Ensemble Methods

What is Ensemble method?

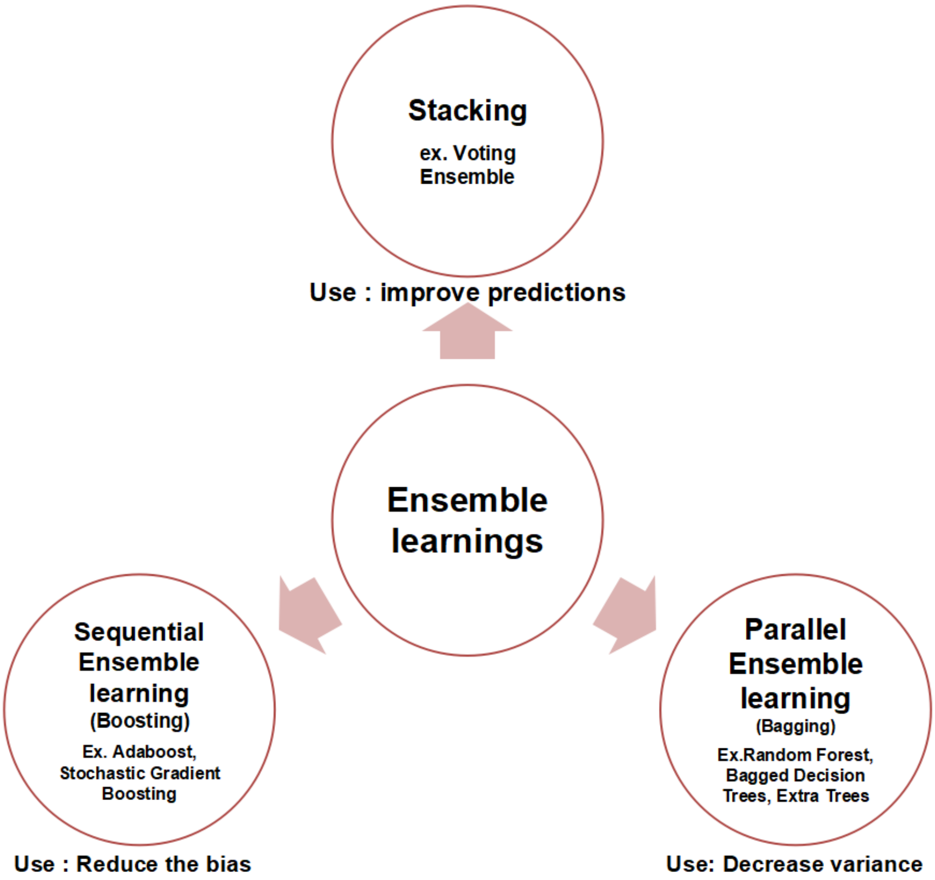
For instance,you ask a complex question to thousands of random people, then aggregate their answers. In many cases you will find that this aggregated answer is better than an expert's answer.

This is called the **wisdom of the crowd**.

Similarly, if you **aggregate** the predictions of a group of predictors (such as classifiers or regressors), you will often get better predictions than with the best individual predictor.

A **group of predictors** is called an **ensemble**; thus, this technique is called **Ensemble Learning**, and an Ensemble Learning algorithm is called an Ensemble method

