Mathematics of Random Forests

1. Probability: Chebyshev inequality

Theorem 1 (Chebyshev inequality): *If X is a random variable with standard deviation σ and mean μ, then for any $\epsilon > 0$,*

$$P(|X - \mu| > \epsilon) \leq \frac{\sigma^2}{\epsilon^2}.$$

Theorem 2 (Bounded convergence theorem): *Given a sequence $h_1(\mathbf{x}), h_2(\mathbf{x}), \ldots$ of fns. with $h_k(\mathbf{x}) \leq M$ for fixed $M > 0$) defined on a space S of finite measure, then*

$$\lim_{k\to\infty} \int_S d\mathbf{x}\; h_k(\mathbf{x}) \underset{k\to\infty}{\to} \int_S d\mathbf{x}\; \lim_{k\to\infty} h_k(\mathbf{x}),$$

i.e., the limit and integration can be interchanged (assuming the limits exist).

Probability background

Recall:

Def. 1 (Indicator function): For any event $A \subset \Omega$ of the sample space, define the *indicator function* (also known as *characteristic function*) of A to be

$$I_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{otherwise} \end{cases} = \begin{cases} 1 & \text{if } A \text{ occurs} \\ 0 & \text{otherwise} \end{cases}.$$

2. Classification trees

Assume we have n patients (samples), and (e.g., mass spectroscopy) feature vectors $\{\mathbf{x}_i\}_{i=1}^n$ with outcomes y_i.

Data:

$$D = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}.$$

Each feature vector

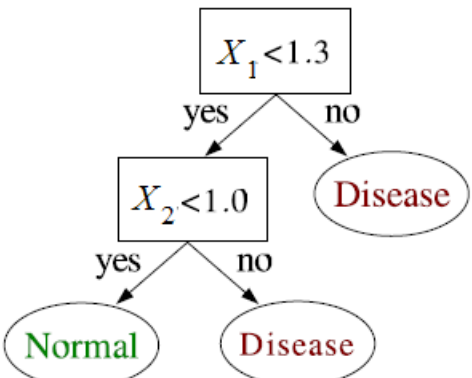
$$\mathbf{x}_k = (x_{k1}, \ldots, x_{kd}).$$

Classification trees

Formal definitions:

Definition 2: A *classification tree* is a decision tree in which each node has a *binary* decision based on whether $x_i < a$ or not for a fixed a (can depend on node).

Classification trees



Classification trees

The top node contains all of the examples $(\mathbf{x}_k, y_k)$, and the set of examples is subdivided among the children of each node according to the classification at that node.

The subdivision of examples continues until every node at the bottom has examples which are in one class only.

At each node, feature x_i and threshold a are chosen to minimize resulting 'diversity' in the children nodes. This diversity is often measured by *Gini criterion*, see below.

Gini criterion

The subdivision continues until every node at the bottom has only one class (disease or normal) in it, assigned as a prediction to input **x**.

Gini Criterion: Define class $C_1 =$ disease; $C_2 =$ normal. How do we measure variation of samples in a node with respect to these two classes?

Suppose there are 2 classes C_1, C_2 and we have examples in set S at our current node.

Now to create child nodes, partition $S = S_1 \cup S_2$.

Gini criterion

(Note each sample S_1, S_2 is partitioned into the two classes C_1, C_2)

Recall $|S| = \#$ objects in set S

Define

$$\widehat{P}(S_j) = \frac{|S_j|}{|S|} = \text{proportion of } S_j \text{ in } S$$

$$\widehat{P}(C_i|S_j) = \frac{|S_j \cap C_i|}{|S_j|} = \text{proportion of } S_j \text{ which is in } C_i.$$

Gini criterion

Define the *variation* $g(S_j)$ in set S_j to be:

$$g(S_j) = \sum_{i=1}^{2} \widehat{P}(C_i|S_j)(1 - \widehat{P}(C_i|S_j)),$$

Note: variation $g(S_j)$ is largest if set S_j is equally divided among C_i. It's smallest when all of S_j is just *one* of the C_i.

We define the variation of this full subdivision of the S_j to be the *Gini index* $= G$ if:

Gini criterion

$$G = \widehat{P}(S_1)g(S_1) + \widehat{P}(S_2)g(S_2)$$

$=$ weighted sum of variations $g(S_1),\ g(S_2)$

3. Random vectors

A *random vector*

$$\mathbf{X} = (X_1, \ldots, X_d)$$

is an array of random variables defined on the same probability space.

