

Decision Trees and Aggregation

Gautam Kamath

Example: Should I wear a jacket?

- (Draw example on board in tree format)
 - Features: temperature, is cloudy?
 - Label: Jacket or no jacket
- (Draw example on board in 2D picture)

Decision Trees

- Simple and intuitive
- “Interpretable”
- Classification or Regression
- Can tend to overfit
- Can handle (some) non-linear functions, but may fail on linear ones
 - (Draw example of diagonal separator, not axis-aligned)
- (Draw example and tree with $-$'s on left, $+$'s on right, but $+$'s on top left and $-$'s on top right)
 - (Do recursive splitting, informally)
- Prediction: Walk down tree, use leaf's prediction

Building a tree

- Start with one node
- Recursively split node
- Split(leaf):
 - Choose variable and threshold
 - Create two leaves and partition of points
- (Draw 1D example with dense +’s on left, dense –’s on right)
- (Show a few example choices of threshold)
- Intuition: Pick split that makes leaves “pure”

How to split?

- Choose a loss function for a node which is small for pure nodes and large for mixed ones
- Split at threshold t which minimizes (weighted) sum of new node costs

$$t^* = \arg \min_t |S_L| \ell(S_L) + |S_R| \ell(S_R)$$

where $S_L = \{(x_i, y_i) : x_i \leq t\}$, $S_R = \{(x_i, y_i) : x_i > t\}$

- Only have to try at most n different values of t
- What loss function to use?

Loss functions? (Draw them for binary)

- Computing $\ell(S)$, where $S = \{(x_i, y_i)\}$ is some subset of examples
- Let $\hat{p}_c$ = fraction of S with label $c = \frac{1}{|S|} \sum_{i \in S} \mathbf{1}\{y_i = c\}$
- (Give example with 4 0's and 2 1's)
- $\hat{y} = \arg \max_c \hat{p}_c$
- Misclassification loss: $\ell(S) = 1 - \hat{p}_{\hat{y}}$
 - If all 0's, $\hat{p}_0 = 1$, $\ell(S) = 1 - 1 = 0$. If 50-50, $\hat{p}_0 = \hat{p}_1 = 1/2$, $\ell(S) = 1/2$
- Entropy: $\ell(S) = - \sum_{classes\ c} \hat{p}_c \log \hat{p}_c$
- Gini index: $\ell(S) = \sum_{classes\ c} \hat{p}_c (1 - \hat{p}_c)$
- In regression setting: $\ell(S) = \min_p \frac{1}{|S|} \sum_{i \in S} (y_i - p)^2 = \frac{1}{|S|} \sum_{i \in S} (y_i - \bar{y}_S)^2$

Which variable to split on? Try all, pick best

$$S_L^{(j)} = \{(x_i, y_i) : x_{ij} \leq t\}, S_R^{(j)} = \{(x_i, y_i) : x_{ij} > t\}$$

$$(j^*, t^*) = \arg \min_{j, t} |S_L^{(j)}| \ell(S_L^{(j)}) + |S_R^{(j)}| \ell(S_R^{(j)})$$

Gini index: $\sum_{classes\ c} \hat{p}_c(1 - \hat{p}_c)$

Split on smokes?

No: $\hat{p}_0 = \frac{3}{4}, \hat{p}_1 = \frac{1}{4}$. Yes: $\hat{p}_0 = \frac{1}{3}, \hat{p}_1 = \frac{2}{3}$.

Cost: $4 \cdot \left(\left(\frac{3}{4} \right) \left(\frac{1}{4} \right) + \left(\frac{1}{4} \right) \left(\frac{3}{4} \right) \right) + 6 \cdot \left(\left(\frac{2}{3} \right) \left(\frac{1}{3} \right) + \left(\frac{1}{3} \right) \left(\frac{2}{3} \right) \right) = 4.16$

Split on age? (Cheat to save time: use 35 as split)

≤ 35 : $\hat{p}_0 = 1, \hat{p}_1 = 0$. > 35 : $\hat{p}_0 = \frac{1}{6}, \hat{p}_1 = \frac{5}{6}$.

