Learning & Association Rules
Part 2

Statistical Data Mining II

Spring 2016

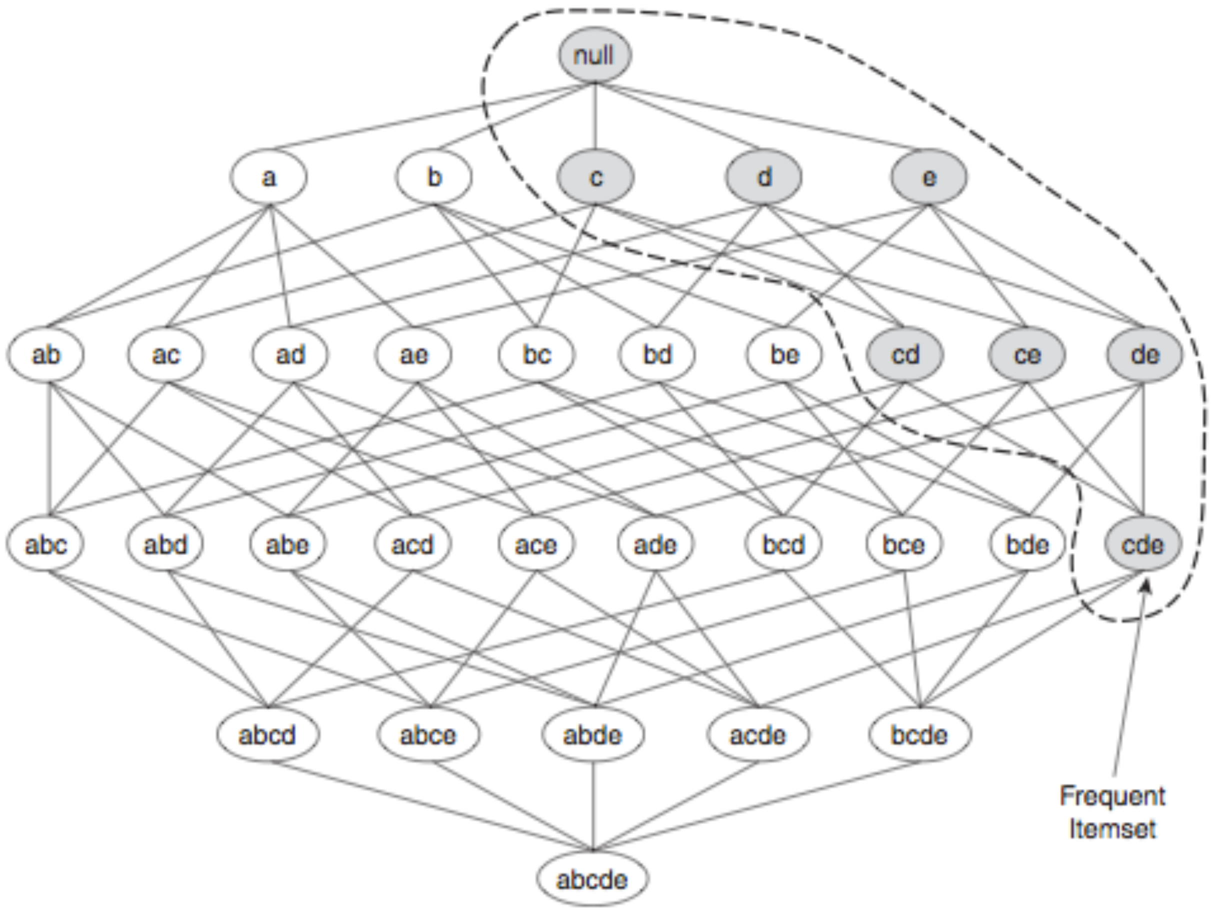
Rachael Hageman Blair

Outline

- Quick Recap
- An Example
- Solved by Association Rule Mining
- Generalizing Association Rules

Recap

Apriori principle: If an item set is frequent, then all of its subsets must also be frequent.



Recap

Significant decrease in computation, for example:

enumerating all itemsets (up to size 3) as candidates will produce

$$\binom{6}{1} + \binom{6}{2} + \binom{6}{3} = 6 + 15 + 20 = 41$$

candidates. With the *Apriori* principle, this number decreases to

$$\binom{6}{1} + \binom{4}{2} + 1 = 6 + 6 + 1 = 13$$

candidates, which represents a 68% reduction in the number of candidate itemsets even in this simple example.

Recap

Some things have to be set:

- Recast the problem as a binary one.
- Threshold on item set confidence $T(K) \geq \mathrm{t}$
- Threshold rules on confidence: $\left\{\mathrm{A} \Rightarrow \mathrm{B} \mid \mathrm{C}(\mathrm{A} \Rightarrow \mathrm{B}) > c\right\}$.
- Also examine "lift": $\Pr\left(A \text{ and } B\right) / \Pr(A)\Pr(B)$

  or in "Market Basket terminology": $L(A \Rightarrow B) = \dfrac{C(A \Rightarrow B)}{T(B)}$.

Example: Voting Records

Data (UCI machine learning data repository):

- Voting records of members of the United States House of Representatives.
- The data is obtained from the 1984 Congressional Voting Records Database. Each transaction contains information about the party affiliation for a representative along with his or her voting record on 16 key issues.

