BAGGING AND BOOSTING

A023119820029

A023119820030

BAGGING

- **Bagging is an ensemble learning technique that helps to improve the performance and accuracy of machine learning algorithms. It is used to deal with bias-variance trade-offs and reduces the variance of a prediction model.**
- **Bagging avoids overfitting of data and is used for both regression and classification models, specifically for decision tree algorithms but can also be applied to other algorithms.**
- **Bagging is also known as Bootstrap aggregating.**

HOW BAGGING WORKS ?

1. **Bootstrapping:** Bagging leverages a bootstrapping sampling technique to create diverse samples. This resampling method generates different subsets of the training dataset by selecting data points at random and with replacement. This means that each time you select a data point from the training dataset, you are able to select the same instance multiple times.
2. **Parallel training:** These bootstrap samples are then trained independently and in parallel with each other using weak or base learners.

3. Aggregation: Finally, depending on the task (i.e. regression or classification), an average or a majority of the predictions are taken to compute a more accurate estimate. In the case of regression, an average is taken of all the outputs predicted by the individual classifiers; this is known as soft voting. For classification problems, the class with the highest majority of votes is accepted; this is known as hard voting or majority voting.

- An example for this algorithm could be Rain forests. Random Forests uses bagging underneath to sample the dataset with replacement randomly.

(A) bagging

step 1

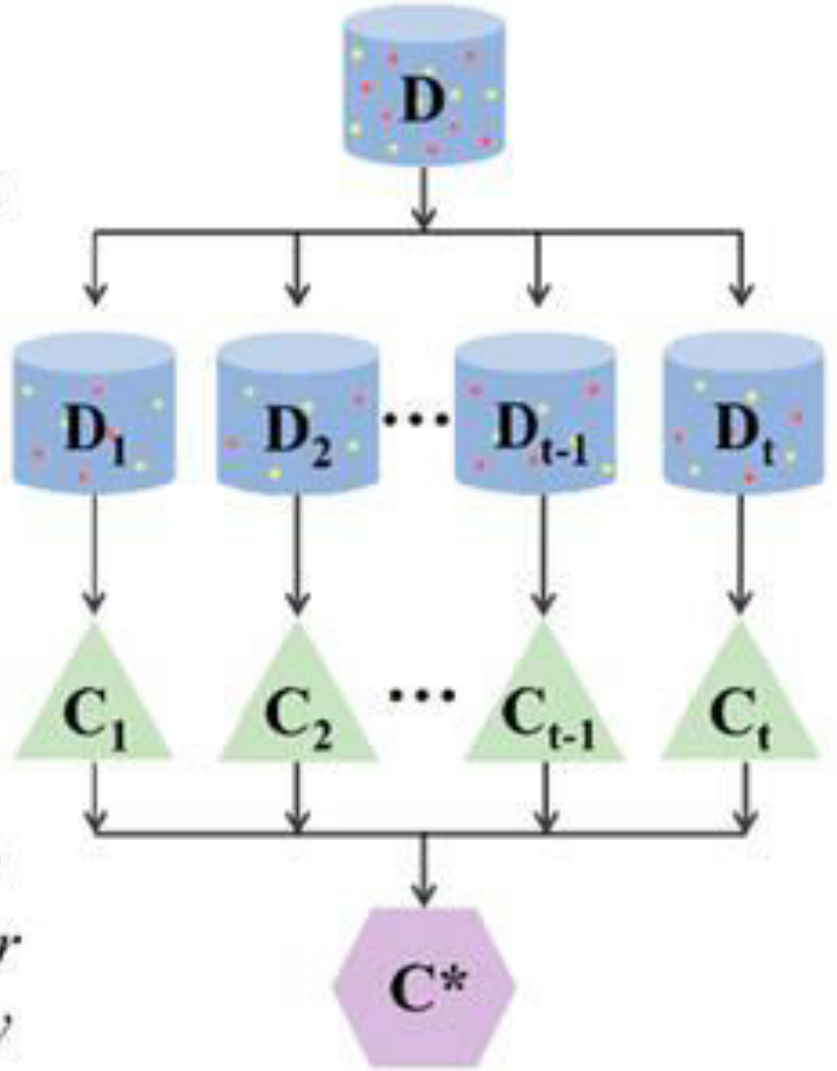
create multiple data sets through random sampling with replacement

