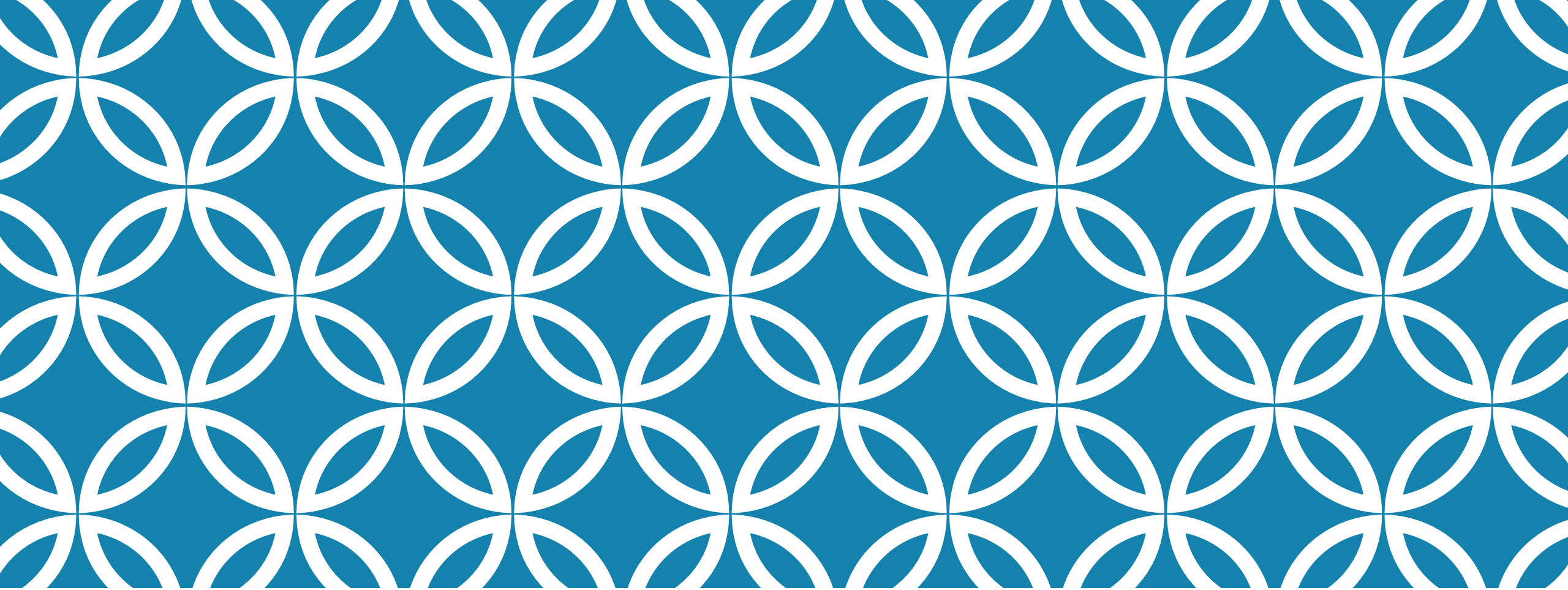


INTRODUCTION TO GRADIENT BOOSTING

A BRANCH OF MACHINE LEARNING

ivy Professional School
ISO 9001 : 2000 Organisation

A PEAK INTO ENSEMBLE METHODS

BRIEF DESCRIPTION

When we try to predict the target variable using any machine learning technique, the main causes of difference in actual and predicted values are **noise, variance, and bias.**

Ensemble helps to reduce these factors (except noise, which is irreducible error)

INTRODUCING BIAS VS VARIANCE TRADE OFF (1/2)

What is Bias?

Bias is the difference between the average prediction of our model and the correct value which we are trying to predict.

Model with high bias pays very little attention to the training data and oversimplifies the model. It always leads to high error on training and test data.

$$\text{bias}[\hat{f}(x)] = \mathbb{E}[\hat{f}(x)] - f(x)$$

INTRODUCING BIAS VS VARIANCE TRADE OFF (2/2)

What is Variance?

Variance is the variability of model prediction for a given data point or a value which tells us spread of our data.

Model with high variance pays a lot of attention to training data and does not generalize on the data which it hasn't seen before. As a result, such models perform very well on training data but has high error rates on test data.