Given **X** as above define its distribution (or *joint distribution* of $X_1, \ldots, X_d$) to be measure μ on $\mathbb{R}^d$ defined by

$$\mu(A) \equiv P(\mathbf{X} \in A),$$

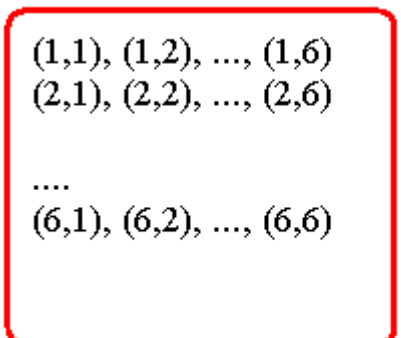
for any $A \in \mathbb{R}^d$ which is measurable.

Random vectors

Ex. 2: Consider rolling 2 dice. Let X_1 be the number on the first die, X_2 the number on the second die. Then the probability space is

$$S = \{\text{all ordered pairs of die rolls}\} =$$

Random vectors



$X_1 =$ first roll; $X_2 =$ second roll,

Random vectors

i.e., if $\omega \in S$ is given by $\omega = (3, 4)$, then

$$X_1(s) = 3; \qquad X_2(\omega) = 4.$$

The random vector (X_1, X_2) satisfies

$$(X_1, X_2)(\omega) = (3, 4).$$

Random vectors

Ex. 3: Let $\mathbf{x} = (x_1, \ldots, x_d)$ be a microarray of a glioma cancer sample in its initial stages. Then each feature x_i is a random variable with some distribution.

For fixed i, let X_i be a model random variable whose probability distribution is the same as the x_i numbers in the microarray.

Then the model random vector $\mathbf{X} = (X_1, \ldots, X_d)$ has a joint distribution which is the same as our microarray samples $\mathbf{x} = (x_1, x_2, \ldots, x_d)$.

Random vectors

For a microarray **x**, let y be the classification of the cancer ($y = +1$ means malignant; $y = -1$ means benign). Then we have another random vector

$$(x_1, \ldots, x_d, y) = (\mathbf{x}, y),$$

with the same distribution as a model vector $(\mathbf{X}, Y)$.

If the distribution of the random vector $(\mathbf{x}, y)$ is given by the model random vector $(\mathbf{X}, Y)$, we write

$$(\mathbf{x}, y) \sim (\mathbf{X}, Y).$$

4. Random forest: formal definition

Assume training set of microarrays

$$D = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$$

drawn randomly from a (possibly unknown) probability distribution $(\mathbf{x}_i, y_i) \sim (\mathbf{X}, Y)$.

Goal: to build a classifier which predicts y from **x** based on the data set of examples D.

Given: ensemble of (possibly weak) classifiers $h = \{h_1(\mathbf{x}), \ldots, h_K(\mathbf{x})\}$.

Random forest: formal definition

If each $h_k(\mathbf{x})$ is a decision tree, then the ensemble is a random forest. We define the parameters of the decision tree for classifier $h_k(\mathbf{x})$ to be $\Theta_k = (\theta_{k1}, \theta_{k2}, \ldots, \theta_{kp})$

(these parameters include the structure of tree, which variables are split in which node, etc.)

Random forest: formal definition

We sometimes write

$$h_k(\mathbf{x}) = h(\mathbf{x}|\, \Theta_k\}.$$

Thus decision tree k leads to a classifier $h_k(\mathbf{x}) = h(\mathbf{x}|\Theta_k)$.

How do we choose which features appear in which nodes of the k^{th} tree? At random, according to parameters Θ_k, which are randomly chosen from a model variable Θ.

Random forest: formal definition

Definition 1. A *random forest* is a classifier based on a family of classifiers $h(\mathbf{x}|\Theta_1), \ldots, h(\mathbf{x}|\Theta_K)$ based on a classification tree with parameters Θ_k randomly chosen from a model random vector Θ.

For the final classification $f(\mathbf{x})$ (which combines the classifiers $\{h_k(\mathbf{x})\}$), each tree casts a vote for the most popular class at input $\mathbf{x}$, and the class with the most votes wins.

Specifically given data $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$: we train a family of classifiers $h_k(\mathbf{x})$.

Random forest: formal definition

Each classifier $h_k(\mathbf{x}) \equiv h(\mathbf{x}|\Theta_k)$ is in our case a predictor of n

$$y = \pm 1 = \text{outcome associated with input } \mathbf{x}.$$

Examples

Example 4: $\Theta =$ parameter of a tree determines a random subset D_Θ of the full data vector D, i.e., we only choose a sub-collection of feature vectors $\mathbf{x} = (x_1, \ldots, x_d)$.

So D_Θ is a subset of

$$D = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\} = \text{full data set.}$$

Thus parameter Θ_k (for classification tree k) determines which subset of full data set D we choose for the tree $h_k(\mathbf{x}) = h(\mathbf{x}|\Theta_k)$.