Cost: $4 \cdot ((0)(1) + (1)(0)) + 6 \cdot \left(\left(\frac{1}{6} \right) \left(\frac{5}{6} \right) + \left(\frac{5}{6} \right) \left(\frac{1}{6} \right) \right) = 1.66$

Age	Smokes	Cancer?
10	No	0
18	Yes	0
25	No	0
35	Yes	0
50	No	1
55	Yes	1
70	Yes	1
80	No	0
85	Yes	1
90	Yes	1

Stopping?

- Depth
- Running time
- Few examples at each leaf
- Leaves are all homogeneous
- Small improvements via a split
 - $\Delta = |S|\ell(S) - (|S_L|\ell(S_L) + |S_R|\ell(S_R))$

Pruning

- Grow tree “fully” without stopping early, regularize over subtrees
- $\min \sum_{\text{leaf } v} \text{"error" in leaf } v + \alpha(\# \text{ of leafs})$
- (Draw different trees at stages, indicate $\alpha = 0$ vs ∞ cases)

Decision Stump

- A three node decision tree (draw)
- Fast, but not very good

Aggregation Methods

Boosting

Boosting

- Given several “weak learners,” can we combine them into a “strong learner”?
- Weak learner: “55% accurate” (slightly better than random guess)
 - Focus on binary classification today
- Strong learner: “90%+ accurate” (a good classifier)
- Iterative process. Train a classifier. “Downweight” points it gets right, “upweight” points it gets wrong. Train classifier on new weighted dataset (draw)
- Bit of a diversion until we get to that...

“Online Learning with Experts”

- Example: Horse racing, with n horses and T races. How to choose which horse to bet on? How to update your bet after each race?
- More general setting, T rounds of the following:
 1. At round t , algorithm specifies weights $p_1^{(t)}, \dots, p_n^{(t)}$ such that $\sum_i p_i^{(t)} = 1$
 - Choose a distribution over the different “experts”
 2. Algorithm experiences “loss” at time t of $\langle p^{(t)}, \ell^{(t)} \rangle$
 - $\ell^{(t)} \in [0,1]^n$ is an adversarially picked loss vector
- Goal: Minimize $\sum_{t=1}^T \langle p^{(t)}, \ell^{(t)} \rangle$
- Try to compete with “best single expert in hindsight”
 - $\min_i \sum_{t=1}^T \ell_i^{(t)}$ -- same as goal when $p_i^{(t)} = 1$ for all t

Hedge Algorithm

Hedge(β), where $\beta \in (0,1)$

1. Initialize $w^{(1)} = [1/n, \dots, 1/n] \in \mathbf{R}^n$
2. For $t = 1, \dots, T$
 1. Set $p^{(t)} = \frac{w^{(t)}}{\sum_i w_i^{(t)}}$ (normalize w into a distribution)
 2. Receive loss $\langle p^{(t)}, \ell^{(t)} \rangle$
 3. Update $w_i^{(t+1)} = w_i^{(t)} \beta^{\ell_i^{(t)}}$ (downweight experts based on loss)

Guarantee: $\sum_{t=1}^T \langle p^{(t)}, \ell^{(t)} \rangle \leq \frac{1}{1-\beta} \left(\log n + \log(1/\beta) \min_i \sum_{t=1}^T \ell_i^{(t)} \right)$

- Must choose β to balance the two costs

Hedge Guarantees

- $\frac{1}{T} \sum_{t=1}^T \langle p^{(t)}, \ell^{(t)} \rangle \leq \frac{1}{T} \min_i \sum_{t=1}^T \ell_i^{(t)} + O\left(\sqrt{\frac{\log n}{T}}\right)$
- LHS: average loss at each step
- RHS: average loss of best expert in hindsight, plus “regret”
- Average regret goes to 0 as $T \rightarrow \infty$
- We can be very competitive with choosing the best expert in hindsight!
- Very powerful framework! Useful in linear programming, game theory, etc.
- Now, how do we use this for standard ML classification...?

