Entropy vs Grini Impurity: Which to use When)

Whenever dataset small (1000,2000 records) -> Use Entropy
Whenever dataset large (IM - 200000, mone) -> Use Gini Impurity

Descin Tree For Numerical Split:

In category features we found the pure split and impure split by counting the target feature values (Yes/no). But what should use do with a continuous feature?

Example:	$\circ_{\mathbf{f_1}}$	0/P	and my surprise made in the
	2.3	Yen	Approach - kilgs 2009
	3.6	Yen	1) Sont the feature values
	4	No	1) Sont the teature values
	2.5	No	2) Set threshold value. 3) Take the minimum value as
	6.4	Yes	
	7.8	No	through 12 and solit (2-3)
	90	Yes	threshold and split (23)
			The second secon

4) Then check purity with entropy

5) Then do further split based on another feature

6) Then calculate total Information gain

7) (2-6) will be confinued for most values (3.6, 4,5.2...)

Yerlono SYer/3No SYer/3No

they to check pun

So, we will get information split from eny every f1 data values, the greaters the gain value, that value we will consider.

Problem: When we have a large dataset of millions of neconds.

Time complexity is extremely high to make decision tree for each continous bake and check in formation gain.

Post Pruring and Pre Pruring Decision Trees: 30711187 projector of

In decision Tree We keep splitting all the nodes till we get all the left node means till the level where we every node is a pure split

Notice and sput of the former value.

Notice of the programme value of the polyter of the polyte

But there is a big problem int in it. When I we split our training data like this, it usually leads to "over fitting" of the model.

ritemental botof stollates Test data Acc II 100/8

\$ 100 mill be continued

for next volues (36) 4, 12

CS CamScanner

To increase the accuracy of Tost am data, we use the two techniques

-> Prost Preuning (Reduce Overstitling)

1 Post Aruning

level 2

Sevel 2

Sevel 2

Sevel 3

Sevel 4

Sevel 4

should be the parameters, sample size,

We can see that, in level dis, one mode has 3 Yes and 2 No. So the probability we we are getting her. Yes is very high because of the high ration so, we don't need to turken split to find pure split as we also ady found the greater probability.

"Steps in post pruning";

- 1) Construct Decision Tree
- 2) Use mx Depth, feature numbers, paramaters, of scikitlearn to prune the true (To cut the brancher)

In post pruring we take the hyper parameter tuning Max Depth.

Prepruning:

In this case we first use hypon parameters tune in the model and it already says how much depth the tree should, what should he the parameters, sample size, features etc. I so before constructing the tree we do hyper purameter tuning and the parameters we get after that, we construct the tree using those.

Steps fore pree preuning

- 1) Do hypercharameter tuning
- 2) Construct Tree with the best parameters

Which one to use when?

Which one to use when?

For small, dataset to use post Pruning

For large Dataset to Use Pre Pruning and of the senance of the large Dataset to Use.

Preserving South of the Pruning and of the senance of the large Dataset to Use.

steps in post pruming ;

1) Construct Decision Truc

2) the more pople, technologies, poromologies of cilies leaves to prime the face (To out the bosoniers)

In post priming we take the hyper proporter timing new papers.