ANALYTICS PATH

Ensemble Learning

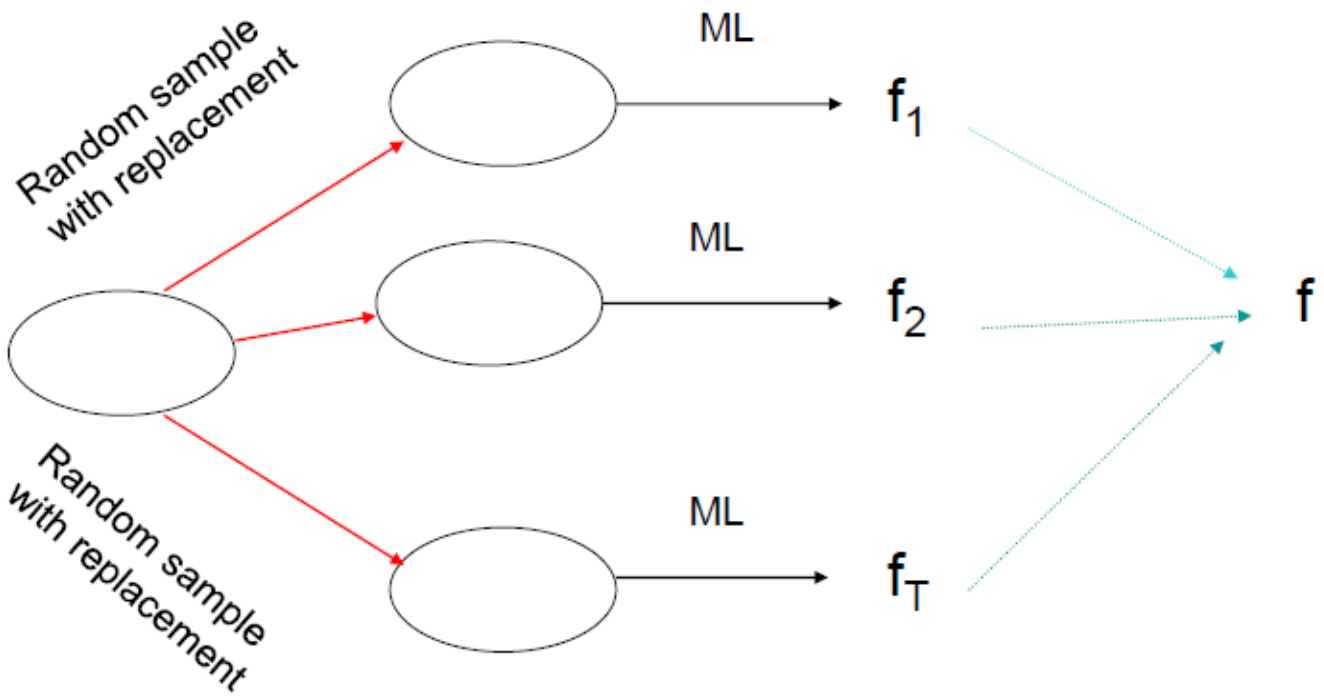
Plot No. 28, 4th Floor, Suraj Trade Center,**Opp. Cyber Towers,** Hitech City, Hyderabad - 500081, Telangana.
India Tel: 040 - 66828899, Mob:+91 7842828899, Email: info@analyticspath.com

Ensembles

- Combine multiple models (weak learners) to produce strong learners

- Advantages
  - Leverage the strengths of different kind of models
  - Reduce overfitting
  - Very small or Very Large data can be handled well
- Disadvantage
  - Speed and Explicability