Most popular Ensemble methods



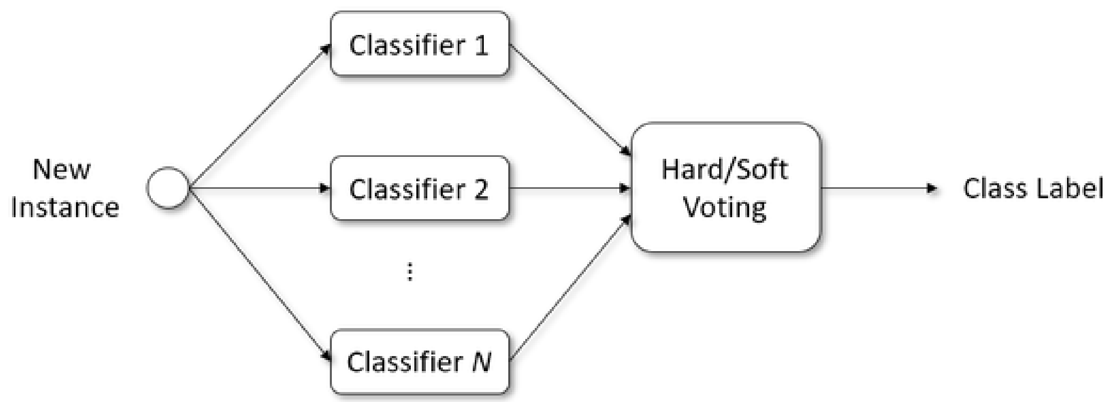
1.Bagging

2.Boosting

3.Stacking

Voting Classifers

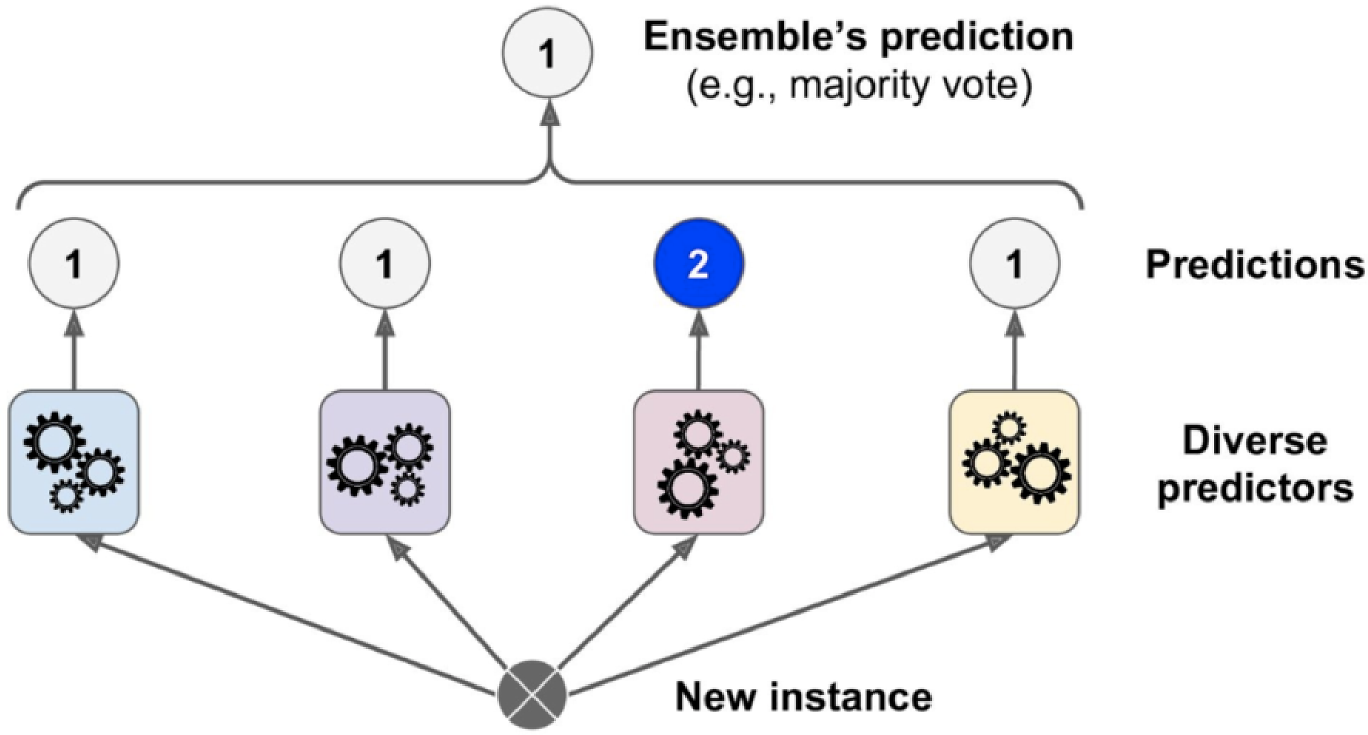
Suppose I have trained a few classifiers, each one achieving about 80% accuracy. You may have a Logistic Regression classifier, an SVM classifier, a Random Forest classifier, a K-Nearest Neighbors classifier, and perhaps a few more.



A very simple way to create an even better classifier is to **aggregate the predictions** of each classifier and **predict the class that gets the most votes**. This majority-vote classifier is called a **hard voting classifier**

