Data Mining 2013
Subgroup Discovery

Ad Feelders

Universiteit Utrecht

October 29, 2013

Bump Hunting

Try to find regions in the input space with relatively high (or low) values for the target variable.

Example applications:

- Identifying interesting market segments
- Identifying high risk groups for specific diseases
- Credit scoring
- Spam detection
- All kinds of *selection* problems

Formal statement of the problem

$x_1, x_2, \ldots, x_p$: input variables, y:target.

Let S_j denote the set of possible values of x_j. The input space is

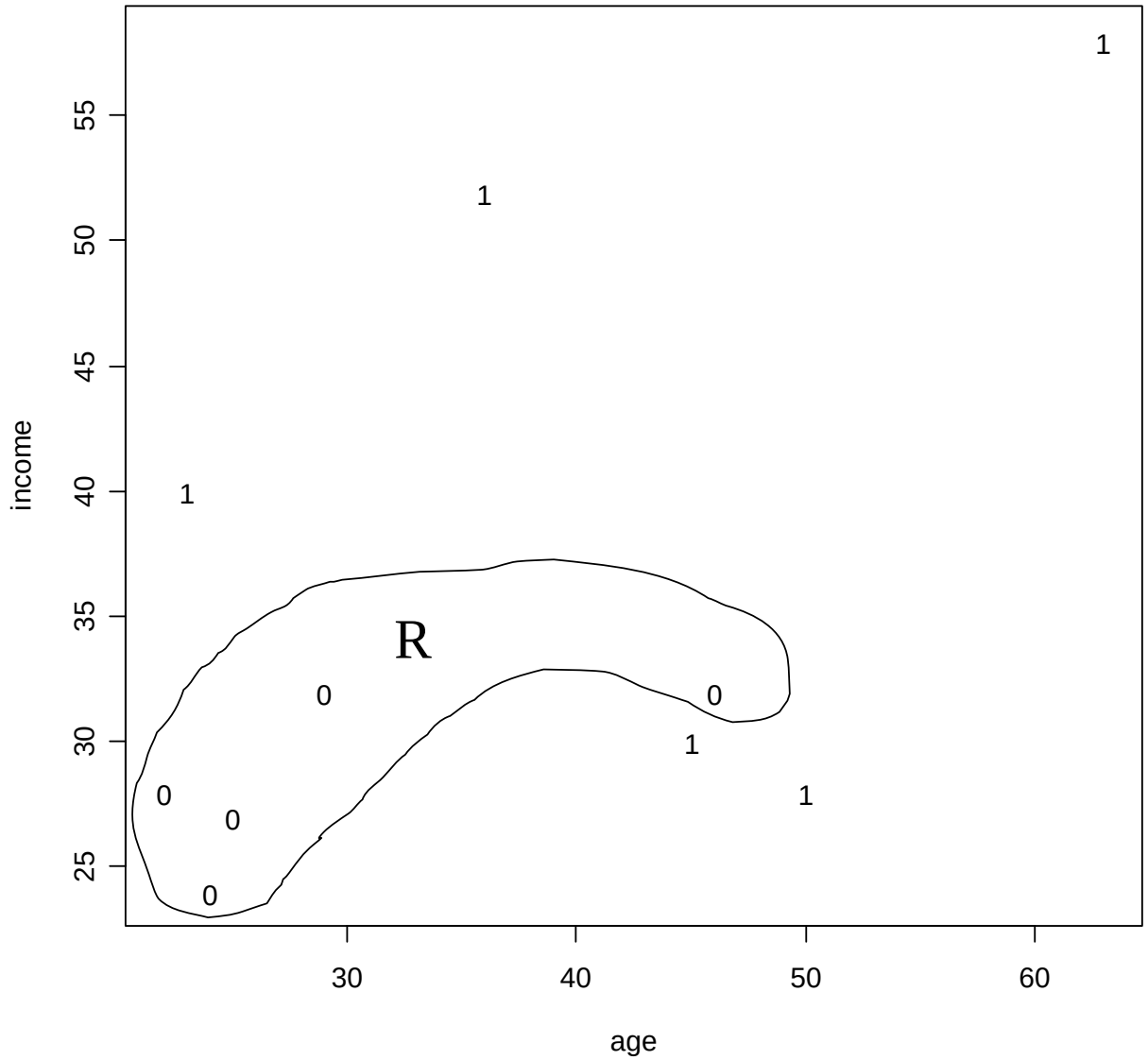
$$S = S_1 \times S_2 \times \ldots \times S_p$$