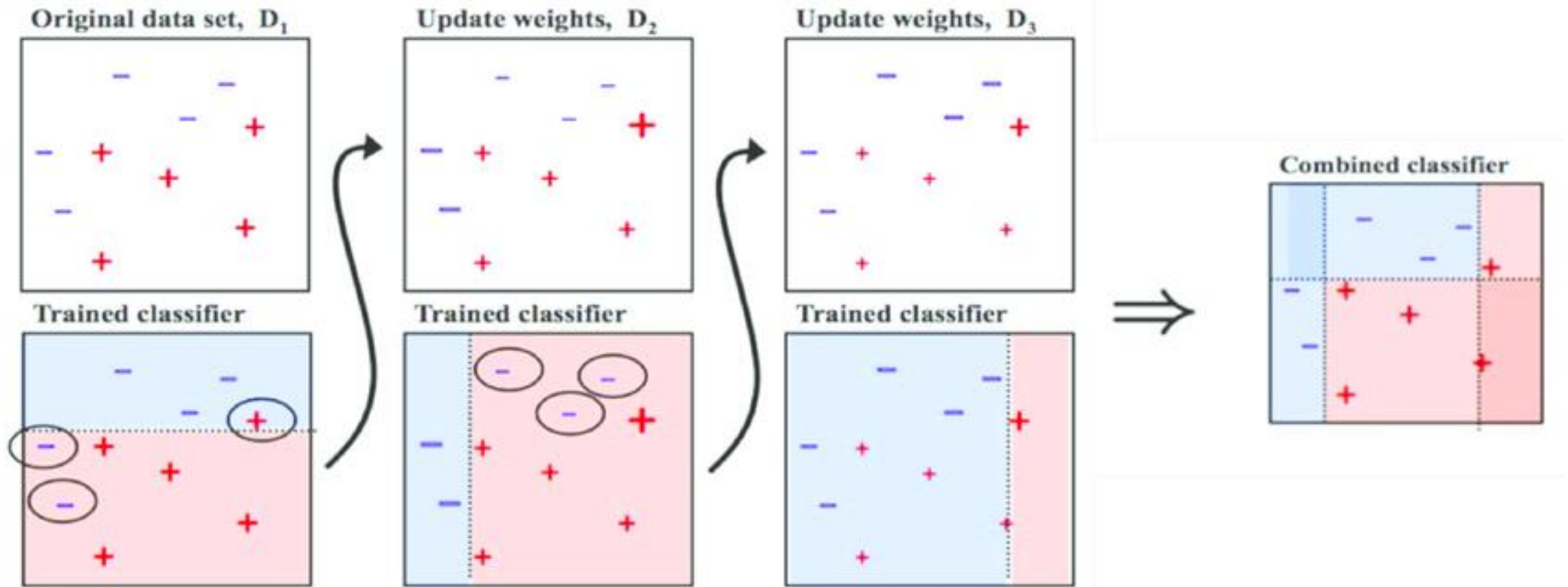
AdaBoost

- Given algorithm WeakLearn that gets 55% accuracy on a training set. Can we boost this to high probability?
 - Nuance: does this well on *every weighting* of the training data
- Wrong way: Run the algorithm many times on the dataset, treat resulting classifiers as “experts.” “Put large weight on good classifiers”
- Right way: Treat the *datapoints* as experts. Put large weight on points that haven’t been learned yet.

AdaBoost

1. Initialize $w^{(1)} = [1/n, \dots, 1/n] \in \mathbf{R}^n$
2. For $t = 1, \dots, T$
 1. Set $p^{(t)} = \frac{w^{(t)}}{\sum_i w_i^{(t)}}$ (normalize w into a distribution)
 2. Run WeakLearn on training set (with weights $p^{(t)}$)
 - Obtain classifier $h^{(t)}$ which maps x datapoints to $[0,1]$ (confidence in classification)
 3. Calculate error $\varepsilon_t = \sum_i p_i^{(t)} |h^{(t)}(x_i) - y_i|$ (should be < 0.5 by WeakLearn guarantees)
 4. Define $\beta_t = \frac{\varepsilon_t}{1-\varepsilon_t}$, if $\varepsilon_t \leq 1/2$ set $w_i^{(t+1)} = w_i^{(t)} \beta_t^{1-|h^{(t)}(x_i)-y_i|}$
 - Note: If ε_t big, then β_t is big. Many errors, so don't downweight points!
3. $h(x) = 1$ if $\sum_{t=1}^T \log\left(\frac{1}{\beta_t}\right) h^{(t)}(x) \geq \frac{1}{2} \sum_{t=1}^T \log\left(\frac{1}{\beta_t}\right)$, 0 else