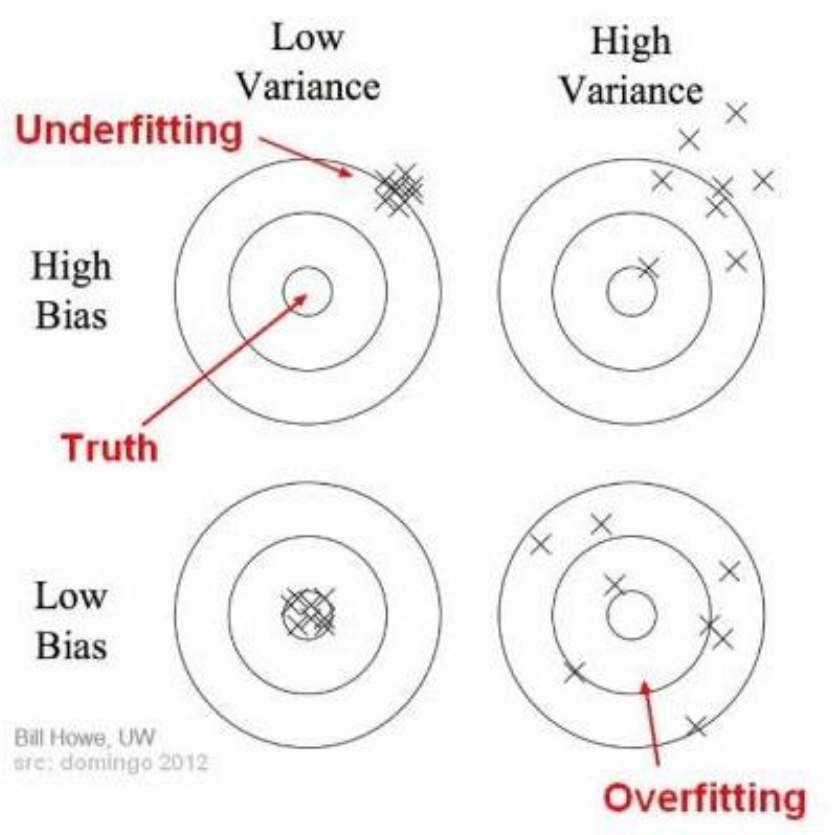
$$\text{var}(\hat{f}(x)) = \mathbb{E}[(\hat{f}(x) - \mathbb{E}[\hat{f}(x)])^2]$$

TOTAL ERROR

Total error = Bias^2 + Variance + Irreducible error

$$\mathbb{E}_x[\mathbb{E}_{\hat{f}}[(y - \hat{f}(x))^2]] = \mathbb{E}_x[\text{bias}[\hat{f}(x)]^2] + \mathbb{E}_x[\text{var}(\hat{f}(x))] + \sigma_\epsilon^2$$

BIAS VS VARIANCE TRADEOFF



2 ENSEMBLE METHODS

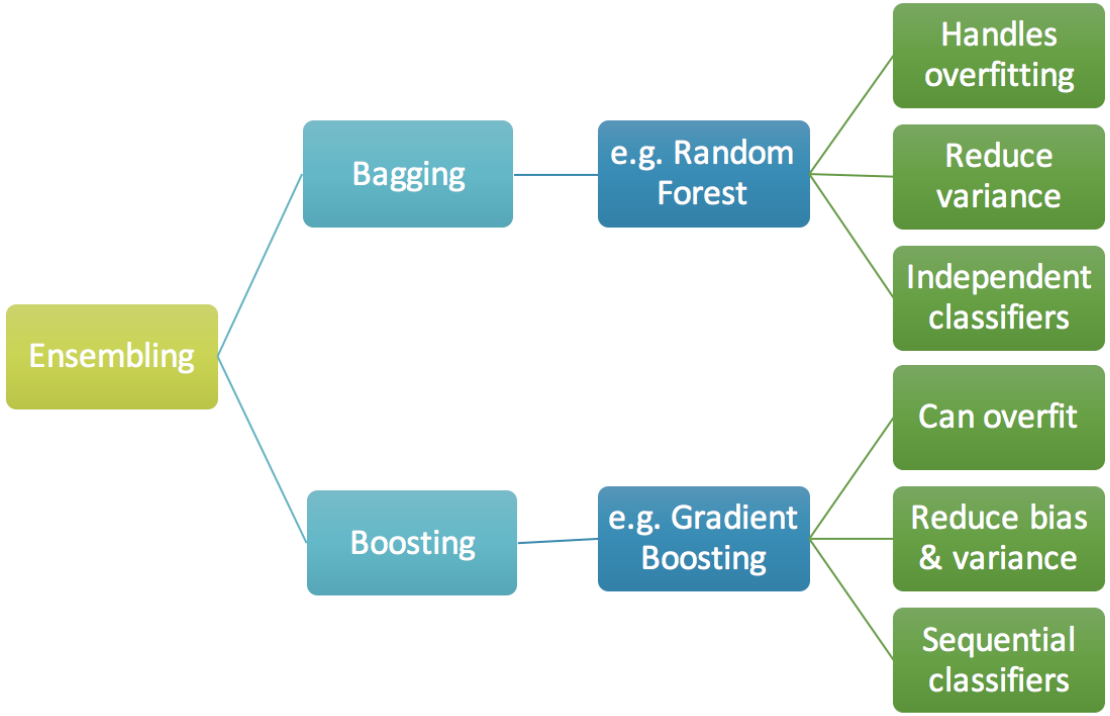
An ensemble is just a collection of predictors which come together (e.g. mean of all predictions) to give a final prediction. The reason we use ensembles is that many different predictors trying to predict same target variable will perform a better job than any single predictor alone..