The objective is to find regions $R \subset S$ for which

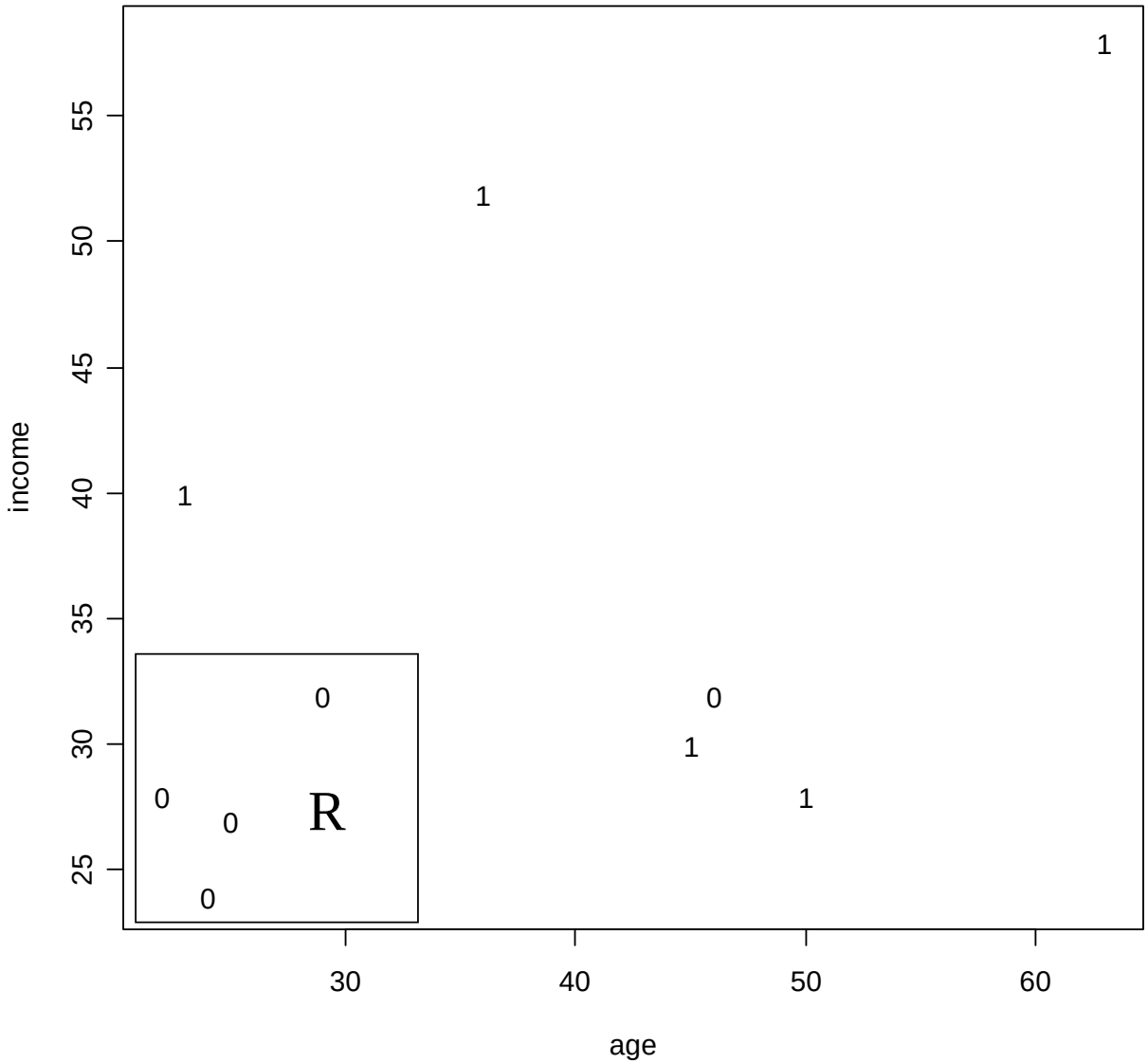
$$\bar{y}_R \gg \bar{y},$$

where $\bar{y}$ is the global mean of y and and $\bar{y}_R$ the mean of y in region R.

Regions in Input Space: Credit Data



Regions Must Have Rectangular Shape



Definition of a box

The regions we are looking for must have *rectangular* shape, hence we call them boxes.

Let $s_i \subseteq S_i$. We define a box

$$B = s_1 \times s_2 \times \ldots \times s_p$$

where $\mathbf{x} \in B \equiv \bigcap_{j=1}^{p} (x_j \in s_j)$.

When $s_i = S_i$, we leave x_i out of the box definition since it may take any value in its domain.

Example Box on Credit Data

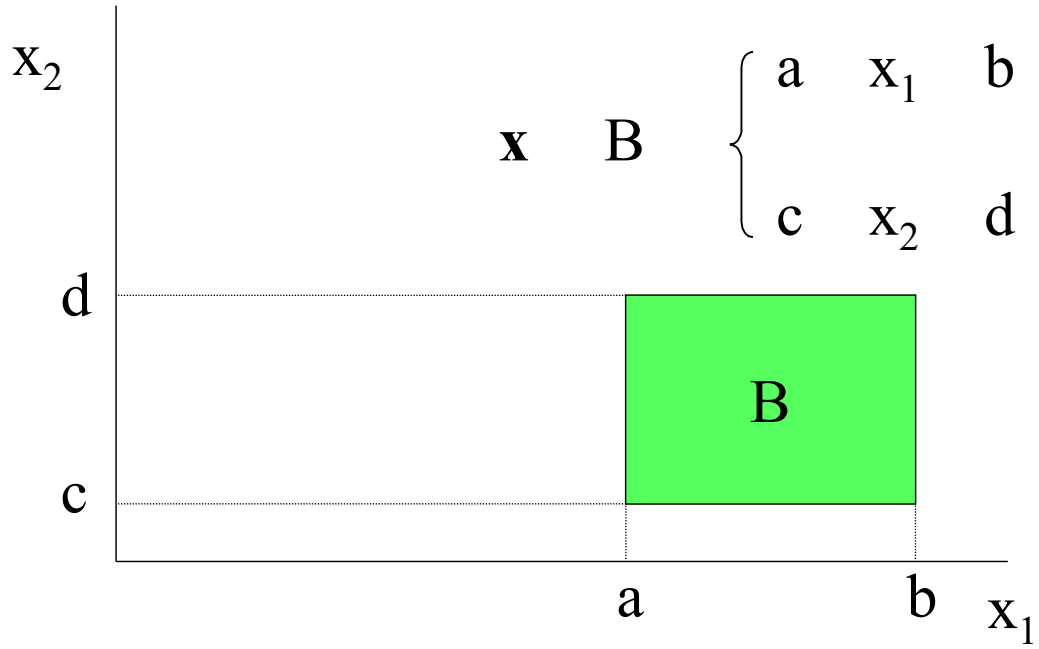
married $=$ yes $\wedge$
age $\geq 33 \wedge$
own house $\in \{$yes,no$\} \wedge$
gender $\in \{$male, female$\} \wedge$
income $\in \mathbb{R}^+$

Is simply written as:

married $=$ yes $\wedge$
age ≥ 33

Example of box on two numeric variables

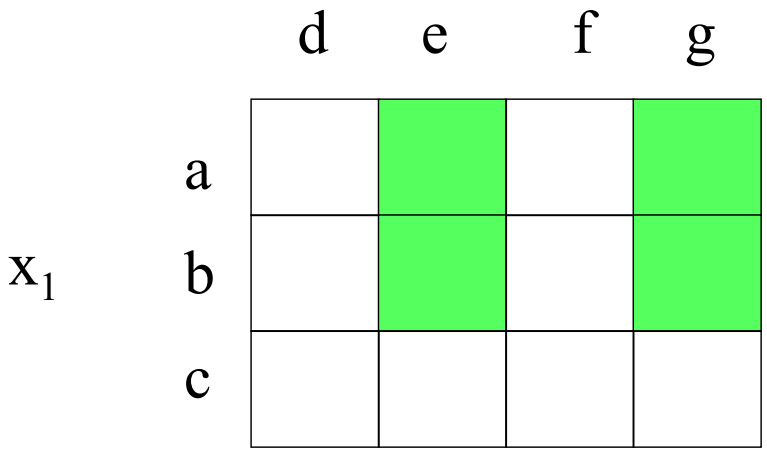
Example of a box defined on two numeric variables, where $\mathbf{x} \in B \equiv x_1 \in [a, b] \cap x_2 \in [c, d]$.



Example of box on two categorical variables

Example of a categorical box where $\mathbf{x} \in B \equiv x_1 \in \{a, b\} \cap x_2 \in \{e, g\}$.

$$\mathbf{x} \quad B \quad \begin{cases} x_1 & a,b\} \\ x_2 & \{e,g\} \end{cases}$$



Boxes may also be defined on combinations of numeric and categorical variables.

