greatlearning
Power Ahead

Ensemble Methods

Week 1 - Bagging

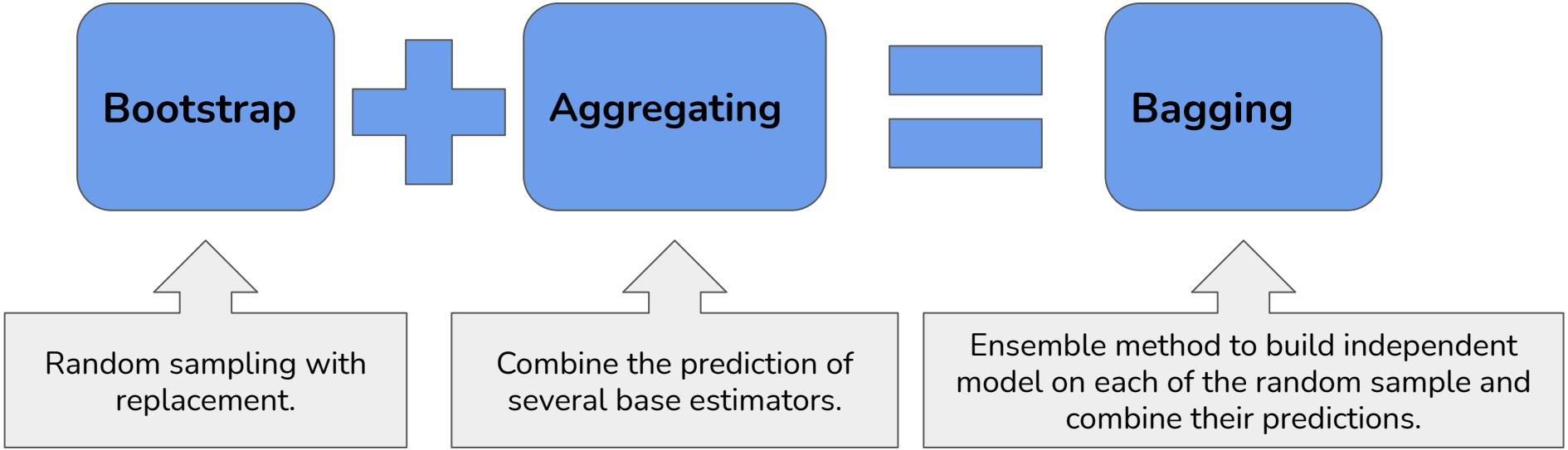
Discussion Questions

1. What is the main motivation behind Ensemble methods?
2. What is Bagging?
3. Why sampling with replacement?
4. How ~63% data gets selected in sampling with replacement?
5. What is Random Forest and what are steps involved in building a random forest?
6. What are the hyperparameters of Random Forest in SKlearn?

What is the main motivation behind Ensemble methods?

- Ensemble is a group of estimators that are used together for prediction in classification as well as regression problems.
- The motivation behind ensemble is the belief that a committee of experts working together are more likely to be accurate than individual experts.
- Ensemble method uses *n* number of base estimators and combines their output to give a final prediction giving better performance and robustness than a single estimator.
- For effective ensemble we have to ensure-
  - ❖ The base estimators are as different from each other as possible.
  - ❖ The errors made by each estimator should be different from each other (independent errors).

What is bagging?



- Each data point has equal probability of selection (1/n) at every stage.
- For classification, the final prediction is the mode of predictions of base estimators.
- For regression, the final prediction is the average of predictions from base estimators.

Why sampling with replacement?

Training data size

- The size of the subsets would decrease if the sampling is without replacement and without repetition you need to throw out a lot of data in order to get reasonable diversity in samples.

Independence b/w samples

- When we sample with replacement, any two sample values are independent. This helps to make sampling sets more independent, which in turn, helps to make base classifiers more independent/ uncorrelated

How ~63% data gets selected in sampling with replacement?

- Assume that we have n data points.
- What's the probability that a specific data point is not selected in n samples with replacement?

$$\left(1-\frac{1}{n}\right)^{n}$$

- If n goes to infinity then this probability is :

$$\frac{1}{e}=0.368$$

- In practice, even for n=100 the calculated probability is approx. 0.366.
- This shows that sampling with replacement guarantees 63% of samples get selected in each iteration.

What is Random Forest and what are steps involved in building a random forest?

- In Random forests, an additional random variation is added into the bagging procedure in order to create greater diversity amongst the resulting models.
- Only a subset of features are selected at random and the best split feature from the subset is used to split each node in a tree, unlike in bagging where all features are considered for splitting a node.
- Each tree is grown as follows:
  - ❖ Sample is drawn from the data set, with replacement and used for growing the trees.
  - ❖ If there are total m input features, m<M features are selected at random.
  - ❖ The best split on these ‘m’ features is used to split the node.
  - ❖ Each tree is grown to the largest extent possible or pre-defined depth.
  - ❖ The new data is predicted by aggregating the predictions of all the trees.

Hyperparameters of Random Forest (sklearn)

- **n_estimators**: The number of trees in the forest, ***default = 100.***
- **max_features***:* The number of features to consider when looking for the best split.
- **class_weight:** Weights associated with classes in the form {class_label: weight}.If not given, all classes are supposed to have weight one.
  For example: If the frequency of class 0 is 80% and the frequency of class 1 is 20% in the data, then class 0 will become the dominant class and the decision tree will become biased toward the dominant classes. In this case, we can pass a dictionary {0:0.2,1:0.8} to the model to specify the weight of each class and the random forest will give more weightage to class 1.
- **bootstrap:** Whether bootstrap samples are used when building trees. If False, the entire dataset is used to build each tree, ***default=True.***
- **max_samples:** If bootstrap is True, then the number of samples to draw from X to train each base estimator. If None (default), then draw N samples, where N is the number of observations in the train data.
- **oob_score:** The out-of-bag (OOB) error is the average error for each observation calculated using predictions from the trees that do not contain that observation in their respective bootstrap sample. This allows the RandomForestClassifier to be fit and validated whilst being trained**, *default=False.***

greatlearning
Power Ahead

Happy Learning !