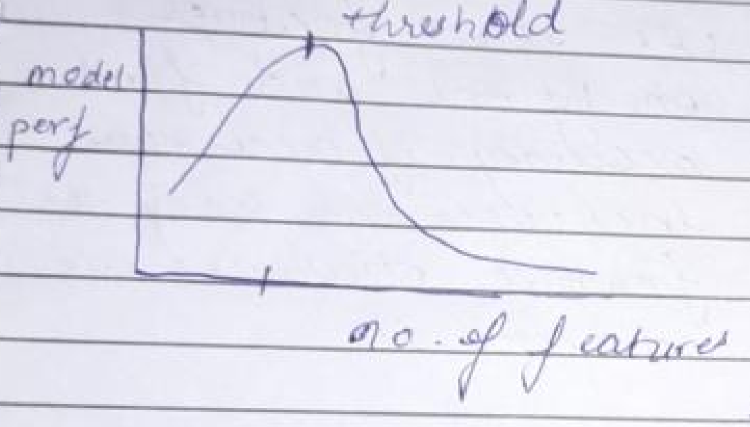
Feature selection

Choosing subset of features from all features.

Why we need feature selection?

Curse of Dimensionality :- If we have lot of features in our dataset, if we further increase no. of features we will end up hampering our model growth. Model will be learning from irrelevant data.



Bad features :-
1. irrelevant
2. redundant

Curse of Dim n
↓
dimensionality red n

feature selection
a, b, c . . . z / outp
↓
5 best features
e, g, z, v, f → o/p

feature extra
a, b, c . . . z
5 feature
az, ecq, dm, a, e
↳ PCA

Feature Selection Methods :-

① Filter Method

We have to filter out rqd features from rest of the features.

Checks relevance of feature with output variable.

All features → select best subset → Algo - ML

↓

CHI square test
ANOVA test
correlation coff

(2) Wrapper Method

└ forward selection └ backward elimination

ex Forward selection

A → ML → accuracy
model

AB → ML → accuracy ✓
↑
model

ABC → ml → accuracy X
X
model

We select features one by one and find accuracy, if accuracy increases we keep the feature otherwise we discard it.

Backward Elimination

ABCDE
↓
model

Test
- ANOVA
- CHI sq
- Pearson coff

Chi-squared → p-value
$P \leq 0.05$ → useful

(3) Embedded Method

It learns the feature while building a model
Takes all PMC of features and check with model performs best.