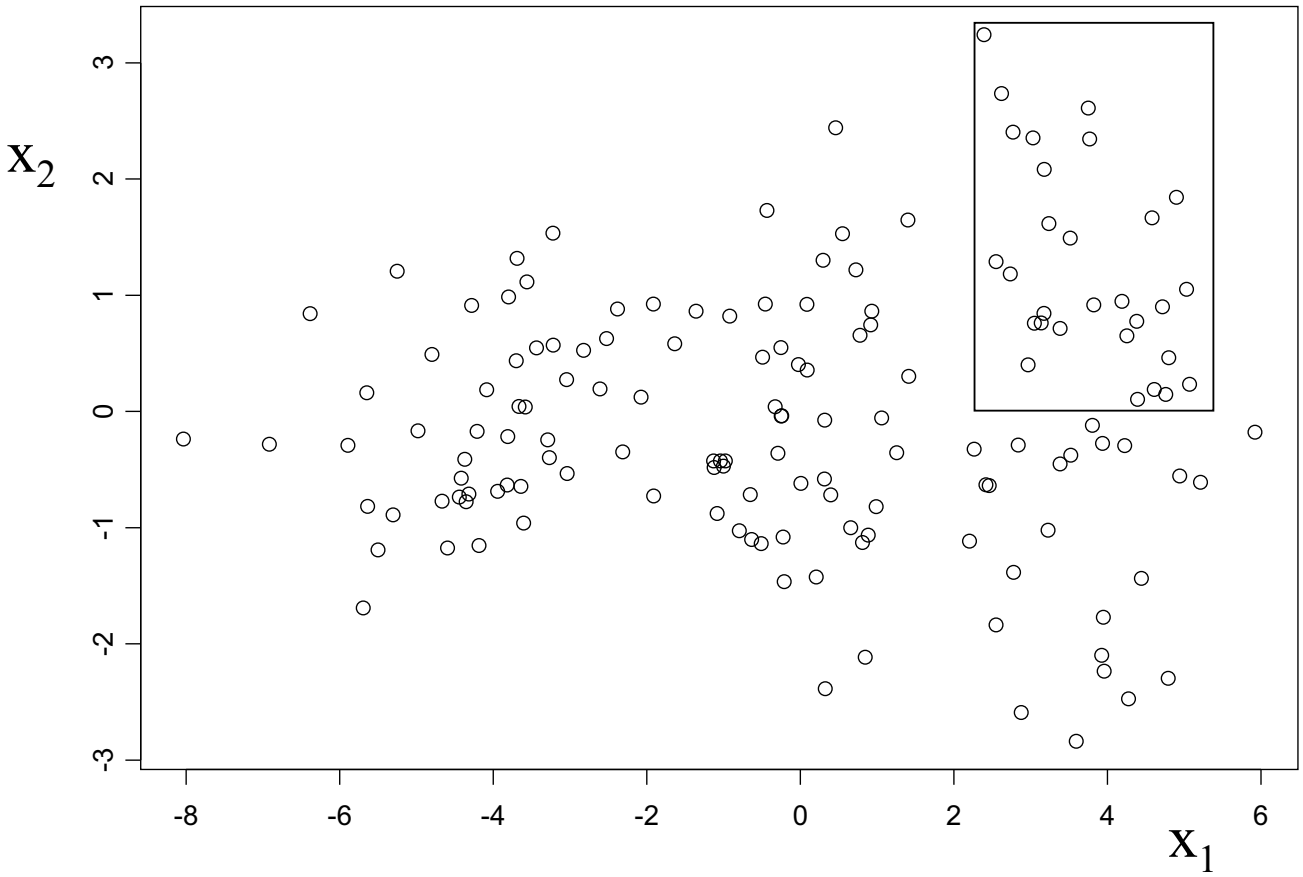
Covering strategy

The same box construction (rule induction) algorithm is applied sequentially to subsets of the data:

- The first box, B_1, is constructed on the entire data set.
- For the construction of the second box, B_2, we remove the data points that fall into B_1.
- In general, B_K is constructed on $\{y_i, \mathbf{x}_i | \mathbf{x}_i \notin \bigcup_{k=1}^{K-1} B_k\}$.

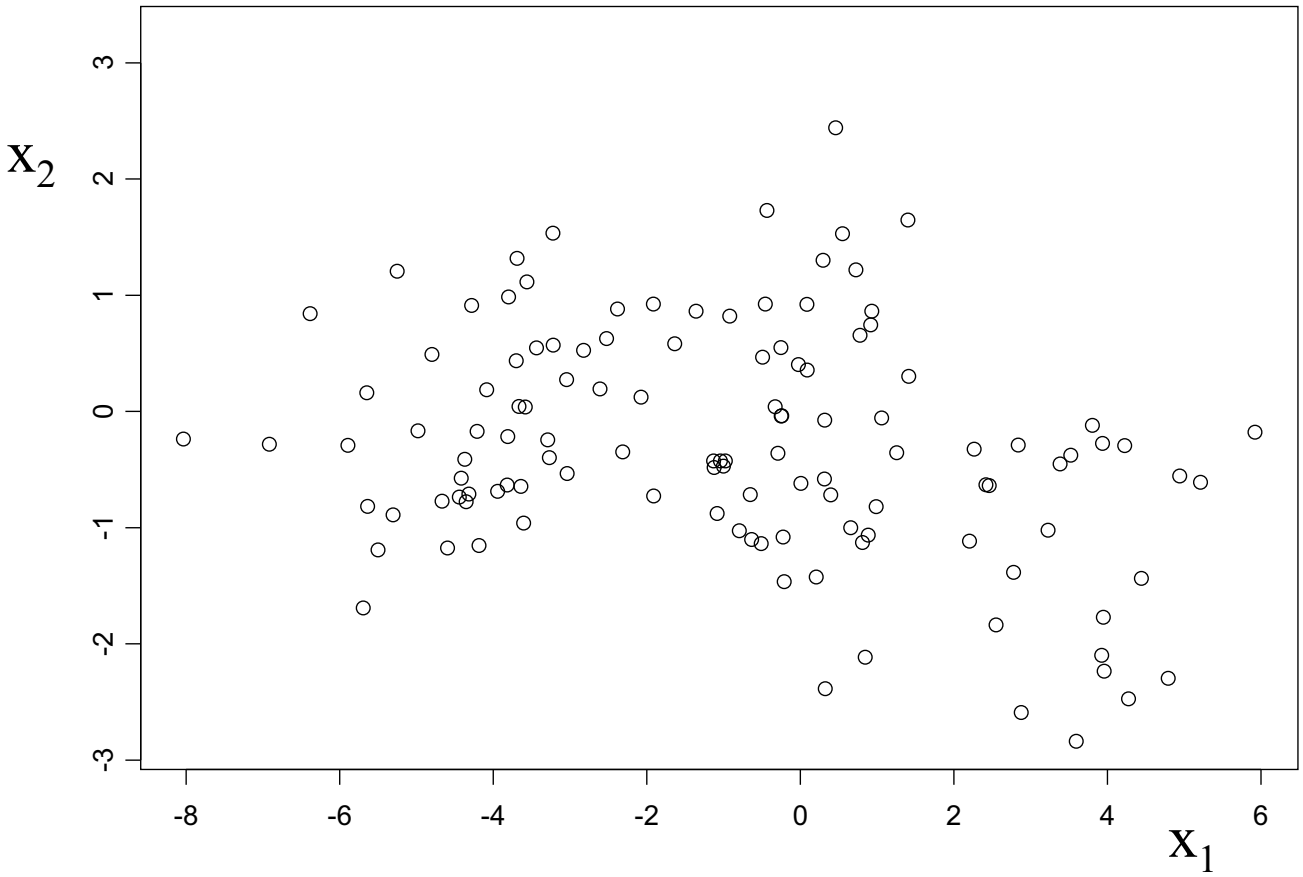
Covering(1)

The first box...



