Margin trees for high dimensional classification

Authors: Robert Tibshirani
Trevor Hastie

Picture: www.ruhr-uni-bochum.de

Presentation: Marian Wieczorek

Motivation

- Classifying elements, described by high dimensional features (more than 10,000)

- Organisation of classes lack interpretability

- High quality methods are slow for more than two classes

- Popular application areas like cancer classification

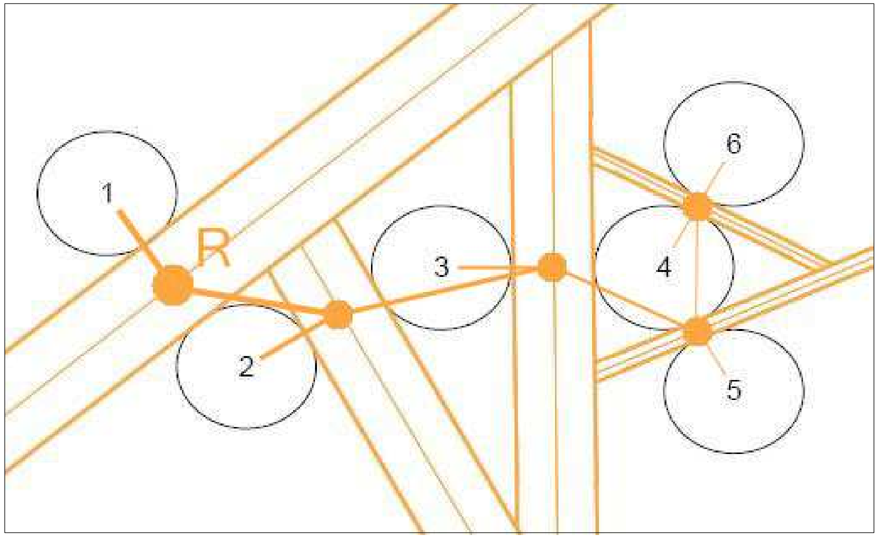
Margin trees

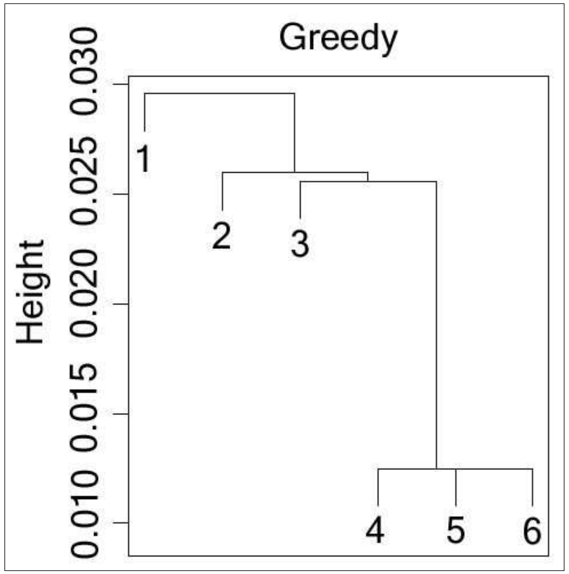
- Approach of creating meaningful abstractions
- Increase performance of the accurate SVM by divide and conquer
  - Combine classes into two groups
  - Calculate classifier for chosen partition
  - Apply procedure on each group

Linkage constraints

- Greedy
  - Top down
  - Choose partition which provides widest margin
  - Requires computation of $O(\sum_{k=1}^{n-1}\binom{n}{k})$ classifiers

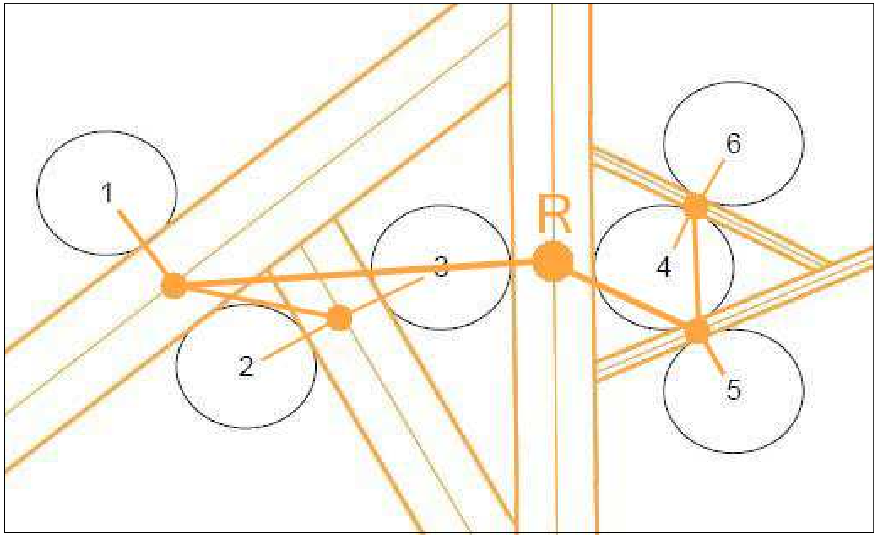
    (own estimation)
  - 14 classes 16,382 possible classifiers