step 2

build multiple learners in parallel

step 3

combine all learners using an averaging or majority-vote strategy



BAGGING ALGORITHM

Let's assume we have a sample of 100 values (x) and we'd like to get an estimate of the mean of the sample.

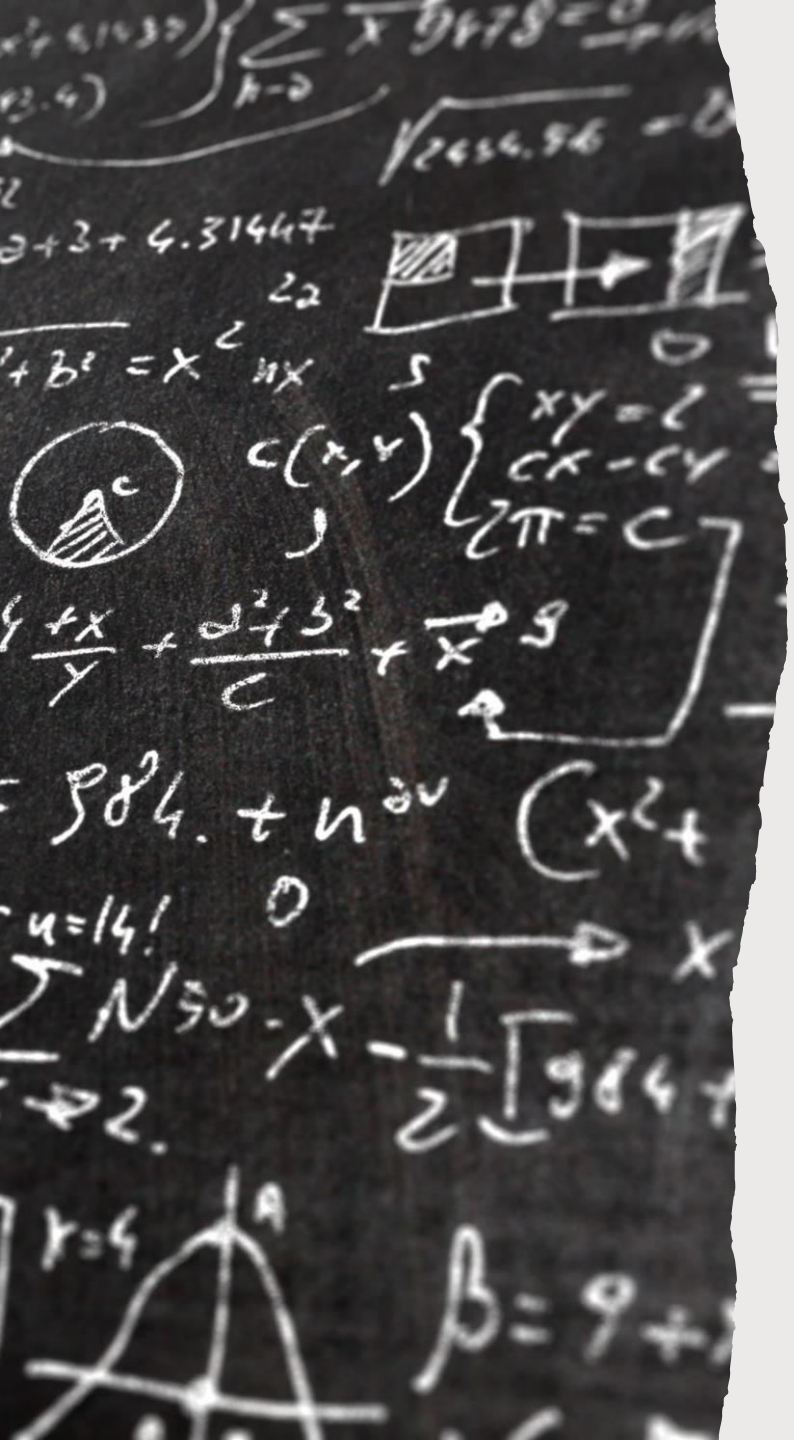
We can calculate the mean directly from the sample as:

