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 $f: <G, H> \rightarrow W$
- Or $P(W \mid G, H)$

<table>
<thead>
<tr>
<th>gender</th>
<th>hours_worked</th>
<th>wealth</th>
<th>$P(W \mid G, H)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>v0:40.5-</td>
<td>poor</td>
<td>0.253122</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rich</td>
<td>0.0245895</td>
</tr>
<tr>
<td></td>
<td>v1:40.5+</td>
<td>poor</td>
<td>0.0421768</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rich</td>
<td>0.0116293</td>
</tr>
<tr>
<td>Male</td>
<td>v0:40.5-</td>
<td>poor</td>
<td>0.331313</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rich</td>
<td>0.0971295</td>
</tr>
<tr>
<td></td>
<td>v1:40.5+</td>
<td>poor</td>
<td>0.134106</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rich</td>
<td>0.105933</td>
</tr>
</tbody>
</table>

| Gender | HrsWorked | P(rich | G,HW) | P(poor | G,HW) |
|--------|-----------|--------|--------|--------|
| F      | <40.5     | .09    | .91    |
| F      | >40.5     | .21    | .79    |
| M      | <40.5     | .23    | .77    |
| M      | >40.5     | .38    | .62    |
How many parameters must we estimate?

- Suppose \( X = \langle X_1, X_2, \ldots, X_n \rangle \), where \( X_i \) and \( Y \) are boolean RV.

- To estimate \( P(Y \mid X_1, X_2, \ldots, X_n) \), how many parameters do we need?

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

- Can Bayes Rule help?

\[
P(Y \mid X) = \frac{P(X \mid Y)P(Y)}{P(X)}
\]
Can we reduce parameters using Bayes Rule?

- Suppose $X = <X_1, X_2, \ldots, X_n>$, where $X_i$ and $Y$ are boolean RV

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

- How many parameters needed to estimate $P(X_1, X_2, \ldots, X_n | Y)$?

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

  - $(\forall i, j, k) \ P(X = x_i \mid Y = y_j, Z = z_k) = P(X = x_i \mid Z = z_k)$
  - Or equivalently, $P(X \mid Y Z) = P(X \mid Z)$
  - $P(\text{Thunder} \mid \text{Rain, Lightning}) = P(\text{Thunder} \mid \text{Lightning})$

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

- Under the conditional independence assumption, then
  - \( P(X_1, X_2 | Y) = \)
Naïve Bayes assumes conditional independence

- In General,
  \[ P(X_1, \ldots, X_n \mid Y) = \prod_{i} P(X_i \mid Y) \]
- How many parameters to describe \( P(X_1, \ldots, X_n \mid Y) \)? \( P(Y) \)?
  - Without conditional independence:
  - 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 \( X^{\text{New}} = <X_1, X_2, \ldots, X_n> \)?
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 $X^{\text{New}}$
  - $Y^{\text{New}} \leftarrow \arg \max_{y_k} \pi_k \prod_i P(X_i^{\text{New}} \mid Y = y_k)$
  - $Y^{\text{New}} \leftarrow \arg \max_{y_k} \prod_i \theta_{ijk}$