There are 435 transactions and 34 items in the data set.

Example: Voting Records

Table 6.3. List of binary attributes from the 1984 United States Congressional Voting Records. Source: The UCI machine learning repository.

1. Republican
2. Democrat
3. handicapped-infants = yes
4. handicapped-infants = no
5. water project cost sharing = yes
6. water project cost sharing = no
7. budget-resolution = yes
8. budget-resolution = no
9. physician fee freeze = yes
10. physician fee freeze = no
11. aid to El Salvador = yes
12. aid to El Salvador = no
13. religious groups in schools = yes
14. religious groups in schools = no
15. anti-satellite test ban = yes
16. anti-satellite test ban = no
17. aid to Nicaragua = yes
18. aid to Nicaragua = no
19. MX-missile = yes
20. MX-missile = no
21. immigration = yes
22. immigration = no
23. synfuel corporation cutback = yes
24. synfuel corporation cutback = no
25. education spending = yes
26. education spending = no
27. right-to-sue = yes
28. right-to-sue = no
29. crime = yes
30. crime = no
31. duty-free-exports = yes
32. duty-free-exports = no
33. export administration act = yes
34. export administration act = no

Table 6.4. Association rules extracted from the 1984 United States Congressional Voting Records.

Association Rule	Confidence
{budget resolution = no, MX-missile=no, aid to El Salvador = yes } $\longrightarrow$ {Republican}	91.0%
{budget resolution = yes, MX-missile=yes, aid to El Salvador = no } $\longrightarrow$ {Democrat}	97.5%
{crime = yes, right-to-sue = yes, physician fee freeze = yes} $\longrightarrow$ {Republican}	93.5%
{crime = no, right-to-sue = no, physician fee freeze = no} $\longrightarrow$ {Democrat}	100%

Example: Survey Data

Dataset: **9,409** questionnaires filled out by shopping mall customers in San Francisco. Utilized answers to the first 14 questions, related to demographics.

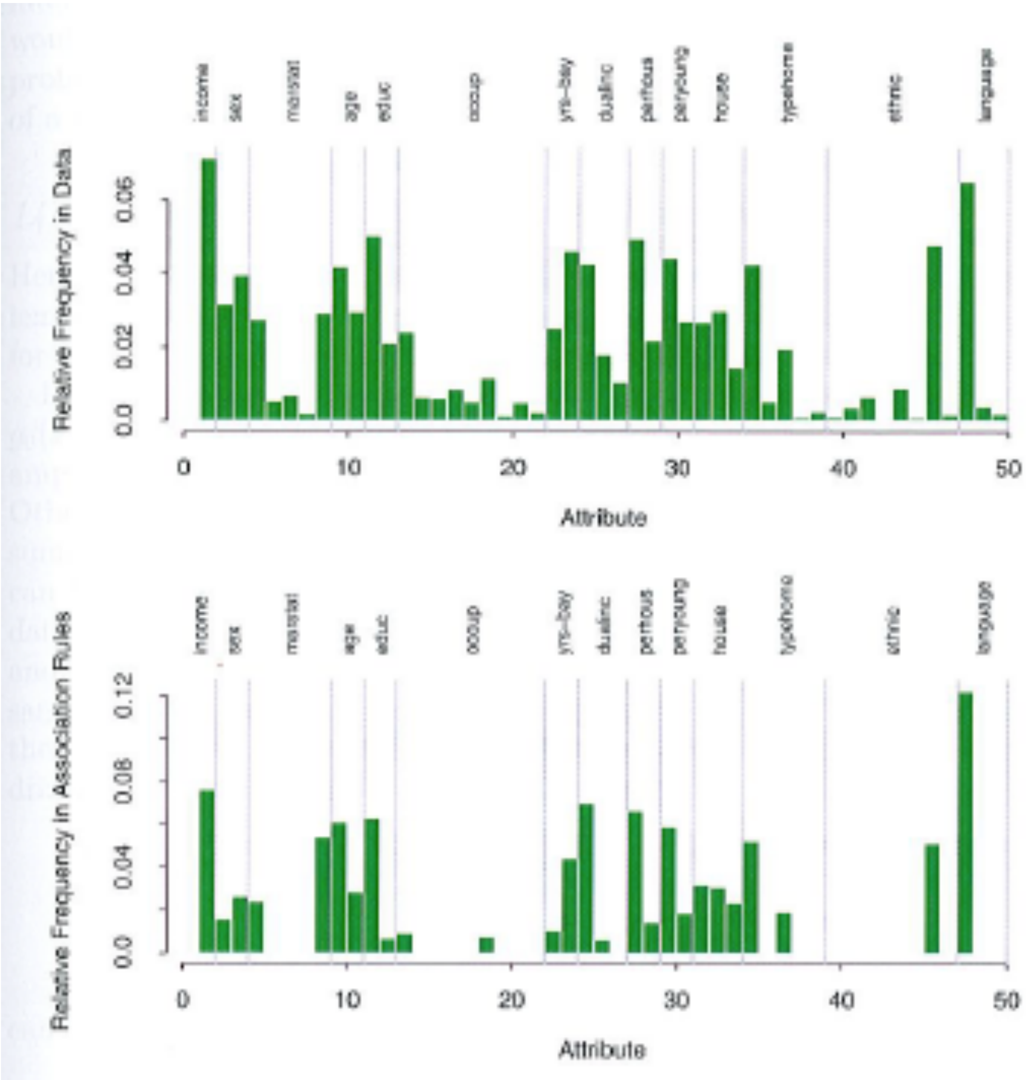
Feature	Demographic	# Values	Type
1	Sex	2	Categorical
2	Marital status	5	Categorical
3	Age	7	Ordinal
4	Education	6	Ordinal
5	Occupation	9	Categorical
6	Income	9	Ordinal
7	Years in Bay Area	5	Ordinal
8	Dual incomes	3	Categorical
9	Number in household	9	Ordinal
10	Number of children	9	Ordinal
11	Householder status	3	Categorical
12	Type of home	5	Categorical
13	Ethnic classification	8	Categorical
14	Language in home	3	Categorical

