UNIVERSITY OF TWENTE.

RANDOM FOREST

ADVANCED COURSE ON MACHINE LEARNING

RANDOM FORESTS
L. BREIMAN (MACHINE LEARNING 2001)

Procedure:

1. Select beforehand a number ***m*** much smaller than the dimensionality **M** of the data.
2. For each new tree draw a new training set, with replacement, for the original training set. This is called *bagging* or *bootstrapping.*
3. In the tree construction select for each node at random ***m*** features and split on the best one.
4. After constructing *sufficient* trees the majority vote over the ensemble is the classification of a new datapoint.

How to estimate ***m*** and how to determine *sufficient*?

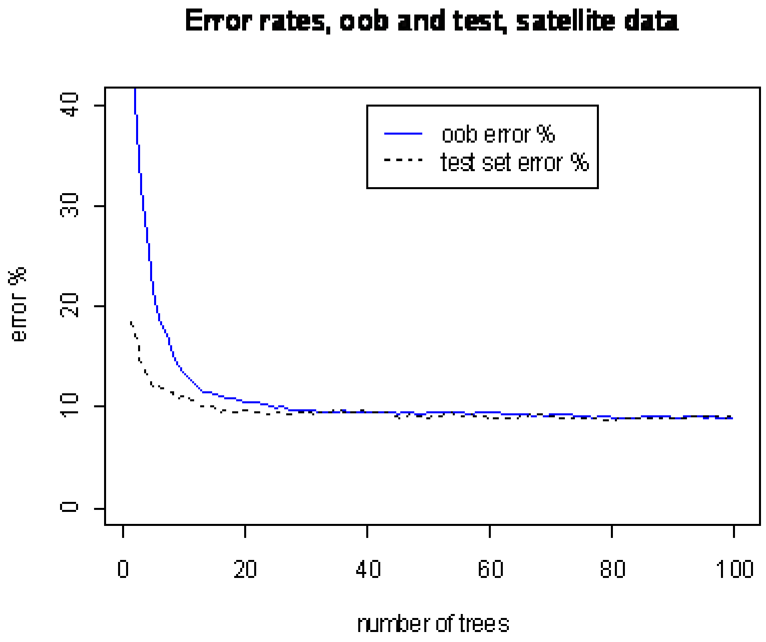
OUT OF BAG ERROR RATE

ADVANTAGE OF BAGGING.

- Due to bagging/bootstrapping approximately 1/3 of the training data is not used for training a tree in the random forest.
- For each data point ***x*** calculate the majority vote over all trees which did not use ***x*** for training (approximately 1/3 of the trees). This the predicted class label for ***x***.
- Calculate the average error rate over the total training set. This is called the out of bag (oob) error rate.
- This oob error rate is a good estimator of the generalization performance of the random forest.

- This implies that random forests do not overfit.

EXAMPLE OF OOB ERROR RATE AND TEST SET ERROR RATE.



HOW TO DETERMINE **m**?

The out-of-bag error rate is used to select **m**.

Here's how:

1. Start with **m** = √ **M**. **M** the dimensionality of the data.
2. Run a few trees, recording the out-of-bag error rate.
3. Increase **m**, decrease **m**, until you are reasonably confident you've found a value with minimum out-of-bag error rate.

HOW TO DETERMINE **sufficient?**

- Once again record the OOB error rate and stop generating trees when it does not decrase anymore.