Bagging

24 October 2018 11:36

Bagging is a type of ensemble learning

$\underset{\sim}{D}$ m classifiers   Committee members

C_1 C_2 C_3 - - - - - C_m

$C_1(x) = \hat{y}_1$ $C_2(x) = \hat{y}_2$ . . . . $C_m(x) = \hat{y}_m$ $\vec{x} \in \mathbb{R}^d$

Predicted label $=$ majority$(\hat{y}_1, \cdots \hat{y}_m) = \hat{y}_B$

Let $p =$ probability that a classifier gives correct label on a x

Let $C_1, \dots C_m$ be independent of each other.

What is the probability that $\hat{y}_B$ is correct?

$$= \sum^{m} \binom{m}{k} p^{k} (1-p)^{m-k} \underset{?}{>} p$$

$k = \frac{m}{2}$

For binary classifiers, this is true when

$p > 0.5$

Two methods of creating classifier committees

- Bootstrap sampling
- Random forests (attribute subsetting)

Bootstrap sampling

$$D = \{(x^1, y^1) \ldots (x^N, y^N)\}$$

Convert this into a point distribution

$$P_D(x,y) = \frac{1}{N} \text{ if } (x,y) \in D$$

$$= 0 \text{ otherwise.}$$

To get one bootstrap sample S of size N from P_D, do this

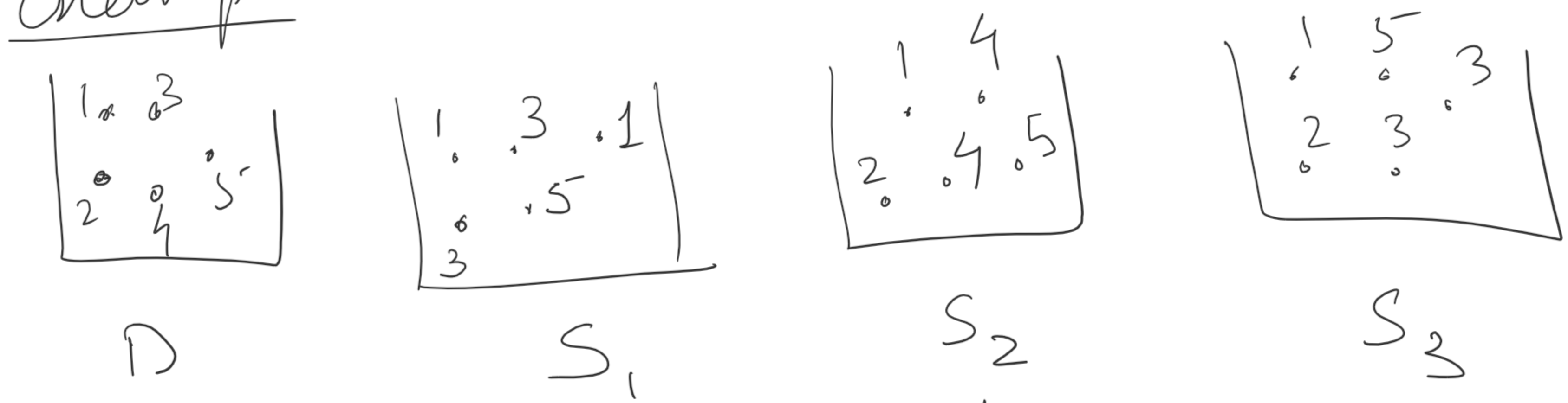
for $i = 1$ to N {

example from D uniformly

select an ... randomly.
add to S.

3

Example:



$B = 3$

$|D| = N$

$C_1 \quad C_2 \quad C_3$

Q: What is the expected number of distinct examples in a bootstrap sample?

A: Expected # of distinct examples = $N *$ probability that an example is selected in any of N trials.

$$\boxed{N * \left(1 - \left(1 - \frac{1}{N}\right)^N\right)}$$

$$\lim_{N \to \infty} \left(1 - \frac{1}{N}\right)^N \to \frac{1}{e}$$

$$\left(1 - \frac{1}{e}\right) \approx 0.63$$

N * [illegible] in any of N trials.

$$\boxed{N * \left(1 - \left(1 - \frac{1}{N}\right)^N\right)}$$

$$\lim_{N \to \infty} \left(1 - \frac{1}{N}\right)^N \longrightarrow \frac{1}{e} \qquad \left(1 - \frac{1}{e}\right) \approx 0.63$$

Bagging is often useful in settings where each classifier is over-fitted on their bootstrap sample. For ex: create a decision tree without pruning.

Bagging smooths the predictions from multiple classifiers and prevents the ill-effects of over-fitting.

II Random Forests

Sample attributes instead of examples:

d attributes in data $X \in R^d$

For each decision tree:

for each node of the tree.

select the best attributes from a randomly sampled set of $\sqrt{d}$ attributes from d.