mean(x) = 1/100 * sum(x)

We know that our sample is small and that our mean has error in it. We can improve the estimate of our mean using the bootstrap procedure:

Create many (e.g. 1000) random sub-samples of our dataset with replacement (meaning we can select the same value multiple times).

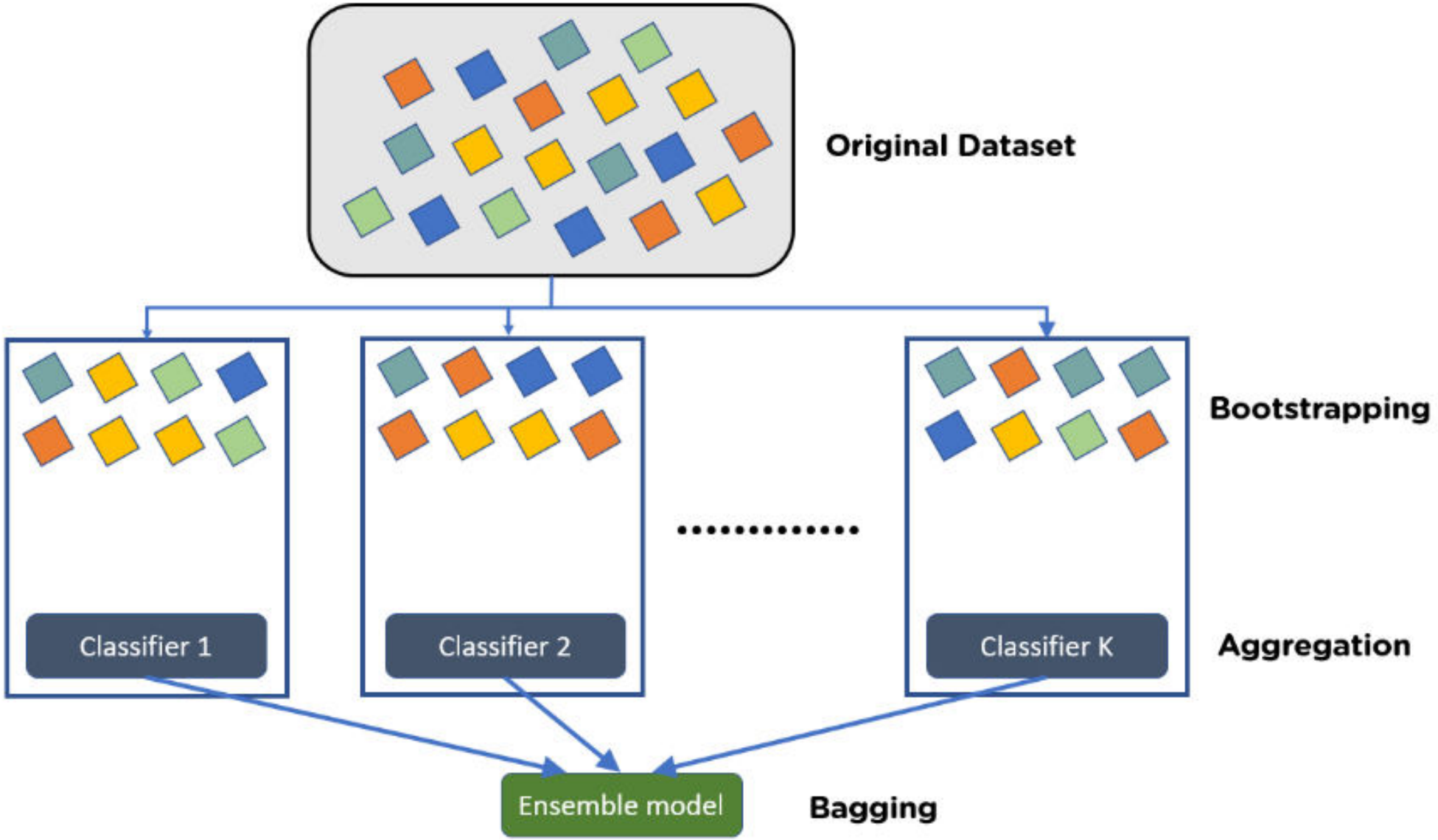
Calculate the mean of each sub-sample.



