

# The bump hunting algorithm

Jerome Friedman and Nicholas Fisher (1998)  
<http://www-stat.stanford.edu/~jhf/ftp/prim.pdf>

# Function Optimization

Often, function approximation is applied in situations for which the actual goal is in some property of the target function.

Let  $S$  be the set of all possible values for the input variable  $x$ .

$$\{x_j \in S_j\}_{j=1}^n$$

The goal is to find a subregion  $R$  of the input domain  $S$  for which

$$\overline{f}_R = \text{avg}_{x \in R} f(x) = \frac{\int_{x \in R} f(x) p(x) dx}{\int_{x \in R} p(x) dx} \gg \overline{f}$$

where  $\overline{f}$  is the average over the entire input space.

# Support

An important property the subregion  $R$  is it's support and straightforward estimates of these quantities will be used:

$$\hat{\beta}_R = \frac{1}{N} \sum_{x_i \in R} 1_{x_i \in R}, \quad \bar{y}_R = \frac{1}{N \hat{\beta}_R} \sum_{x_i \in R} y_i$$

where  $N$  is the size of the dataset.

# PRIM : Patient Rule Induction Method

- PRIM is a Bump Hunting Algorithm that partitions the input variable space into box shaped regions.
- PRIM searches for sub-regions where the target variable has the maxima by bump-hunting
- Two major steps
  - Top-down peeling off
  - Bottom-up pasting

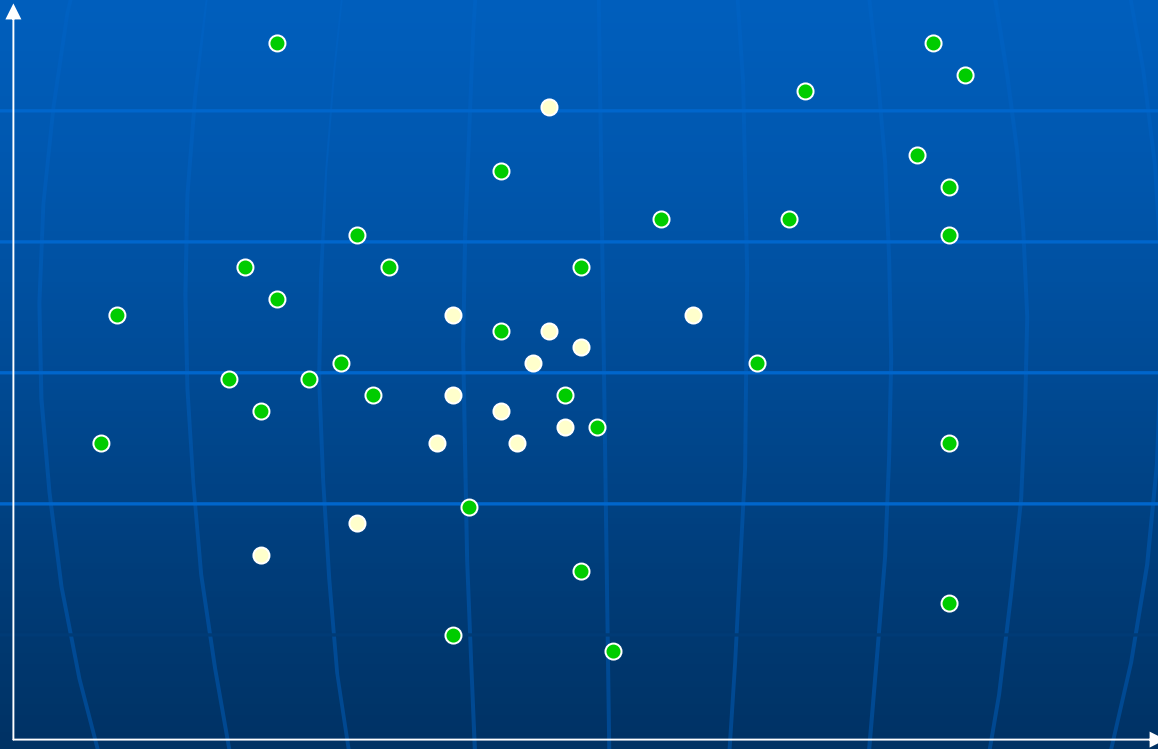
PRIM generates a collection of rules (final boxes) sequentially.

