Random forests and averaging classifiers

Gábor Lugosi

ICREA and Pompeu Fabra University

Barcelona

joint work with

Gérard Biau (Paris 6)
Luc Devroye (McGill, Montreal)

Leo Breiman

Binary classification

$(\mathbf{X}, \mathbf{Y}) \in \mathbb{R}^{\mathbf{d}} \times \{-\mathbf{1}, \mathbf{1}\}$ observation/label pair

$\mathbf{g_n}(\mathbf{X}) = \mathbf{g_n}(\mathbf{X}, \mathbf{D_n}) \in \{\mathbf{0}, \mathbf{1}\}$ classifier, based on

$\mathbf{D_n} = (\mathbf{X_1}, \mathbf{Y_1}), \ldots, (\mathbf{X_n}, \mathbf{Y_n})$, i.i.d. training data, distributed as $(\mathbf{X}, \mathbf{Y})$.

$\mathbf{L}(\mathbf{g_n}) = \mathbb{P}\{\mathbf{g_n}(\mathbf{X}) \neq \mathbf{Y} | \mathbf{D_n}\}$ loss of $\mathbf{g_n}$.

a posteriori probability $\eta(\mathbf{x}) = \mathbb{P}\{\mathbf{Y} = \mathbf{1} | \mathbf{X} = \mathbf{x}\}$.

Bayes classifier: $\mathbf{g}^*(\mathbf{x}) = \mathbb{1}_{\eta(\mathbf{x}) \geq \mathbf{1}/\mathbf{2}}$.

Bayes risk: $\mathbf{L}^* = \mathbf{L}(\mathbf{g}^*)$.

$\{\mathbf{g_n}\}$ is *consistent* if $\mathbf{L}(\mathbf{g_n}) \to \mathbf{L}^*$ in probability.

Local averaging

Historically the first non-parametric classification rules.

Histogram, **k**-nearest neighbor, kernel classifiers.

Fix and Hodges (1951-52),
Cover and Hart (1967),
Glick (1973),
Devroye and Wagner (1976),
Stone (1977),
Gordon and Olshen (1978),
Devroye and Györfi (1983).

Stone's 1977 theorem

Local averaging classifiers:

$$\mathbf{g_n(x) = 1} \quad \text{iff} \quad \sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{Y_i W_{ni}(x) \geq 0}$$

where $\mathbf{W_{ni}(x) = W_{ni}(x, X_1, \ldots, X_n) \geq 0}$

and $\sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{W_{ni}(x) = 1}$.

Stone's 1977 theorem

Consistency holds if

(i) $\lim_{n\to\infty} \mathbb{E}\left\{\max_{1\le i\le n} W_{ni}(X)\right\} = 0.$

(ii) For all $a > 0$,

$$\lim_{n\to\infty} \mathbb{E}\left\{\sum_{i=1}^{n} W_{ni}(X)\mathbb{1}_{\|X_i - X\| > a}\right\} = 0.$$

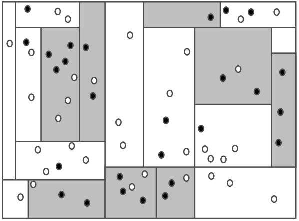
(iii) There is a $c > 0$ such that, for every $f \ge 0$,

$$\mathbb{E}\left\{\sum_{i=1}^{n} W_{ni}(X) f(X_i)\right\} \le c\mathbb{E}f(X).$$

Tree classifiers

Histograms based on data-dependent partitions.

Partition is constructed by recursive splitting.



See Breiman, Freedman, Olshen, and Stone (1984), Devroye, Györfi, and Lugosi (1996) for surveys.

Consistency of tree classifiers

Many versions suggested in the literature are inconsistent.

General consistency theorems:

Assume the partition depends on $\mathbf{X_1, \ldots, X_n}$ only. Let $\mathbf{A(X)}$ denote the cell containing $\mathbf{X}$ and $\mathbf{N(X) = \sum_{i=1}^{n} \mathbb{1}_{X_i \in A(X)}}$.