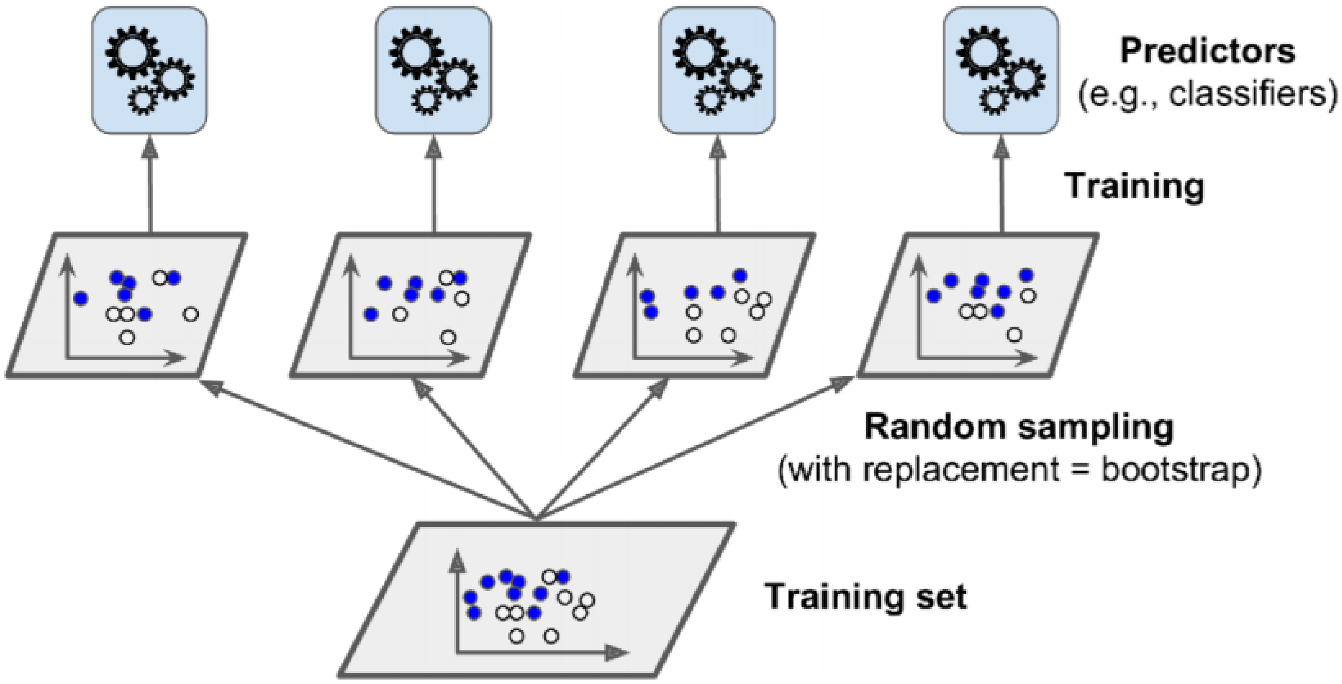
In **soft voting**, every individual classifier **provides a probability value** that a specific data point belongs to a particular target class.

The predictions are **weighted by the classifier's** importance and summed up. Then the target label with the **greatest sum of weighted probabilities wins the vote.**



Somewhat surprisingly, this voting classifier often achieves a higher accuracy than the best classifier in the ensemble. In fact, even if each classifier is a weak learner (mean- ing it does only slightly better than random guessing), the ensemble can still be a strong learner (achieving high accuracy), provided there are a sufficient number of weak learners and they are sufficiently diverse.

Bagging and Pasting



One way to get a **diverse set of classifiers** is to use **very different training algorithms.**

Another approach is to use the same **training algorithm** for every **predictor**, but to **train them on different random subsets** of the **training set**.

When **sampling** is **performed with replacement**, this method is called bagging (short for bootstrap aggregating) **bold text**.

When **sampling is performed without replacement**, it is called **pasting**.

In statistics, **resampling with replacement** is called **bootstrapping**

Once all **predictors are trained**, the **ensemble can make a prediction** for a new instance by **simply aggregating the predictions of all predictors**.

The **aggregation** function is typically the **statistical mode** (i.e., the **most frequent prediction**, just like a hard voting classifier) for **classification**, or the **average for regression.**

Each **individual predictor** has a **higher bias** than if it were **trained on the original training set**, but **aggregation reduces both bias and variance.**

Generally, the **net result** is that the **ensemble** has a **similar bias** but a **lower variance than a single predictor trained on the original training set**

Out-of-Bag Evaluation

With bagging, **some instances may be sampled several times** for any given predictor, while** others may not be sampled at all**.

By **default** a **BaggingClassifier** samples m training instances **with replacement** (bootstrap=True), where m is the size of the training set.

