Bayesian Classification

http://css.engineering.uiowa.edu/~comp/

Bayesian Classification: Why?

- **Probabilistic learning**: Computation of explicit probabilities for hypothesis, among the most practical approaches to certain types of learning problems.
- **Incremental**: Each training example can incrementally increase/decrease the probability that a hypothesis is correct. Prior knowledge can be combined with observed data.
- **Probabilistic prediction**: Predict multiple hypotheses, weighted by their probabilities.
- **Benchmark**: Even if Bayesian methods are computationally intractable, they can provide a benchmark for other algorithms.
Bayesian Theorem

- Given training data $D$, posteriori probability of a hypothesis $h$, $P(h|D)$ follows the Bayes theorem

\[
P(h|D) = \frac{P(D|h)P(h)}{P(D)}
\]

- MAP (maximum posteriori) hypothesis

\[
h_{MAP} = \arg\max_{h \in H} P(h|D) = \arg\max_{h \in H} P(D|h)P(h).
\]

- Practical difficulty: Initial knowledge of probabilities is required, significant computational cost

Example

Training data set $D$ describes computers by their speed and price.

$h = \text{is hypothesis that } D \text{ is a computer (i.e., } D \text{ belongs to a specific category } C \text{ )}$

$P(D|h) = \text{probability that } D \text{ is a computer given the speed and price}$

$P(D|h)$ is difficult to compute
Naïve Bayes Classifier (1)

- simplified assumption: features are conditionally independent:

\[
P(h|D) = \frac{P(D|h)P(h)}{P(D)} \quad P(C_j|V) \propto P(C_j) \prod_{i=1}^{n} P(v_i|C_j)
\]
Naive Bayesian Classifier (2)

- Given a training set, we can compute the probabilities

<table>
<thead>
<tr>
<th>Outlook</th>
<th>P</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>sunny</td>
<td>2/9</td>
<td>3/5</td>
</tr>
<tr>
<td>overcast</td>
<td>4/9</td>
<td>0</td>
</tr>
<tr>
<td>rain</td>
<td>3/9</td>
<td>2/5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Humidity</th>
<th>P</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>high</td>
<td>3/9</td>
<td>4/5</td>
</tr>
<tr>
<td>normal</td>
<td>6/9</td>
<td>1/5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Temperature</th>
<th>P</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>hot</td>
<td>2/9</td>
<td>2/5</td>
</tr>
<tr>
<td>mild</td>
<td>4/9</td>
<td>2/5</td>
</tr>
<tr>
<td>cool</td>
<td>3/9</td>
<td>1/5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Windy</th>
<th>P</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>true</td>
<td>3/9</td>
<td>3/5</td>
</tr>
<tr>
<td>false</td>
<td>6/9</td>
<td>2/5</td>
</tr>
</tbody>
</table>

Bayesian classification

\[ P(h|D) = \frac{P(D|h)P(h)}{P(D)} \]

- The classification problem may be formalized using a-posteriori probabilities:
  - \( P(C|X) \) = probability that the sample tuple \( X = <x_1, \ldots, x_k> \) is of class \( C \)
  - E.g., \( P(\text{class}=N \mid \text{outlook}=\text{sunny}, \text{windy}=\text{true}, \ldots) \)
  - Idea: assign sample \( X \) class label \( C \) such that \( P(C|X) \) is maximal
Estimating a-posteriori probabilities

- **Bayes theorem:**
  
  \[ P(C|X) = \frac{P(X|C) \cdot P(C)}{P(X)} \]

  - \( P(X) \) is constant for all classes
  - \( P(C) \) = relative freq of class C samples
  - C such that \( P(C|X) \) is maximum = C such that \( P(X|C) \cdot P(C) \) is maximum
  - Problem: computing \( P(X|C) \) is infeasible!

Naïve Bayesian Classification

- Naïve assumption: attribute independence
  
  \[ P(x_1, \ldots, x_k|C) = P(x_1|C) \cdot \ldots \cdot P(x_k|C) \]

  - If i-th attribute is categorical:
    \( P(x_i|C) \) is estimated as the relative freq of samples having value \( x_i \) as i-th attribute in class C
  - If i-th attribute is continuous:
    \( P(x_i|C) \) is estimated thru a Gaussian density function
  - Computationally easy in both cases
Play-tennis example: estimating $P(x_i|C)$

