Ensemble Learning - II

J Junaid Mahmud

Sampling

> 💡 Sampling is the selection of a subset or a statistical sample of individuals from within a statistical population.

With Replacement

> 💡 If each member of a population is replaced after it is picked, then that member has the possibility of being chosen more than once. When sampling is done with replacement, then events are considered to be independent, meaning the result of the first pick will not change the probabilities for the second pick.

Suppose you pick three cards with replacements. The first card you pick out of the 52 cards is the **Q of spades**. You put this card back, reshuffle the cards and pick a second card from the 52-card deck. It is the **ten of clubs.** You return, reshuffle and pick a third card from the 52-card deck. This time it is **Q of spades again.** Your picks are {Q of spades, ten of clubs, Q of spades}. You have picked the Q of spades twice. You pick each card from the 52-card deck.

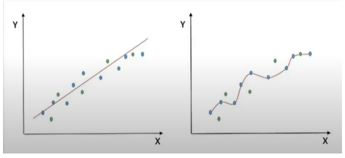
Without Replacement

> 💡 When sampling is done without replacement, each member of a population may be chosen only once. In this case, the probabilities for the second pick are affected by the result of the first pick.

Suppose you pick three cards without replacement. The first card you pick is **K of hearts, with** a **probability** of **1/52.** You keep this card aside and pick another one with a **probability of 1/51.** You put this aside and pick another with a **probability of 1/50.** You can not pick the same card twice as you have picked cards without replacement.

Bias

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



Bias

- Here the line is the final prediction.
- The measurement of **how far the points are from the line** is called **bias.** The distance of the points from the prediction line.
- For image 1, it's higher. For image 2, it's lower. Since the points are closer to the line.
- If bias is lower, error is lower. Bias is higher, error is higher.

> **Q. Is bias and loss the same**?

> They are synonymous. But bias is a different measure. Not all the losses work the same way. For **linear regression**, loss can be similar to bias. But for others, it may not be the same as bias.

> How efficient a model is, can be justified with two things. One of them is **bias.**

> Although there is a similarity in the definition of bias and loss. But if the loss is less, the model gets better. But if the bias is less, the model may not get better.

Low Bias

> 💡 The model will closely match the training dataset.
>
> Low-bias model will capture the dataset trend perfectly. It is considered an **overfitting model** with a **very low error rate.**

High Bias

> 💡 The model will not match the training dataset closely.
>
> High-bias model will not be able to capture the dataset the dataset trend. It is considered as the
> **underfitting model which has a high error rate.**

Variance

> 💡 Variance is the variability of a model and how much it is sensitive to another subset of the training dataset.

Low Variance

💡 The model is less sensitive to changes in the training data and can produce consistent estimates of the target function with different subsets of data from the same distribution.

Case of Underfitting.

High Variance

💡 The model is very sensitive to changes in the training data and can result in significant changes in the estimates of the target function when trained on different subsets of data from the same distribution.

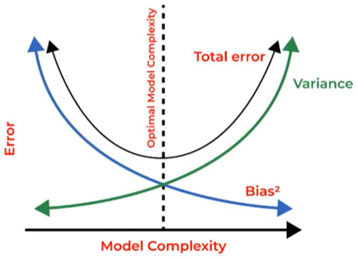
Case of Overfitting.

> Q. Why high variance is **overfitting** and low variance is **underfitting**?
> ⇒ For example: the following
> 1st iteration 80%, 2nd iteration 93%. Mostly the result is over 90%. So this is overfitting.

Bias	Variance	Outcome
Low	Low	Ideal Scenario
High	High	Can not capture trend, high sensitivity to change in training data
Low	High	Overfitting
High	Low	Underfitting

- I can somehow accept low bias, but not high bias.
- Same for low variance. However high variance is not acceptable.
- Low bias is kind of **overfitting.** Low variance is a kind of **overfitting.** So, balanced.

