Chapter 8 – Naïve Bayes

Data Mining for Business Intelligence
Shmueli, Patel & Bruce

© Galit Shmueli and Peter Bruce 2010
Characteristics

Data-driven, not model-driven

Make no assumptions about the data
Naïve Bayes: The Basic Idea

For a given new record to be classified, find other records like it (i.e., same values for the predictors)

What is the prevalent class among those records?

Assign that class to your new record
Usage

- Requires categorical variables
- Numerical variable must be binned and converted to categorical
- Can be used with very large data sets
- Example: Spell check programs assign your misspelled word to an established “class” (i.e., correctly spelled word)
Exact Bayes Classifier

Relies on finding other records that share **same predictor values** as record-to-be-classified.

Want to find “probability of belonging to class C, given specified values of predictors.”

Even with large data sets, may be hard to find other records that **exactly match** your record, in terms of predictor values.
Solution – Naïve Bayes

• Assume independence of predictor variables (within each class)

• Use multiplication rule

• Find same probability that record belongs to class C, given predictor values, without limiting calculation to records that share all those same values
Calculations

1. Take a record, and note its predictor values
2. Find the probabilities those predictor values occur across all records in C1
3. Multiply them together, then by proportion of records belonging to C1
4. Same for C2, C3, etc.
5. Prob. of belonging to C1 is value from step (3) divide by sum of all such values C1 ... Cn
6. Establish & adjust a “cutoff” prob. for class of interest
Example: Financial Fraud

Target variable: Audit finds fraud, no fraud

Predictors:
- Prior pending legal charges (yes/no)
- Size of firm (small/large)
<table>
<thead>
<tr>
<th>Charges?</th>
<th>Size</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>small</td>
<td>truthful</td>
</tr>
<tr>
<td>n</td>
<td>small</td>
<td>truthful</td>
</tr>
<tr>
<td>n</td>
<td>large</td>
<td>truthful</td>
</tr>
<tr>
<td>n</td>
<td>large</td>
<td>truthful</td>
</tr>
<tr>
<td>n</td>
<td>small</td>
<td>truthful</td>
</tr>
<tr>
<td>n</td>
<td>small</td>
<td>truthful</td>
</tr>
<tr>
<td>y</td>
<td>small</td>
<td>fraud</td>
</tr>
<tr>
<td>y</td>
<td>large</td>
<td>fraud</td>
</tr>
<tr>
<td>n</td>
<td>large</td>
<td>fraud</td>
</tr>
<tr>
<td>y</td>
<td>large</td>
<td>fraud</td>
</tr>
</tbody>
</table>
Exact Bayes Calculations

**Goal:** classify (as “fraudulent” or as “truthful”) a small firm with charges filed

There are 2 firms like that, one fraudulent and the other truthful

\[ P(\text{fraud} | \text{charges}=y, \text{size}=\text{small}) = \frac{1}{2} = 0.50 \]

Note: calculation is limited to the two firms matching those characteristics
Naïve Bayes Calculations

Same goal as before

Compute 2 quantities:
- Proportion of “charges = y” among frauds, times proportion of “small” among frauds, times proportion frauds = \( \frac{3}{4} \times \frac{1}{4} \times \frac{4}{10} = 0.075 \)
- Prop “charges = y” among frauds, times prop. “small” among truthsfuls, times prop. truthsfuls = \( \frac{1}{6} \times \frac{4}{6} \times \frac{6}{10} = 0.067 \)

\[
P(\text{fraud} \mid \text{charges, small}) = \frac{0.075}{(0.075 + 0.067)} = 0.53
\]
Naïve Bayes, cont.

- Note that probability estimate does not differ greatly from exact

- All records are used in calculations, not just those matching predictor values

- This makes calculations practical in most circumstances

- Relies on assumption of independence between predictor variables within each class
Independence Assumption

- Not strictly justified (variables often correlated with one another)
- Often “good enough”
Advantages

- Handles purely categorical data well
- Works well with very large data sets
- Simple & computationally efficient
Shortcomings

- Requires large number of records

- Problematic when a predictor category is not present in training data
  
  Assigns 0 probability of response, ignoring information in other variables
On the other hand...

- Probability rankings are more accurate than the actual probability estimates
  Good for applications using lift (e.g. response to mailing), less so for applications requiring probabilities (e.g. credit scoring)
Summary

• No statistical models involved

• Naïve Bayes (like KNN) pays attention to complex interactions and local structure

• Computational challenges remain