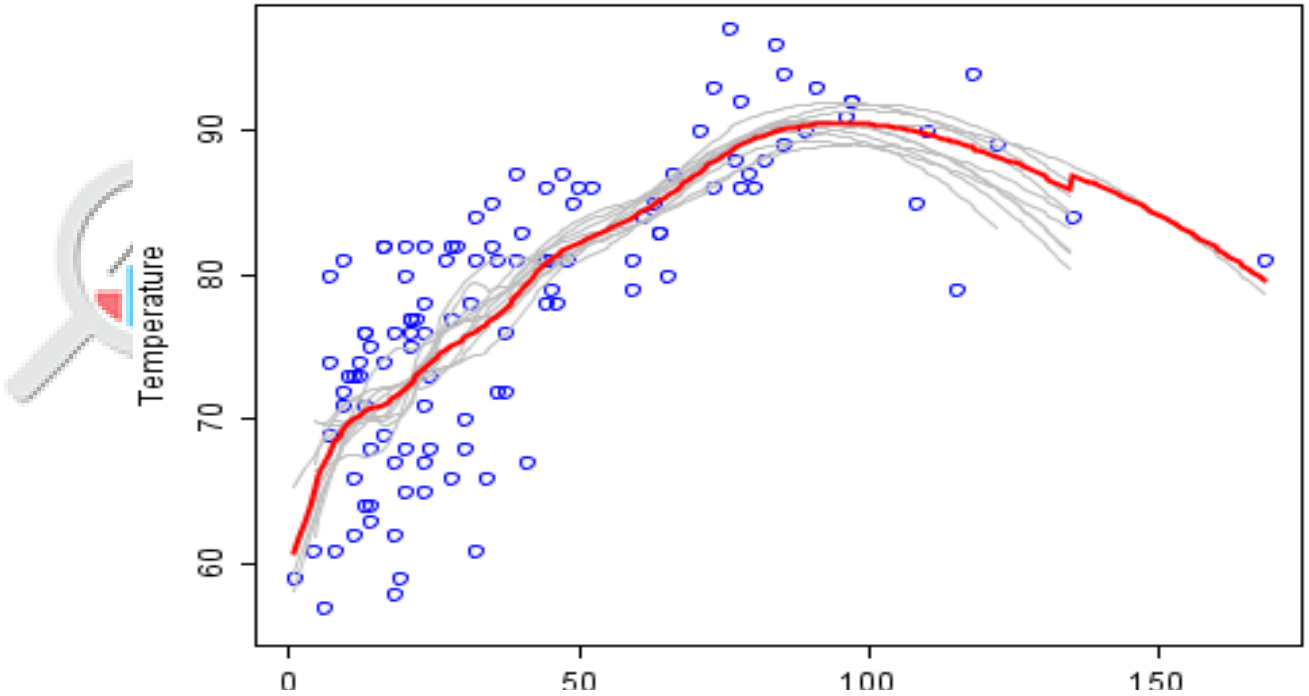
Bagging


Random sample with replacement
ML
ML
ML
f_1
f_2
f_T
f
Random sample with replacement

Bagging (Bootstrap Aggregating)

- Bootstrap the samples and create multiple sets
  - Pick randomly with replacement
- Run a base learner and generate one weak classifier for each set
- Vote on test samples

Bagging - Example



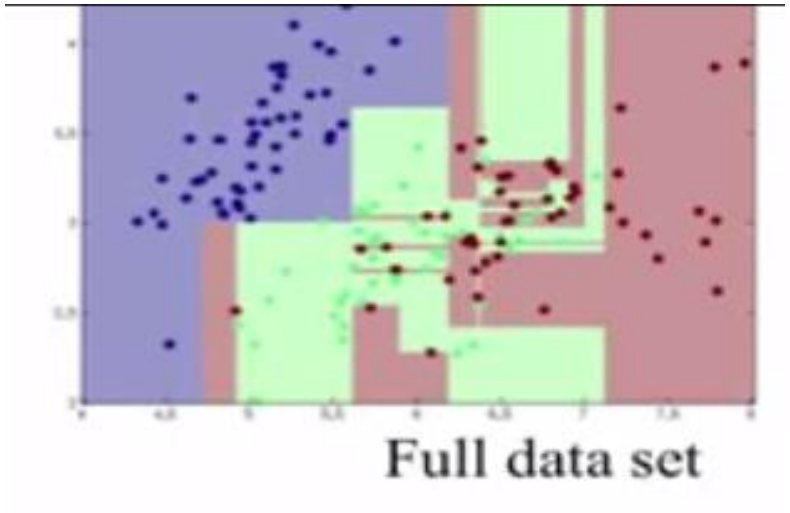
Plot No. 28, 4th Floor, Suraj Trade Center,**Opp. Cyber Towers,** Hitech City, Hyderabad - 500081, Telangana.
India Tel: 040 - 66828899, Mob:+91 7842828899, Email: info@analyticspath.com

Bagging Contd…

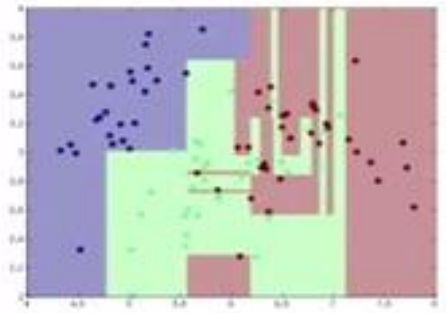
- Bagging is good for unstable learners as it reduces variance and overfitting
  - How do I generate a large number of unstable learners
    - Choose records randomly
- Rules, decision trees
  - should work wonders