In R package:

```
> library(ElemStatLearn)
> ?marketing
```

An Example

Results: The algorithm found a total of 6288 association rules, involving less than 5 predictors with support of at least 10%.



An Example

Some Association Rules with “high support, confidence and lift”:

- Support 25%, confidence 99.7% and lift 1.03:

$$\begin{bmatrix} \text{number in household} & = & \mathit{1} \\ \text{number of children} & = & \mathit{0} \end{bmatrix}$$
$$\Downarrow$$
$$\text{language in home} = \mathit{English}$$

- Support 13.4%, confidence 80.8%, and lift 2.13:

$$\begin{bmatrix} \text{language in home} & = & \mathit{English} \\ \text{householder status} & = & \mathit{own} \\ \text{occupation} & = & \{\mathit{professional/managerial}\} \end{bmatrix}$$
$$\Downarrow$$
$$\text{income} \geq \mathit{\$40{,}000}$$

- Support 26.5%, confidence 82.8% and lift 2.15:

$$\begin{bmatrix} \text{language in home} & = & \mathit{English} \\ \text{income} & < & \mathit{\$40{,}000} \\ \text{marital status} & = & \mathit{not\ married} \\ \text{number of children} & = & \mathit{0} \end{bmatrix}$$
$$\Downarrow$$
$$\text{education} \notin \{\mathit{college\ graduate},\ \mathit{graduate\ study}\}$$

More “hands-on” examples coming…

Freeware (http://www.borgelt.net/apriori.html) standalone, C, developed by software engineer.

Alternative package: “arules” -> calls the C implementation by Christian Borgelt.

Preprocessing: Removing observations with missing values, each ordinal predictor was cut at its median and coded by two dummy variables, the resulting dataset is 6876 (obs) x 50 (dummy variables)

Generalized Association Rules

- Relax this idea of operating with huge databases.

- **<u>Focus:</u>** Related and important problem:
  Identify high-density regions of the model space.

<u>First pass</u>*: Cast an unsupervised problem as a supervised problem.*

↓

<u>Second pass</u>*: Generalize Association Rules in this context.*

Unsupervised disguised as Supervised

Let $g(x)$ be the unknown data probability density to be estimated.

Let $g_0(x)$ be the reference density (e.g., uniform over range of variables)

A sample, $x_1, x_2, \ldots, x_N$, is drawn from $g(x)$. (real data)

A sample of size N_0 can be drawn from $g_0(x)$. (simulated data)

- We can pool the sample, and assign weights, to create a sample drawn from the mixture density: $\left(g(x) + g_0(x)\right)/2$.
  Assign $w = N_0 / N + N_0$ to those drawn from $g(x)$.
  Assign $w_0 = N / N + N_0$ to those drawn from $g_0(x)$.

Unsupervised disguised as Supervised

<u>We can assign the following class labels</u>:

$Y = 1$ to each sample point drawn from $g(x)$, and

$Y = 0$ to those points drawn from $g_0(x)$, then:

$$\mu(x) = E(Y \mid x) = \frac{g(x)}{g(x) + g_0(x)}$$

$$= \frac{g(x)/g_0(x)}{1 + g(x)/g_0(x)}.$$

Which can be estimated via supervised learning using the combined sample:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_{N+N_0}, y_{N+N_0}).$$

Unsupervised disguised as Supervised

We can calculate $\hat{g}(x)$ from this estimate:

$$\hat{g}(x) = g_0(x) \frac{\hat{\mu}(x)}{1 - \hat{\mu}(x)}.$$

(akin to logistic regression).

*A good framework for this translation, logistic regression.
Other methods work as well.