If $\mathbf{diam(A(X)) \to 0}$ and $\mathbf{N(X) \to \infty}$ in probability then the classifier is consistent (Devroye, Györfi, and Lugosi, 1996).

For general partitions consistency holds under an additional combinatorial condition (Lugosi and Nobel, 1993).

Random forests

Tree classifiers are unstable.

Breiman (2001) suggests “bootstrap” randomization in building trees:

(1) choose a cell at random
(2) choose $\mathbf{m} < \mathbf{d}$ coordinates at random
(3) cut at a point (and direction) giving the largest decrease in empirical error.

Repeat until every cell is pure.

Repeat the random tree classifier a zillion times and take majority vote.

Additional randomization is achieved by bootstrap sampling.

Simpler version

Breiman asked if this classifier was consistent. Performs well in practice.

Simpler version:

(1) choose a cell at random
(2) choose a coordinate at random
(3) cut at a random point.

Repeat $\mathbf{k}$ times.

Repeat the random tree classifier a zillion times and take majority vote.

Local averaging rule with weights
$\mathbf{W_{ni}(X)} \sim \mathbb{P}_{\mathbf{Z}}\{\mathbf{X}, \mathbf{X_i} \text{ are in same cell}\}$.

Averaged classifiers

Let $\mathbf{g_n(X, Z, D_n) = g_n(X, Z)}$ be a randomized classifier.

Probability of error:

$$\mathbf{L(g_n) = \mathbb{P}_{(X,Y),Z}\{g_n(X, Z, D_n) \neq Y\}} \ .$$

Averaged classifier: $\mathbf{\overline{g}_n(x) = \mathbb{1}_{\mathbb{E}_Z g_n(x,Z) \geq 1/2}}$

Main lemma:

If $\mathbf{g_n}$ is consistent then $\mathbf{\overline{g}_n}$ is also consistent.

Averaging "stabilizes."

Consistency of simple version

We obtain consistency without computing the weights $\mathbf{W_{ni}(X)}$.

Assume $\mathbf{X}$ is supported in $[0,1]^d$.

Then $\overline{g}_n$ is consistent whenever $k \longrightarrow \infty$ and $k/n \longrightarrow 0$ as $k \longrightarrow \infty$.

It suffices to prove consistency of the randomized “base” classifier.

It is enough to show $\mathrm{diam}(A(X,Z)) \to 0$ and $N(X,Z) \to \infty$ in probability.

Consistency of simple version

$\mathbf{N(X, Z) \to \infty}$ and $\mathbf{diam(A(X, Z)) \to 0}$, in probability, are both easy to show.

Interestingly, for $\mathbf{d > 1}$, $\mathbf{\sup_x diam(A(x, Z)) \not\to 0}$.

If $\mathbf{d \geq 3}$, the number of cells with diameter $\mathbf{1}$ (in sup norm) is a supercritical branching process.

A scale invariant version

(1) choose a cell at random
(2) choose a coordinate at random
(3) cut at a random data point.

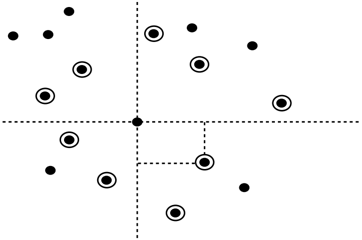
Repeat $\mathbf{k}$ times.

Repeat the random tree classifier a zillion times and take majority vote.

If the distribution of $\mathbf{X}$ has non-atomic marginals in $\mathbb{R}^{\mathbf{d}}$, then $\overline{\mathbf{g}}_{\mathbf{n}}$ is consistent whenever $\mathbf{k} \longrightarrow \infty$ and $\mathbf{k}/\mathbf{n} \longrightarrow \mathbf{0}$ as $\mathbf{k} \longrightarrow \infty$.

Breiman's original random forest

Lin and Jeon (2006) point out that any random forest classifier that cuts down to pure cells is a weighted *layered* nearest neighbor rule.



No such rule can be consistent if the distribution of $\mathbf{X}$ is concentrated on a diagonal.

Randomizing inconsistent classifiers

Averaging consistent randomized classifiers preserves consistency.

The converse is not true: averaging inconsistent classifiers may lead to consistency.

