Polynomial Regression

- It is the form of linear Regression in which the relationship b/w the Independent variable (x) & dependent variable (y) is modeled as an n^{th} degree polynomial.
- Polynomial regression fits a nonlinear Relationship b/w the value of x & the Corresponding Conditional mean of y, denoted by $E(y|x)$

Why Poly-Regs?

- For Curvilinear hypothesis relationships.
- Inspection of Residuals (y axis), if we try to fit a linear model to Curved data, a Scatterplot of Residuals (y axis) on the predictor (x axis) will have patches of many +ve residuals in the middle. Hence, it is not appropriate.
- All Indp. variables being Indpt. is not satisfied in Poly Regs.

Uses:-

to know,

- Growth Rate of tissues
- Progression of disease epidemics
- Distribution of Carbon Isotopes in lake sediments.

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_m x^m$$

$$\Rightarrow y = \beta_0 + \sum_{i=1}^{m} \beta_i x_i^q + \text{(FP)} \rightarrow \text{Polynomial function.}$$

Advantage (using degree of polynomial as hyperparameter)

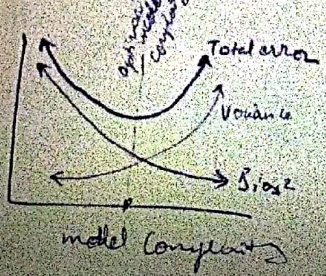
- Broad Range of function can be fit under it
- Polynomial Basically fits wide range of Curvature.
- Provides Best approximation of Relationship b/w dependent and Independent variable.

- High Bias indicates underfit
- To overcome underfitting, we need to Inc the complexity of the model
  i.e RMSE ↓ & R^2 ↑
- degree is (low) then error (↑)

Disadvantage

- Too Sensitive to Outliers
- Presence of 1 or more outliers can affect the results.
- Few availability of tools for detection of outliers in nonlinear regg than that for linear Regs.
- Chances of overfitting

- High variance means high degree of Polynomial.
  ⇒ training error is low for small no of training eg.
  ⇒ training eg ↑ error ↓
  (1)
- For high variance ↑ no of samples in the training set will dec the gap b/w Cross validation & training error



Polynomial Regression # 5 minute Egg.

- Here linear model is used on Non-linear data set.
- Relationship b/w dependent & Independent variable is non linear in nature
- Simple linear regg. eqn we know,

$y = \beta_0 + \beta_1 x_1$ ← linear regg.

$\rightarrow y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$ (It is 1 degree polynomial)

i.e

$\Rightarrow \boxed{y = \beta_0 + \sum_{i=1}^{n} \beta_i x_i}$ — (i)

- Check Independent's power, we'll get to know degree of polynomial

- 0 degree polynomial: $\boxed{y = \text{Constant}}$

1 degree polynomial: $y = mx^{(1)} + c$

2 degree polynomial: $y = ax^{(2)} + bx + c$ ∪

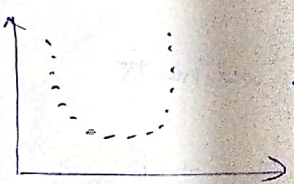
$$\boxed{y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \cdots + \beta_n x^n}$$

$$\Rightarrow \boxed{y = \beta_0 + \sum_{i=1}^{n} \beta_i x_i^{\#} + \boxed{F_p}}$$

(ii) ←

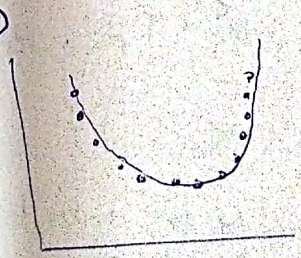
Polynomial function, it tries to add variables

we know, for the below we can't apply linear model ∴ if we want to apply linear model in this i.e in (i) we normally had this as a linear/multiple regg.



← isme humne 2 degree ki zeroorat hai :: 2 degree polynomial ∪ hota hai ; i.e parabolic function k form m hota hai

Standard eqn i.e $[x_1, x_2]$, Now for (ii) we see this F_p, F_p here will inc the power of variable ?i.e $[x_1, x_2, x_1^2, x_2^2]$, this F_p is the polynomial function, jisne add or power ko inc krne ka kaam kiya hai.



F_p → Kitne variable chahiye & un variables ki degree kitni chahiye taki linear model ko hum use kr ske for non linear data set.

SVR

SVM:-

- Stands for Support vector machine, is a classifier.
- Classifiers perform classification, predicting discrete categorical labels.

SVR

- Stands for Support Vector Regression, is a Regressor.
- Regressor performs Regression, predicting Continuous Ordered Variable.

SVR & SVM use very similar algorithms, but predict different type of variables.

In Simple regression we try to minimise the error rate. (pred - actual)

In SVR — fit the error within a certain threshold. (boundary)

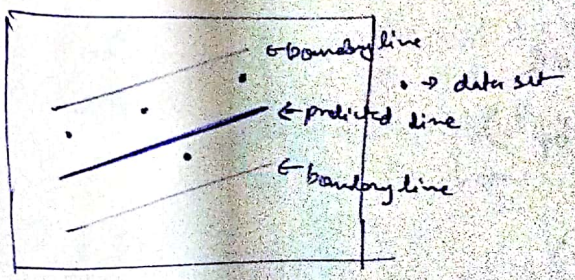
Important terms in SVR:-

Kernel: function used to map a lower dimensional data into a higher dimensional data. (RBF - Radial Basis Function kernel)

Hyperplane: In SVM, it is the seperation line b/w the data classes.
(predicted line) In SVR, it is the line that will help us predict the continuous (or target value.