<table>
<thead>
<tr>
<th>Outlook</th>
<th>Temperature</th>
<th>Humidity</th>
<th>Windy</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>sunny</td>
<td>hot</td>
<td>high</td>
<td>false</td>
<td>N</td>
</tr>
<tr>
<td>sunny</td>
<td>hot</td>
<td>high</td>
<td>true</td>
<td>N</td>
</tr>
<tr>
<td>overcast</td>
<td>hot</td>
<td>high</td>
<td>false</td>
<td>P</td>
</tr>
<tr>
<td>rain</td>
<td>mild</td>
<td>high</td>
<td>false</td>
<td>P</td>
</tr>
<tr>
<td>rain</td>
<td>cool</td>
<td>normal</td>
<td>false</td>
<td>P</td>
</tr>
<tr>
<td>rain</td>
<td>cool</td>
<td>normal</td>
<td>true</td>
<td>N</td>
</tr>
<tr>
<td>overcast</td>
<td>cool</td>
<td>normal</td>
<td>true</td>
<td>P</td>
</tr>
<tr>
<td>sunny</td>
<td>mild</td>
<td>high</td>
<td>false</td>
<td>N</td>
</tr>
<tr>
<td>sunny</td>
<td>cool</td>
<td>normal</td>
<td>false</td>
<td>P</td>
</tr>
<tr>
<td>rain</td>
<td>mild</td>
<td>normal</td>
<td>false</td>
<td>P</td>
</tr>
<tr>
<td>overcast</td>
<td>mild</td>
<td>high</td>
<td>true</td>
<td>P</td>
</tr>
<tr>
<td>overcast</td>
<td>hot</td>
<td>normal</td>
<td>false</td>
<td>P</td>
</tr>
<tr>
<td>rain</td>
<td>mild</td>
<td>high</td>
<td>true</td>
<td>N</td>
</tr>
</tbody>
</table>

$P(p) = 9/14$

$P(n) = 5/14$

P(overcast|p) = 4/9
P(overcast|n) = 0
P(rain|p) = 3/9
P(rain|n) = 2/5

P(true|n) = 3/5
P(true|p) = 3/9

P(false|n) = 2/5
P(false|p) = 6/9

P(high|n) = 4/5
P(high|p) = 2/9

P(normal|n) = 2/5
P(normal|p) = 6/9

P(cool|n) = 1/5
P(cool|p) = 4/9

P(mild|n) = 2/5
P(mild|p) = 2/9

P(rain|n) = 2/5
P(rain|p) = 3/9

P(overcast|n) = 0
P(overcast|p) = 4/9

P(sunny|n) = 3/5
P(sunny|p) = 2/9

P(hot|n) = 2/5
P(hot|p) = 2/9

P(cool|p) = 3/9
P(cool|p) = 3/9

P(mild|p) = 4/9
P(mild|p) = 4/9

P(overcast|p) = 4/9
P(overcast|n) = 0

P(sunny|p) = 2/9
P(sunny|n) = 3/5

P(high|p) = 2/9
P(high|p) = 2/9

P(normal|p) = 6/9
P(normal|n) = 2/5

P(true|p) = 3/9
P(true|n) = 3/5

P(false|p) = 6/9
P(false|n) = 2/5

Classifying $X$

- An unseen sample $X = \langle\text{rain, hot, high, false}\rangle$

$P(X|p) \cdot P(p) = P(\text{rain}|p) \cdot P(\text{hot}|p) \cdot P(\text{high}|p) \cdot P(\text{false}|p) \cdot P(p) = 3/9 \cdot 2/9 \cdot 3/9 \cdot 6/9 \cdot 9/14 = 0.010582$

$P(X|n) \cdot P(n) = P(\text{rain}|n) \cdot P(\text{hot}|n) \cdot P(\text{high}|n) \cdot P(\text{false}|n) \cdot P(n) = 2/5 \cdot 2/5 \cdot 4/5 \cdot 2/5 \cdot 5/14 = 0.018286$

- Sample $X$ is classified in class $n$ (don’t play)
The independence hypothesis…

- … makes computation possible
- … yields optimal classifiers when satisfied
- … but is seldom satisfied in practice, as attributes (variables) are often correlated.

Attempts to overcome this limitation:

- **Bayesian networks**, that combine Bayesian reasoning with causal relationships between attributes
- **Decision trees**, that reason on one attribute at the time, considering most important attributes first

---

Bayesian Belief Networks (1)

The conditional probability table for the variable LungCancer

<table>
<thead>
<tr>
<th></th>
<th>LC</th>
<th>~LC</th>
</tr>
</thead>
<tbody>
<tr>
<td>(FH, S)</td>
<td>0.8</td>
<td>0.2</td>
</tr>
<tr>
<td>(FH, ~S)</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>(~FH, S)</td>
<td>0.7</td>
<td>0.3</td>
</tr>
<tr>
<td>(~FH, ~S)</td>
<td>0.1</td>
<td>0.9</td>
</tr>
</tbody>
</table>

Bayesian Belief Networks
Bayesian Belief Networks (2)

- Bayesian belief network allows a *subset* of the variables conditionally independent
- A graphical model of causal relationships
- Several cases of learning Bayesian belief networks
  - Given both network structure and all the variables: easy
  - Given network structure but only some variables
  - When the network structure is not known in advance

Reference

- J. Han, M. Kamber; *Data Mining*, 2001; Morgan Kaufmann Publishers: San Francisco, CA.