This may be the case with Breiman's original random forest if $\mathbf{X}$ has a density.

We work out a stylized example.

A randomized nearest neighbor rule

For $\mathbf{x} \in \mathbb{R}$, let $\mathbf{X_{(1)}(x), X_{(2)}(x), \ldots, X_{(n)}(x)}$ be $\mathbf{X_1, \ldots, X_n}$ ordered according to distances to $\mathbf{x}$.

Let $\mathbf{U_1, \ldots, U_n}$ be i.i.d. uniform $\mathbf{[0, m]}$.

Let $\mathbf{g_n(x, Z) = Y_{(i)}(x)}$ if and only if

$$\mathbf{\max(i, U_i) \le \max(j, U_j)} \quad \text{for } \mathbf{j = 1, \ldots, n}$$

$\mathbf{X_{(i)}(x)}$ is the *perturbed nearest neighbor* of $\mathbf{x}$.

$\mathbf{\overline{g}_n(x) = \mathbb{1}_{\mathbb{E}_Z g_n(x,Z) \ge 1/2}}$ is the *averaged perturbed nearest neighbor classifier*.

Consistency

The averaged perturbed nearest neighbor classifier is consistent if $\mathbf{m} \longrightarrow \infty$ and $\mathbf{m}/\mathbf{n} \longrightarrow \mathbf{0}$.

Proof: $\overline{\mathbf{g}}_{\mathbf{n}}$ is a local averaging classifier with

$$\begin{aligned} \mathbf{W}_{\mathbf{ni}}(\mathbf{x}) &= \mathbb{P}_{\mathbf{Z}}\{\mathbf{X}_{(\mathbf{i})}(\mathbf{x}) \text{ is the perturbed nearest neighbor of } \mathbf{x}\} \\ &= \cdots \text{can be written explicitly} \end{aligned}$$

Stone's theorem may be used.

Bagging

In bagging, suggested by Breiman (1996), bootstrap samples are generated from the original data set.

Let $q_n \in [0, 1]$. In a bootstrap sample $D_n(Z)$ each (X_i, Y_i) is present with probability q_n.

Given a classifiers $\{g_n\}$, let

$$g_n(X, Z, D_n) = g_N(X, D_n(Z)) \ ,$$

By drawing many bootstrap samples, one obtains the averaged classifier $\overline{g}_n(x, D_n) = \mathbb{1}_{\mathbb{E}_Z g_N(x, D_n(Z)) \geq 1/2}$.

If $nq_n \to \infty$ as $n \to \infty$ then the bagging classifier is consistent.

Bagging the 1-NN classifier

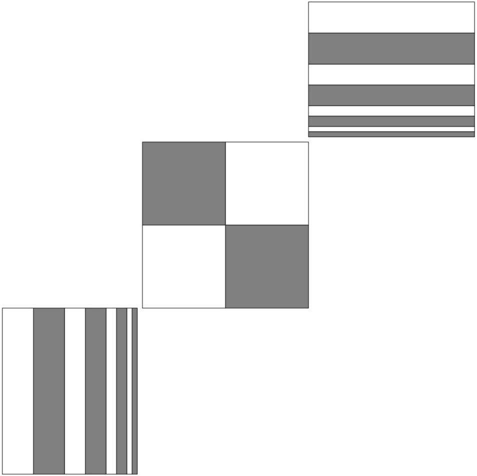
It may help to choose much smaller values of $\mathbf{q_n}$.

The $\mathbf{1}$-nearest neighbor rule is not consistent unless either $\mathbf{L^* = 0}$ or $\mathbf{L^* = 1/2}$.

However, the bagging averaged $\mathbf{1}$-nearest neighbor classifier is consistent for all distributions of $\mathbf{(X, Y)}$ if and only if $\mathbf{q_n} \longrightarrow \mathbf{0}$ and $\mathbf{nq_n} \longrightarrow \infty$.

Greedy trees

Greedy trees like Breiman's may be inconsistent for another reason:



Questions

Is Breiman's original random forest consistent if $\mathbf{X}$ has a density?

In what situations does randomizing and averaging help?

Random forests