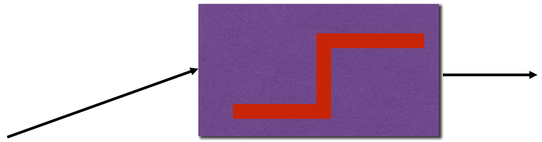
Ensembles

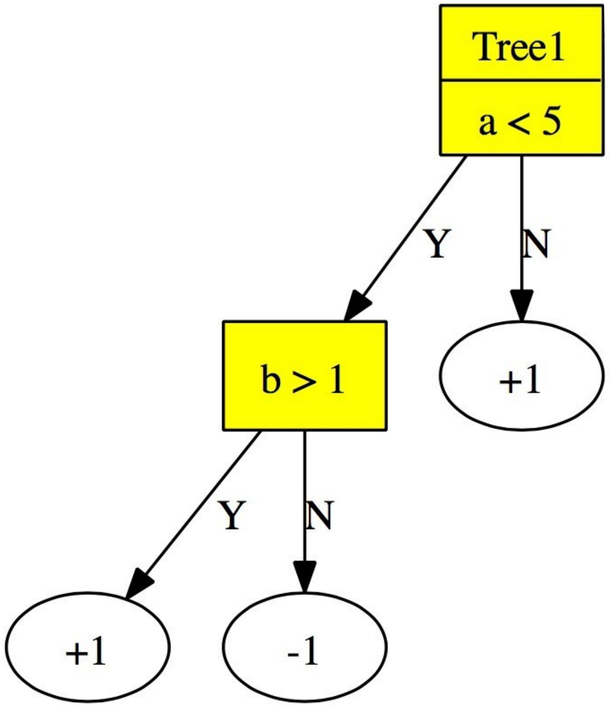
What are ensembles

- Ensembles are predictors defined as an average/vote over “base” or “weak” predictors.
- Weak learner is faced with a variant of the original prediction problem.
- Ensembles come in two main flavors:
  - Boosting based Ensembles
  - Bootstrap based Ensembles.