Unsupervised disguised as Supervised

We can calculate $\hat{g}(x)$ from this estimate:

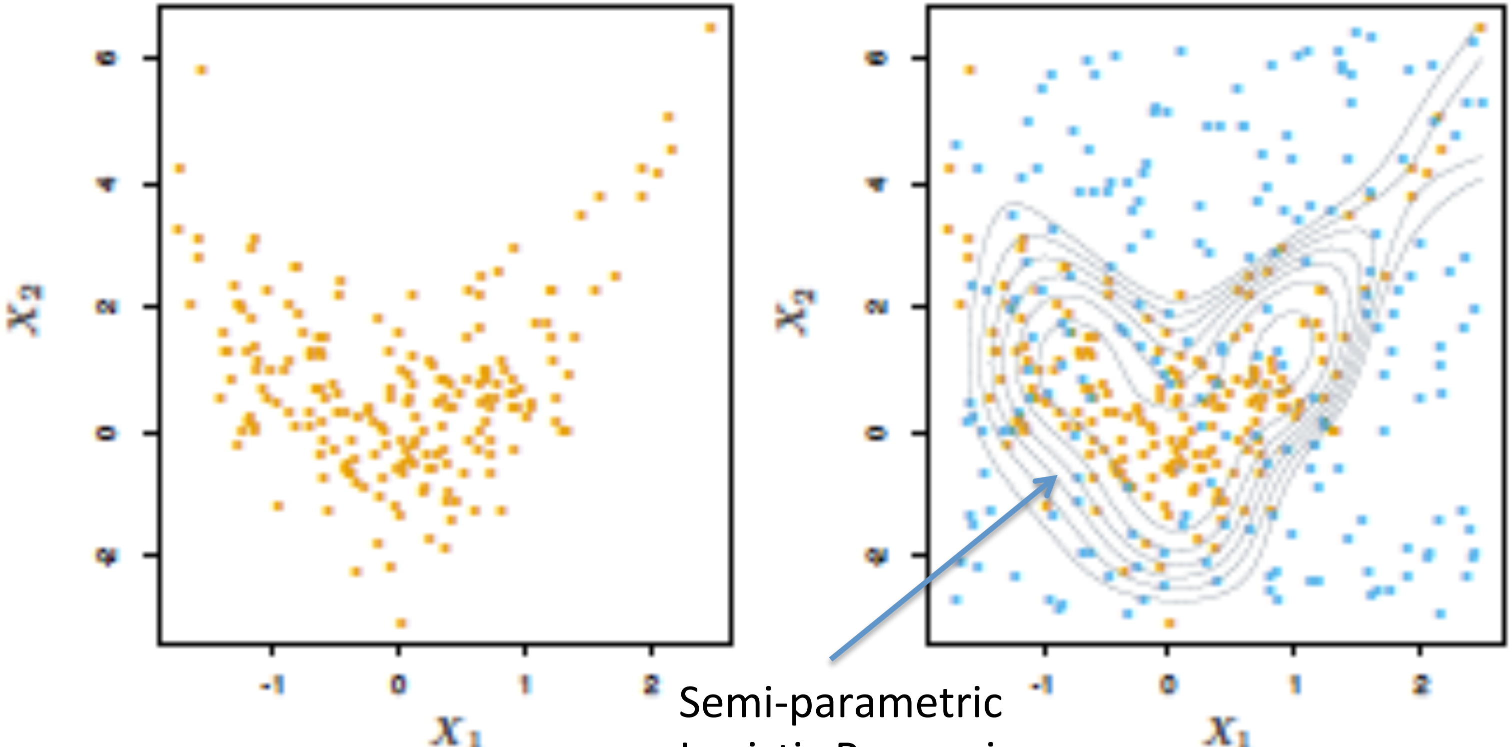
$$\hat{g}(x) = g_0(x)\frac{\hat{\mu}(x)}{1-\hat{\mu}(x)}.$$

(akin to logistic regression).

*A good framework for this translation, logistic regression.
Other methods work as well.

*Through the manipulation of the conditional mean estimate, we have acquired an estimate of underlying density.

Unsupervised disguised as Supervised



Semi-parametric
Logistic Regression

Training set Y=1
200 points

Reference set Y=0 (blue)
200 points

*Ideas can be extended for any reference density, choice determines accuracy.

Unsupervised disguised as Supervised

- If the goal is accuracy then $g_0(x)$ should be chosen carefully.
- Typically accuracy is not the goal in this case.
- Note $\mu(x)$ and $g(x)$ are both monotonic functions of the ratio $g(x)/g_0(x)$.

We are thinking in terms of "ratios", "contrasts", by that, we can think of "departures", in this case, how far off is the reference density from the true density.

departures from uniformity -> $g_0(x)$ might be uniform
departures from normality -> $g_0(x)$ would be Gaussian
departures from independence -> $g_0(x)=\prod_{j=1}^{p} g_j(x_j)$, where $g_j(x_j)$ is the marginal data of X_j.

Unsupervised disguised as Supervised

Not commonplace:

- The size of the generated data must be at least as large as the data sample, $N_0 \geq N$.
- May be in an underdetermined system.
- Monte Carlo sample may require substantial computation.

*These obstacles are vanishing in some cases, due to computational machinery.

*Promising in terms of reopening as a research area

Unsupervised disguised as Supervised

Recall: Market Basket Analysis

We want to find regions to maximize

Variable values (support)

$$\Pr\left[\bigcap_{j=1}^{p}\left(X_j \in s_j\right)\right]$$

Conjunctive Rule

We had to simplify this, in order to tackle "real world problems".

We can cast this problem in a supervised framework.