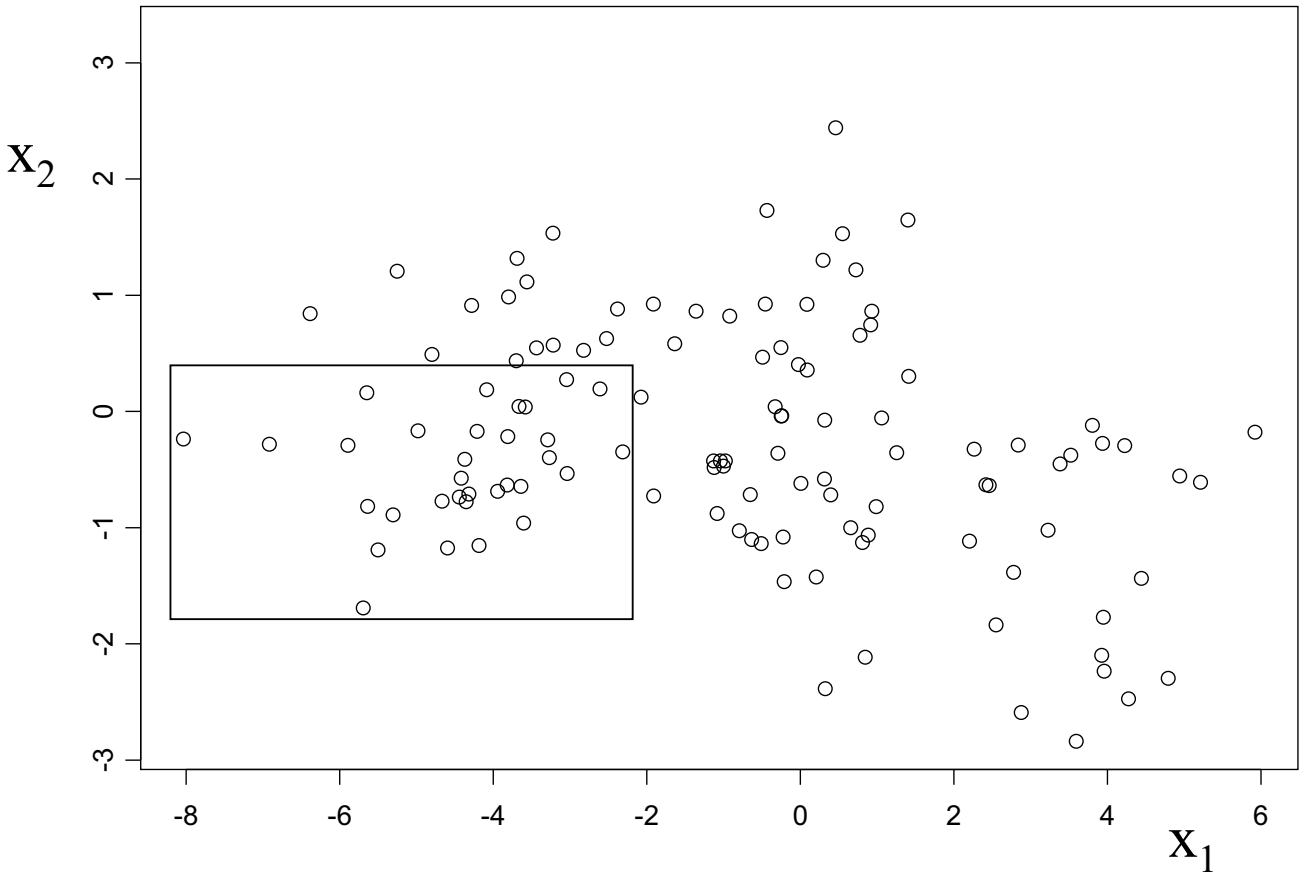
Covering(2)

Data for construction of the second box...



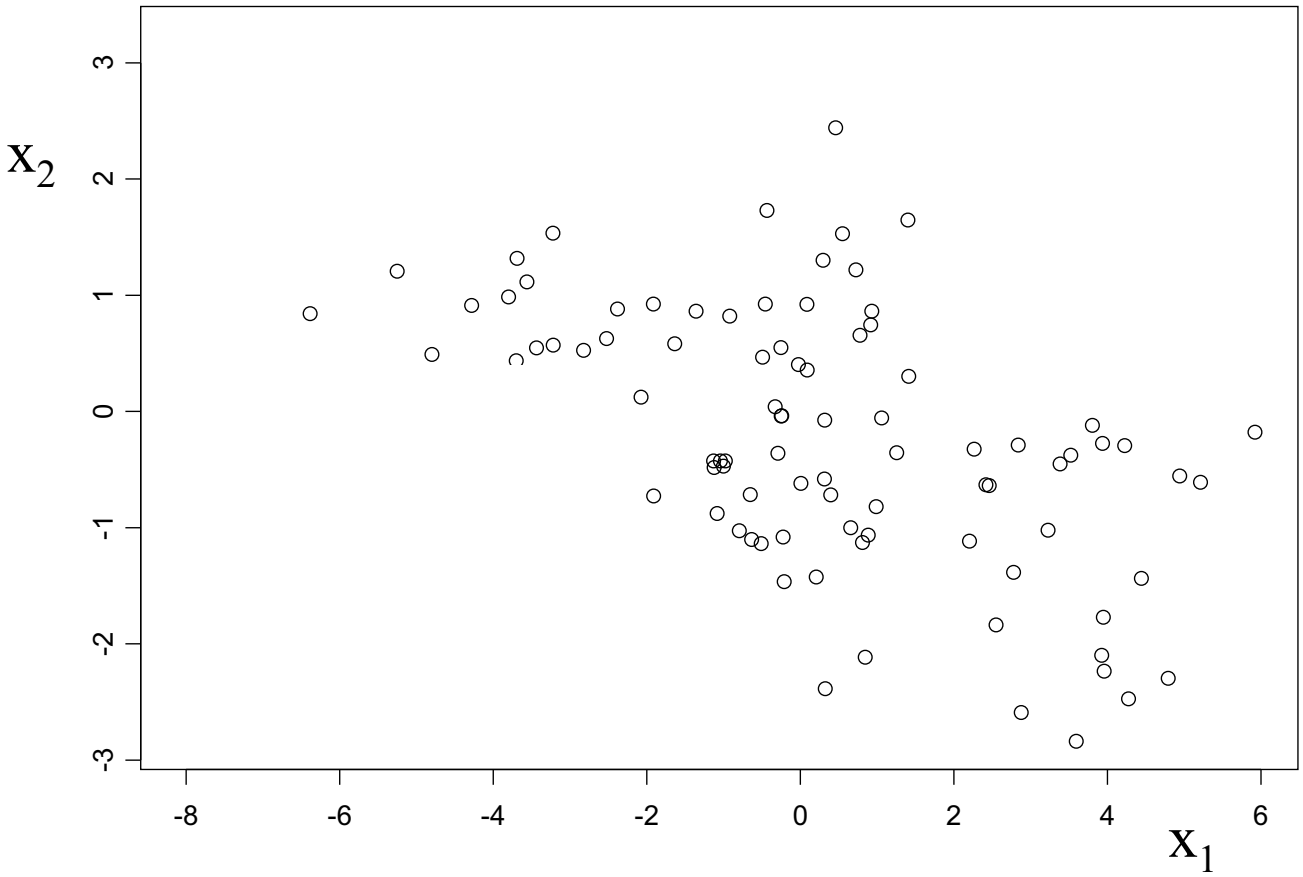
Covering(3)

The second box...