An Ensemble of trees

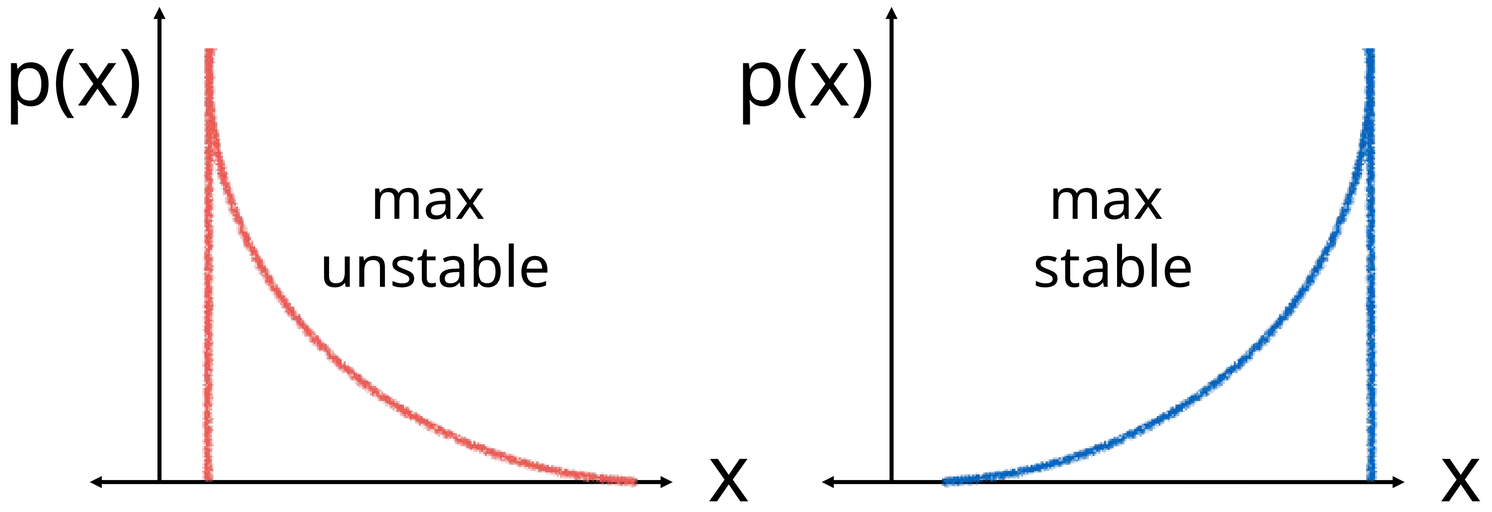


+1/-1



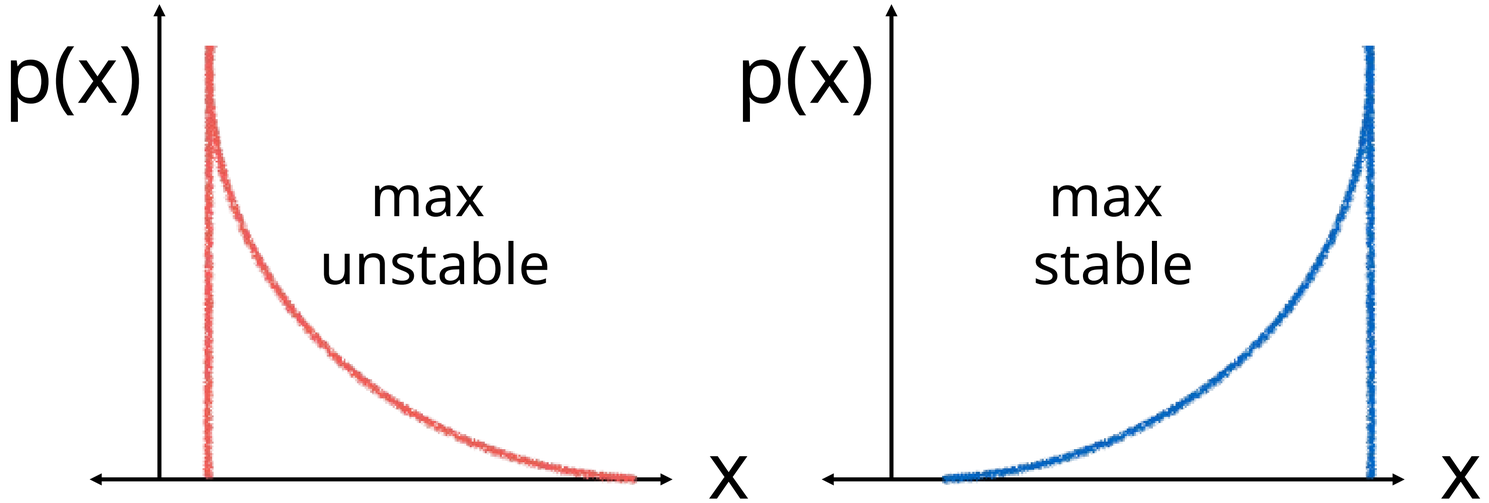
Stability of statistics

- Quantities we want to estimate: mean, std, median, min, max
- Stable estimator: varies little from sample to sample.
- Median is the most stable, mean is less stable, std still less.
- The variation of max depends on the distribution.
- Direct estimation of stability: use several independent datasets. (requires a lot of data).



The Bootstrap

- A method for estimating out-of-sample variation
- Instead of collecting truly independent samples, we create semi-independent samples from the given sample.
- **S** : Original sample of size N.
- Bootstrap sample: select N examples from **S** independently at random with replacement.
- Use bootstrap samples as if they were independent samples.



1990

An Introduction to the Bootstrap

Bradley Efron
Department of Statistics
Stanford University
and
Robert J. Tibshirani
Department of Preventative Medicine and Biostatistics
and Department of Statistics, University of Toronto

CHAPMAN & HALL/CRC
Boca Raton London New York Washington, D.C.

Bagging = bootstrap aggregation

- Decision trees have high data variation.
  - i.e. the generated tree is sensitive to small changes in the training set.
- To reduce the variation, we take a majority vote over several runs, each using an independent random resample of the training data.
- **Bootstrap:** Running an algorithm over random resampling.
- Trees can be learned in parallel
- The result is a reduction in variation with no increase in the bias.

Random Forests

- Based on bagging trees.
- Additional randomization: before choosing which leaf to split and how, choose a random subset of the features.
- Decreases the correlation between different trees.
- Speeds up the learning process.
- All trees get equal weight (1.0)
- All trees can be learned in parallel.

Gradient Tree Boosting

- The trees are trained sequentially, one after the other.
- Each tree is trained using a **weighted** training set. The weights represent the gradient of the loss function.
- Each tree receives a different weight.
- Stochastic gradient boosting: use random resampling of the training set a.k.a. Bagging.

Summary

- Ensemble learning in Spark ML
- Bagging = Bootstrap Aggregation
- Random Forests
- Boosted Gradient trees.