Generalized Association Rules

Reformulation

Find subsets of the integers $J \subset \{1,2,\ldots,p\}$ and corresponding value subsets $s_j, j \in J$ for the corresponding variables X_j, such that:

$$\hat{\Pr}\left[\bigcap_{j\in J}\left(X_j \in s_j\right)\right] = \frac{1}{N}\sum_{i=1}^{N} I\left(\bigcap_{j\in J} x_{ij} \in s_j\right)$$

is large.

The set: $x_{ij} \in s_j$ is called a "generalized" item set.

*Note – subsets corresponding to quantitative variables, have to be diced up into intervals, and categorical variables can also be broken up.

Generalized Association Rules

Favoritism

- That is, we look for sets that are more frequent than would be expected if all joint values were uniformly distributed.

In general favors the item sets whose marginal constitutes are individually frequent:

$$\frac{1}{N}\sum_{i=1}^{N} I\left(\bigcap_{j\in J} x_{ij} \in s_j\right)$$

is large.

Generalized Association Rules

- Choose a reference distribution, draw a sample, assigning a binary output Y. The goal is to use the training data to find the regions:

$$R = \bigcap_{j \in J} X_j \in s_j$$

for which the target function $\mu(x) = E\left(Y \mid x\right)$ is relatively large. May also require data support of regions

$$T\left(R\right) = \int_{x \in R} g(x)\, dx$$

is not too small.

Generalized Association Rules

- Choose a reference distribution, draw a sample, assigning a binary output Y. The goal is to use the training data to find the regions:

$$R = \bigcap_{j \in J} X_j \in s_j$$

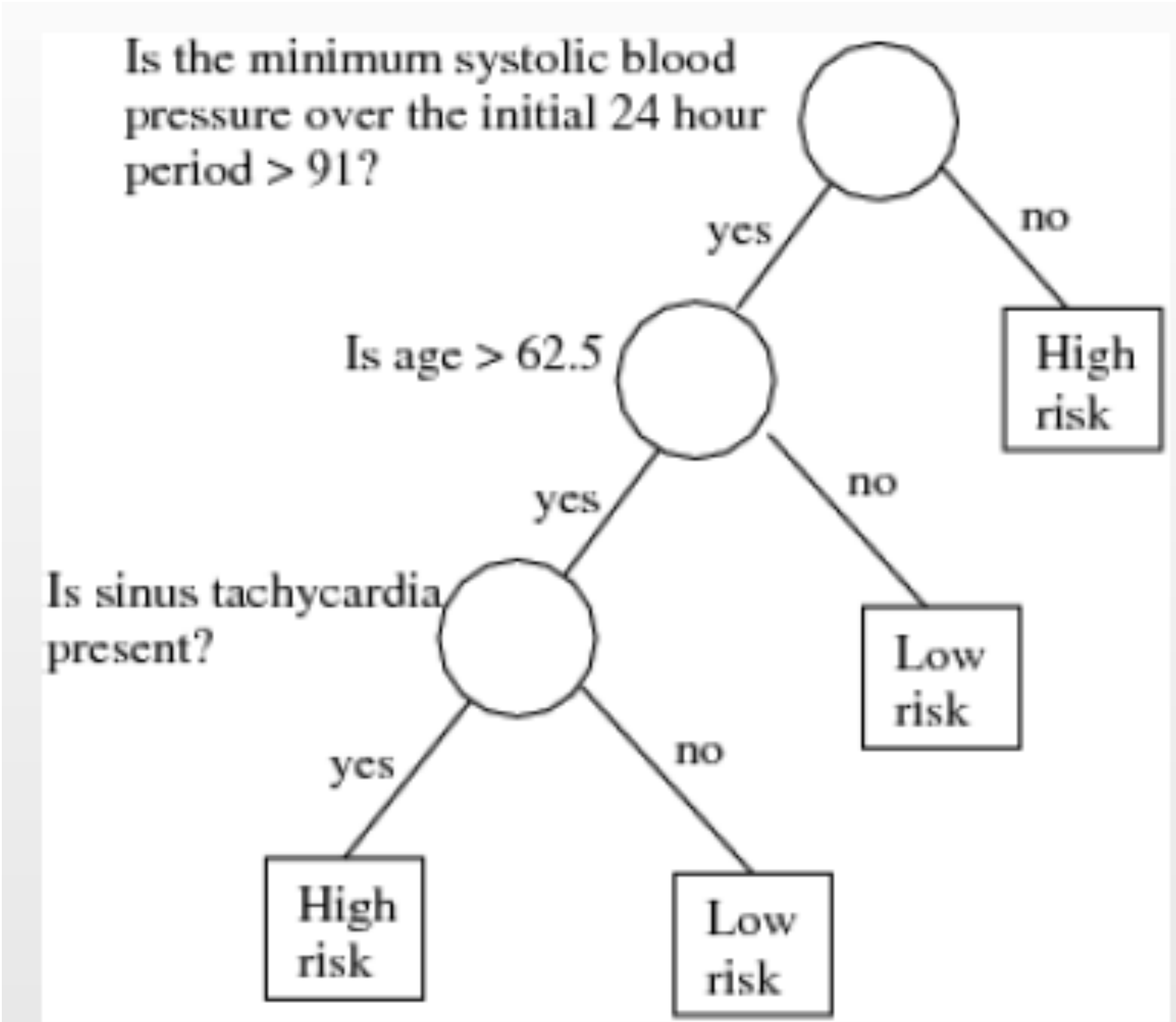
for which the target function $\mu(x) = E\left(Y \mid x\right)$ is relatively large. May also require data support of regions