Ensemble classified into Bagging and Boosting.

Bagging is a simple ensembling technique in which we build many *independent* predictors/models/learners and combine them using some model averaging techniques. (e.g. weighted average, majority vote or normal average)

Boosting is an ensemble technique in which the predictors are not made independently, but sequentially.

2 ENSEMBLE METHODS



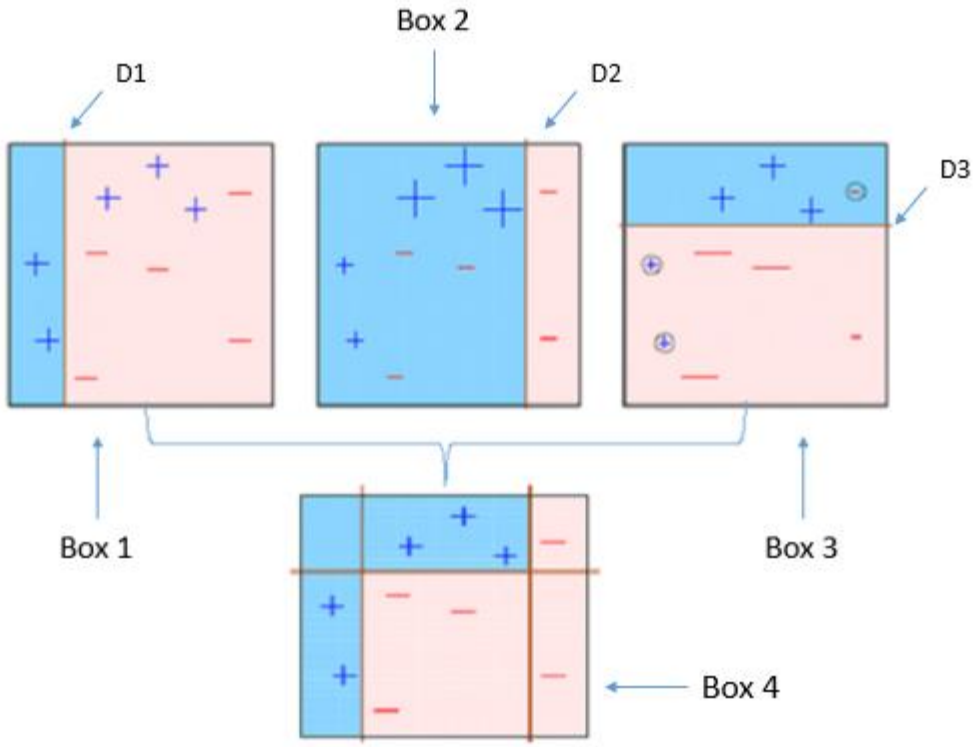
WHAT IS GRADIENT BOOSTING?

WHAT IS GRADIENT BOOSTING

Boosting is a sequential technique which works on the principle of an ensemble. It combines a set of weak learners and delivers improved prediction accuracy.

At any instant t, the model outcomes are weighed based on the outcomes of previous instant t-1. The outcomes predicted correctly are given a lower weight and the ones miss-classified are weighted higher.

HOW GRADIENT BOOSTING WORKS?



HOW GRADIENT BOOSTING WORKS?

1. Box 1: The first classifier (usually a decision stump) creates a vertical line (split) at D1. It says anything to the left of D1 is + and anything to the right of D1 is -. However, this classifier misclassifies three + points.

