Feature Reduction

Reducing number of features makes the model faster

Performance of Model may improve when only important features are provided to the model.

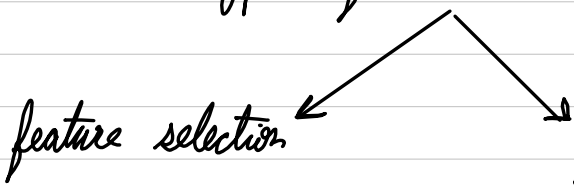
Makes data easier to visualize

Data is easier to store & obtain

Noise is removed

Reduces complexity of Model

Types of Dimentionality reduction

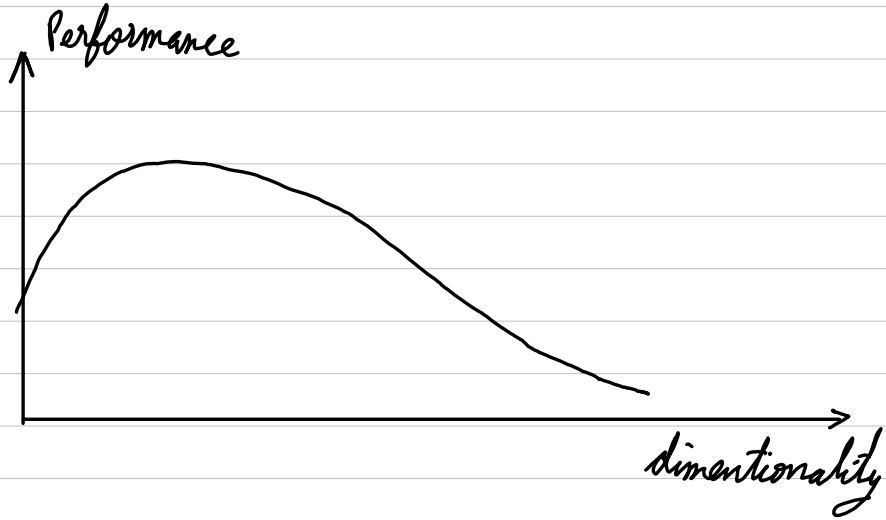


Select from existing dimentions

Make new varables of k dimentions

Curse of Dimentionality

Increasing number of features increases performance upto a point. Then the performance starts to degrade



Occams' razor: Remove all that is unnecessary

Unnecessary features → Remove (Dimentionality Reduction)

Unnecessary mapping → Remove (Regularization)