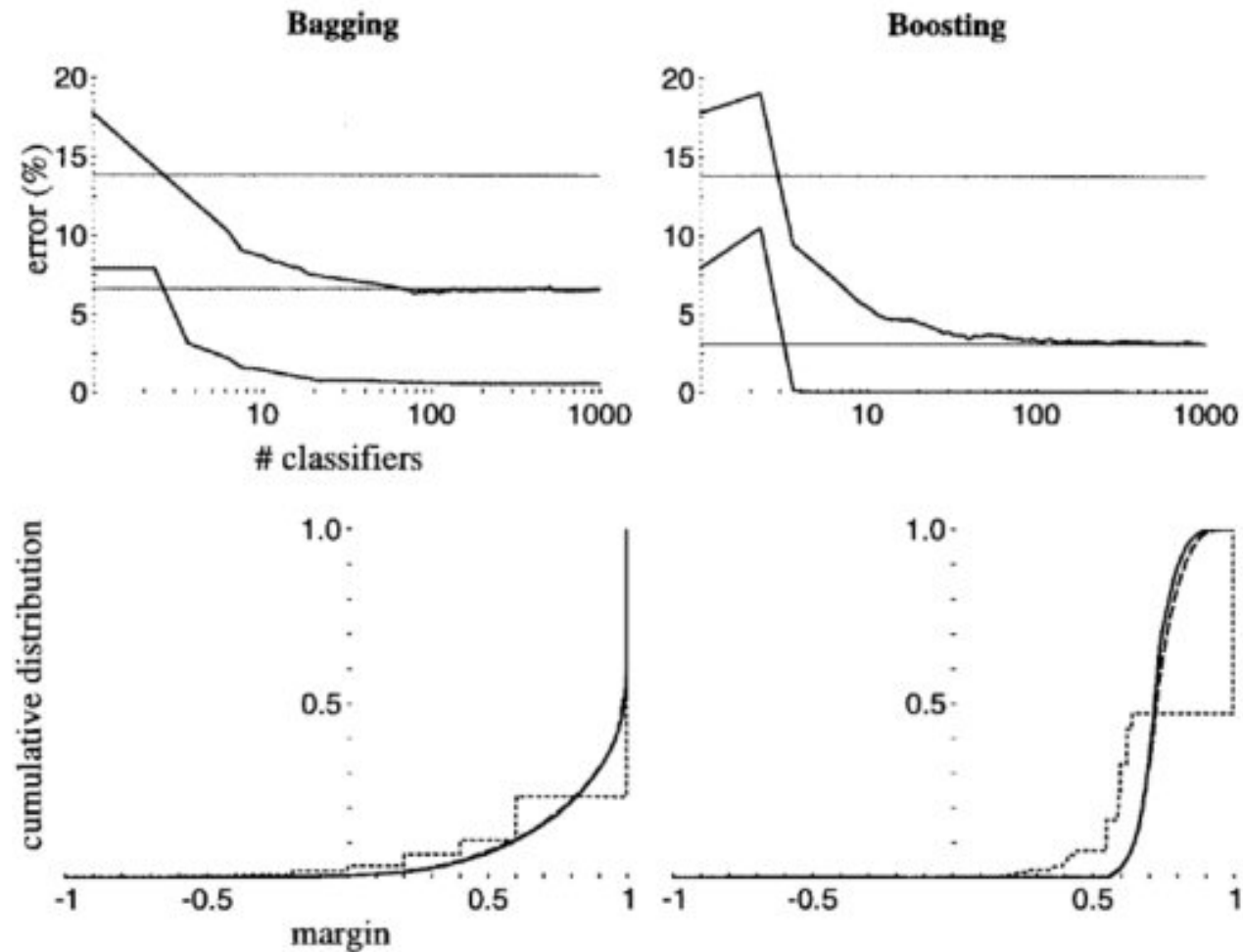
Illustration of AdaBoost



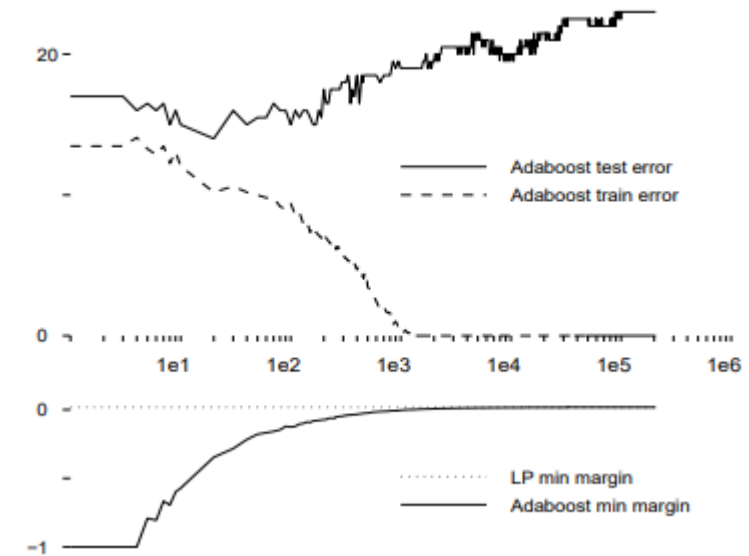
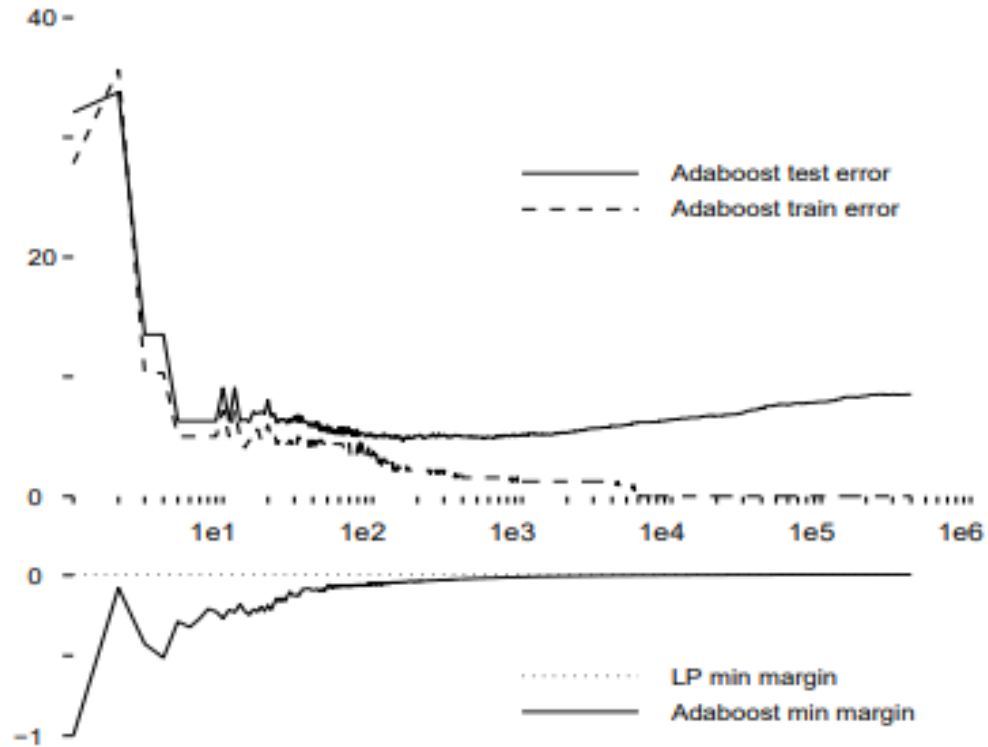
Training Error

- Training error: $\frac{1}{n} \sum_{i=1}^n \mathbf{1}\{h(x_i) \neq y_i\} \leq 2^T \prod_{t=1}^T \sqrt{\varepsilon_t(1 - \varepsilon_t)}$
- Suppose we say $\varepsilon_t \leq \frac{1}{2} - \gamma$ (bound on error of t -th classifier)
- Then training error $\leq \exp(-2T\gamma^2)$ (decreases exponentially fast)
- But we really want good test error...
- An alternate perspective: consider exponential loss $\sum e^{-y_i h(x_i)}$
 - Can be seen as a type of optimization of the exponential loss
- Generalize well when using a simple base classifier

