On The Robustness of Decision Tree Learning Under Label Noise

https://arxiv.org/abs/1605.06296 and Hastie 2009

Goal: theoretical investigation of decision trees & label noise
↳ show for large samples, split rule under noisy data same for noise free

$S = (X, Y)^n$ ~ ideal noise free iid

S^η where $\tilde{y}_{x_i} = y_{x_i}$ w/ prob $(1-\eta_{x_i})$
$\tilde{y}_{x_i} = -y_{x_i}$ w/ prob η_{x_i}
(noise, true)

symmetric/uniform if $\eta_x = \eta \ \forall x$
- η unknown to learning algo

Split rule & impurity

$$C(f) = G(v) - (q\,G(v_L) + (1-q)G(v_R))$$

say p = pos, $q = (1-p)$ = frac negative
then Gini $= 2pq$
$C_{entropy} = -p\log p - q\log q$

Noise Tolerance:
$\operatorname{argmin}_{f \in F} C(f) = \operatorname{argmin}_{f \in F} C^\eta(f)$

Under symmetric label noise,
$p^\eta = p(1-\eta) + q\eta = p(1-2\eta) + \eta$

Theorem 3: splitting criterion based on gini impurity is noise tolerant if $\eta < 0.5$

Lemma 7: if a leaf node v has n samples, under symmetric label noise w/ $\eta < 0.5$, majority voting will not fail w/ probability at least $1-\delta$ when $n \geq 2/\rho^2(1-2\eta)^2 \cdot \ln(\frac{1}{\delta})$ where ρ is the difference between the fraction of pos & negative samps in the noise free case.

Sample Complexity of RF

For a classifier:
$\text{error}_{gen} = \text{error}_{bias} + \text{error}_{var} + \sigma^2_{noise}$

For a classifier:

$$error_{gen} = error_{bias} + error_{var} + \sigma^2_{noise}$$

Under symmetric label noise, err_{bias} is same for a single decision tree & random forest.
Claim: $error_{gen}$ controlled by $error_{var}$.

If ρ = pairwise correlation and var σ^2 for each tree, then random forest w/ N trees has variance:

$$error_{var} = \rho\sigma^2 + \frac{1-\rho}{N}\sigma$$

Could one say that the different weighting of kernel polytopes decreases the correlation between trees and decreases the variance?

Hastie 2009

Notation:

trees $b=1$ to B

Classification: $\hat{C}^B_{rf}(x) = \text{maj vote}\{\hat{C}_b(x)\}_1^B$

Regression: $\hat{f}^B_{rf}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$

Notes

- expectation of B i.i.d bagged trees is the same for any one
- bias of bagged trees is the same for any. so improvement must be through variance reduction

An avg of B iid vars all w/ variance σ^2 has variance $\frac{1}{B}\sigma^2$. If vars are identically distributed and have pairwise correlation ρ, then variance of the avg is:

$$\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2$$

as $B \to \infty$, second term disappears

∴ reducing the correlation btw trees w/o increasing the variance too much