Classification - An Overview over Standard and Credal Classification Approaches with an Emphasis on Instability and Interpretability

Carolin Strobl
Department of Statistics, LMU

Durham 2008
Classification - Types of...

- Unsupervised
  - Cluster analysis

- Supervised
  - Nonparametric
    - Classification trees
    - Credal classification trees (Abellan & Moral)
  - Parametric
    - Logistic regression
    - Discriminant analysis
    - Naive Bayes classifier
    - Naive credal classifier (Zaffalon)

References
Classification - Types of...

- Classification
  - Unsupervised
    - Cluster analysis
  - Supervised
    - Nonparametric
      - Classification trees
        - Credal classification trees (Abellan & Moral)
    - Parametric
      - Logistic regression
      - Discriminant analysis
      - Naive Bayes classifier
        - Naive credal classifier (Zaffalon)

Classification trees
Ensemble Methods
TWIX
Selecting extra cutpoints
Choice of $s_{\text{MAX}}$
Summary
References
Supervised classification

predict $Y$, e.g., $\in \{1, 2\}$ from $X_1, \ldots, X_p$ of any type

class attributes
response predictors
dependent variable independent variables
covariates
## Supervised classification

<table>
<thead>
<tr>
<th>$X_2$</th>
<th>$a$</th>
<th>$X_1$</th>
<th>$b$</th>
<th>$c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a$</td>
<td>$P(Y = 2</td>
<td>X_1 = a, X_2 = a)$</td>
<td>$P(Y = 2</td>
<td>X_1 = b, X_2 = a)$</td>
</tr>
<tr>
<td>$b$</td>
<td>$P(Y = 2</td>
<td>X_1 = a, X_2 = b)$</td>
<td>$P(Y = 2</td>
<td>X_1 = b, X_2 = b)$</td>
</tr>
<tr>
<td>$c$</td>
<td>$P(Y = 2</td>
<td>X_1 = a, X_2 = c)$</td>
<td>$P(Y = 2</td>
<td>X_1 = b, X_2 = c)$</td>
</tr>
<tr>
<td>$d$</td>
<td>$P(Y = 2</td>
<td>X_1 = a, X_2 = d)$</td>
<td>$P(Y = 2</td>
<td>X_1 = b, X_2 = d)$</td>
</tr>
<tr>
<td>$e$</td>
<td>$P(Y = 2</td>
<td>X_1 = a, X_2 = e)$</td>
<td>$P(Y = 2</td>
<td>X_1 = b, X_2 = e)$</td>
</tr>
</tbody>
</table>
Problem

cannot estimate each $P(Y = 2|X_1 = x_1, X_2 = x_2)$ from data set of limited size

especially not for many predictors, many categories…

⇒ need a shortcut
Examples for shortcuts

- Naive Bayes

\[
P(Y = 2|X_1, \ldots, X_p) = \frac{P(X_1, \ldots, X_p|Y = 2) \cdot P(Y = 2)}{P(X_1, \ldots, X_p|Y = 2) \cdot P(Y = 2) + P(X_1, \ldots, X_p|Y = 1) \cdot P(Y = 1)}
\]

priors \(\Rightarrow\) marginal class frequencies

shortcut: assume predictors are uncorrelated

\[
P(X_1, \ldots, X_p|Y = 2) \Rightarrow \prod_{j=1}^{p} P(X_j|Y = 2)
\]
Classification trees

shortcut: don’t look at full-factorial crosstab

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>a</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X₁</td>
<td>a</td>
<td>b</td>
<td></td>
</tr>
<tr>
<td>X₂</td>
<td></td>
<td>c</td>
<td></td>
</tr>
<tr>
<td></td>
<td>d</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>e</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

but build more parsimonious partition if sufficient

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>a</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X₁</td>
<td>a</td>
<td>b</td>
<td></td>
</tr>
<tr>
<td>X₂</td>
<td>c</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>d</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>e</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

by adding one predictor at a time recursively
Supervised classification

in general: learn some classification rule

\[ \hat{y} = \hat{f}(x_1, \ldots, x_p) \Rightarrow \hat{P}(Y = 2|X_1 = x_1, \ldots, X_p = x_p) \]

predict class membership or posterior probability

from a learning sample, where the values of \( Y \) and \( X_1, \ldots, X_p \) are known

later: predict unknown values of \( Y \) for new sample, where only \( X_1, \ldots, X_p \) are known
Supervised classification

aim: learn systematic aspects of association between $Y$ and $X_1, \ldots, X_p$

from randomly drawn learning sample
Supervised classification

aim: learn systematic aspects of association between $Y$ and $X_1, \ldots, X_p$

from randomly drawn learning sample

but do not adapt too closely to random variation in the learning data!
Classification - Types of...