Feature extraction

Yeilds New data set from old dataset

Prevent redundancy

Prevent irrelevancy

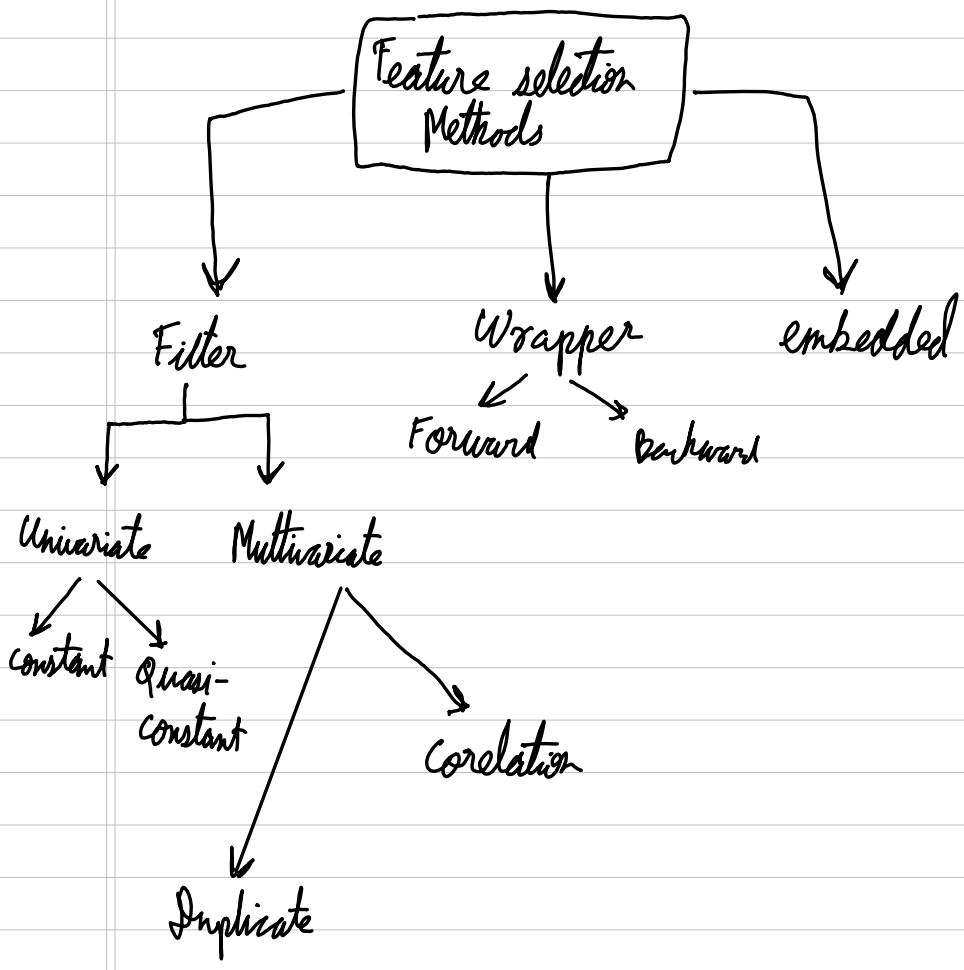
eg PCA

Feature selection

Features are selected & omitted

New space is not formed

More interpretable than Feature extraction



Filter Method

→ select features independant of any model

→ Rely only on characteristics of variables

→ Inexpensive methods very fast

→ Remove irrelevant & constant redundant features

Univariate filter → treat every feature indipendantly

A) Constant features

eg. gender of all patients is female, then no point keeping it in dataset

B) Quasi Constant

One value occupies majority records
eg. age > 50 for 98% patients then remove age.

Multivariate filter → Relation to other features

A) Duplicated features

B) Corellation filters

If two variables are highly corelated among themselves then they are redundant

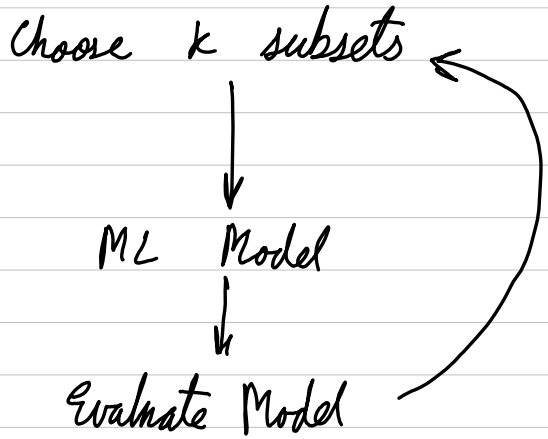
Use pearson corelation coficiant to calculate it

Wrapper Methods

Greedy search approach

Evaluate all combinations by evaluation creiterion like accuracy

Cross Validation based methods



High computational time required

High Overfitting

Feature selection : Forward selection

Start with 0 parameters

Take a feature & check

Take another feature & check

⋮

all features

Take 2 features & check

⋮

Check for combinations

Backward selection

start with all features

Remove features & check

Embedded Methods

Perform Feature selection during model training

learning algorithm takes advantage of its own variable selection process

More accurate like Wrapper

Fast like filter

less prone to overfitting

Train Model

↓

select features based on importance

↓

Remove Non important features

example RIDGE & LASSO Regularization