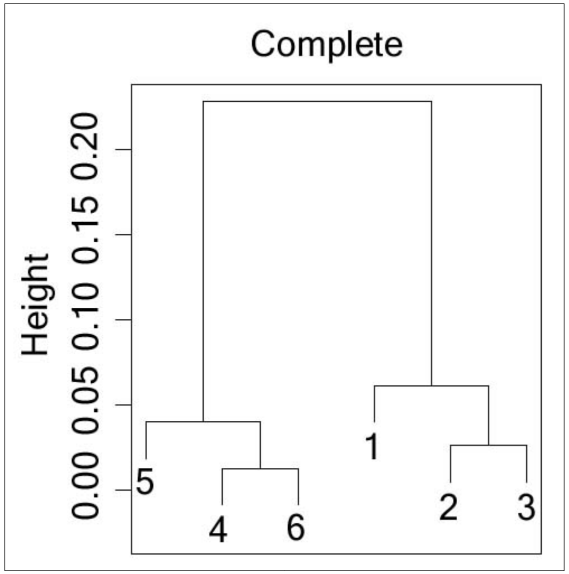




Linkage constraints

- Complete linkage
  - Bottom up
  - At first each class is in its own group
  - Combine groups having least widest margin
  - Requires computation of only O(n^2) classifiers





Organisation of classes

- Greedy does not care about distances between classes in the same group
  - Produces stringy trees by splitting off single classes

- Complete linkage produces groups having same size but differs in shapes (Brian T. Luke)
  - Might be more interpretable because of balanced hierarchy

Kill two birds with one stone?

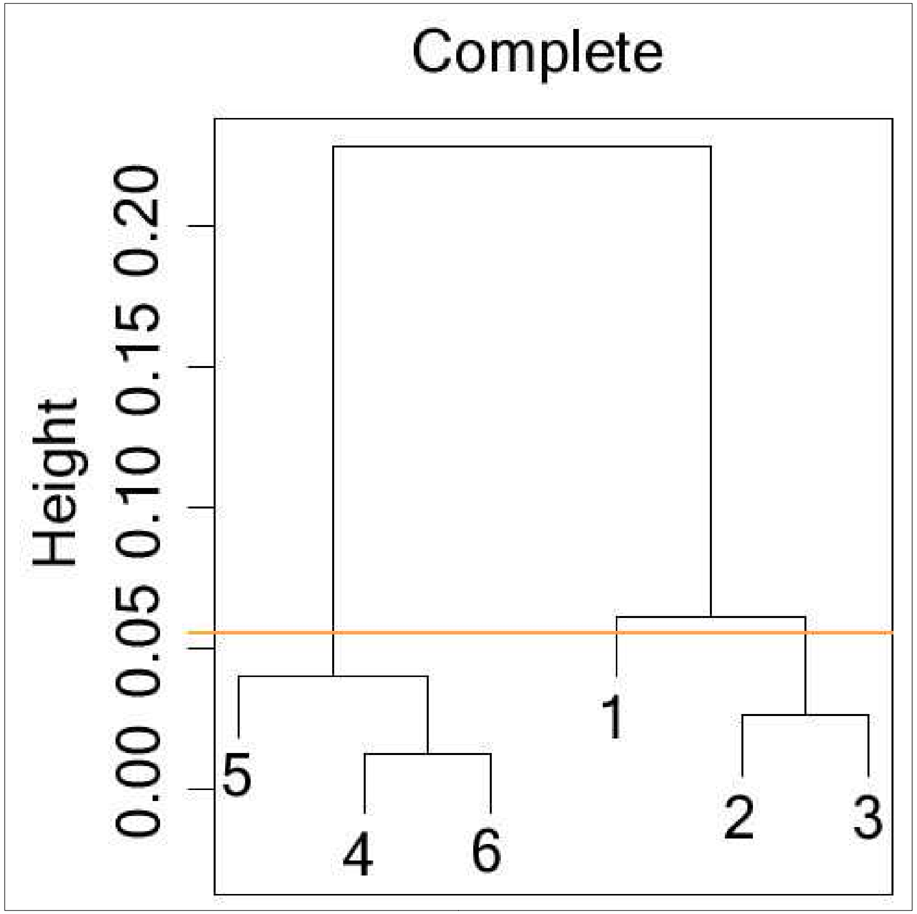
- Complete linkage trees turn out to be competitive to the greedy tree's robustness
- Construced computationally fast
- Mostly balanced Fast classification