Overfitting with AdaBoost

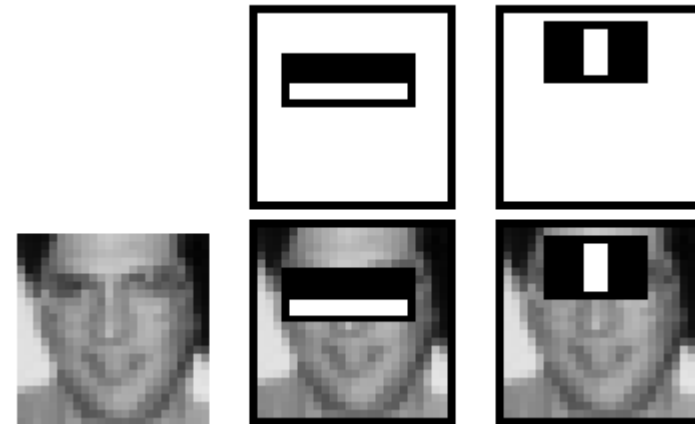


Overfitting with AdaBoost

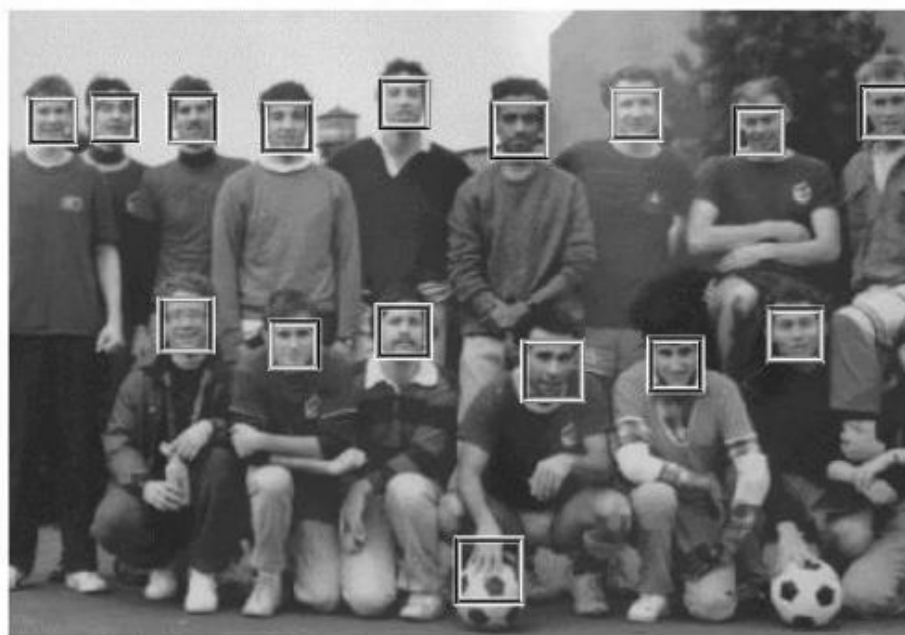


Face detection application (Viola, Jones '01)

- Start with very very simple classifiers
- Use boosting to combine into something better



Face detection application (Viola, Jones '01)



Bagging

Bagging: Bootstrap Aggregating

- Bootstrap sampling + aggregation
- Example: estimate μ given $X_1, \dots, X_n \sim N(\mu, \sigma^2)$
- Simple solution: use empirical mean $\hat{\mu} = \frac{1}{n} \sum X_i$
 - Note $E[\hat{\mu}] = \mu, \text{Var}[\hat{\mu}] = \text{Var}\left[\frac{1}{n} \sum X_i\right] = \frac{1}{n^2} \text{Var}[\sum X_i] = \frac{1}{n^2} \cdot n \cdot \text{Var}[X_1] = \frac{\sigma^2}{n}$.
- Variance may be very large...
- If we have Bn points from $N(\mu, \sigma^2)$, can form $S_1 = \{X_1, \dots, X_n\}, S_2 = \{X_{n+1}, \dots, X_{2n}\}, \dots, S_B = \{X_{(B-1)n+1}, \dots, X_{Bn}\}$
- $\hat{\mu}^{(j)} = \frac{1}{n} \sum_{z \in S_j} z$ and $\hat{\mu}^{(avg)} = \frac{1}{B} \sum_{j \in [B]} \hat{\mu}^{(j)}$
- $E[\hat{\mu}^{(avg)}] = \mu, \text{Var}\left[\frac{1}{B} \sum_{j \in [B]} \hat{\mu}^{(j)}\right] = \frac{1}{B^2} \cdot B \cdot \frac{\sigma^2}{n} = \frac{\sigma^2}{nB}$

Bootstrap Sampling

- Averaging over B independent datasets reduces variance by factor B
 - But needs B times more data...
- Idea: cheat and just reuse parts of the same dataset!
 - Not independent, but still seems to work
- Bootstrap sampling $\approx$ sampling with replacement