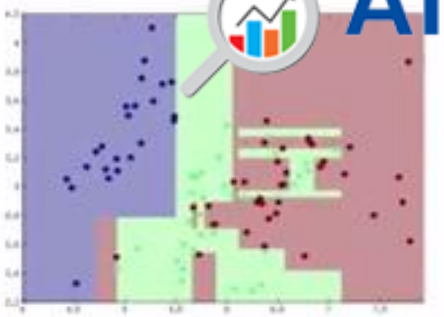
Bagging

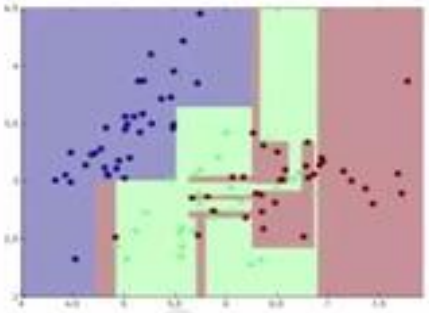
Simulates "equally likely" data sets we could have observed instead, & their classifiers

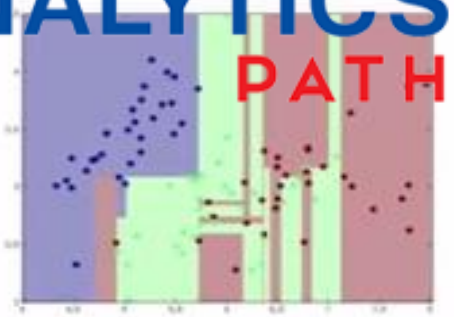


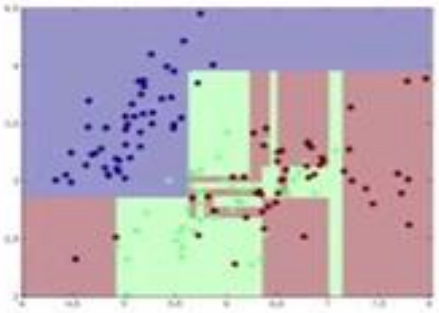
Full data set



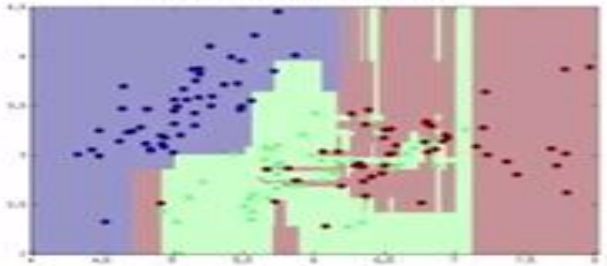




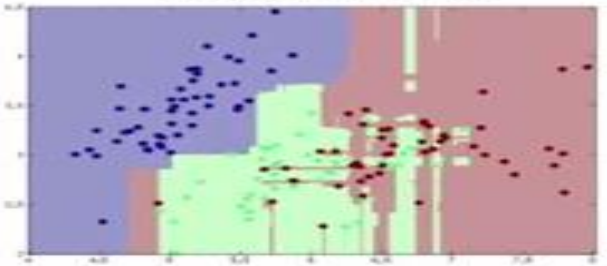




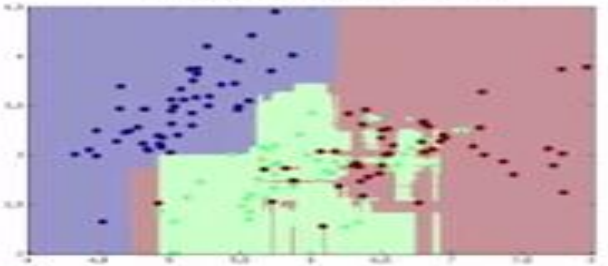
Avg of 5 trees



Avg of 25 trees



Avg of 100 trees



Plot No. 28, 4th Floor, Suraj Trade Center,**Opp. Cyber Towers,** Hitech City, Hyderabad - 500081, Telangana.
India Tel: 040 - 66828899, Mob:+91 7842828899, Email: info@analyticspath.com

A lot of data??

- With a lot of data, we usually learn the same classifier most of the times. So how do we handle this?
- How do we model when we have a lot of dimensions?

ANALYTICS PATH

Random Forest

A variant of the Bagging concept

Plot No. 28, 4th Floor, Suraj Trade Center,**Opp. Cyber Towers,** Hitech City, Hyderabad - 500081, Telangana.
India Tel: 040 - 66828899, Mob:+91 7842828899, Email: info@analyticspath.com

Random Forest

ANALYTICS PATH

- Select a large number of data sets through bagging (size is same in all)
- Use m input variables at each node of each tree. m should be much less than M (total attributes).
- Each tree is fully grown and not pruned
- Mode (for classification) or average for regression of all the trees is used as prediction.

Random Forests

- One of the best Machine Learning Algorithms
- Regression and Classification problems can be solved
- Right number of trees and right number of attributes to be used are to be selected