$$T\left(R\right) = \int_{x \in R} g(x)\,dx$$

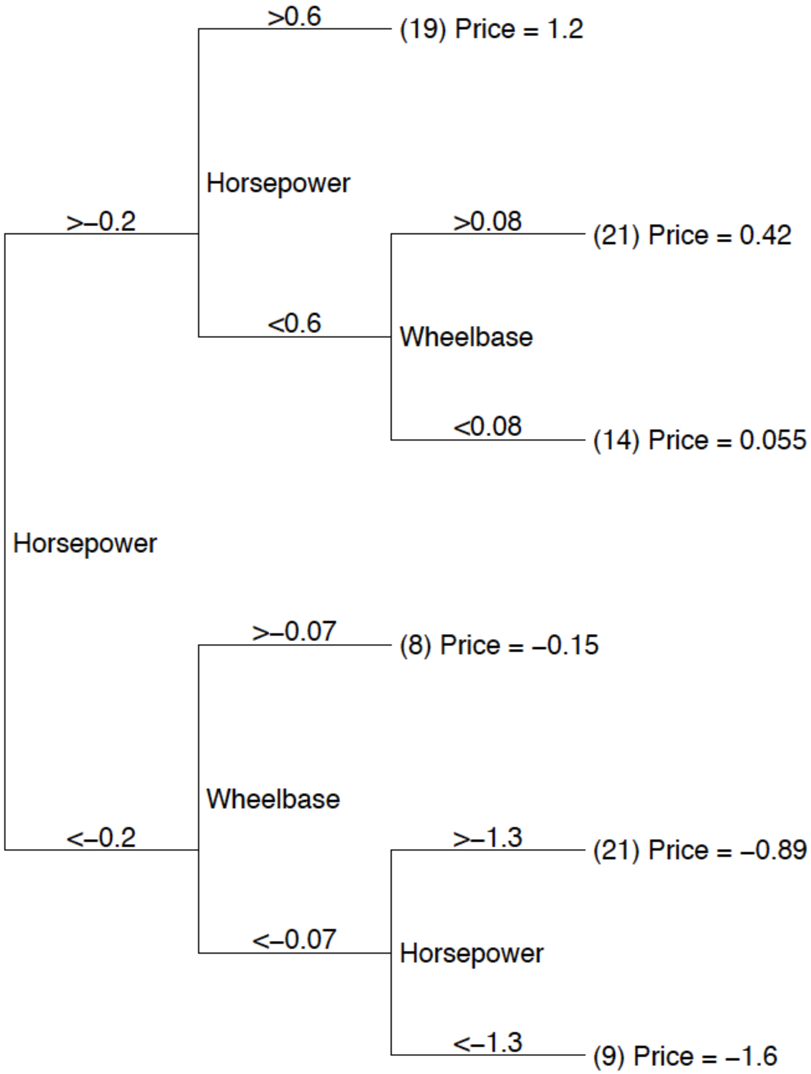
is not too small.

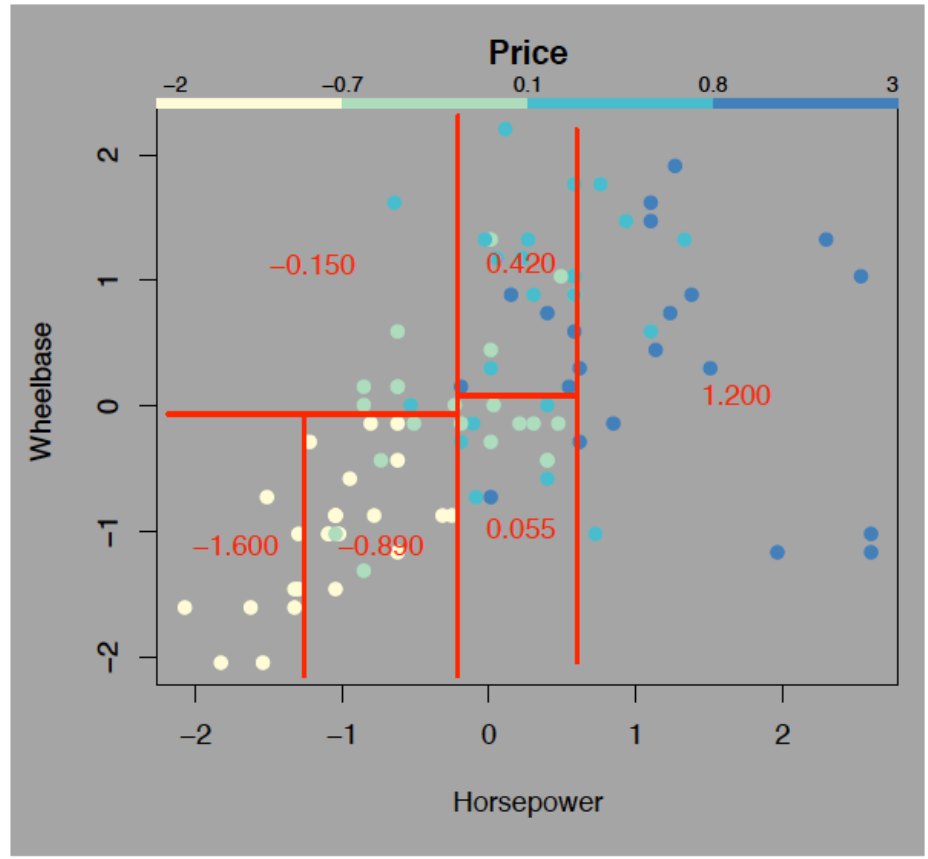
Defined by conjunctive rules, certain Methods are idea, e.g., CART applied to pooled data. Look over terminal nodes, disjoint by construction.