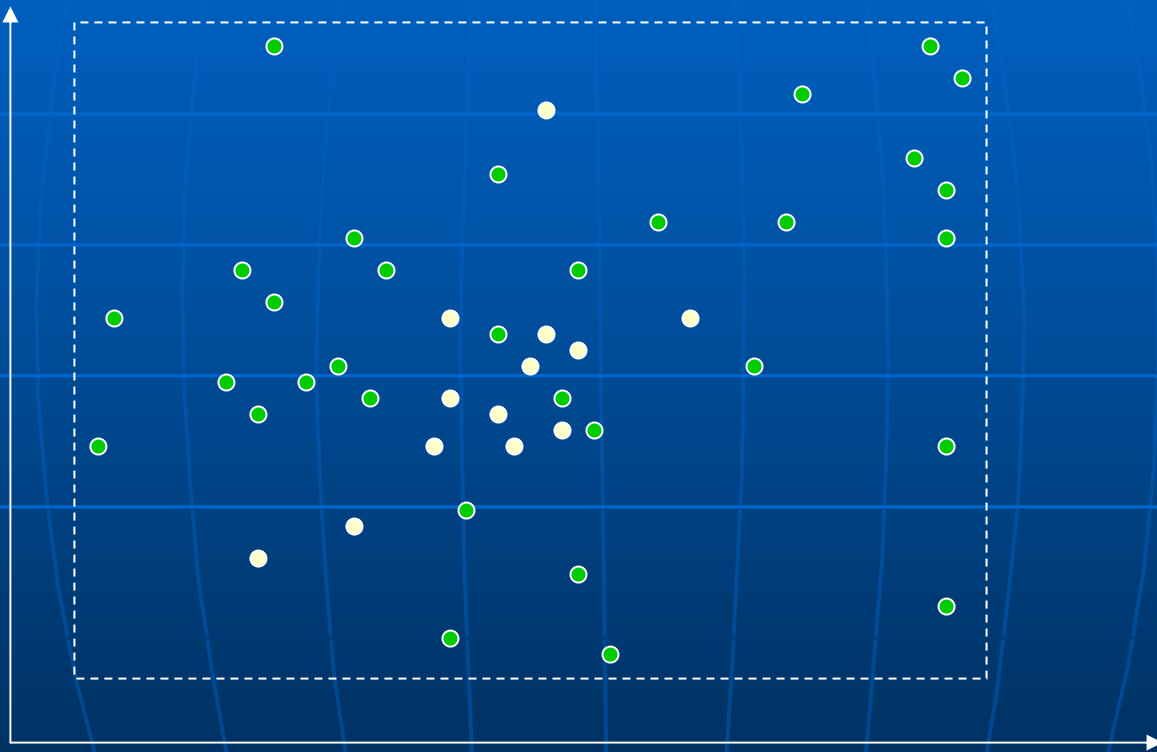
# PRIM Bump Hunting: algorithm

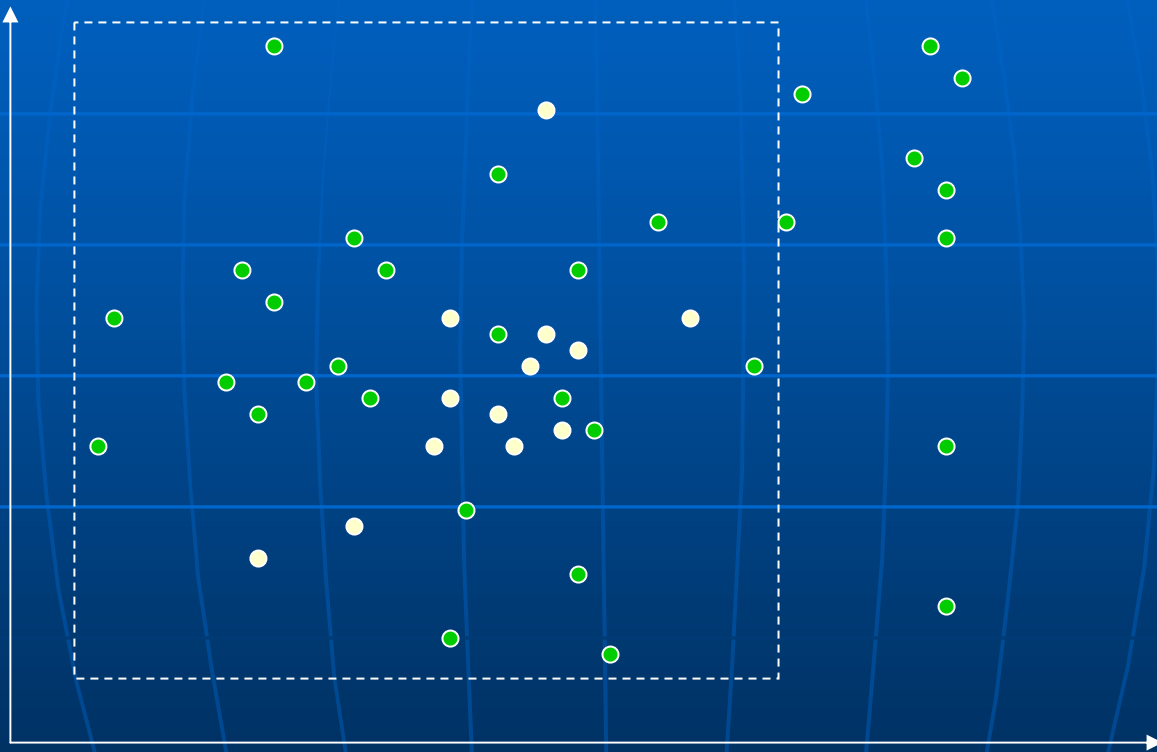
1. Start with all of the training data and a maximal box containing all the data.
2. Shrink the box by compressing one face, so as to peel off the proportion  $\alpha$  of observations having either the highest values of a predictor  $X_j$  or the lowest. Choose the peeling that produces the highest response mean in the remaining box.
3. Repeat 2 until minimal number of observations remain in the box.
4. Expand the box along any face, as long as the resulting box mean increases.