and

- Yields an exact algorithm for the greedy criterion

Base of exact algorithm for greedy criterion

- Margin between elements of different groups is greater or equal than margin between those groups
- Cut in complete linkage tree at height M implies less margins in subtrees



Exact algorithm for greedy criterion

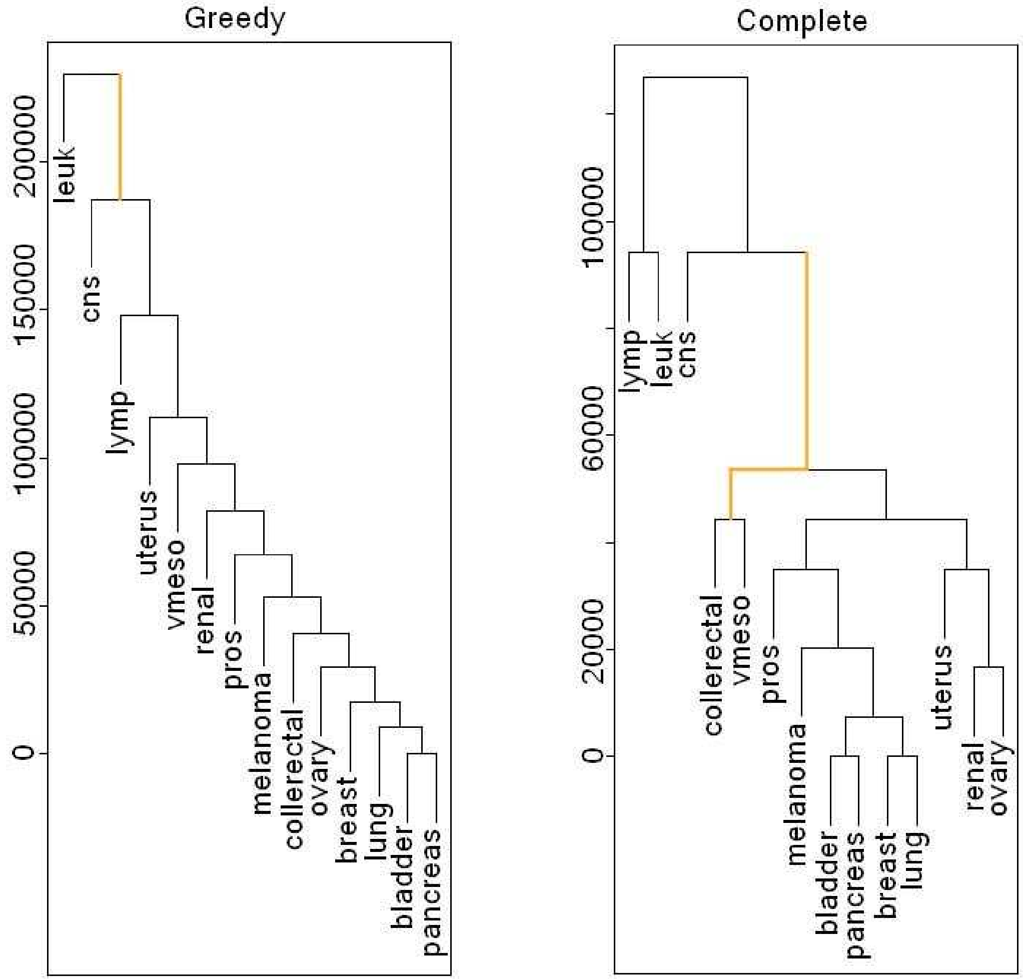
- Greedy criterion according to a monoton decreasing sequence of margins
  - Build complete linkage tree
  - Determine the widest margin achieved by one versus the rest classifiers and the margin of complete linkage tree
  - Cut tree at found height, collapse nodes and proceed with subtrees
- Terminates in $O(n^2+n\cdot n+n)$ □ $O(n^2)$ (own estimation)

Experiment

- Microarray cancer data set of Ramaswamy
  - Samples: 198 tumours (144 for training and 54 for testing)
  - Features: 16,063 genes
  - Classes: 14 types
- Comparison of all-pairs SVM, exact greedy, complete linkage and nearest centroid classifiers

Experiment

- Approximation of greedy tree can fail performing wide margins!



