Ensemble techniques / Random forest/Ada boost

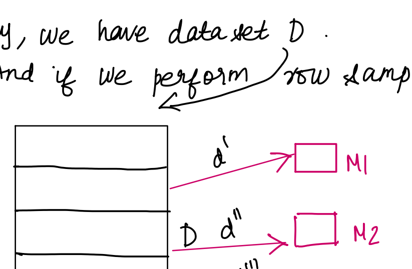
Sunday, 29 December 2024 14:47

Ensemble techniques:

Can we use multiple algo to solve a pbm?

Bagging | Boosting

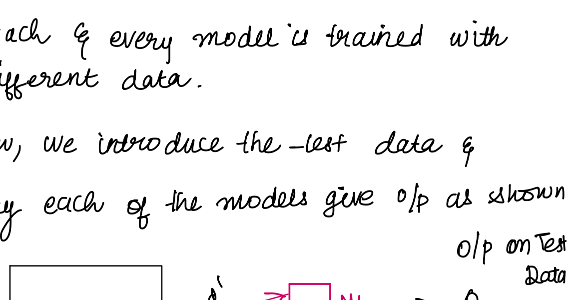
Say, we have data set D.
And if we perform row sampling to d', d'', d''' etc



We create multiple models M_1, M_2 etc of different algorithms using samples of data from D like d', d'' etc.

Each & every model is trained with different data.

Now, we introduce the test data & say each of the models give o/p as shown



Here we use majority voting to decide the o/p

This type of aggregation is called Bootstrap aggregation

For Regression pbm, average is taken as the o/p

Bagging - parallel combntn of models.

Boosting:

Its a sequential combination

→ M1 → M2 → M3 → M4 → o/p

All these are weak learners but combined together

Bagging	Boosting
① Random Forest classifier, Regression	① AdaBoost
	② Gradient Boost
	③ Xg boost

Random Forest Classifier & Regression:

Pbm with Decision Tree is that it tends to overfit. But how?

Happens when the tree becomes too deep leading to overly specific splits.

① splitting untill pure nodes - ie data points in a node belong to the same class or have same value. This results in branches that capture noise.

② Handling Noise: If the data contains noise the tree creates splits to accomodate these anomalies

③ Regularization is not done eg: max depth, min samples / leaf.

DT - Low Bias, High Variance } convert it to

Low Bias, Low Variance }

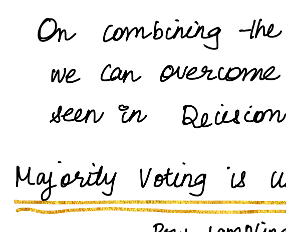
In Random Forest:

We have multiple models in parallel combination but all the models are decision Trees.

On combining the o/p of the models, we can overcome the pbm of overfitting seen in Decision Tree.

Majority Voting is used to get the o/p.

Row sampling + feature sampling



Both row sampling & feature sampling is done.

① Is normalisation rqrd for R.F?
No.
For Algorithms like KNN: Yes bcoz distance is calculated Euclidean or Manhattan.

② Random Forest not impacted by outliers.

We can also have custom Bagging by using the models we want.

Boosting techniques:

Say we have the below dataset:

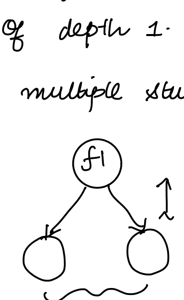
f1	f2	f3	f4	o/p	weight:
—					1/7
—					1/7
—					1/7
—					1/7
—					1/7
—					1/7
					1/7

1. We need to give weights to the rows so that on adding them up it equals to 1.

2. Now to decide which feature to start with, we consider Information gain & Entropy

3. We use feature 1 to create a decision tree of depth 1. - a stump

   such multiple stumps are created



4. We provide the entire data to the stump & train them.

   Say only 1 record o/p was predicted wrong by the stump.

5. Total Error $= \frac{1}{7}$

6. Check the performce of stump:

$$= \frac{1}{2}\log_e\left(\frac{1-TE}{TE}\right)$$

$$= \frac{1}{2}\log_e\left(\frac{1-1/7}{1/7}\right)$$

$$= 0.895$$

7. Update the weights:

   The weights are updated such that the correctly classified ones weights are reduced & the wrongly classified ones weights are increased. This ensures that it is passed on to the next weak learner model.

   New sample wt:

   Correct: weight $\times e^{-P_s} = \frac{1}{7} \times e^{-0.895} = 0.05$

   Wrong weight updates: weight $\times e^{P_s} = 0.349$.

New weight	Normlsd wt (divide by Σ New wt)
0.05	0.07
0.05	0.07
0.05	0.07
0.349 ↑	0.537
0.05	0.07
0.05	0.07
0.05	0.07.

Buckets: are created

ie [0 - 0.07]
[0.07 - 0.14]
[0.14 - 0.21]
0.349 [0.21 - 0.747] } This record has the highest bucket size
[0.747 - 0.751]

The model randomly creates numbers between 0 & 1 & whichever bucket the no falls in those records are picked. Hence records that are classified wrongly will have higher weights inturn leading to wider buckets & hence high Chance that it will get picked & go to second tree & contd.

We may have min 100 models..

Finally we pass the test data & use majority voting to find the o/p (In case of classifctn)

& In case of regression, avg o/p is considered as o/p.

White box Vs Black Box Model:
- Based on explainability.
Linear Regression - white
D.T - white
Random Forest - Black
ANN - Black box model.