Bias-Variance Tradeoff



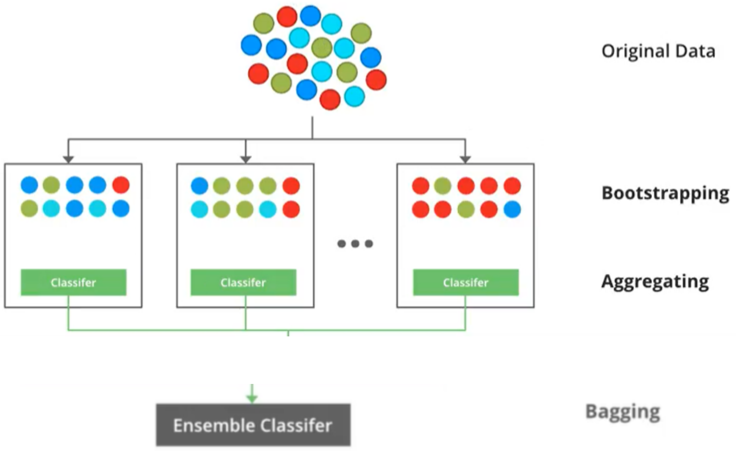
Model Complexity:

- Number of parameters is high.
- Where error is low, bias is low. Error is high, variance is high.
- If any parameter is extreme, the machine learning model does not work well. So everything should be balanced.

Bagging

> 💡 Bootstrap Aggregating
> It decreases the variance and helps to avoid overfitting.

- Multiple subsets are created from the original dataset with equal tuples, selecting observations with replacement. So some will have repeated data.
- A base model is created on each of these subsets.
- Each model is learned in parallel with each training set and independent of each other.
- The final predictions are determined by combining the predictions from all the models.

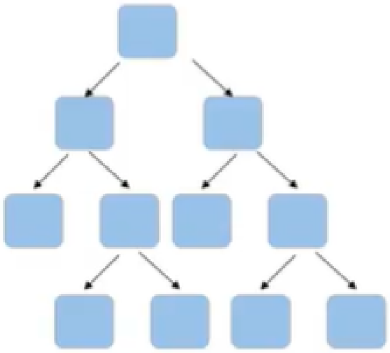


Random Forest

> 💡 It is an ensemble method.

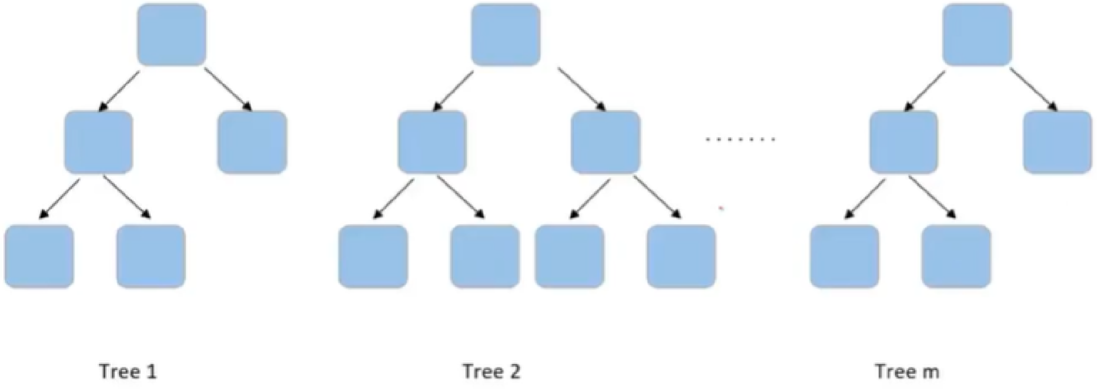
- Random bootstrap samples - Tree Level Sampling. We sample the data here. (Rows)
- Random Sampling of Feature - Node Level Sampling. We sample the features here. (Columns)
- Root Node.
- Node.
- Leaf.

