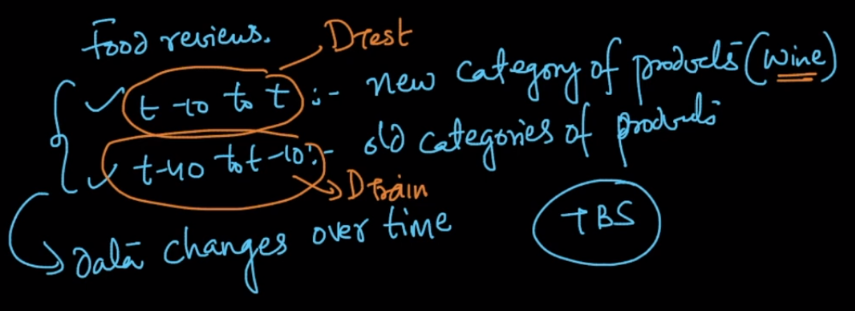
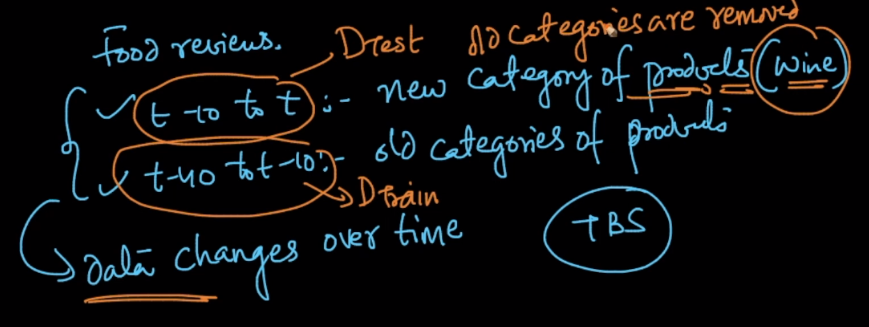
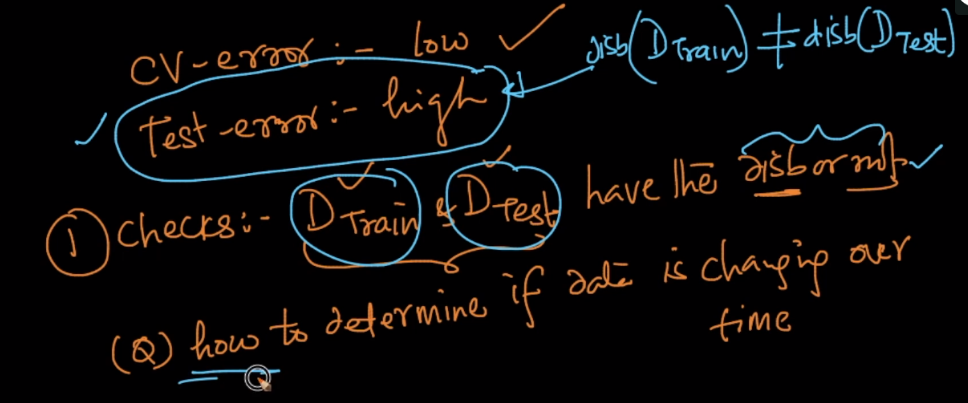


If we do train test split on basis of time or time base splitting, then it may be possible that in future some new features are added and old features are removed, then in such case distribution of training data and test data changes, as some features are not present in train, or some features that are not in test are in train, then in such case distribution of both train and test data would be different.

This will lead to random results on new querying point. So how to check if train and test have same distribution or not.





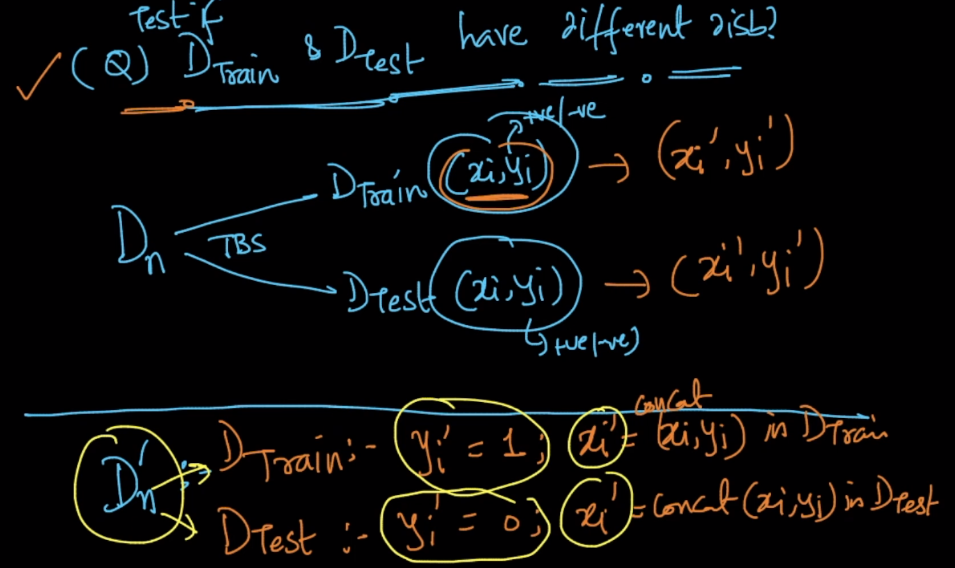


We’ll create a new train and test data as,

new\_train: features(xi) = concat(xi, yi) in old\_train and class label as 1;