Covering(4)

Data for construction of the third box...



Covering: when do we stop?

Box construction continues until

- there is no box in the remaining data with
  - sufficient *support*, and
  - sufficiently high target mean
- the user wants to stop!

In computing the *support* of B_K we count the data points that fall into B_K (but not into any of the previous boxes) and divide by the number of observations of the entire data set.

Box construction (rule induction)

- Given the data (or a subset of the data), produce a box with target mean as large as possible
- Not feasible to consider all possible boxes
- Apply heuristic search to find a good box

Patient Rule Induction with PRIM

The box construction strategy of PRIM consists of two phases:

1. Patient successive top-down refinement, followed by
2. bottom-up recursive expansion.

Top-down peeling

- Begin with a box B that covers all the (remaining) data
- At each step a small subbox b within the current box B is removed
- Remove subbox $b*$ that yields the largest target mean within $B - b*$

Candidates for removal: numeric variables

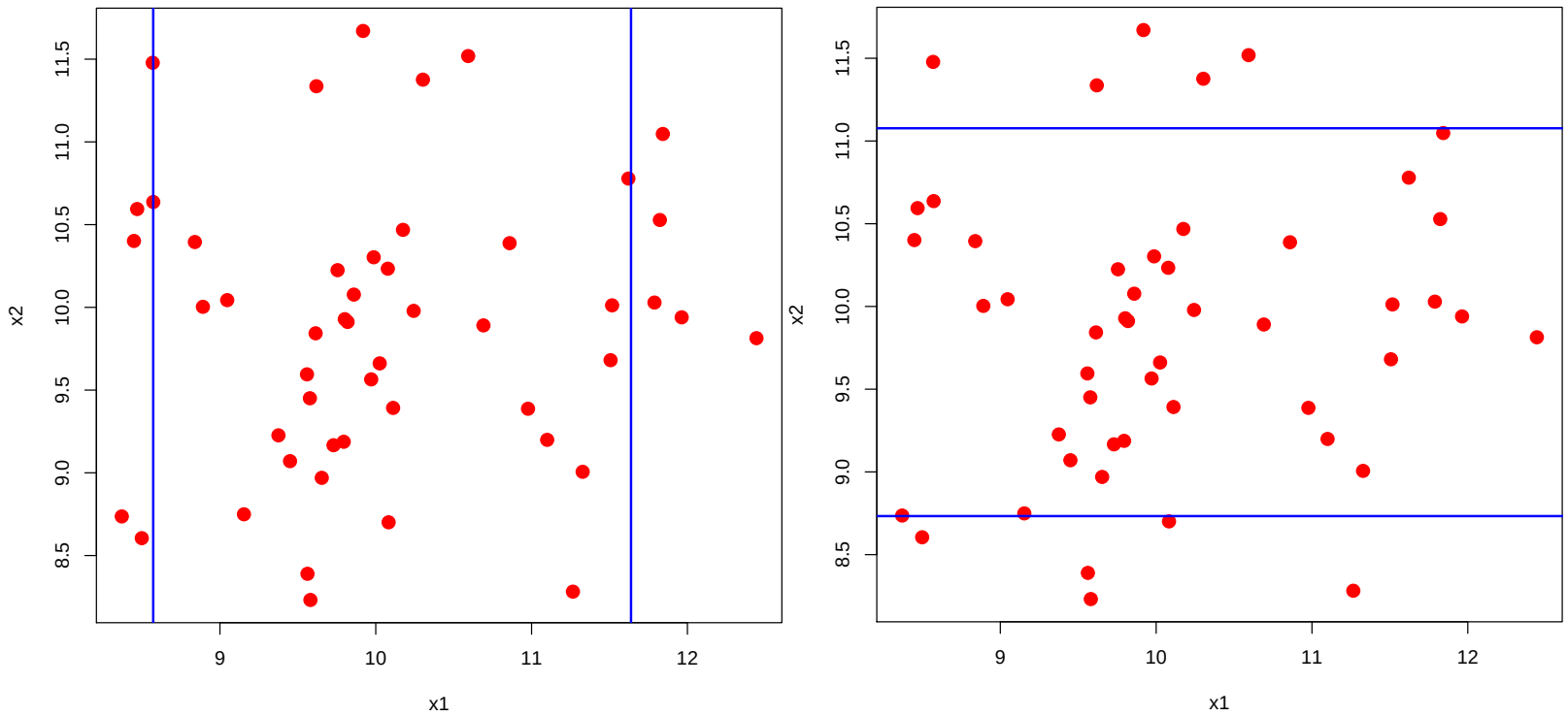
Candidate subboxes for numerical variable x_j:

$$b_{j-} = \{\mathbf{x} | x_j < x_{j(\alpha)}\}$$
$$b_{j+} = \{\mathbf{x} | x_j > x_{j(1-\alpha)}\}$$

with $x_{j(\alpha)}$ the α-quantile of x_j in the current box, i.e. $p(x_j < x_{j(\alpha)}) = \alpha$.

Typically $\alpha \leq 0.1$, so in each step only a small part of the data points is peeled off (hence the term *patient* rule induction).

Candidates for removal: numeric variables



The current box contains 50 data points.

With $\alpha = 0.1$, we can peel off 5 data points at a time.

For each variable, we can peel off the 5 lowest or 5 highest values.

Candidates for removal: categorical variables

Candidate subboxes for categorical variable x_j:

$$b_{jm} = \{\mathbf{x} | x_j = s_{jm}\}, \quad s_{jm} \in S_j$$

For each value in the domain of x_j we can peel off the subbox with that value.

- Harder to control the amount of data that is removed.
- We will ignore alpha for categorical variables.

Top-down peeling: pseudocode

Top-down peeling

Repeat

$C(b) \leftarrow$ set of candidates for removal

$b* \leftarrow \arg\max_{b \in C(b)} \bar{y}_{B-b}$

$B \leftarrow B - b*$

$\beta_B \leftarrow$ support of B

Until $\beta_B \leq \beta_0$

Return B

Top-down peeling: example

Record	age	married?	own house	income	gender	y
1	22	no	no	28,000	male	0
2	46	no	yes	32,000	female	0
3	24	yes	yes	24,000	male	0
4	25	no	no	27,000	male	0
5	29	yes	yes	32,000	female	0
6	45	yes	yes	30,000	female	1
7	63	yes	yes	58,000	male	1
8	36	yes	no	52,000	male	1
9	23	no	yes	40,000	female	1
10	50	yes	yes	28,000	female	1

Peeling Options

Take $\alpha = \frac{1}{3}$ and $\beta_0 = 0.4$.

Possible peelings on age:

1. b_- : age < 25. $B - b_-$: age ≥ 25. $\bar{y} = \frac{4}{7}$.
2. b_+ : age > 45. $B - b_+$: age ≤ 45. $\bar{y} = \frac{3}{7}$.

	age	y
1	22	0
2	23	1
3	24	0
4	25	0
5	29	0
6	36	1
7	45	1
8	46	0
9	50	1
10	63	1

Peeling Options

Take $\alpha = \frac{1}{3}$ and $\beta_0 = 0.4$.
Possible peelings on income:

1. b_- : income < 28. $B - b_-$: income ≥ 28. $\bar{y} = \frac{5}{8}$.
2. b_+ : income > 32. $B - b_+$: income ≤ 32. $\bar{y} = \frac{2}{7}$.

	income	y
1	24	0
2	27	0
3,4	28	0,1
5	30	1
6,7	32	0,0
8	40	1
9	52	1
10	58	1

Peeling Options

The best peeling action is to peel off *married=no*, which leaves us with the married people and $\bar{y} = \frac{4}{6}$.