Note a Decision Stump is a Decision Tree model that only splits off at one level, therefore the final prediction is based on only one feature.

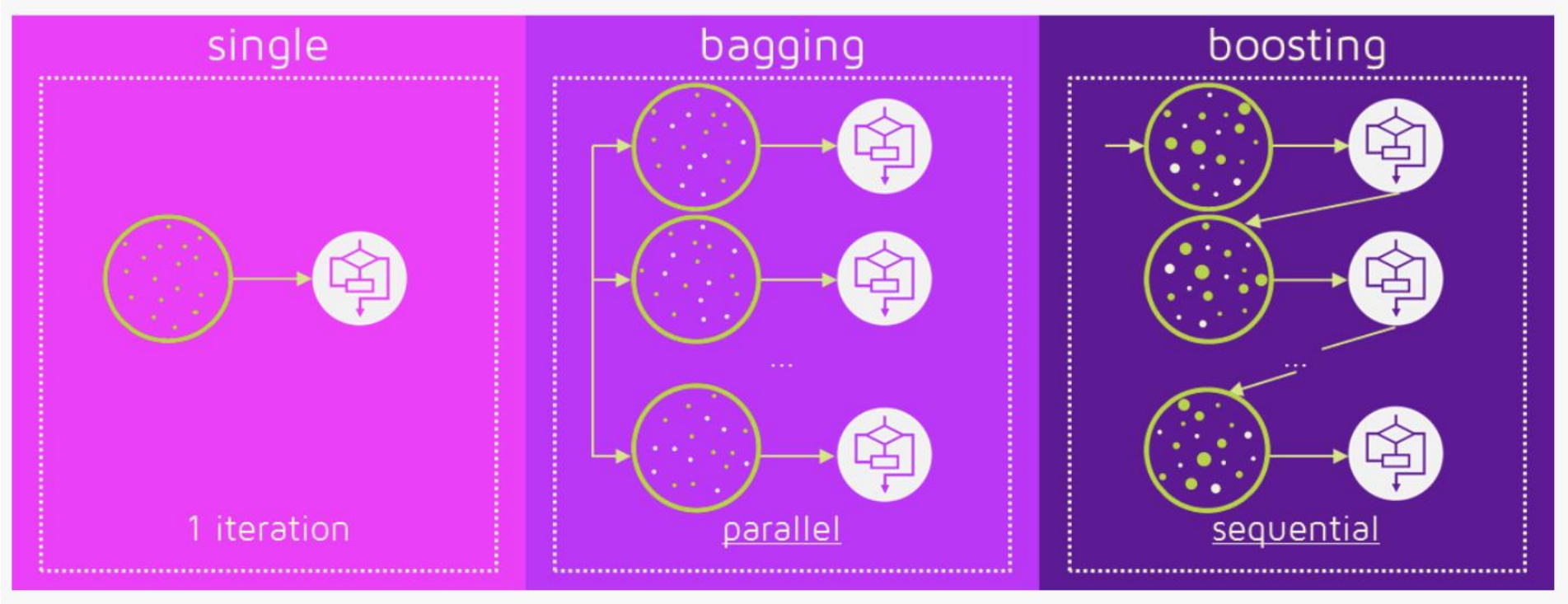
2. Box 2: The second classifier gives more weight to the three + misclassified points (see the bigger size of +) and creates a vertical line at D2. Again it says, anything to the right of D2 is - and left is +. Still, it makes mistakes by incorrectly classifying three - points.

3. Box 3: Again, the third classifier gives more weight to the three - misclassified points and creates a horizontal line at D3. Still, this classifier fails to classify the points (in the circles) correctly.

4. Box 4: This is a weighted combination of the weak classifiers (Box 1,2 and 3). As you can see, it does a good job at classifying all the points correctly.

HOW GRADIENT BOOSTING WORKS?

The observations have an unequal probability of appearing in subsequent models and ones with the highest error appear most.



Professional School
ISO 9001 : 2000 Organisatio

INUTUTION BEHINED GRAIDENT BOOSTING

THE INTUITION

We want our predictions, such that our loss function (MSE) is minimum. By using gradient descent and updating our predictions based on a learning rate, we can find the values where MSE is minimum.

The intuition behind gradient boosting algorithm is to repetitively leverage the patterns in residuals and strengthen a model with weak predictions and make it better.
Once we reach a stage that residuals do not have any pattern that could be modeled, we can stop modeling residuals (otherwise it might lead to overfitting).
Algorithmically, we are minimizing our loss function, such that test loss reach its minima.

THANK YOU