> Q. Why Random Forest is called a **Random Forest?**
>
> ⇒
>
> We use random subsets of data to decide with different decision trees. Like a forest.



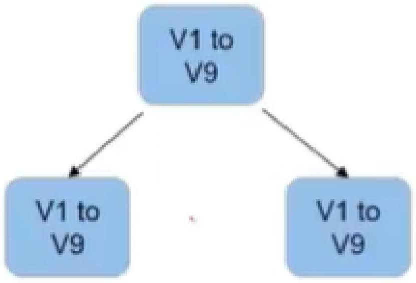
Hyper-parameters

n_estimators = Number of trees



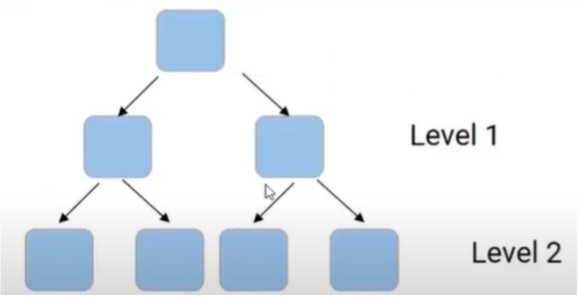
max_features

- square root of the total number of features.
- log of the total number of features.



max_depth

- How many levels are there in a tree?
- The more level, the more split.

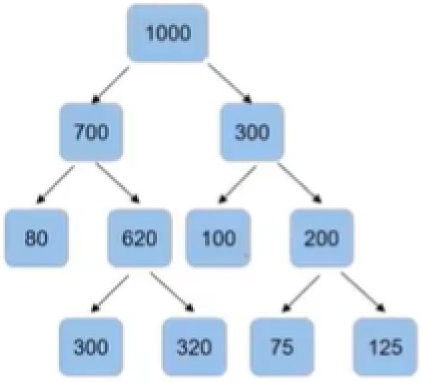


min_samples_split

- If it’s set to 100 (for example), we will not split a node if it has 100 or less number of samples.

min_samples_leaf

- If set to 100, then we will not consider the node where samples are less than 100.



Gini Impurity

> Gini impurity tells us what the probability of misclassifying is in an observation

$$Gini\ Impurity = 1 - Gini$$

$$Gini = (p_1^2 + p_2^2 + p_3^2 + ... + p_n^2)$$

Entropy

$$E = -\sum_{i=1}^{n} p_i \ log_2\ (p_i)$$

Information Gain

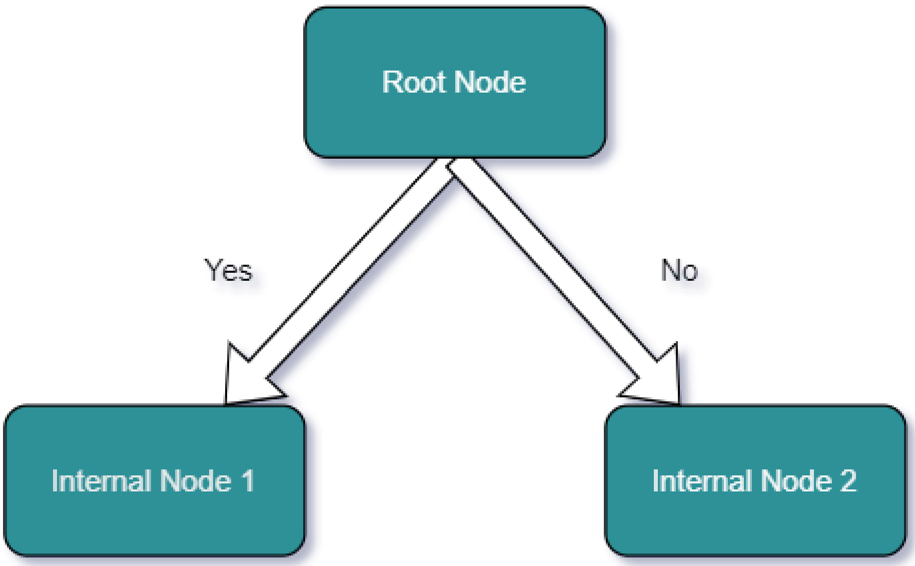
$$Information\ Gain = Entropy_{\ parent} - Entropy_{\ children}$$

- The higher is the information gain for a feature, the higher the chance of that feature to be selected for the decision tree.