CART

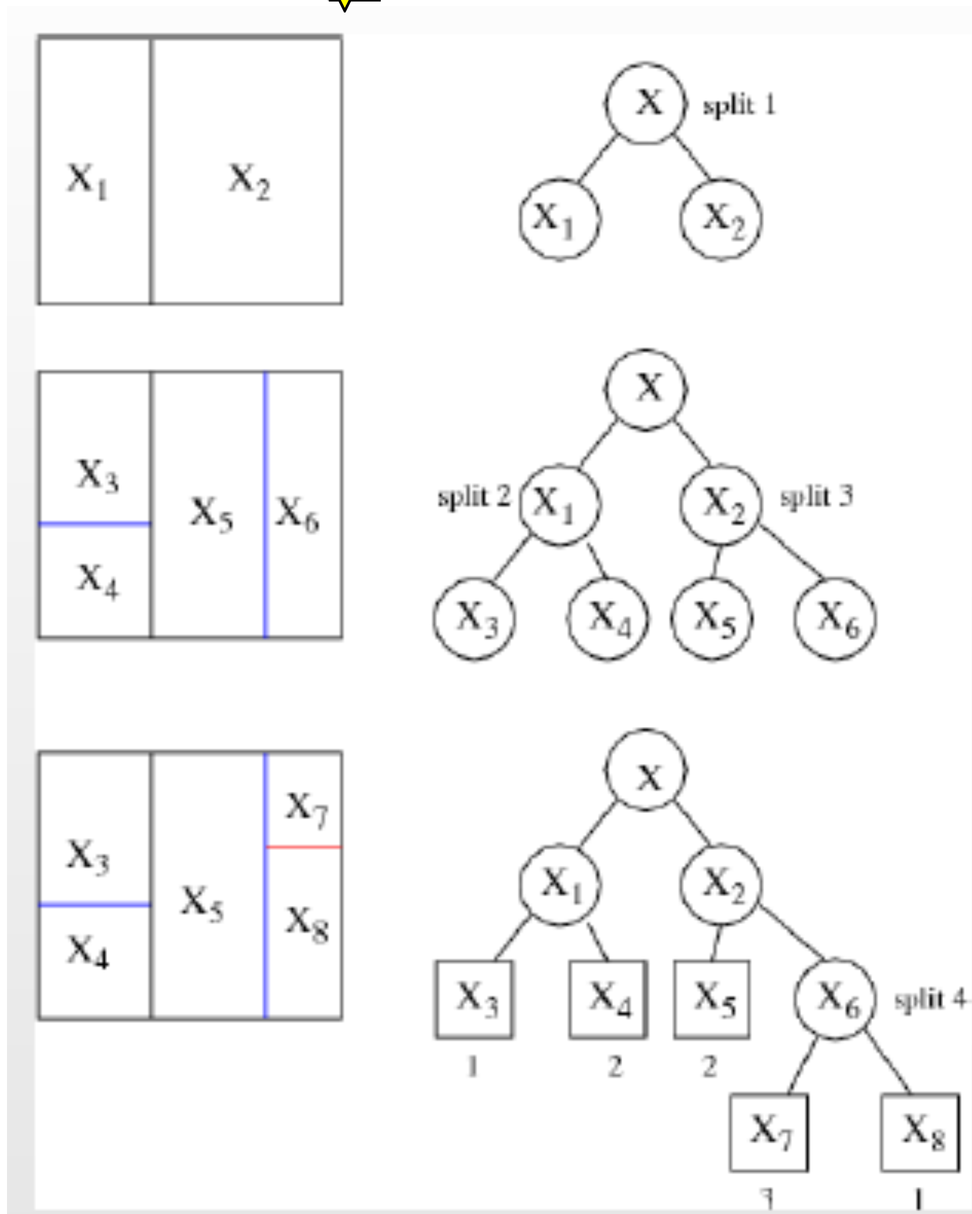

Is the minimum systolic blood pressure over the initial 24 hour period > 91?
yes
no
High risk
Is age > 62.5
yes
no
Low risk
Is sinus tachycardia present?
yes
no
High risk
Low risk