- Classification
  - Unsupervised: Cluster analysis
  - Supervised
    - Nonparametric: Classification trees
    - Parametric: Logistic regression, discriminant analysis, naive Bayes classifier
  - Credal classification trees (Abellan & Moral)
  - Naive credal classifier (Zaffalon)

References
Classification - Types of...

- Classification
- Classification trees
  - Ensemble Methods
  - TWIX
  - Selecting extra cutpoints
  - Choice of $s_{\text{MAX}}$
  - Summary
  - References

Unsupervised
- Cluster analysis

Supervised
- Nonparametric
  - Classification trees
    - Credal classification trees
      - (Abellan & Moral)

- Parametric
  - Logistic regression
  - Discriminant analysis
  - Naive Bayes classifier
  - Naive credal classifier
    - (Zaffalon)
Classification - Types of...

Classification

Unsupervised
- Cluster analysis

Supervised
- Nonparametric
  - Classification trees
    - Credal classification trees
      - (Abellan & Moral)

- Parametric
  - Logistic regression
  - Discriminant analysis
  - Naive Bayes classifier
  - Naive credal classifier
    - (Zaffalon)

Choice of $s_{\text{MAX}}$

References
Classification trees

- **k-ary**
  - C 4,5
    - instable (⇒ overfitting)
  - credal classification trees
    (Abellan & Moral)
    - more robust, less overfitting
    - limited to categorical predictors
    - not parsimonious
    - variable selection bias (can be avoided)

- **binary**
  - CART
    - cutpoint selection in each variable
    - more flexible, parsimonious
    - instable (⇒ overfitting)
  - ensemble methods
    - combine set of trees
    - stable, better prediction
    - smooth function
    - not well interpretable

References
Classification trees: k-ary splitting

\[ X_3 = 1, X_1 = 1 \]
\[ X_3 = 1, X_1 = 2 \]
\[ X_3 = 1, X_1 = 3 \]
Credal classification trees

Abellan & Moral

- apply Impricise Dirichlet Model (IDM) locally in each node
Credal classification trees

Abellan & Moral

- apply Imprecise Dirichlet Model (IDM) locally in each node
  1. use upper entropy for conservative variable selection (⇒ avoid overfitting)

References

Credal classification trees

Abellan & Moral

- apply Impricise Dirichlet Model (IDM) locally in each node
  1. use upper entropy for conservative variable selection
     (⇒ avoid overfitting)
  2. make credal prediction in final nodes
Credal classification trees

Abellan & Moral

- apply Impricise Dirichlet Model (IDM) locally in each node
  1. use upper entropy for conservative variable selection (⇒ avoid overfitting)
  2. make credal prediction in final nodes

+ more robust, less overfitting
- limited to categorial predictors
- not parsimonious
- variable selection bias (can be avoided) (Strobl, 2005)
Classification trees: binary splitting

\[
\begin{align*}
X_3 &= 1, \quad X_2 = 2 \\
X_3 &= 1, \quad X_2 = 1 \lor X_2 = 3
\end{align*}
\]
Classification trees: binary splitting

\[ X_3 = 1, \ X_1 \leq 2.5 \]

\[ X_3 = 2 \]

\[ X_3 = 1, \ X_1 > 2.5 \]
Variable and cutpoint selection

for starting node $C$ and daughter nodes $C_{L,c_j}$ and $C_{R,c_j}$

- select the best cutpoint $c_j^*$ within the range of each predictor variable $X_j$ with respect to impurity reduction

$$c_j^* = \arg \max_{c_j} \Delta J (C, C_{L,c_j}, C_{R,c_j})$$

- choose the variable $X_{j^*}$ with best criterion value in its best cutpoint

$$X_{j^*} = \arg \max_j \left\{ \max_{c_j} \Delta J (C, C_{L,c_j}, C_{R,c_j}) \right\}$$

or stop (!)
Cutpoint selection

\[ y = 2 \]
\[ y = 1 \]
\[ X_j \]
Problems of classification trees

I. variable selection bias
   preference for, e.g., variables with many categories

solution:

▶ unbiased variable selection approaches (Loh and Shih, 1997; Kim and Loh, 2001; Dobra and Gehrke, 2001; Shih, 2004; Lausen et al., 2004; Strobl, 2005; Hothorn et al., 2006; Strobl et al., 2007)

▶ simplest solution: choose variable first, then cutpoint
Problems of classification trees

II. instability to small changes in the training sample

- small changes in the training sample
  ⇒ different best cutpoint
- different best cutpoint in one node
  ⇒ completely different tree

solution:
create ensembles (sets) of trees
Problems of classification trees

II. instability to small changes in the training sample

- small changes in the training sample
  ⇒ different best cutpoint
- different best cutpoint in one node
  ⇒ completely different tree

solution:
create ensembles (sets) of trees

- bagging (Breiman, 1996, 1998), random forests
  (Breiman, 2001)
Problems of classification trees

II. instability to small changes in the training sample

- small changes in the training sample
  ⇒ different best cutpoint
- different best cutpoint in one node
  ⇒ completely different tree

solution:
create ensembles (sets) of trees

- bagging (Breiman, 1996, 1998), random forests
  (Breiman, 2001)
- TWIX (Potapov, 2006)
Bootstrap sampling

population

sample

bootstrap samples
Bagging and random forests
Bagging and random forests

+ aggregated results are more stable
+ decision boundaries are smoothed
+ high prediction accuracy
~ random
  - not interpretable
TWIX

Classification
Classification trees
Ensemble Methods
TWIX
Selecting extra cutpoints
Choice of $s_{\text{MAX}}$
Summary
References
Our aim

make TWIX adaptive
Our aim

make TWIX adaptive

interpretability

stability, accuracy

single tree  ←  adaptive TWIX  →  bagging, random forest
Mountain plots

Classification
Selecting extra cutpoints
Adaptive selection
Results
Choice of $s_{\text{MAX}}$
Summary
References
Current selection rules

- select the best $m$ cutpoints
- select the best $m$ cutpoints that are local maxima

but: exponential growth in $m$!
Adding virtual observations

taking literally the question:

“would a different cutpoint be optimal, if we had slightly different data?”
Adding virtual observations

taking literally the question:

“would a different cutpoint be optimal, if we had slightly different data?”

⇒ add virtual observations and see what happens...

if the same cutpoint is still optimal: fine

but if not: split in the other one as well
we have already decided that we want to split in $X_j$ and want to select all cutpoints that may be suited for further splitting.

\[ y = 2 \]

\[ y = 1 \]
Unknown value of $X_j$

for each candidate cutpoint:
Unknown value of $X_j$

for each candidate cutpoint:

- add a virtual observation with an unknown value of $X_j$
Unknown value of $X_j$

for each candidate cutpoint:

- add a virtual observation with an unknown value of $X_j$
- see what happens if you assign it
Unknown value of $X_j$

for each candidate cutpoint:

- add a virtual observation with an unknown value of $X_j$
- see what happens if you assign it
  - to the left node (value of $X_j \leq$ cutpoint)
Unknown value of $X_j$

for each candidate cutpoint:

- add a virtual observation with an unknown value of $X_j$
- see what happens if you assign it
  - to the left node (value of $X_j \leq$ cutpoint)
  - to the right node (value of $X_j >$ cutpoint)
Unknown value of $X_j$

for each candidate cutpoint:

- add a virtual observation with an unknown value of $X_j$
- see what happens if you assign it
  - to the left node (value of $X_j \leq$ cutpoint)
  - to the right node (value of $X_j >$ cutpoint)
- overall: add $s_{max}$ observations
  - $s_L$ to left and $s_{max} - s_L$ to right node
  - with $s_L$ in 1, \ldots, $s_{max}$
  - and $s_{max}$ in 1, \ldots, $s_{MAX}$
Unknown value of $Y$

for each combination of observations:

- see what happens if you assign, e.g., the $s_L$ observations currently assigned to the left node holding the first $i$ observations of the sample ordered with respect to $X_j$
Unknown value of $Y$

for each combination of observations:

▶ see what happens if you assign, e.g, the $s_L$ observations currently assigned to the left node holding the first $i$ observations of the sample ordered with respect to $X_j$

▶ to class 1 or
Unknown value of $Y$

for each combination of observations:

- see what happens if you assign, e.g., the $s_L$ observations currently assigned to the left node holding the first $i$ observations of the sample ordered with respect to $X_j$
  - to class 1 or
  - to class 2
Unknown value of $Y$

for each combination of observations:

- see what happens if you assign, e.g., the $s_L$ observations currently assigned to the left node holding the first $i$ observations of the sample ordered with respect to $X_j$
  - to class 1 or
  - to class 2
  - probabilities for class 2

$$[\pi_L(i, s_L), \overline{\pi}_L(i, s_L)] = \left[ \frac{n_2(i)}{i + s_L} , \frac{n_2(i) + s_L}{i + s_L} \right]$$
Unknown value of $Y$

for each combination of observations:

- see what happens if you assign, e.g., the $s_L$ observations currently assigned to the left node holding the first $i$ observations of the sample ordered with respect to $X_j$
  - to class 1 or
  - to class 2
  - probabilities for class 2

$$[\pi_L(i, s_L), \overline{\pi}_L(i, s_L)] = \left[ \frac{n_2(i)}{i + s_L}, \frac{n_2(i) + s_L}{i + s_L} \right]$$

- coincide with upper and lower probabilities from IDM with hyperparameter $s_L$
Use upper entropy distribution

from these probability intervals:

▶ derive the distribution that corresponds to the most conservative evaluation of the impurity criterion $\Delta I$
Use upper entropy distribution

from these probability intervals:

- derive the distribution that corresponds to the most conservative evaluation of the impurity criterion $\Delta \mathcal{I}$
- i.e. the distribution that is closest to the uniform distribution over the classes
Use upper entropy distribution from these probability intervals:

- derive the distribution that corresponds to the most conservative evaluation of the impurity criterion $\Delta \mathcal{I}$
- i.e. the distribution that is closest to the uniform distribution over the classes
- upper entropy distribution $\pi_L^*(i, s_L)$
Use upper entropy distribution

from these probability intervals:

- derive the distribution that corresponds to the most conservative evaluation of the impurity criterion $\Delta \mathcal{I}$
- i.e. the distribution that is closest to the uniform distribution over the classes
- upper entropy distribution $\pi^*_L(i, s_L)$
- can be derived by means of the upper entropy algorithm of Abellán and Moral (2003)
Use upper entropy distribution

from these probability intervals:

- derive the distribution that corresponds to the most conservative evaluation of the impurity criterion $\Delta \mathcal{I}$
- i.e. the distribution that is closest to the uniform distribution over the classes
- upper entropy distribution $\pi_L^*(i, s_L)$
- can be derived by means of the upper entropy algorithm of Abellán and Moral (2003)
- one distribution $\Rightarrow$ one cutpoint
Extra cutpoints

make an extra split in any cutpoint that is optimal in any combination of virtual observations
Extra cutpoints

make an extra split in any cutpoint that is optimal in any combination of virtual observations

if partition is very instable $\Rightarrow$ a lot
Extra cutpoints

make an extra split in any cutpoint that is optimal in any combination of virtual observations

if partition is very instable ⇒ a lot

if partition is very stable ⇒ not too many
Extra cutpoints

make an extra split in any cutpoint that is optimal in any combination of virtual observations

if partition is very instable ⇒ a lot

if partition is very stable ⇒ not too many

⇒ complexity of ensemble depends on data
## Extra cutpoints

<table>
<thead>
<tr>
<th>variable</th>
<th>position $i$ necessary for change of cutpoints</th>
<th>$s^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_8$</td>
<td>33</td>
<td>65 58</td>
</tr>
<tr>
<td>$X_1$</td>
<td>31</td>
<td>38  3</td>
</tr>
<tr>
<td>$X_4$</td>
<td>38</td>
<td>32  5</td>
</tr>
</tbody>
</table>
Mountain plots

