Hugin & Weka for Learning from Data
Using Bayesian Networks

Ashwin Patthi
University of South Carolina
Predicting Probability distribution of unknown variables from data/cases

Resulting learning model can be used as classifier

Knowledge discovery
Classification Task

Classifying variable $y = x_0$ called the class variable given set of variables $X = x_1, x_2, \ldots, x_k$ called attribute variables.
Classifier is learned from a dataset $D$ consisting of samples over $(x, y)$ on network $B_x$ over probability distribution $U$.

$h : X \rightarrow y$
Learning using Hugin

Uses EM (Estimation Maximization) Algorithm (Batch learning)
This is used only when structure is available
Experience table must be provided for nodes whose conditional probabilities are to inferred
Performs number of iteration on cases
Computes log-likelihood and attempts to maximize it
 Stops when two successive log-likelihood is less than tolerance
Learning on Chest Clinic data

- Visit to Asia?
- Smoker?
- Tuberculosis?
- Lung cancer?
- Bronchitis?
- Tuberculosis or cancer?
- Positive X-ray?
- Dyspnoea?
Assumptions

No prior knowledge on any distribution

Set probability distribution to 1 except for “Tuberculosis or Cancer”

Also experience value to 10 (or some low value)
Data set (asia.data)

First line is header

Each record is a case

N/A = Not available
Run the Learning Algo
Resulting Marginal Probabilities

- **Visit to Asia?**
  - Yes: 45.39
  - No: 54.61

- **Smoker?**
  - Yes: 42.62
  - No: 57.38

- **Chest Clinic Present?**
  - Yes: 6.67
  - No: 93.33

- **Positive X-ray?**
  - Yes: 13.08
  - No: 86.92

- **Smoker?**
  - Yes: 49.90
  - No: 50.10

- **Tuberculosis or Cancer?**
  - Yes: 8.88
  - No: 91.12

- **Tuberculosis?**
  - Yes: 2.37
  - No: 97.63

- **Dyspnoea?**
  - Yes: 2.48
  - No: 97.52
The resulting model with computed conditional probabilities can be used as classifier to predict the new unknown conditional probabilities.
Collection of Machine Learning Algorithms for data mining tasks
Contains tools for data pre-processing, classification, clustering, visualization etc.
Bayesian Network classifier and editor is one of them
Basic Assumptions

All variables are discrete finite variables
If continuous can convert to discrete using class filters.attribute.Discretize

No instances have missing values. If found can be filled by attribute.ReplaceMissingValues
Input to Bayesian Net Classifier
## Results

![Image of Weka classifier output]

### Classifier Options

- **Test options**
  - Use training set
  - Supplied test set
  - Cross-validation
  - Percentage split

### Classifier Output

- **LogScore Bayes:** -26771.703539888724
- **LogScore BDeu:** -26783.728842977427
- **LogScore MDL:** -26799.68957920934
- **LogScore ENTROPY:** -26730.91202641952
- **LogScore AIC:** -26745.91202641952

**Time taken to build model:** 3.4 seconds

### Stratified Cross-Validation

<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correctly Classified Instances</td>
<td>9483</td>
</tr>
<tr>
<td>Incorrectly Classified Instances</td>
<td>517</td>
</tr>
<tr>
<td>Kappa statistic</td>
<td>0.6676</td>
</tr>
<tr>
<td>Mean absolute error</td>
<td>0.0753</td>
</tr>
<tr>
<td>Root mean squared error</td>
<td>0.223</td>
</tr>
<tr>
<td>Relative absolute error</td>
<td>39.7663</td>
</tr>
<tr>
<td>Root relative squared error</td>
<td>72.5024</td>
</tr>
<tr>
<td>Total Number of Instances</td>
<td>10000</td>
</tr>
</tbody>
</table>

### Detailed Accuracy By Class

<table>
<thead>
<tr>
<th>Class</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class 1</td>
<td>94.83%</td>
</tr>
<tr>
<td>Class 2</td>
<td>5.17%</td>
</tr>
</tbody>
</table>
Pros & Cons of Hugin

Pros:
Can generate data with missing values based on available network
Can predict prob. distribution for all variables

Cons:
No feature for handling missing values
No feature to predict class of all the cases at once
Pros and Cons of WEKA

Pros:
Feature to handle missing values
Can predict class of unknown data sets all at once

Cons:
Can’t generate data with missing values
Can’t predict probability distribution
1. Bouckaert, Remco R. Bayesian network classifiers in weka. Department of Computer Science, University of Waikato, 2004

2. Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1

3. More information on Hugin tool

4. More information on WEKA tool
Questions ?