Bootstrap Example

- Given dataset of size n , create B datasets of size n , where each is constructed by drawing n samples (with replacement) from original
- Example: Given dataset X_1, X_2, X_3, X_4, X_5
- $S_1 = \{X_3, X_4, X_1, X_1, X_4\}, S_2 = \{X_5, X_5, X_3, X_1, X_2\}, \dots, S_B = \dots$
- Again, use $\hat{\mu}^{(j)} = \frac{1}{n} \sum_{z \in S_j} z$ and $\hat{\mu}^{(avg)} = \frac{1}{B} \sum_{j \in [B]} \hat{\mu}^{(j)}$
- $E[\hat{\mu}^{(avg)}] = \mu, \text{Var} \left[\frac{1}{B} \sum_{j \in [B]} \hat{\mu}^{(j)} \right] = ? ?$
- Can't compute variance as before, since we lost independence
- Intuition: sampling is already a random process. Bootstrap sampling re-simulates that random process and averages over runs.

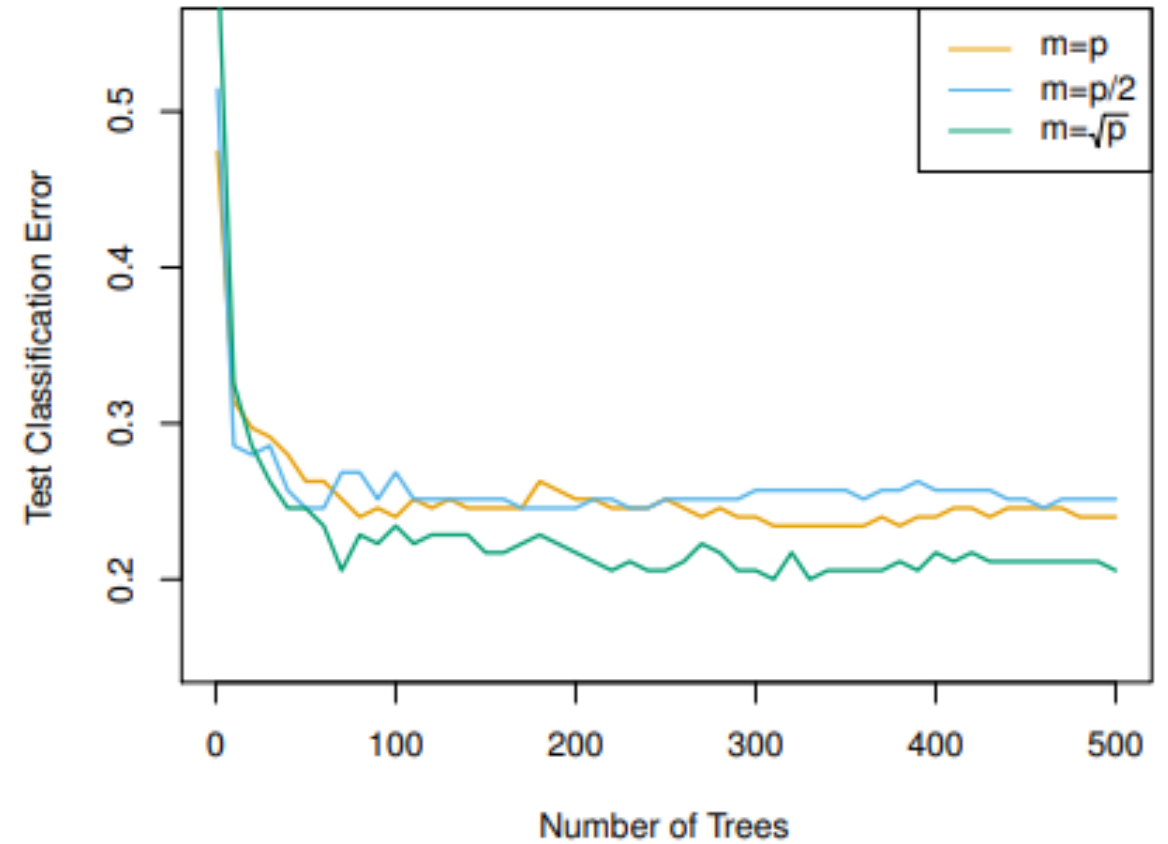
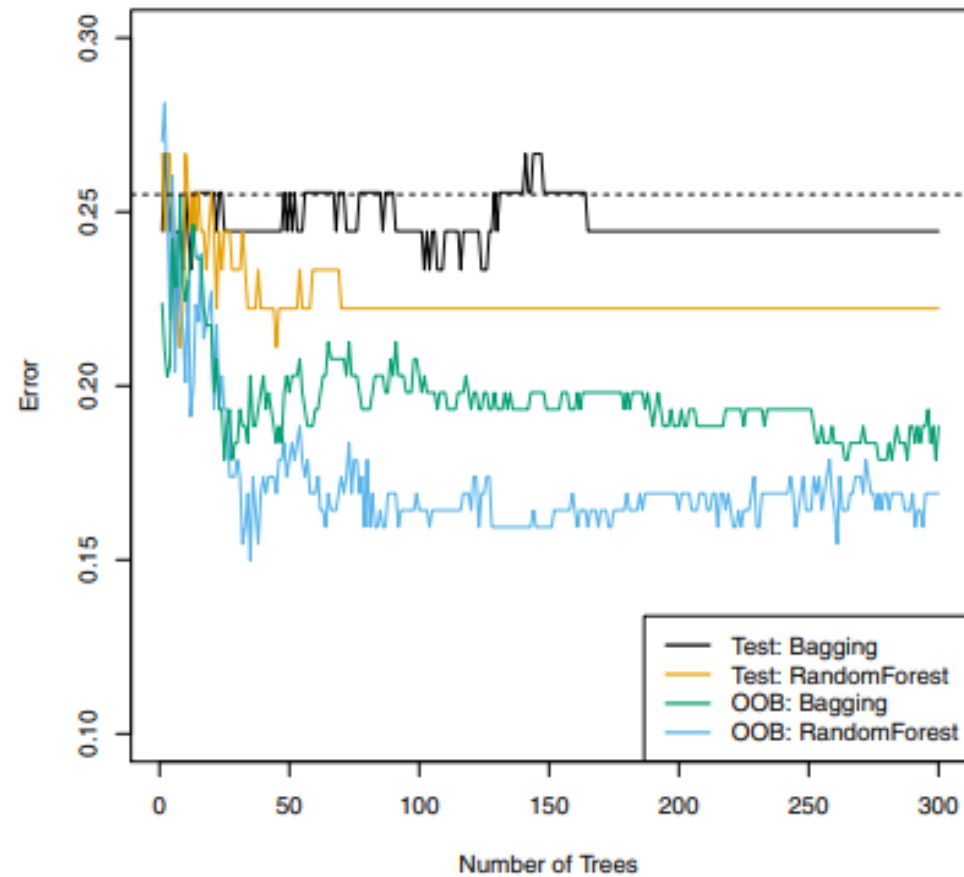
Bagging in ML