Example

	Online Courses	Background	Working Status	Exam Result
1	Y	Math	NW	Pass
2	N	Math	W	Fail
3	Y	Math	W	Fail
4	Y	CS	NW	Pass
5	N	Other	W	Fail
6	Y	Other	W	Fail
7	Y	Math	NW	Pass
8	Y	CS	NW	Pass
9	N	Math	W	Pass
10	N	CS	W	Pass
11	Y	CS	W	Pass
12	N	Math	NW	Pass
13	Y	Other	W	Fail
14	N	Other	NW	Fail
15	N	Math	W	Fail

- **Exam Result** is the target here.



$$Entropy_{parent-node} = -\left[\frac{8}{15} * log_2\left(\frac{8}{15}\right) + \frac{7}{15} * log_2\left(\frac{7}{15}\right)\right]$$
$$= 0.9968$$

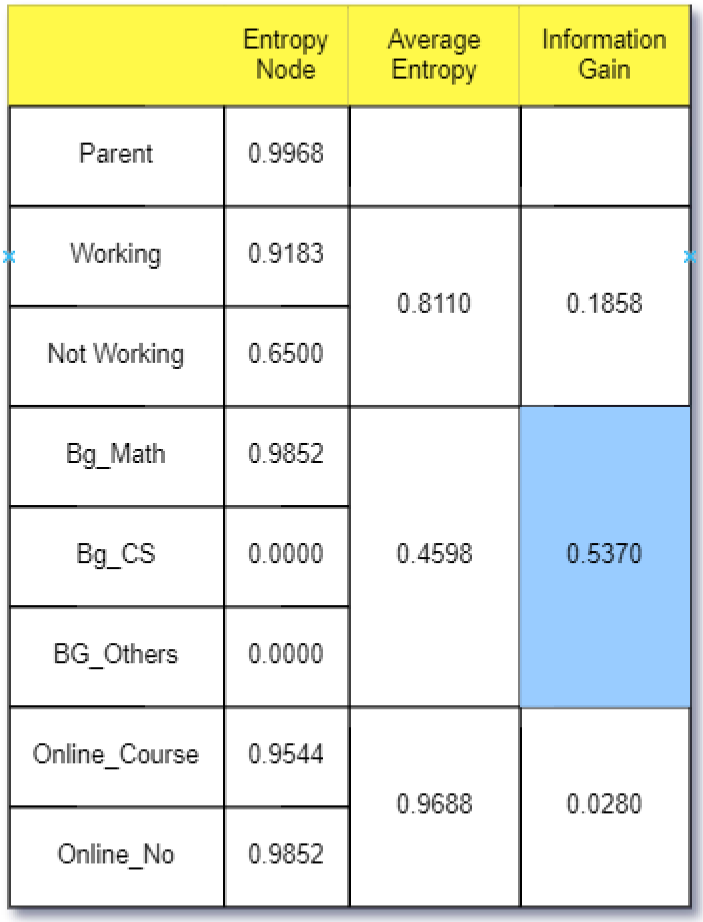
$$Feature\ Name:\ Working\ Status$$
$$Total\ Sample\ =\ 15$$
$$W_{total}\ =\ 9, W_{pass}\ =\ 3, W_{fail}\ =\ 6$$
$$NW_{total}\ =\ 6, NW_{pass}\ =\ 5, NW_{fail}\ =\ 1$$

$$Entropy_W = -\left[\frac{3}{9} * log_2\left(\frac{3}{9}\right) + \frac{6}{9} * log_2\left(\frac{6}{9}\right)\right]$$
$$= 0.9183$$

$$Entropy_{NW} = -\left[\frac{5}{6} * log_2\left(\frac{5}{6}\right) + \frac{1}{6} * log_2\left(\frac{1}{6}\right)\right]$$
$$= 0.6500$$

$$Information\ Gain = Entropy_{parent} - Entropy_{children}$$
$$= 0.9968 - 0.8110$$
$$= 0.1858$$

- Since **Background** feature has higher information gain, we will use this for splitting.