3. Calculate the average of all of our collected means and use that as our estimated mean for the data.

- For example, let’s say we used 3 resamples and got the mean values 2.3, 4.5 and 3.3. Taking the average of these we could take the estimated mean of the data to be 3.367.

- This process can be used to estimate other quantities like the standard deviation and even quantities used in machine learning algorithms, like learned coefficients.


Original Dataset
Bootstrapping
Classifier 1
Classifier 2
Classifier K
Aggregation
Ensemble model
Bagging

BENEFITS

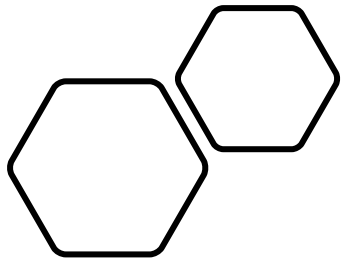
Ease of implementation: Python libraries such as scikit-learn (also known as sklearn) make it easy to combine the predictions of base learners or estimators to improve model performance.

Reduction of variance: Bagging can reduce the variance within a learning algorithm. This is particularly helpful with high-dimensional data, where missing values can lead to higher variance, making it more prone to overfitting and preventing accurate generalization to new datasets.

CHALLENGES

Loss of interpretability: While the output is more precise than any individual data point, a more accurate or complete dataset could also yield more precision within a single classification or regression model.

Computationally expensive: Bagging slows down and grows more intensive as the number of iterations increase. Thus, it’s not well-suited for real-time applications. Clustered systems or a large number of processing cores are ideal for quickly creating bagged ensembles on large test sets.



- **Less flexible:** As a technique, bagging works particularly well with algorithms that are less stable. One that are more stable or subject to high amounts of bias do not provide as much benefit as there’s less variation within the dataset of the model.