Experiment

- Similar error rates 20, 18, 18 and 35 (for nearest centroids)

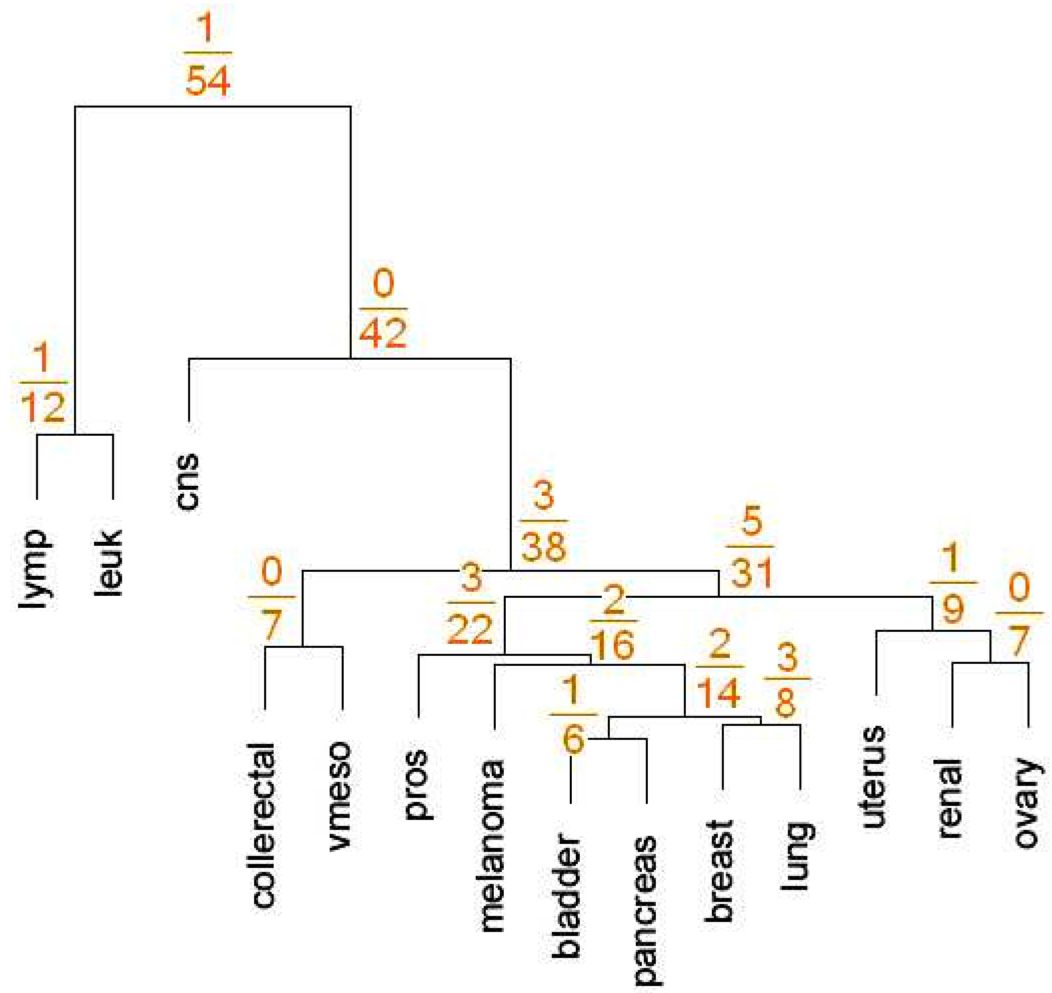
	Margin Trees	
	Greedy	Complete
SVM	10	10
Greedy	0	2

- Table shows the number of times each classifier disagreed on the test set

- 90% overlap in true-positives and disagreement almost only on false-positives

Experiment

- It emerges that the error rate increases close to the leafs



Feature selection

- Nodes at the top seem to be less arbitrative

- Some features might have no effect on classification

- Reducing features would be beneficial, because
  - groups would be more interpretable
  - classification could become faster