Then the best peeling action is on age:

	age	y
1	24	0
2	29	0
3	36	1
4	45	1
5	50	1
6	63	1

So we get the rule: married $=$ yes $\wedge$ age ≥ 36 with $\bar{y} = 1$ and $\beta = 0.4$.

Bottom-up pasting

- In top-down peeling we only look one step ahead: we choose the peel that leads to the best subbox.
- This means that box boundaries are determined without knowledge of later peels.
- Therefore the final box can sometimes be improved by readjusting its boundaries.
- In bottom-up pasting we enlarge the final box until the next paste will cause $\bar{y}$ in the box to decrease.

Preventing Overfitting

- Which box from the sequence to select?
- Beware of overfitting!
- Partition the available data in a training set and a test set.
- Select the box with the highest target mean on the test data.

Example: British Family Expenditure Survey

Unit of analysis: Household.
Data on the budget share spent on

- food,
- clothing,
- alcohol,
- and so on,

as well as data on total household expenditure (totexp), total net household income (inc), age of household head (age), and number of children in the household (nk).

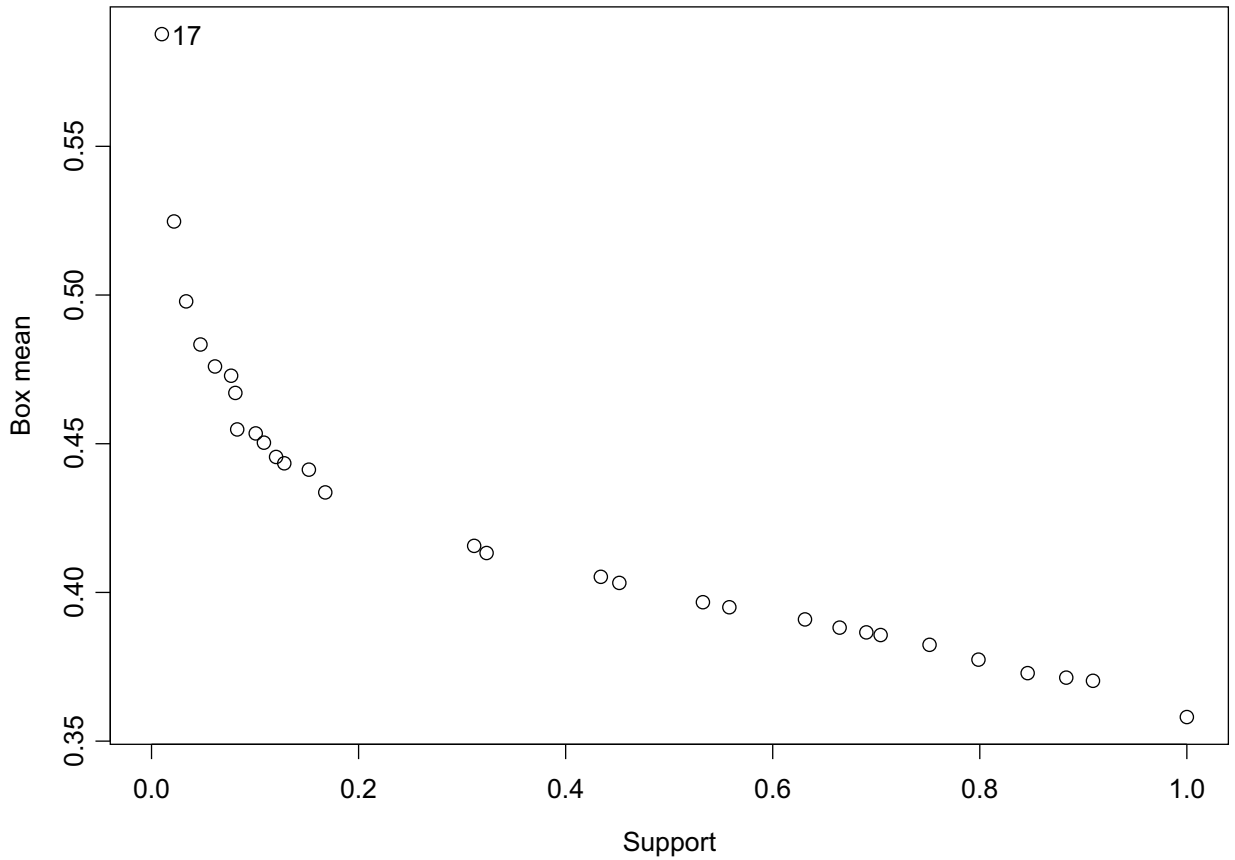
Example: British Family Expenditure Survey

With bump hunting we may for example look for profiles of households that spend a relatively large share of their budget on food.

On average the households in the sample spend about 36% of their budget on food.

Which groups spend (a much) bigger part of their budget on food?

Peeling Trajectory: box mean on test sample



Rule 17: if totexp $<$ 45 and age $>$ 33 and inc $<$ 135, then wfood $=$ 58% (against 36% for average household).

Post-Processing: Removing redundant variables

We can simplify a box by removing variables from its definition. Example:

if totexp $<$ 65 and age $>$ 36 and 105 $<$ inc $<$ 135, then wfood $=$ 48% (support $\approx$ 0.7%)

	Remove variable (+below)	
Definition	Mean	Support
totexp $<$ 65.00	0.3558	0.9901
age $>$ 36.50	0.4240	0.1578
105.0 $<$ inc $<$ 135.0	0.4377	0.0375

Subboxes in Data Surveyor

Like in PRIM each eligible sub-box is defined on a single variable, but in a more greedy manner:

1. Numeric variable x_j:

$$B'_{jcd} = \{\mathbf{x} \in B | x_j \in [c, d]\}, \quad c, d \in S_j \text{ and } c < d.$$