5. Steps 1-4 give a sequence of boxes, with different numbers of observations in each box. Use cross-validation to choose a member of the sequence. Call the box  $B_1$ .
6. Remove the data in box  $B_1$  from the dataset and repeat steps 2-5 to obtain a second box, and continue to get as many boxes as desired:  $B_1, B_2, \dots, B_k$ . Each box is defined by a set of rules involving a subset of predictors like :

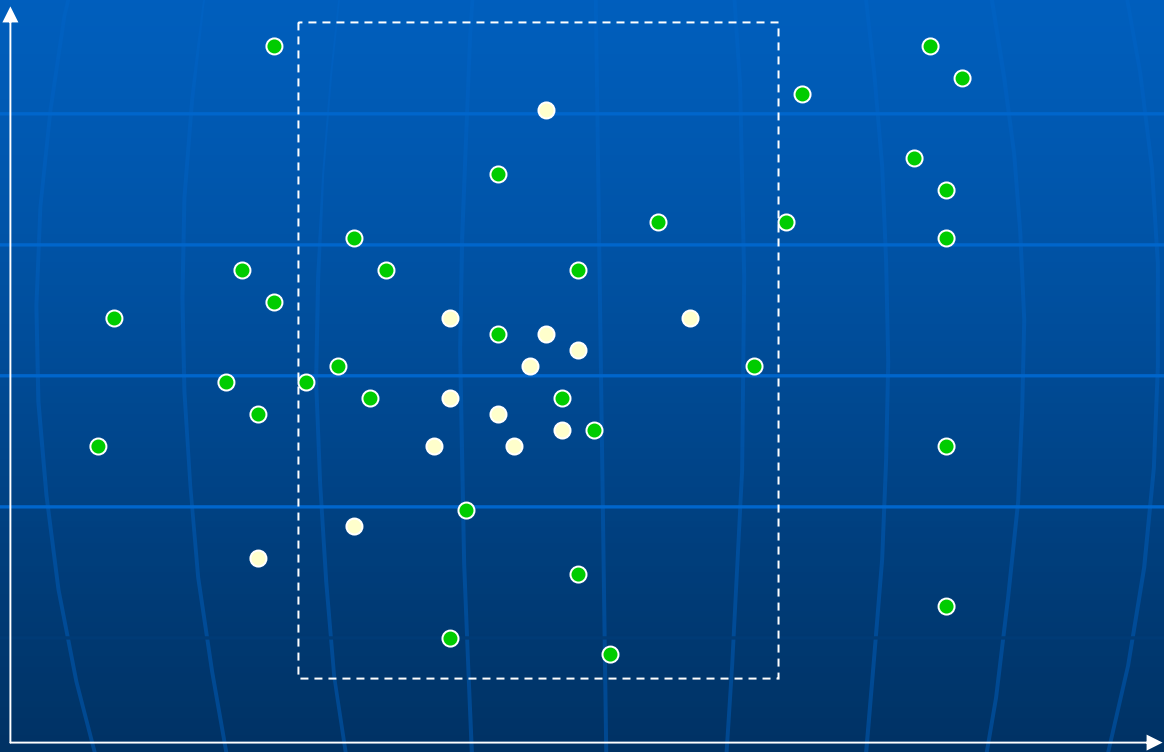
$$(a_1 \leq X_1 \leq b_1) \text{ and } (a_2 \leq X_2 \leq b_2)$$

