\theta_{ijk} = P(X = x_{ij} \mid Y = y_k)$

- Classify $X^\text{New}$
  - $Y^\text{New} \leftarrow \arg \max_{y_k} P(Y \mid y_k) \prod_{i} P(X_i^\text{New} \mid Y = y_k)$
  - $Y^\text{New} \leftarrow \arg \max_{y_k} \pi_k \prod_{i} \theta_{ijk}$