17.9.1

Bootstrap Aggregation

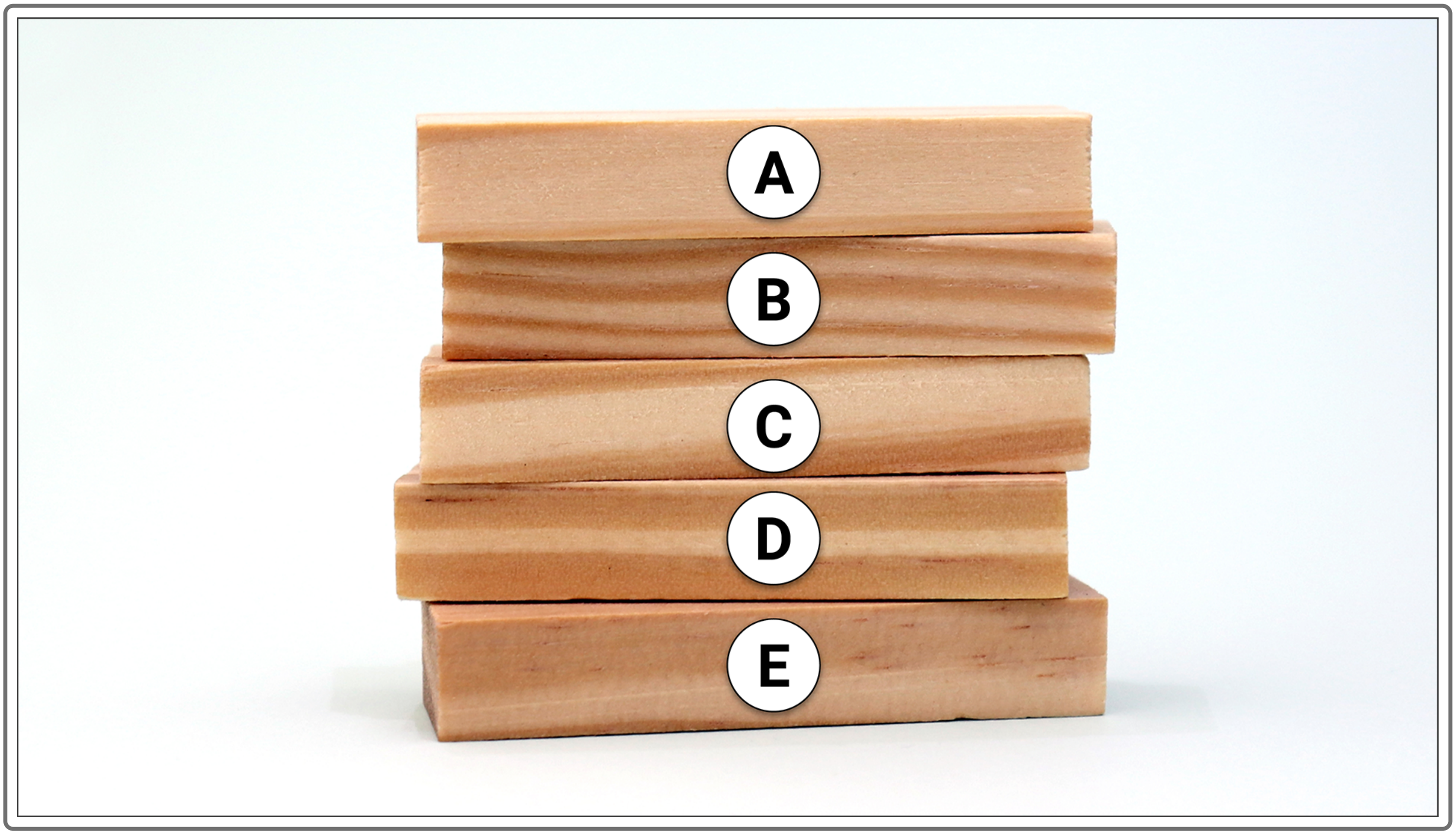
Bootstrap aggregation, or "bagging," is an ensemble learning technique that combines weak learners into a strong learner. In fact, you have already seen a model that uses bootstrap aggregation as part of its algorithm: the random forest model.

Jill reminds you that decision trees are prone to overfitting, meaning that the algorithm's predictions are excessively tailored to the specific dataset. When there's overfitting, a model's performance will suffer when it encounters a new dataset. One way to try to overcome this problem is with bootstrap aggregation. Let's look at how it works in more detail.

Bootstrap aggregation, also called bagging, is a machine learning technique used to combine weak learners into a strong learner. Bagging is composed of two parts: bootstrapping and aggregation.

Bootstrapping

Bootstrapping is a sampling technique in which samples are randomly selected, then returned to the general pool and replaced, or put back into the general pool. As a concrete example, picture your dataset as a bag containing five wooden blocks, each labeled with the letter A, B, C, D, or E.