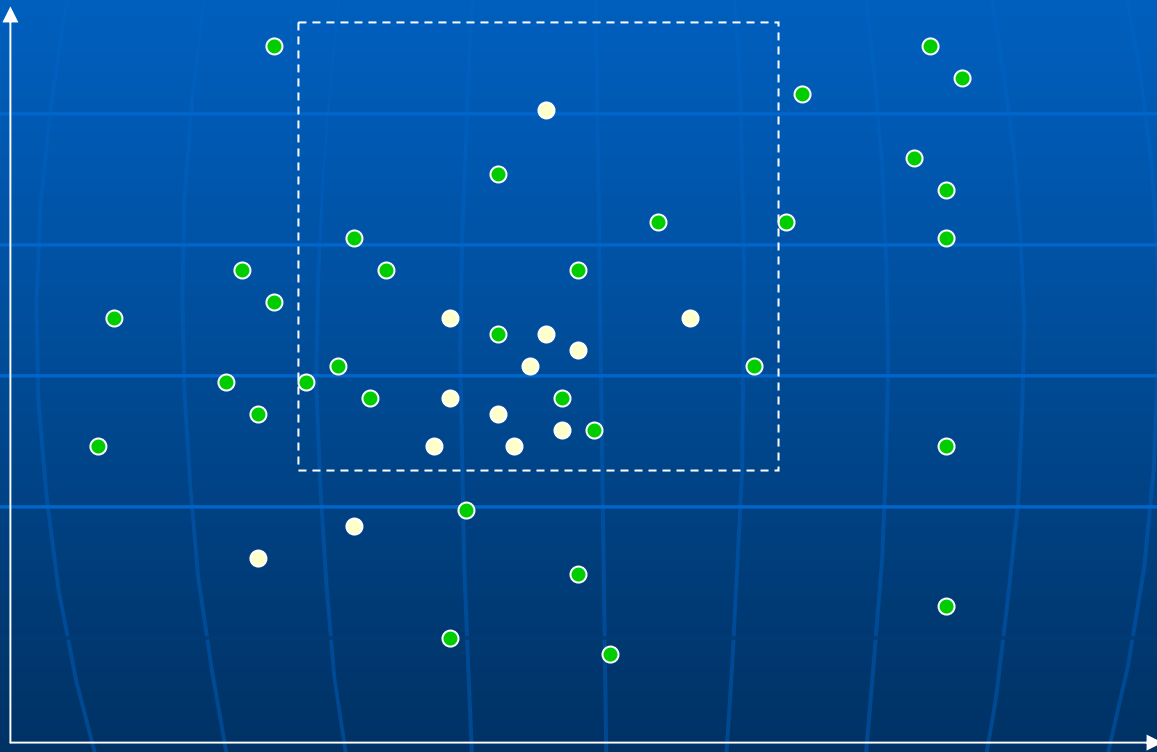
# “PRIM in action”