Examples

Then the ensemble of classifiers (now an RF) consists of trees, each of which sees a different subset of the data.

Example 3: Θ determines subset $\mathbf{x}_\Theta$ of the full set of features $\mathbf{x} = (x_1, \ldots, x_d)$. Then

$h(\mathbf{x}|\Theta_k) =$ classification tree using subset $\mathbf{x}_\Theta$ of entries of full feature vector $\mathbf{x}$

[dimension reduction]

Examples

In data mining (where dimension d is very high) this situation is common.

5. General ensemble methods - properties

Given a *fixed* ensemble

$$h = (h_1(\mathbf{x}), \ldots, h_K(\mathbf{x}))$$

of classifiers with random data vector $(\mathbf{x}, y)$:

If A is any outcome for a classifier $h_k(\mathbf{x})$ in the ensemble, we define

$\widehat{P}(A) =$ proportion of classifiers h_k $(1 \leq k \leq K)$ for which event A occurs

$=$ *empirical probability* of A.

Ensemble methods: definitions

Define empirical *margin function* by:

$$\widehat{m}(\mathbf{x}, y) \equiv \widehat{P}_k(h_k(\mathbf{x}) = y) - \max_{j \neq y} \widehat{P}_k(h_k(\mathbf{x}) = j), \qquad (1)$$

$=$ *average margin* of the ensemble of classifiers

$=$ extent to which average number of votes for correct class exceeds the average number of votes for the next-best class

Ensemble methods: definitions

$\approx$ confidence in the classifier.

Definition 2: The *generalization error* of the classifier ensemble h is

$$e = P_{\mathbf{x},y}(\widehat{m}(\mathbf{x}, y) < 0).$$

[subscript $\mathbf{x}, y$ indicates prob. measured in $\mathbf{x}$, y space, i.e., $(\mathbf{x}, y)$ is viewed as the random variable].

Ensemble methods: definitions

Theorem 1: *As $K \to \infty$ (i.e., as the number of trees increases),*

$$e \underset{K \to \infty}{\longrightarrow} P_{\mathbf{x},y}\left[P_\Theta(h(\mathbf{x},\Theta) = y) - \max_{j \neq y} P_\Theta(h(\mathbf{x},\Theta) = j) < 0\right] \tag{2}$$

note $P_{\mathbf{x},y}$ denotes prob. as $\mathbf{x}, y$ varies - similarly for P_Θ)

Ensemble methods: definitions

Proof: Note

$$\widehat{P}_k(h_k(\mathbf{x}) = j) = E_k[I(h_k(\mathbf{x}) = j)] \equiv \frac{1}{K}\sum_{k=1}^{K} I[h_k(\mathbf{x}) = j]$$

where E_k = average over k; $h_k(\mathbf{x}) \equiv h(\mathbf{x}|\Theta_k)$.

Recall $I(A) = 1$ if A occurs and 0 otherwise.

Claim for our random sequence $\Theta_1, \Theta_2, \ldots$, and $\forall$ data vectors **x**, it suffices to show

Ensemble methods: definitions

$$\frac{1}{K}\sum_{k=1}^{K} I[h(\mathbf{x}|\Theta_k) = j] \rightarrow P_{\Theta}(h(\mathbf{x}|\Theta) = j). \qquad (3)$$

Why? Let $g_K(\mathbf{x}, y) \equiv$ RHS of (1), $g(\mathbf{x}, y) =$ quantity in bracket of (2)

Then note for each $\mathbf{x}, y$ if (3) holds we would have

$$g_K(\mathbf{x}, y) \rightarrow g(\mathbf{x}, y). \qquad (4)$$

Note also

Ensemble methods: definitions

$$P_{\mathbf{x},y}(g_K(\mathbf{x},y) < 0) = E_{\mathbf{x},y}[I(g_K(\mathbf{x},y) < 0)],$$

so by bounded convergence theorem and (4)

$$P_{\mathbf{x},y}(g_K(\mathbf{x},y) < 0) \underset{K \to \infty}{\longrightarrow} P(g(\mathbf{x},y) < 0).$$

thus proving theorem. Thus we must only prove (3).

To prove (3): for fixed training set **x** and tree with parameter Θ, set of **x** with $h(\mathbf{x}|\Theta) = j$ is a union of boxes

Ensemble methods: definitions

$$B \equiv I_1 \times \ldots \times I_d = \{\mathbf{x} | x_i \in I_i\}$$

for fixed collection of intervals $\{I_i\}_{i=1}^d$.

Assuming a finite number of models Θ for $h(\mathbf{x}|\Theta)$ (e.g., finite number of sample subsets, finite number of feature space subsets).

Then $\exists$ a finite number K of such unions of boxes, call them $\{S_k\}_{k=1}^K$.