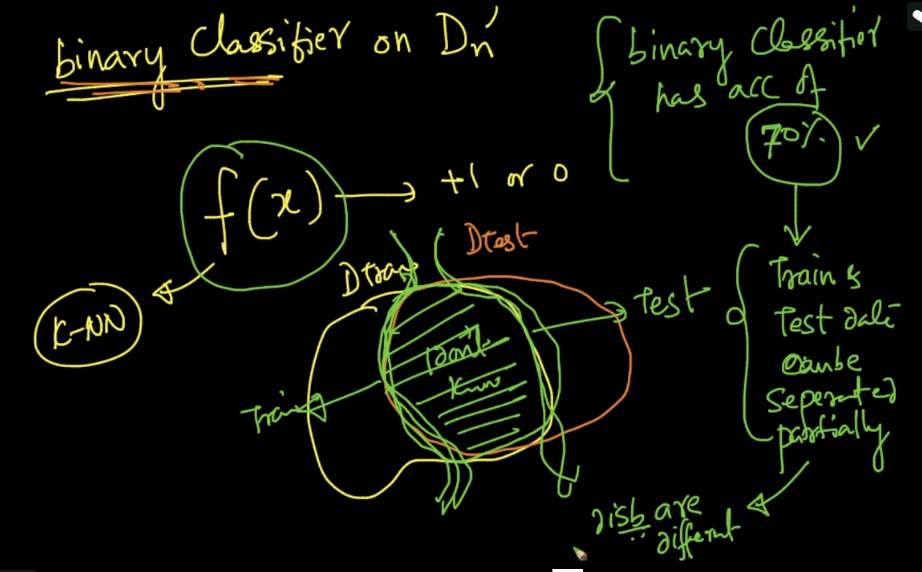
new\_test: features(xi) = concat(xi, yi) in old\_test and class label as 0;

now we’ll merge both, and create two samples from them and apply model on that.

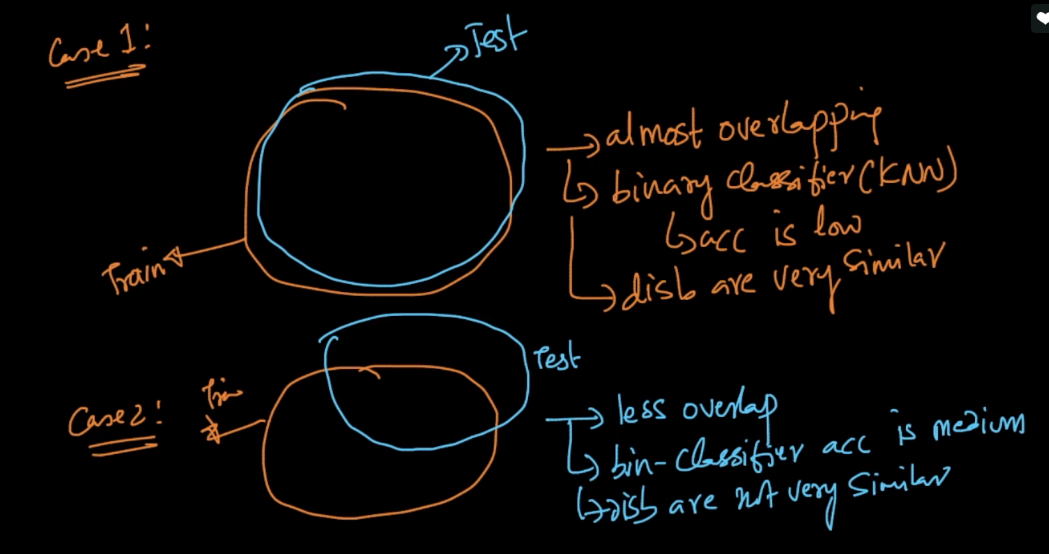


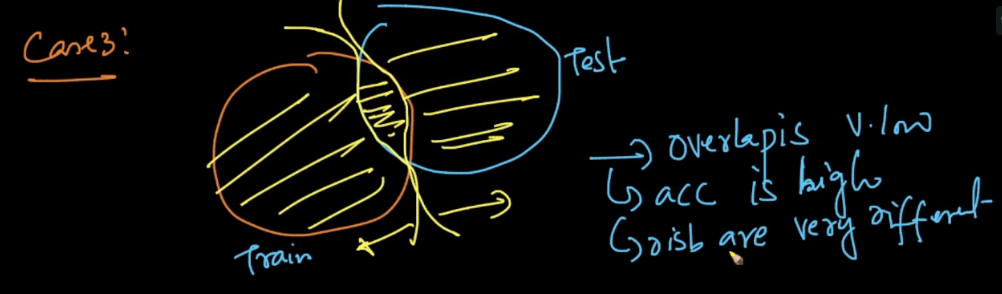
Now if, test accuracy is high that means they are very less overlapping that means they both have different distribution, because x\_test will be predicted 0 more no of times, and if they have to be same x\_test should be predicted 1 more no. of times.

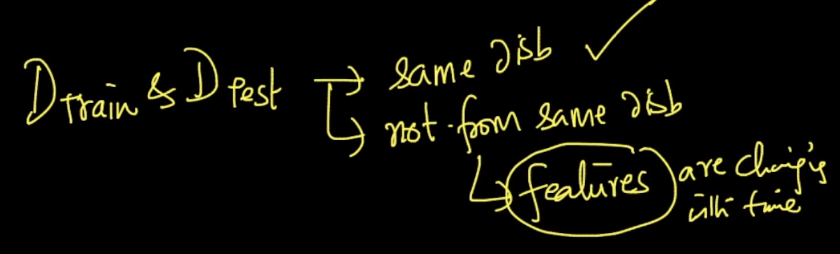
And if test accuracy is low that means they are very high overlapping that mean they both have same distribution.



Three cases for this can be observed, which is hown below.







If train and test data don’t have same distribution and features might be changing with time, and in such case there is need to change, design and build new features.