	Entropy Node	Average Entropy	Information Gain
Parent	0.9968		
Working	0.9183	0.8110	0.1858
Not Working	0.6500		
Bg_Math	0.9852	0.4598	0.5370
Bg_CS	0.0000		
BG_Others	0.0000		
Online_Course	0.9544	0.9688	0.0280
Online_No	0.9852		

Calculating Gini for Background

$$Gini\ _{math} = 1 - \left(\frac{4}{7}\right)^2 - \left(\frac{3}{7}\right)^2 = 0.4897$$

$$Gini\ _{CS} = 1 - \left(\frac{4}{4}\right)^2 - \left(\frac{0}{5}\right)^2 = 0$$

$$Gini\ _{Others} = 1 - \left(\frac{0}{4}\right)^2 - \left(\frac{4}{4}\right)^2 = 0$$

$$- - -$$

$$So,$$

$$Gini\ _{bkgrd} = \frac{7}{15} * 0.4897 + \frac{4}{15} * 0 + \frac{4}{15} * 0$$
$$= 0.2286$$

Calculating Gini for Work Status

$$
\begin{aligned}
Gini_{\,Working} &= 1 - \left(\frac{6}{9}\right)^2 - \left(\frac{3}{9}\right)^2 = 0.44 \\
Gini_{\,notWorking} &= 1 - \left(\frac{5}{6}\right)^2 - \left(\frac{1}{6}\right)^2 = 0.278 \\
&---\\
&So, \\
Gini_{\,workStatus} &= \frac{9}{15} * 0.44 + \frac{6}{15} * 0.278 \\
&= 0.378
\end{aligned}
$$

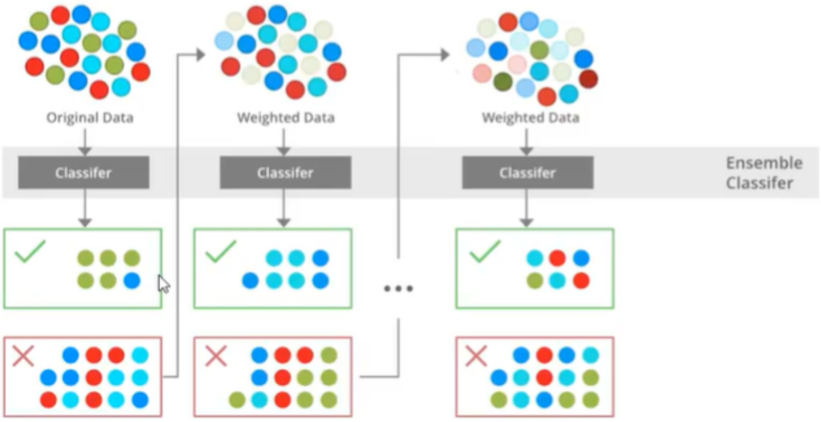
Calculating Gini for Online

$$
\begin{aligned}
Gini_{\,Online} &= 1 - \left(\frac{5}{8}\right)^2 - \left(\frac{3}{8}\right)^2 = 0.4688 \\
Gini_{\,notOnline} &= 1 - \left(\frac{3}{7}\right)^2 - \left(\frac{4}{7}\right)^2 = 4898 \\
&---\\
&So, \\
Gini_{\,Online} &= \frac{8}{15} * 0.4688 + \frac{7}{15} * 0.4898 \\
&= 0.479
\end{aligned}
$$

Boosting

Build a model by using weak models in series.