- Some methods are inherently high variance
- Decision trees
 - Learn decision tree on 2 halves of the same dataset $\rightarrow$ (very?) different trees
 - DTs choose argmin over splits – small # samples could flip comparable splits
- Use bootstrap aggregating to reduce variance
 1. Bootstrap sample B datasets of size n
 2. Run some learning algorithm on each, get classifiers $\hat{f}^{(1)}, \dots, \hat{f}^{(B)}$
 3. Aggregate $\hat{f}^{(1)}, \dots, \hat{f}^{(B)}$
 - How? Regression $\hat{f}(x) = \frac{1}{B} \sum \hat{f}^{(j)}(x)$.
 - Classification $\hat{f}(x) = \text{majority vote of } \hat{f}^{(j)}(x)$

Random Forests

- Bagging on decision trees
 - Add randomness to reduce correlations between trees' predictions.
- Standard decision trees: When choosing which feature to split on, look at all d features and pick the “best” one
 - Downside: if one feature is very informative, will be used in all B datasets
- Random forests: When choosing which feature to split on, look at a random subsample of $m \ll d$ features and pick the “best” one
 - Say, $m = \sqrt{d}$
 - Resample for each split

Random Forests



Bagging and Boosting

- “Simple” way to improve performance
- Generic, flexible with any base learner
- Loses some interpretability
- Bagging: can be done in parallel
- Boosting: inherently sequential (thus takes lots of time)