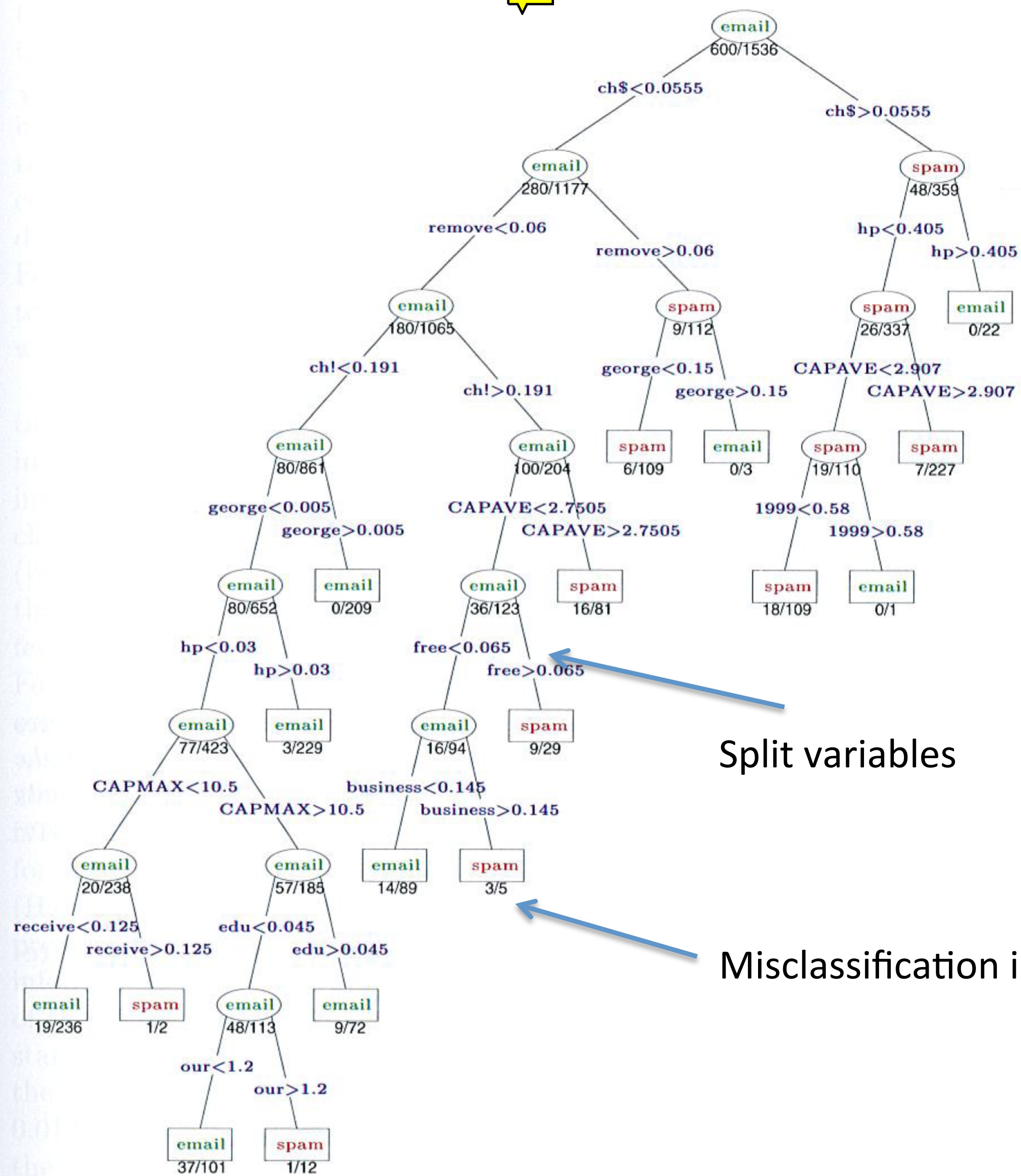
CART: A tree structure classification rule for car price





A visualization





Split variables

Misclassification in the test data

CART

Divide and Conquer.

- Partition the space into subspaces where the models and interactions are manageable.

- Decision Trees: Use recursive partitioning to subdivide the space until simple models can be fit.

Generalized Association Rules

- CART applied to pooled data will produce a decision tree that attempts to model the target over the entire space by a disjoint set of regions (terminal nodes).
- Each region is defined by a rule of the form $R = \bigcap_{j \in J} X_j \in s_j$. which are simply the terminal nodes.
- Terminal nodes with high average y values are candidates for high-support general item sets, $\bar{y}_t = ave(y_t \mid x_t \in t)$.
- The support is given by:

$$T(R) = \bar{y}_t \cdot \frac{N_t}{N + N_0}.$$

Number of pooled observations Within the region represented by the terminal node.

- General association rules can be mined out of the high-support regions, and ranked according to confidence and lift.

Generalized Association Rules

Example of an association rule CART derived:

Association rule 2: Support 25%, confidence 98.7% and lift 1.97.

$$\left[\begin{array}{rcl} \text{age} & \leq & 24 \\ \text{occupation} & \notin & \{\textit{professional, homemaker, retired}\} \\ \text{householder status} & \in & \{\textit{rent, live with family}\} \end{array}\right]$$
$$\Downarrow$$
$$\text{marital status} \in \{\textit{single, living together-not married}\}$$

Association rule 3: Support 25%, confidence 95.9% and lift 2.61.

$$\left[\begin{array}{rcl} \text{householder status} & = & \textit{own} \\ \text{type of home} & \neq & \textit{apartment} \end{array}\right]$$
$$\Downarrow$$
$$\text{marital status} = \textit{married}$$