2. Categorical variable x_j:

$$B'_{jm} = \{\mathbf{x} \in B | x_j = s_{jm}\}, \quad s_{jm} \in S_j.$$

Data Surveyor: beam search

Beam Search

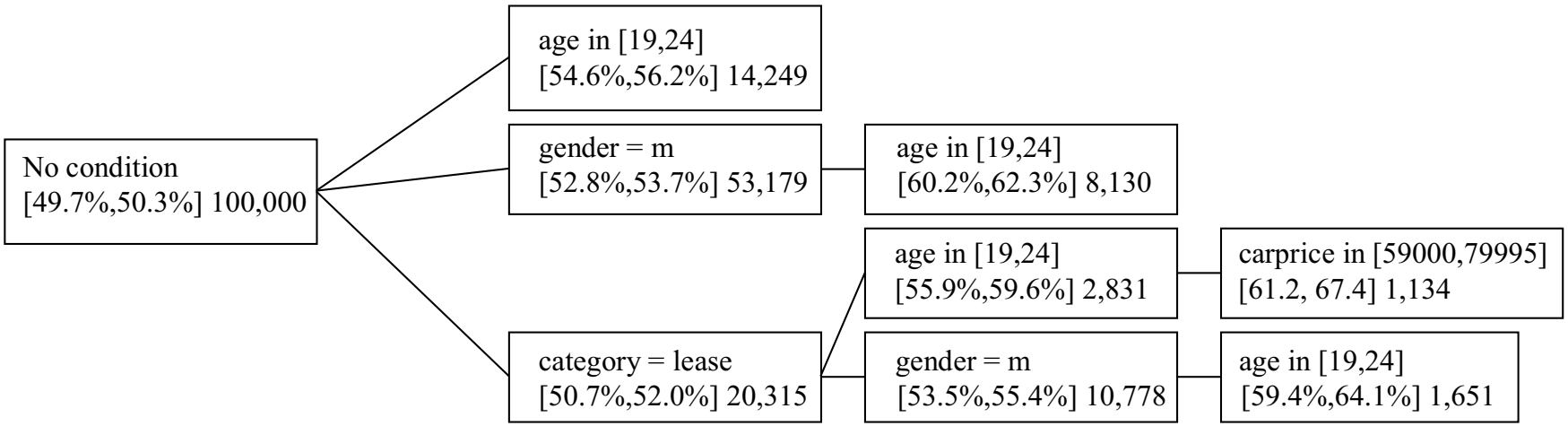
Beamset $\leftarrow$ {initial box}
Repeat
  all-subboxes $\leftarrow \varnothing$
  For each box B_i in Beamset do
    $C(B_i) \leftarrow$ set of candidate subboxes of B_i
    all-subboxes $\leftarrow$ all-subboxes $\cup$ $C(B_i)$
  Beamset $\leftarrow$ best w subboxes from all-subboxes
Until no improvement possible or depth $= d$
Return Beamset

Data Surveyor

The w subboxes are chosen as follows:

1. Mean value of the target in the subbox should be *significantly* higher.
2. It is required that the subbox has support of at least β_0.
3. From those, the w boxes with the largest target mean are chosen.

Example of beam search



Beam width $= 3$, depth $= 3$ and minimum support $= 1\%$

Example Application

Applying PRIM and logistic regression for selecting high-risk subgroups in very elderly ICU patients, B. Nannings, A. Abu-Hanna, E. de Jonge, International Journal of Medical Informatics, 2008.

Motivation: Elderly patients have in general a worse prognosis than younger patients, but it is unknown which patients are at very high risk.

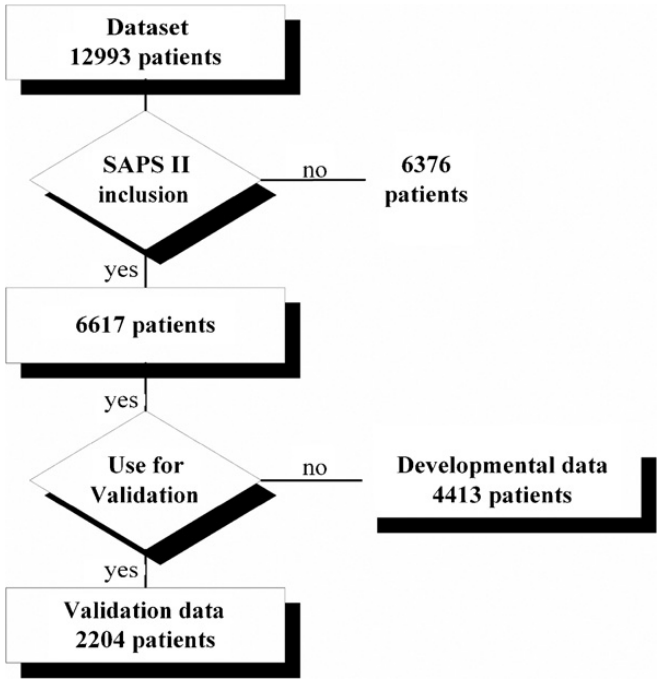
Why interested in high-risk subgroups?

- reveal determinants that provide insight into the patient sub-populations
- some determinants may be risk factors that can be acted upon
- enrollment of high risk patients into studies on therapeutic interventions
- support decisions on (withholding) treatment

Example Application

Data: all 12,993 admissions of patients 80 years and older between January 1995 and October 2005 originating from all 33 adult ICUs participating in the National Intensive Care Evaluation (NICE).

SAPS II Exclusion: no readmissions, no cardio-surgical patients, no patients with burns.



PRIM analysis performed

- The mortality in the complete data set is 34.5%.
- Searched for largest subgroups with mortality at least 85% on development set and held-out set.
- Each subgroup should include at least 3% of data in the development set.
- Run PRIM with different parameter settings (different values of α) to obtain different subgroups.
- Stop when union of subgroups covers 10% of the development set.

Best subgroup found with PRIM

The best subgroup found was:

- 24 h urine production $< 0.83\ \ell$,
- mechanical ventilation at 24 h after admission,
- lowest systolic blood pressure during the first 24 h < 75 mmHg,
- lowest pH during the first 24 h < 7.3, and
- medical or unscheduled surgical reason for admission.

The mortality of this group was 94.8% on the development set and 91.8% on the validation set.

The support of the subgroup on the validation set was 2.8%.

Exceptional Model Mining (EMM)

Extension of subgroup discovery, where we fit a (simple) model to the subgroup and its complement.

A subgroup is interesting if the model fitted to it is substantially different from the global model.

How to quantify the difference between models?

EMM: partitioning of attributes

In EMM we distinguish between

- Attributes Z used for *defining* subgroups.
- Attributes X (and possibly Y) for modeling.

These are not allowed to overlap.

Example: Simple Linear Regression

Given observations x_i, y_i we choose a and b such that the sum of squared errors

$$\sum_{i=1}^{n}(y_i - (a + bx_i))^2$$

is minimized. The fitted regression line is

$$\hat{y} = a + bx$$

The coefficient b (the slope of the line) indicates by how much y is expected to change in case x increases with one unit.

Predicting House Prices

We expect that the house price increases with lot size, so in the model

$$\text{price} = a + b \times \text{lot size}$$

we expect b to be positive. In fact, on *Windsor Housing Data* we get

$$\text{price} = 34,140 + 6.6 \times \text{lot size}$$

Hence we expect price to increase with 6.6 Canadian Dollars when lot size increases with one square foot.