- Assign an equal weight to each of the data points.
- Provide this as input to the model and identify the wrongly classified data point.
- Increase the weight of the wrongly classified data points.
- Increase the weight of the wrongly classified data points and decrease the weight of correctly classified data points. And the normalize the weights of all data points.
- If (got required results) End else Goto step 1.
- End


Original Data
Weighted Data
Weighted Data
Classifer
Classifer
Classifer
Ensemble Classifer
...

AdaBoost

AdaBoost

Adaptive Boosting, is a machine learning algorithm that combines the predictions of multiple weak classifiers to improve the accuracy of binary classification models.

Row No	Gender	Age	Income	Illness	Sample Weights
1	Male	41	40000	Yes	1/5
2	Male	54	30000	No	1/5
3	Female	42	25000	No	1/5
4	Female	40	60000	Yes	1/5
5	Male	46	50000	Yes	1/5

$$Performance\ of\ the\ stump\ = \frac{1}{2}\log_e \frac{1 -\ Total\ Error}{TotalError}$$

$$\begin{aligned}\alpha &= \frac{1}{2}\log_e \left(\frac{1-\frac{1}{5}}{\frac{1}{5}}\right)\\ \alpha &= \frac{1}{2}\log_e \left(\frac{0.8}{0.2}\right)\\ \alpha &= \frac{1}{2}\log_e \left(4\right)\\ \alpha &= \frac{1}{2} * (1.38)\\ \alpha &= 0.69\end{aligned}$$

$$New\ Sample\ Weight = old\ weight * e^{\pm Amount\ of\ say(\alpha)}$$

$$\begin{aligned}New\ Sample\ Weight &= \frac{1}{5} * e^{-0.69}\\ &= 0.2 * 0.502\\ &= 0.1004\end{aligned}$$

$$\begin{aligned}New\ Sample\ Weight &= \frac{1}{5} * e^{-0.69}\\ &= 0.2 * 1.994\\ &= 0.3988\end{aligned}$$

Iteration 1

Update the sample weights

Row No	Gender	Age	Income	Illness	Sample Weights	New Sample Weights
1	Male	41	40000	Yes	1/5	0.1004
2	Male	54	30000	No	1/5	0.1004
3	Female	42	25000	No	1/5	0.1004
4	Female	40	60000	Yes	1/5	**0.3988**
5	Male	46	50000	Yes	1/5	0.1004

Then normalize the weights

Row No	Gender	Age	Income	Illness	Sample Weights	New Sample Weights
1	Male	41	40000	Yes	1/5	0.1004/0.8004 = 0.1254
2	Male	54	30000	No	1/5	0.1004/0.8004 = 0.1254
3	Female	42	25000	No	1/5	0.1004/0.8004 = 0.1254
4	Female	40	60000	Yes	1/5	**0.3988/0.8004 = 0.4982**
5	Male	46	50000	Yes	1/5	0.1004/0.8004 = 0.1254

Now group them

Row No	Gender	Age	Income	Illness	Sample Weights	New Sample Weights	Buckets
1	Male	41	40000	Yes	1/5	0.1254	0 to 0.125
2	Male	54	30000	No	1/5	0.1254	0.1254 to
3	Female	42	25000	No	1/5	0.1254	0.2508 to
4	Female	40	60000	Yes	1/5	**0.4982**	**0.3762 to**
5	Male	46	50000	Yes	1/5	0.1254	0.8744 to

Iteration 2

Row No	Gender	Age	Income	Illness
1	**Female**	40	**60000**	**Yes**
2	Male	54	30000	No
3	Female	42	25000	No
4	**Female**	40	**60000**	**Yes**
5	**Female**	40	**60000**	**Yes**

- Here 1, 4, and 5 are the same. They are taken from the previous wrong classification, as it had more weight.