Define

Ensemble methods: definitions

$$\phi(\Theta) = k \text{ if } \{\mathbf{x} : h(\mathbf{x}|\Theta) = j\} = S_k$$

Let

$$N_k = \#\text{ times } \phi(\Theta_m) = S_k.$$

Then

$$\frac{1}{M}\sum_{m=1}^{M} I(h(\mathbf{x}|\Theta_m) = j) = \frac{1}{M}\sum_{k} N_k I(\mathbf{x} \in S_k).$$

By the law of large numbers,

Ensemble methods: definitions

$$N_k = \frac{1}{M}\sum_{m=1}^{M} I(\phi(\Theta_m) = k)$$

converges with probability 1 to

$$E_\Theta[I(\phi(\Theta) = k)] = P_\Theta(\phi(\Theta) = k).$$

Thus

$$\frac{1}{M}\sum_{m=1}^{M} I[h(\mathbf{x}|\Theta_m) = j] \rightarrow \sum_k P_\Theta(\phi(\Theta) = k) I(\mathbf{x} \in S_k)$$

$$= P_{\Theta}(h(\Theta, \mathbf{x}) = j),$$

proving (3) and completing proof.

6. Random forests as ensembles

Instead of fixed ensemble $\{h_k(\mathbf{x})\}_{k=1}^{K}$ of classifiers, consider RF model:

We have $h(\mathbf{x}|\Theta)$; Θ specifies classification tree classifier $h(\mathbf{x}|\Theta)$

We have a fixed (known) probability distribution for Θ determining variety of trees

Random forests as ensembles

Definition 3: The *margin function* of an RF is:

$$m(\mathbf{x}, y) = P_\Theta(h(\mathbf{x}|\Theta) = y) - \max_{j \neq Y} P_\Theta(h(\mathbf{x}|\Theta) = j).$$

The *strength* of the forest (or any family of classifiers) is

$$s = E_{\mathbf{x},y} m(\mathbf{x}, y).$$

The *generalization error* is (Chebyshev inequality)

$$e = P_{\mathbf{x},y}(m(\mathbf{x}, y) < 0) \leq P_{\mathbf{x},y}(|m(\mathbf{x}, y) - s| \geq s) \leq \frac{V(m)}{s^2},$$

giving a (weak) bound.

Random forests as ensembles

Better bounds can be obtained in

Breiman (2001), Random Forests, in *Machine Learning.*

Random forests as ensembles

7. Some sample applications (Breiman):

Some performance statistics on standard databases:

Forest random input (random feature selection) - single feature at a time (percentage error rates)

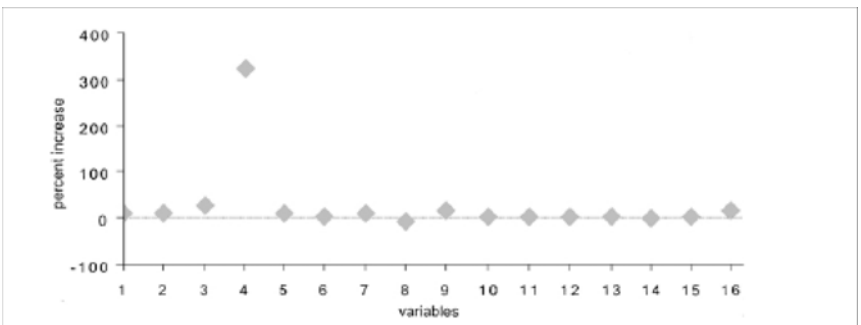
Random forests as ensembles

Data set	Adaboost	Selection	Forest-RI single input	One tre
Glass	22.0	20.6	21.2	36.9
Breast cancer	3.2	2.9	2.7	6.3
Diabetes	26.6	24.2	24.3	33.1
Sonar	15.6	15.9	18.0	31.7
Vowel	4.1	3.4	3.3	30.4
Ionosphere	6.4	7.1	7.5	12.7
Vehicle	23.2	25.8	26.4	33.1
German credit	23.5	24.4	26.2	33.3
Image	1.6	2.1	2.7	6.4
Ecoli	14.8	12.8	13.0	24.5
Votes	4.8	4.1	4.6	7.4
Liver	30.7	25.1	24.7	40.6
Letters	3.4	3.5	4.7	19.8
Sat-images	8.8	8.6	10.5	17.2
Zip-code	6.2	6.3	7.8	20.6
Waveform	17.8	17.2	17.3	34.0
Twonorm	4.9	3.9	3.9	24.7
Threenorm	18.8	17.5	17.5	38.4
Ringnorm	6.9	4.9	4.9	25.7

Breiman, 2001

Random forests as ensembles

Variable importance - determined by accuracy decrease with noise in variable:



Breiman, 2001