Hard thresholding

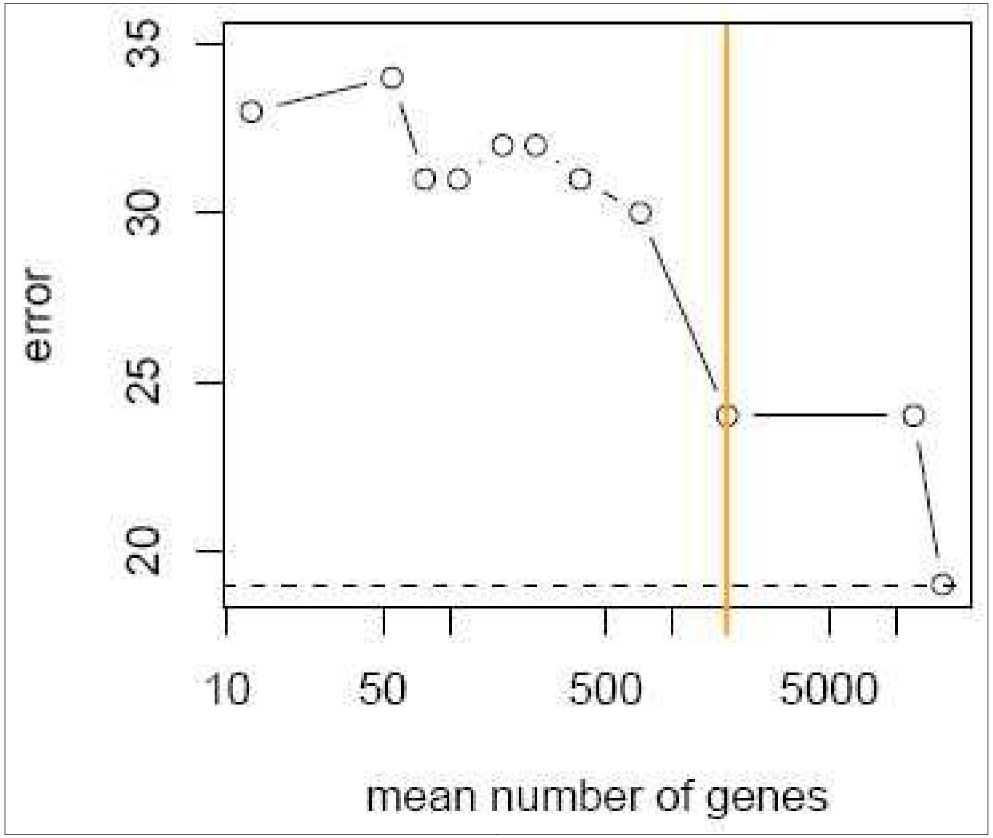
- Sort coefficients of weight vector which defines orientation of a margin in descending order
- Choose a number n_k for the k-th split
- Set first n_k coefficients to zero
- Adjust only position of reduced margin
- How to choose n_k?

Limit of deviation

- Introduce a new variable α
- α is limit of deviation between unmodified and trimmed weight vector
- At each level n_k is chosen individually
  - Fewer features at the top
  - More features close to the leafs
- Use tenfold cross-validation to estimate α

Quality of hard thresholding

- Experiments point out that features can be reduced to < 12.5% without too much loss of accuracy
- Recursive feature elimination only small advantage for already little number of features



Quality of hard thresholding

- Preserves interpretability
  - Remaining coefficients are subset of total featurevector
  - Usually reduced coefficients not easily predictable with common methods

Discussion

- Experiments show that margin trees are competitive to accurate methods like
  - Multiclass support vector machine
  - Nearest centroid methods
- Provide meaningful hierarchy and interpretable feature reduction
- Leave the door open for other classification strategies

Discussion

- Nonlinear separable class distribution impede feature reduction
- Number of training samples is supposed to be less than number of features. Else:
  - Not linearly separable
  - One class might be splitted in to leaves
- There are several related methods to their work with asserts and drawbacks

Further reading

- Sources of paper:
  stat.stanford.edu/~hastie/Papers/margintree.pdf
- Agglomerative clustering:
  fconyx.ncifcrf.gov/~lukeb/agclust.html
- Nonlinear support vector machines:
  www2.tuebingen.mpg.de/agbs/sc06/wiki/slides_nonlinear_svms.pdf