Imagine that you draw samples of the dataset from this bag three times. Since each sampling is the same size as the original dataset, you must draw five blocks for each sample. To do so, you grab a wooden block randomly from the bag, and after noting which block you drew, you replace it, meaning that you put it back into the bag. Because you return the block to the bag, it's possible to draw the same block again in the next draw. You repeat the process until you have a sample whose size is the same as the original dataset. The result might appear as follows:

Sample 1: A, A, A, B, D

Sample 2: A, B, B, C, E

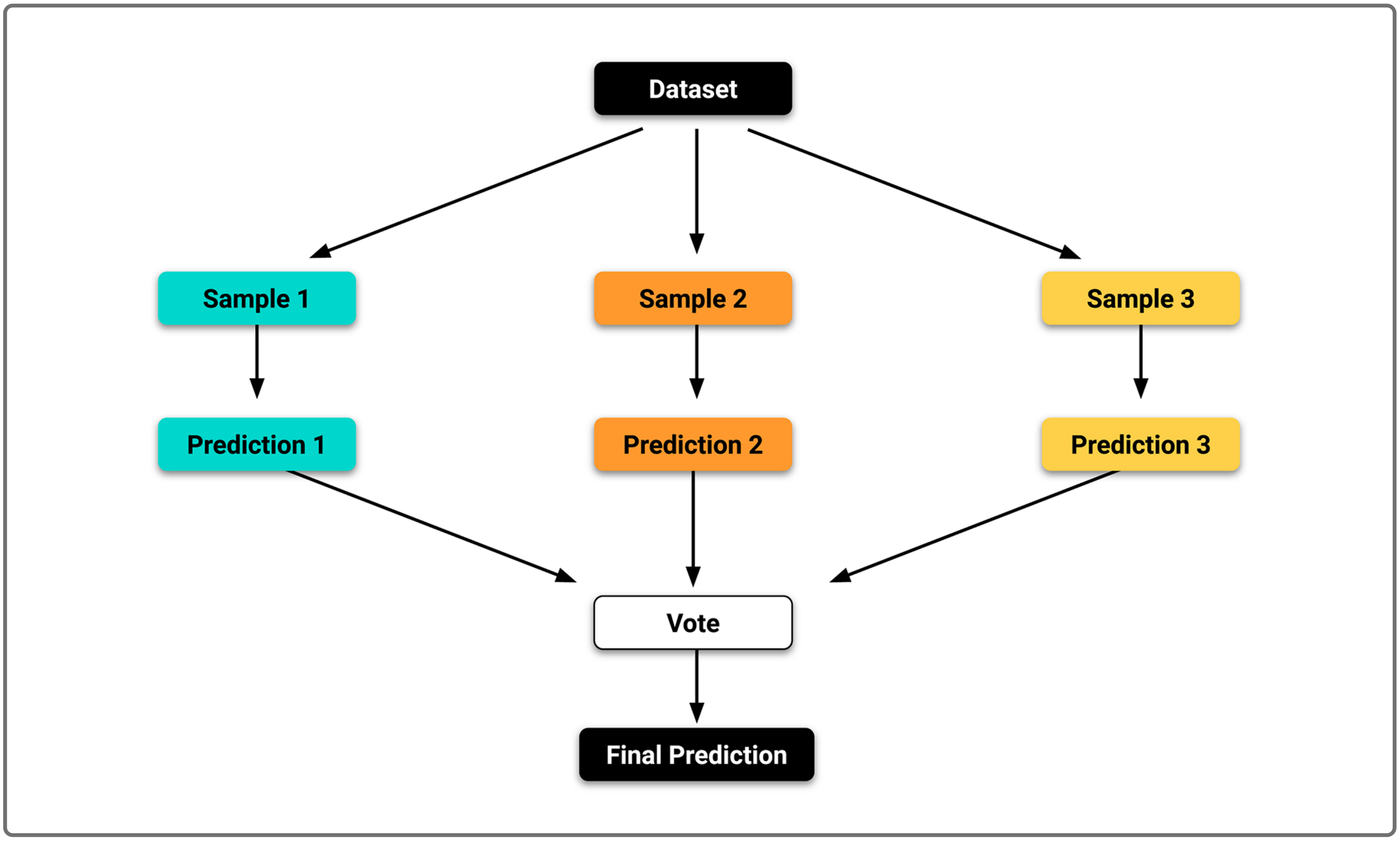
Sample 3: B, C, D, D, E

In our example above, each sample contains multiple occurrences of the same block. Sample 1 drew the letter A three times, Sample 2 drew the letter B twice, and Sample 3 drew the letter D twice. In other words, each observation (letter) may occur repeatedly in any given sample. In real life, this means that if your dataset were a Pandas DataFrame, a given row may occur multiple times in a sample.

In summary, bootstrapping is simply a sampling technique with which a number of samples are made, and in which an observation can occur multiple times.

Aggregation

In the aggregation step, different classifiers are run, using the samples drawn in the bootstrapping stage. Each classifier is run independently of the others, and all the results are aggregated via a voting process. Each classifier will vote for a label (a prediction). The final prediction is the one with the most votes.



A dataset consists of the following items: 1, 2, 3, 4, and 5. Which of the following could be a bootstrap sample?

○ 1, 2, 3, 4, 5

○ 1, 1, 2, 4, 4

○ Both A and B.

Check Answer

Finish ▶