BOOSTING ALGORITM

A023119820029

AGENDA

INTRODUCTION

BAGGING VS BOOSTING

TYPES

BENEFITS AND CHALLENGES

APPLICATIONS

INTRODUCTION

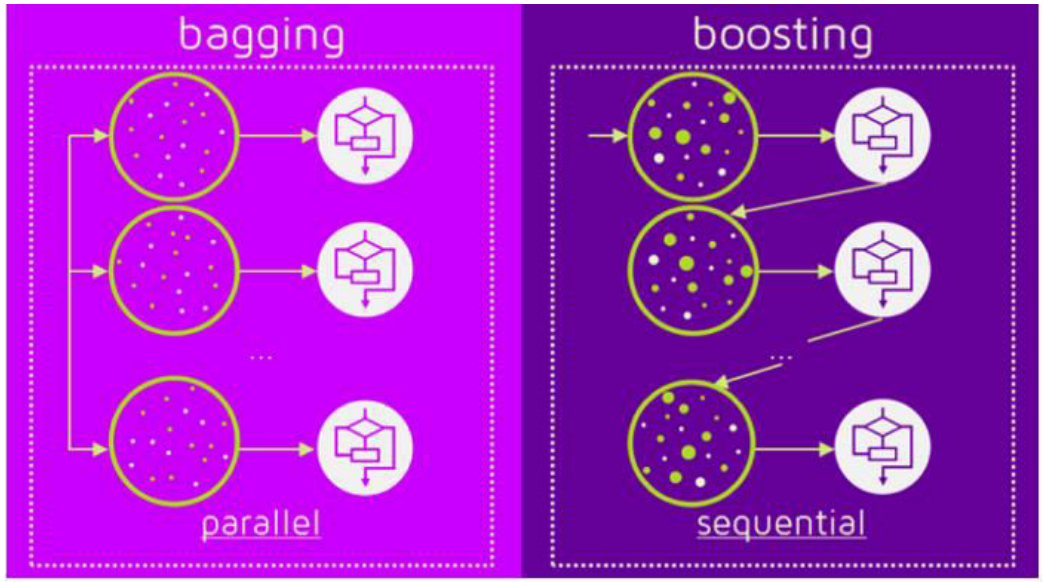
Boosting is an ensemble learning method that combines a set of weak learners into a strong learner to minimize training errors. In boosting, a random sample of data is selected, fitted with a model and then trained sequentially—that is, each model tries to compensate for the weaknesses of its predecessor.

HOW BOOSTING WORKS

Boosting Algorithms combines each weak learner to create one strong prediction rule. To identify the weak rule,
there is a base Learning algorithm (Machine Learning). Whenever the Base algorithm is applied, it creates new prediction rules using the iteration process. After some iteration, it combines all weak rules to create one single prediction rule.
To choose the right distribution follows the below-mentioned steps:
Step 1: The base Learning algorithm combines each distribution and applies equal weight to each distribution.
Step 2: If any prediction occurs during the first base learning algorithm, then we pay high attention to that prediction error.
Step 3: Repeat step 2 until the limit of the Base Learning algorithm has been reached or high accuracy.
Step 4: Finally, it combines all the weak learner to create one strong prediction tule.

BAGGING VS BOOSTING

BAGGING AND BOOSTING ARE TWO MAIN TYPES OF ENSEMBLE LEARNING METHODS. THE MAIN DIFFERENCE BETWEEN THESE LEARNING METHODS IS THE WAY IN WHICH THEY ARE TRAINED. IN BAGGING, WEAK LEARNERS ARE TRAINED IN PARALLEL, BUT IN BOOSTING, THEY LEARN SEQUENTIALLY