Classification
Selecting extra cutpoints
Adaptive selection
Results
Choice of $s_{\text{MAX}}$
Summary
References
Simulation study

vary influence of predictor (coefficient $\beta$ in a logistic regression model)

vary sample size $n$ and number of virtual observations $s_{MAX}$
## Average number of different cutpoints

<table>
<thead>
<tr>
<th>n</th>
<th>β</th>
<th>2</th>
<th>5</th>
<th>10</th>
<th>20</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>0.5</td>
<td>1.65</td>
<td>2.21</td>
<td>3.16</td>
<td>6.29</td>
<td>13.21</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>1.52</td>
<td>2.09</td>
<td>2.76</td>
<td>4.11</td>
<td>13.15</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>1.36</td>
<td>1.89</td>
<td>2.66</td>
<td>3.47</td>
<td>12.61</td>
</tr>
<tr>
<td></td>
<td>2.0</td>
<td>1.36</td>
<td>1.95</td>
<td>2.27</td>
<td>3.29</td>
<td>12.18</td>
</tr>
<tr>
<td>200</td>
<td>0.5</td>
<td>1.39</td>
<td>1.85</td>
<td>2.55</td>
<td>4.40</td>
<td>12.37</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>1.33</td>
<td>1.84</td>
<td>2.33</td>
<td>3.00</td>
<td>9.45</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>1.38</td>
<td>1.72</td>
<td>2.27</td>
<td>2.91</td>
<td>6.31</td>
</tr>
<tr>
<td></td>
<td>2.0</td>
<td>1.29</td>
<td>1.79</td>
<td>2.29</td>
<td>3.04</td>
<td>4.66</td>
</tr>
</tbody>
</table>
Choice of $s_{\text{MAX}}$

- certain percentage of the sample size
Choice of $s_{MAX}$

- certain percentage of the sample size
  - for example: about 5% of the original data set might consist of faulty observations, erroneous measurements...
Choice of $s_{MAX}$

- certain percentage of the sample size
  - for example: about 5% of the original data set might consist of faulty observations, erroneous measurements...
  - Hampel (1980) cites historical data with up to 40% severe errors
Choice of $s_{\text{MAX}}$

- certain percentage of the sample size
  - for example: about 5% of the original data set might consist of faulty observations, erroneous measurements...
  - Hampel (1980) cites historical data with up to 40% severe errors
- number of subjects who did not participate in the study
Choice of $s_{\text{MAX}}$

- Certain percentage of the sample size
  - For example: about 5% of the original data set might consist of faulty observations, erroneous measurements...
  - Hampel (1980) cites historical data with up to 40% severe errors

- Number of subjects who did not participate in the study

- Could also be chosen relative to the sample size in the current node
Summary

single trees are nice to interpret but instable
Summary

single trees are nice to interpret but instable

ensembles give better predictions but are hard to interpret
single trees are nice to interpret but instable

ensembles give better predictions but are hard to interpret

adaptively selecting the number of cutpoints can
Summary

single trees are nice to interpret but instable

ensembles give better predictions but are hard to interpret

adaptively selecting the number of cutpoints can

▶ reduce the complexity
Summary

single trees are nice to interpret but instable

ensembles give better predictions but are hard to interpret

adaptively selecting the number of cutpoints can

- reduce the complexity
- so that a single tree may result as a special case
Summary

single trees are nice to interpret but instable

ensembles give better predictions but are hard to interpret

adaptively selecting the number of cutpoints can

▶ reduce the complexity
▶ so that a single tree may result as a special case
▶ otherwise serve as a diagnostic
Summary

single trees are nice to interpret but instable

ensembles give better predictions but are hard to interpret

adaptively selecting the number of cutpoints can

▶ reduce the complexity
▶ so that a single tree may result as a special case
▶ otherwise serve as a diagnostic

open question: aggregation of (interval-valued) results