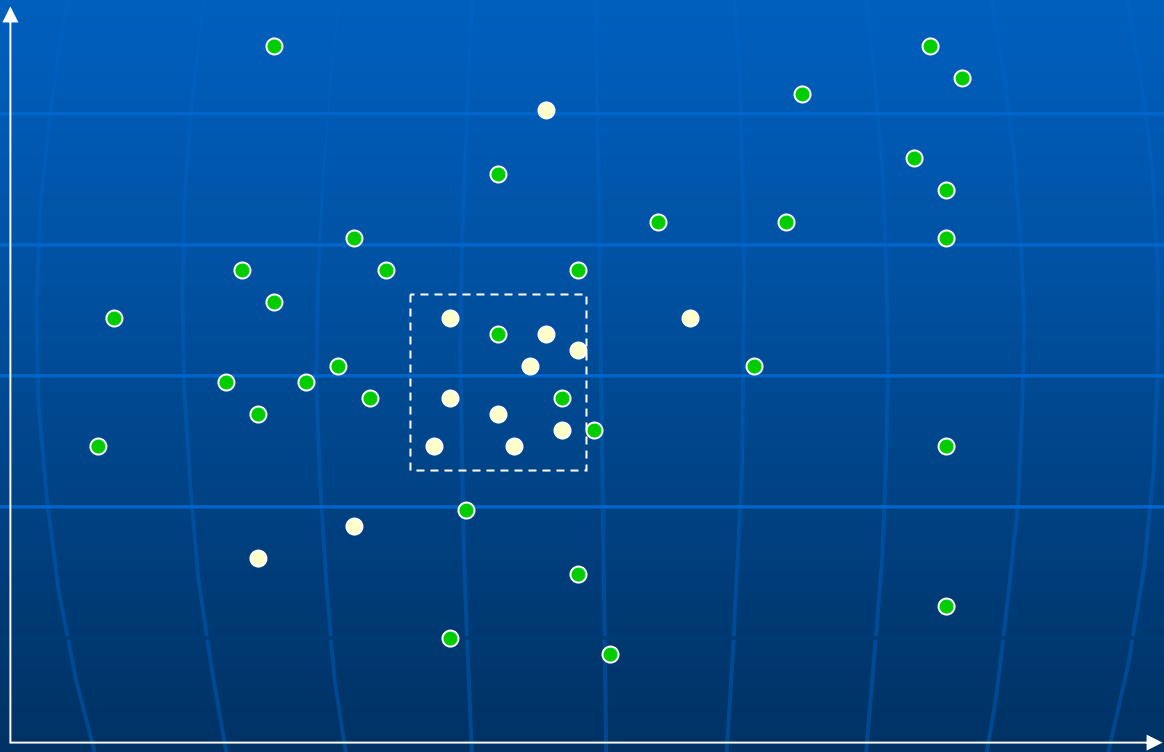












# Example: Marketing Data

## Have a pet

- $y=1$  (have a pet);  
 $y'=.15$
- $B_1: y_1'=.80 \quad \beta_1=.17$ 
  - Age < 45
  - Education < 15 years
  - Live in Bay area > 3
  - Home either house or mobile.
  - Ethnic class: Native American, East Indian, White, missing
- $B_2: y_2'=.76 \quad \beta_2=.08$ 
  - Number of children (< 18) in home > 0
  - Household status: Own, live with parents, missing.
  - Ethnic class: Native American, East Indian, White, missing
- PRIM identified two market segments covering 25% of the sample for which the odds of owning a pet are roughly 5/1.

# Example: Marketing Data

## Number of flights per year

- N=9409 questionnaires (502 questions)
- Questions included: sex, marital status, age, income, type of home, ethnic classification, # in household, etc.
- Examine the frequency of air travel by number of round trip flights per year.
- PRIM Induction gives the following
  - $y$ =number of flights per year (mean  $y=1.7$ )
  - $B_1: y_1' = 4.2 \quad \beta_1 = .08$ 
    - Education > 16 years
    - Occupation in professional/managerial
    - Income > \$50,000
    - Number of children < 18 in home  $\leq 1$ .
  - $B_2: y_2' = 3.2 \quad \beta_2 = .07$ 
    - Education > 12 years
    - $18 < \text{age} < 35$
    - Income > \$30,000
    - Married/Dual income

## Bump Hunting : comments

- Easy to interpret.
- No variable transformations.
- No imputation for missing values needed.
- Practically no prior assumptions on data.

# Improvements

- Hybrid decision trees.
- Genetic algorithm.
  - The Bump Hunting Method Using the Genetic Algorithm with the Extreme-Value Statistics by YUKIZANE et al. IJICE Transactions on Information and Systems 2006.
- Use existing variable selection method to pick variables and followed by systematic search using binary, deciles, etc.

# Sample Lift chart

Max KS is 0.592