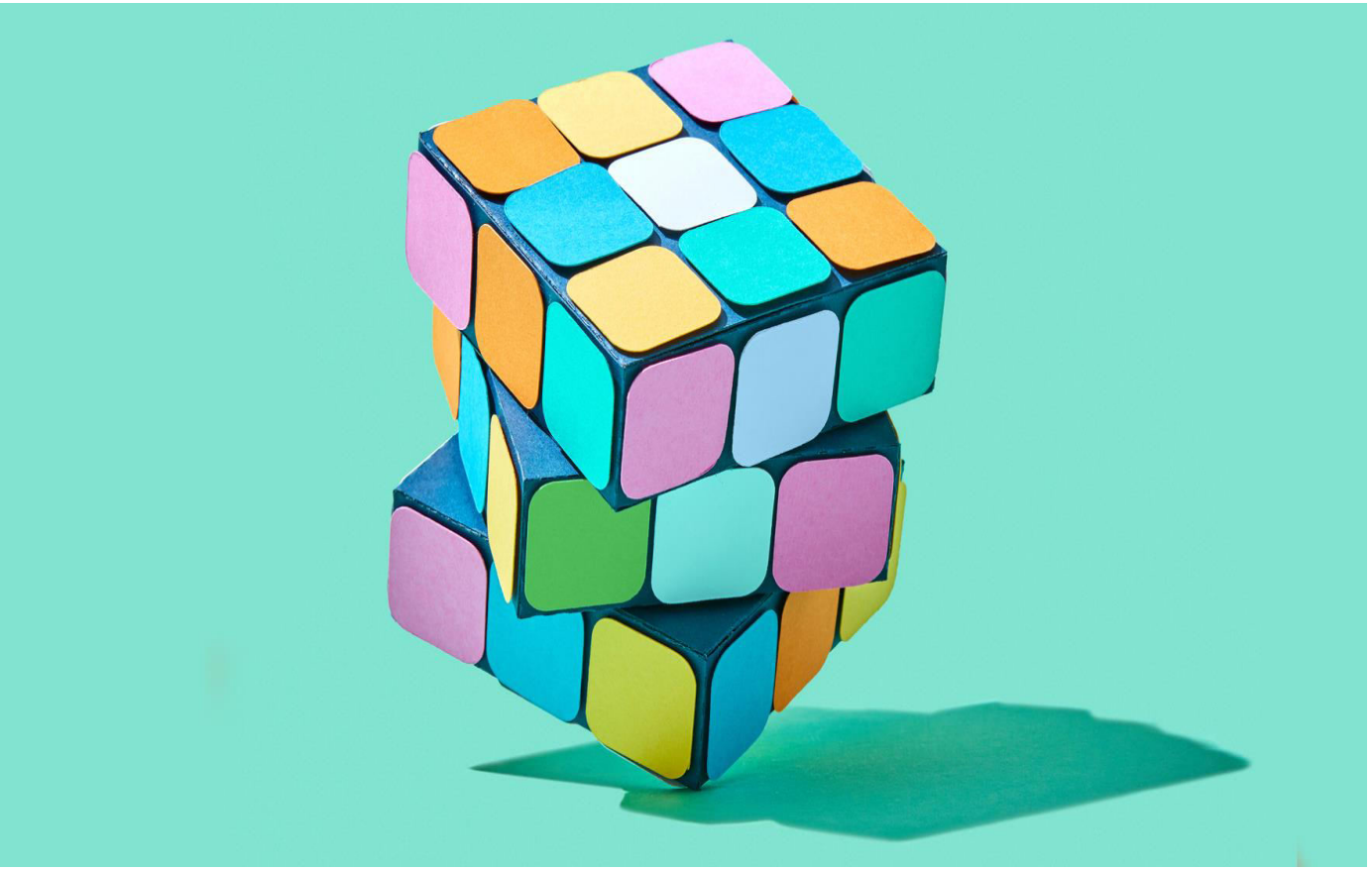
TYPES

1. Adaptive boosting
2. Gradient boosting
3. Extreme gradient boosting

BENEFITS

1. Ease of Implementation
2. Reduction of bias
3. Computational Efficiency

CHALLENGES



- One disadvantage of boosting is that it is sensitive to outliers since every classifier is obliged to fix the errors in the predecessors. Thus, the method is too dependent on outliers.

- Another disadvantage is that the method is almost impossible to scale up. This is because every estimator bases its correctness on the previous predictors, thus making the procedure difficult to streamline.

WHAT'S NEXT

LOOKING AHEAD

APPLICATIONS

Healthcare

- Boosting is used to lower errors in medical data predictions, such as predicting cardiovascular risk factors and cancer patient survival rates
- Applying boosting to multiple genomics platforms can improve the prediction of cancer survival time.

IT

- Gradient boosted regression trees are used in search engines for page rankings
- Viola-Jones boosting algorithm is used for image retrieval.

Finance

- Boosting is used with deep learning models to automate critical tasks, including fraud detection, pricing analysis.

SUMMARY

Boosting algorithms represent a different machine learning perspective: turning a weak model to a stronger one to fix its weaknesses


THANK YOU