Predicting House Prices

Now consider the subgroup

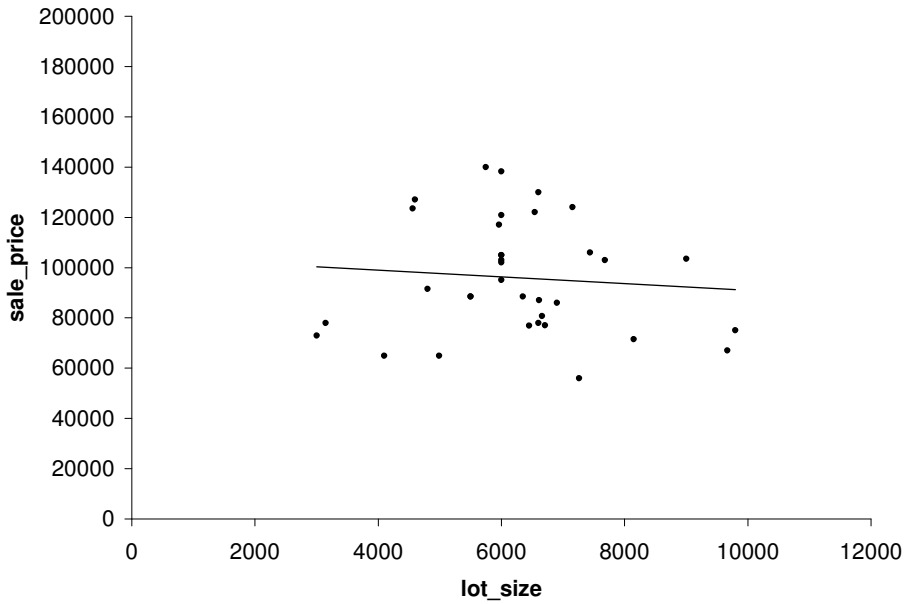
$$drive = 1 \wedge rec.room = 1 \wedge nbath > 1$$

(houses with a driveway and a recreational room and at least two bathrooms). Fitting a regression model to this subgroup gives

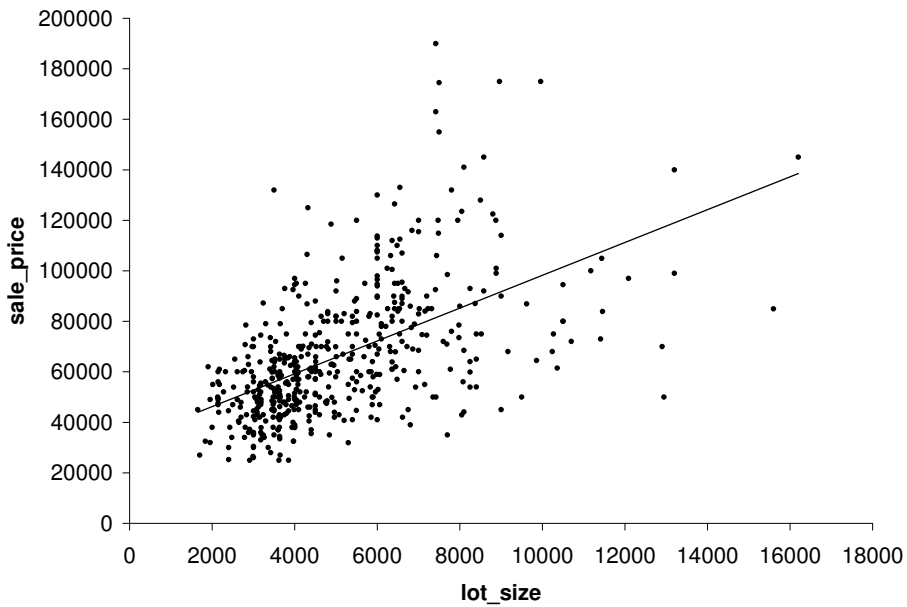
$$\text{price} = 104,350 - 1.34 \times \text{lot size}$$

Price decreases with lot size?

Predicting House Prices



Subgroup



Complement

Measuring the quality of a subgroup

Define a binary variable s, where

$$s_i = \begin{cases} 1 & \text{if row } i \text{ is in the subgroup} \\ 0 & \text{otherwise} \end{cases}$$

Fit the model

$$\text{price}_i = a + b \times \text{lot size}_i + c \times s_i + d \times (s_i \times \text{lot size}_i)$$

If coefficient d is significantly different from zero, then the slope in the subgroup is significantly different from the slope in its complement, since

$$\text{price}_i = \begin{cases} a + b \times \text{lot size}_i & \text{if } s_i = 0 \\ (a + c) + (b + d) \times \text{lot size}_i & \text{if } s_i = 1 \end{cases}$$

Quality Of Subgroup

```
>hprice.lm <- lm(sale.price ~ s+lot.size+s:lot.size,
                  data=cbind(hprice.dat,s=s))
>summary(hprice.lm)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 33230.3046  2434.0494  13.652  < 2e-16 ***
s           71119.3888 16062.7762   4.428 1.15e-05 ***
lot.size        6.5015     0.4407  14.754  < 2e-16 ***
s:lot.size     -7.8414     2.5021  -3.134  0.00182 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Quality of subgroup is 1 − pvalue for coefficient of `s:lot.size`.

Example: Giffen Behavior

- The demand for a product will usually decrease as its price increases.
- According to economic textbooks, there are circumstances however, for which we should expect to see an upward sloping demand curve.
- The common example is that of poor families that spend most of their income on a relatively inexpensive staple food (e.g. rice or wheat) and a small part on a more expensive type of food (e.g. meat).
- If the price of the staple food rises, people can no longer afford to supplement their diet with the more expensive food, and must consume more of the staple food.

The dataset we analyze was collected in different counties in the Chinese province Hunan, where rice is the staple food. The price changes were brought about by giving vouchers to randomly selected households to subsidize their purchase of rice.

Example: Giffen Behavior

The global model estimated is:

$$\%\Delta \text{staple}_{i,t} = \alpha + \beta \%\Delta p_{i,t} + \sum \gamma \%\Delta Z_{i,t}$$

where

- $\%\Delta \text{staple}_{i,t}$ denotes the percent change in household i's consumption of rice,
- $\%\Delta p_{i,t}$ is the percent change in the price of rice due to the subsidy (negative for $t = 2$ and positive for $t = 3$), and
- $\%\Delta Z_{i,t}$ is a vector of percent changes in other control variables including income and household size.

For each household, two changes are observed: the change between periods 2 and 1 ($t = 2$), capturing the effect of giving the subsidy; and the change between periods 3 and 2 ($t = 3$) capturing the effect of removing the subsidy.

Example: Giffen Behavior

- The coefficient of primary interest is β. If $\beta > 0$ we observe Giffen behavior.
- The other variables are included in the model to control for other possible influences on demand, so that the effect of price can be reliably estimated.
- For the extremely poor, one might not observe Giffen behavior, because they consumed rice almost exclusively anyway, and therefore have no other possibility than buying less of it in case of a price increase.
- The Initial Staple Calorie Share (ISCS) was also measured, and the hypothesis is that families with a high value for ISCS do not display Giffen behavior.
- At depth 1, the best subgroup we found was ISCS ≥ 0.87 with $\hat{\beta} = -0.96$ (against $\hat{\beta} = 0.22$ for the complete data base). The size of this subgroup is $n = 106$.
- This confirms the conclusion that Giffen behavior does not occur for families that almost exclusively consume rice anyway.
- This conclusion can also be reached by defining subgroups on *income per capita* rather than ISCS. For example, the 4th ranked subgroup we found was

  Income per Capita $\leq$ *64.67, with a slope of* -0.41,

  and the 6th ranked subgroup was

  Income per Capita $\geq$ *803.75, with a slope of* 0.79.

Conclusion

- Covering strategy leads to an *ordered* list of rules.
- May be used to construct a classifier (decision list).
- May be used to find groups with high target mean.
- EMM finds subgroups where *models* deviate, rather than a target variable.