This means that only **about 63%** of the **training instances** are sampled on average for each predictor

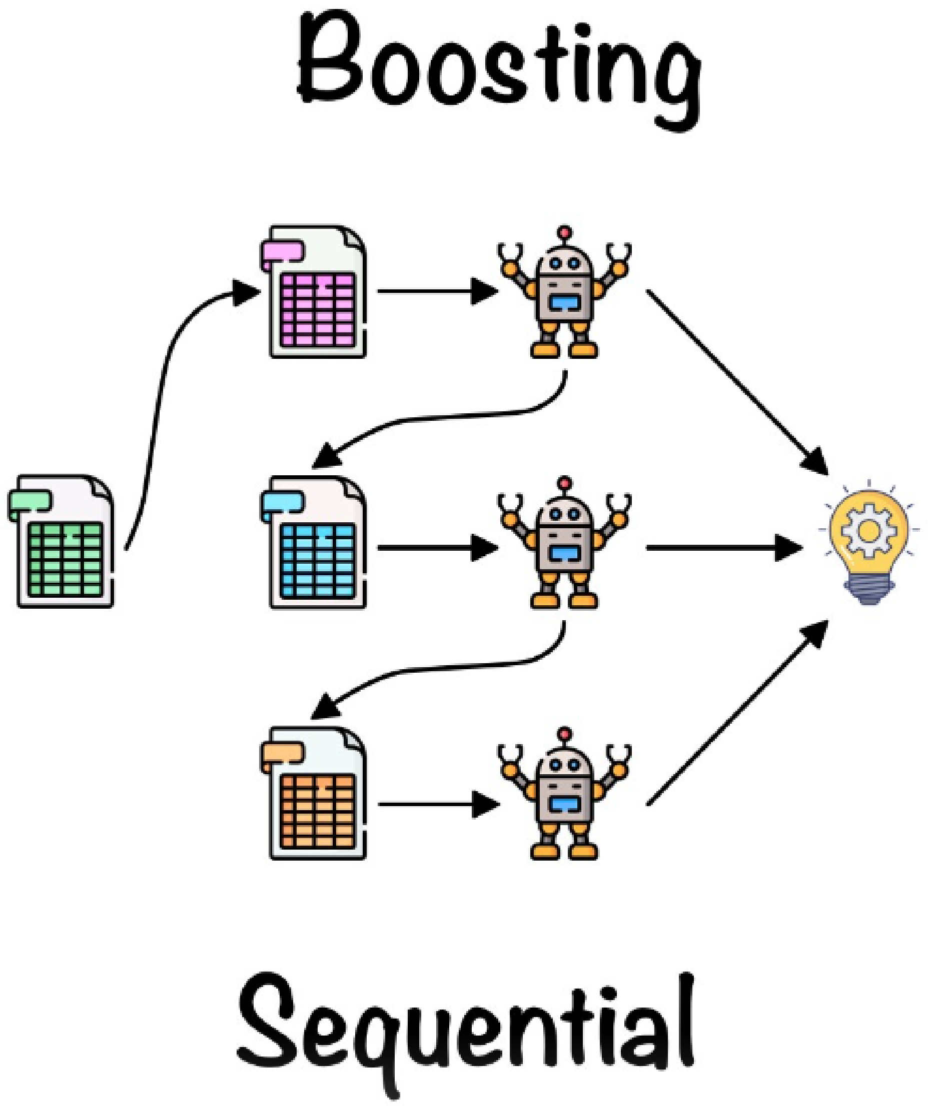
As m grows, this ratio approaches 1 – exp(–1) ≈ 63.212%

The remaining **37%** of the training instances that are not sampled are **called out-of-bag (oob)** instances. Note that they are not the same 37% for all predictors.

Since a **predictor never sees the oob instances during training**, it can be evaluated on these instances , without the **need for a separate validation set**. You can **evaluate** the **ensemble** itself by** averaging out the oob evaluations of each predictor**.

Boosting

Boosting (originally called **hypothesis boosting**) refers to any **Ensemble method** that can **combine several weak learners into a strong learner**.



The general idea of ** most boosting methods** is to **train predictors sequentially**, each trying to correct its predecessor.

There are many **boosting methods available**, but by far the most popular are **AdaBoost(short for Adaptive Boosting) and Gradient Boosting**.

Advantages and Disadvantages of ensemble methods

credits : https://medium.com/@aravanshad

Advantages

1.Intuitively, ensembles allow the different needs of a difficult problem to be handled by hypotheses suited to those particular needs.

2.Mathematically, ensembles provide an extra degree of freedom in the classical bias/variance tradeoff, allowing solutions that would be difficult (if not impossible) to reach with only a single hypothesis.

3.They're unlikely to overfit.

Disadvantages

1.The model that is closest to the true data generating process will always be best and will beat most ensemble methods. So if the data come from a linear process, linear models will be much superior to ensemble models.

2.Ensemble methods are usually computationally expensive. Therefore, they add learning time and memory constrains to the problem.