Random Forests:
Proximity, Variable Importance and Visualization

Random Forests

Repeat the following steps many times:

- Take a bootstrap sample of the data.
- Fit a tree[*] to the bootstrap sample.

Vote (or average) the trees to determine the prediction.

[*] At each node, independently, split on the best of *m* randomly-chosen variables.

Why Random Forests?

- Accuracy
- Interpretability using:

  proximities

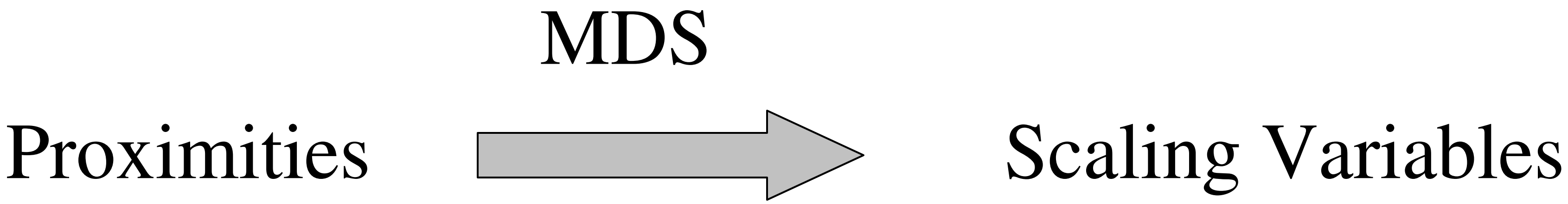
  variable importance:

  - *overall*
  - *casewise*
  - *classwise*

WHAT ARE THESE?

Proximities and scaling

Proximity: each time two items end up in the same terminal node, increase their proximity by 1/(number of items in the node)



MDS

Proximities

Scaling Variables

Variable importance

For each tree, look at the out-of-bag data:

- Randomly permute the values of variable *j*.
- Pass oob data down the tree, save the classes.

For case *i* and variable *j* find:

oob error rate with variable *j* permuted – oob error rate without permutation

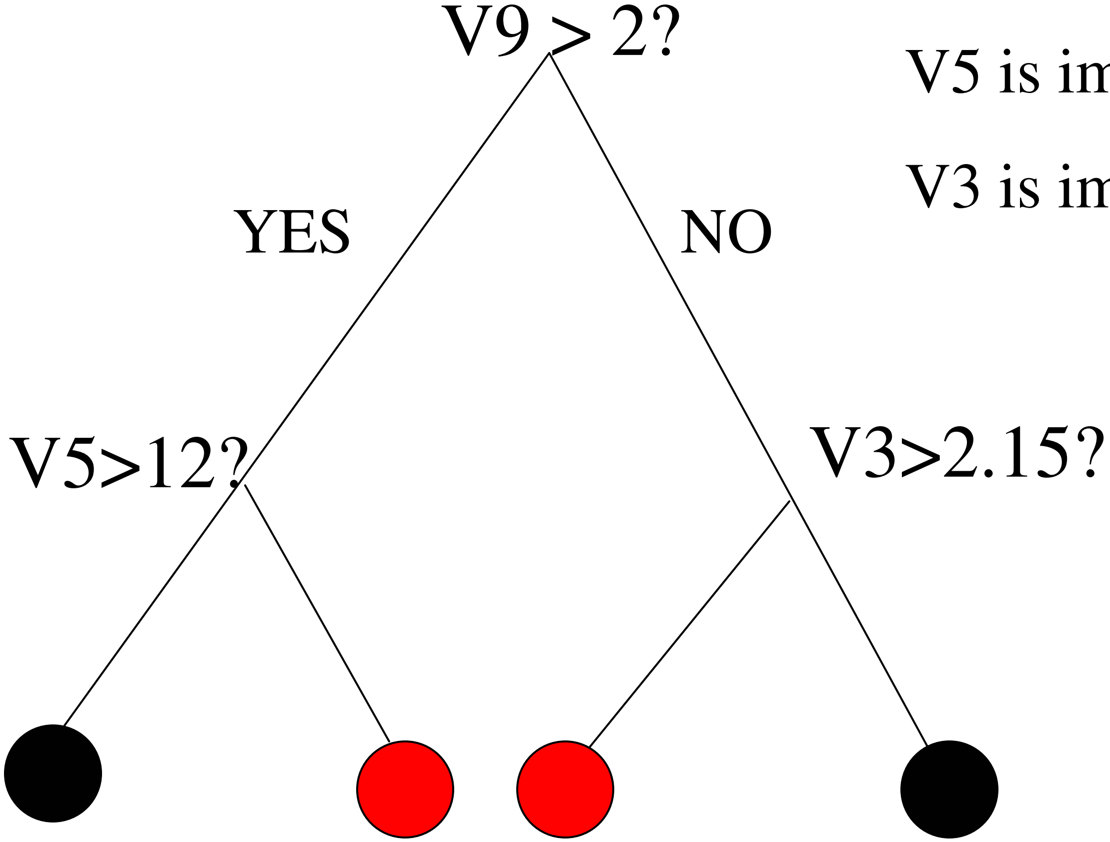
Average for overall/classwise variable importance.

Finding variable importance for a class 2 case:

oob in tree:	No permutation	Permute variable 1	…	Permute variable p
1	2	2	…	1
3	2	2	…	2
10	2	1	…	1
17	2	2	…	2
19	1	1	…	1
23	2	2	…	1
…	...	…	…	…
992	2	2	…	2
% Error	10%	11%	…	35%

CASEWISE variable importance

Different variables are important in different regions of the data



V9 > 2?

YES

NO

V5>12?

V3>2.15?

V5 is important if V9 > 2

V3 is important if V9 ≤ 2

Why Random Forests?

- Accuracy
- Interpretability using:

  proximities

  variable importance:

  - *overall*
  - *casewise*
  - *classwise*

HOW DO WE USE THEM?

Visualizing proximities

- at-a-glance information about which classes are close, which classes differ
- find clusters within classes
- find easy/hard/unusual cases
- see how clusters or unusual points differ
- see which variables are locally important

  (eg which help separate one class out of several)

RAFT

RAndom Forests graphics Tool

- java-based, stand-alone application
- uses output files from the fortran code
- download RAFT from

www.stat.berkeley.edu/users/breiman/

RandomForests/cc_graphics.htm

COMMERCIAL VERSION: Salford Systems

Raft uses VisAD

www.ssec.wisc.edu/~billh/visad.html

and ImageJ http://rsb.info.nih.gov/ij/

Case study I : Brain Cancer Microarrays

Pomeroy et al. Nature, 2002.

Dettling and Bühlmann, Genome Biology, 2002.

42 cases, 5,597 genes, 5 tumor types:

- 10 medulloblastomas BLUE
- 10 malignant gliomas PALE BLUE
- 10 atypical teratoid/rhabdoid tumors (AT/RTs) GREEN
- 4 human cerebella ORANGE
- 8 PNETs RED

Case study II : Autism

data courtesy of J.D.Odell and R. Torres, USU

154 subjects (308 chromosomes)

7 variables, all categorical (up to 30 categories)

2 classes:

- normal (69 subjects) BLUE
- autistic (85 subjects) RED