Boundary line:- In SVM there are two lines other than hyperplane which creates a margin.
The support vector can be on the Boundary line or outside it.
The boundary line seperates the two classes.
In SVR it is same concept.

Support Vectors:- These are the data points which are closest to the Boundary.
The dist. of the pts is minimum or least.



Random Forest

- It is an ensemble classifier that consists of many decision trees and outputs the class that is the mode of the class's output by individual trees.

- It combines Breiman's 'bagging' idea & the random selection of features.

Decision trees

=> These are individual learners that are combined, then commonly used

=> One type of decision tree is called CART — Classification & Regression tree.

-> **CART** divides feature space into sets of disjoint rectangular regions

- Regions should be pure w.r.t. response variable
- Simple model is fit in each region — majority vote for classification, constant value for regression.

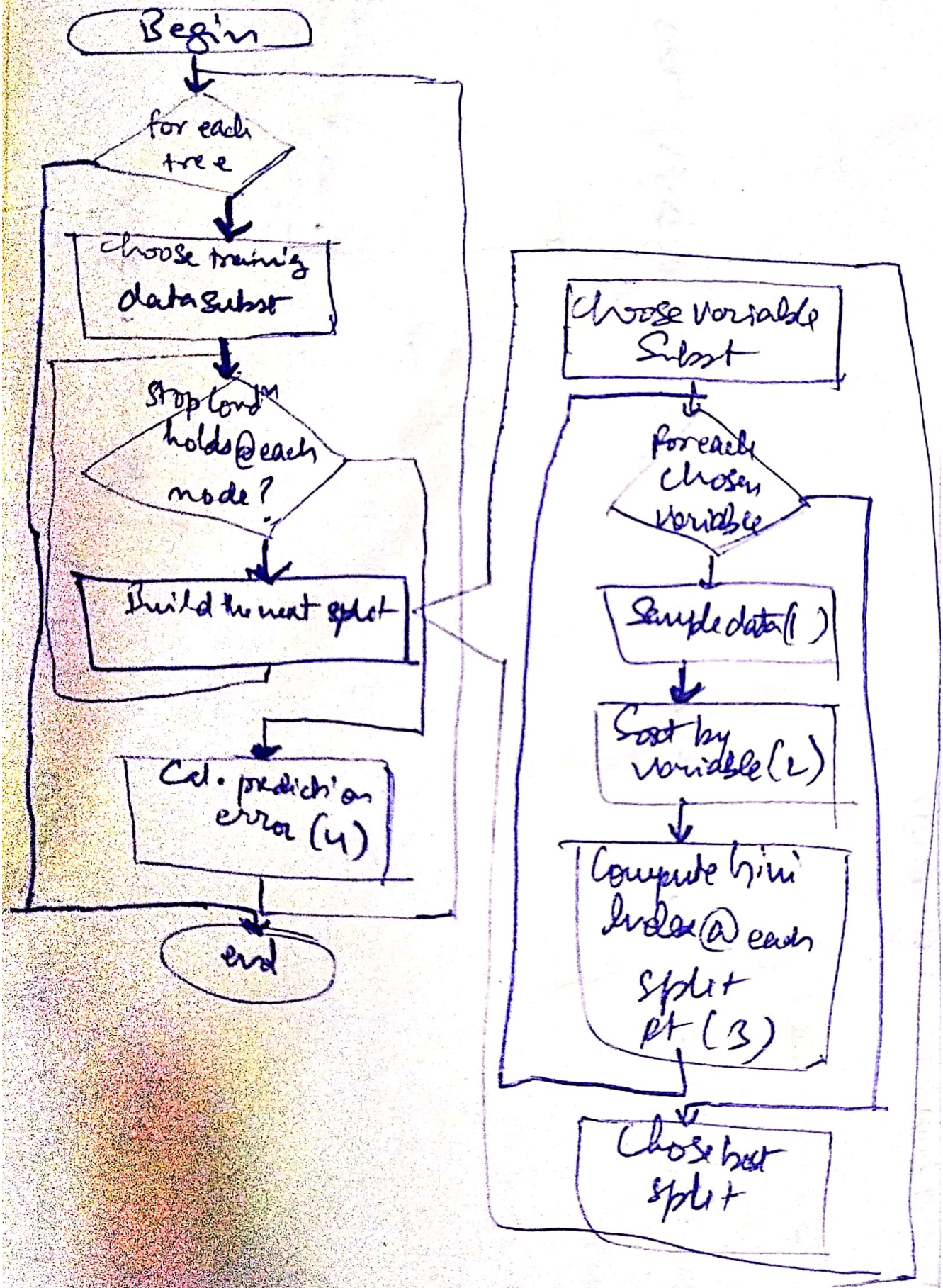
-> Decision tree involves greedy, recursive partitioning.

- Simple dataset with two predictors

TI	PE	Response
1.0	M2	Good
2.0	M1	bad
⋮	⋮	⋮
4.5	M5	?

- Greedy, recursive partitioning along TI & PE

Algorithm Flowchart



→ Splits are chosen acc. to purity measure (e.g. sq. error (regression), Gini index etc)

→ N is selected

Build trees until the error no longer dec

→ M is selected

Try to recommend defaults, half of them & twice of them & pick the best

Advantages of Random Forest

(1) Most Accurate learning algorithm available, Produces highly accurate classifier.
(2) Runs efficiently on large databases.
(3) Can handle thousands of input variable without variable deletion.
(4) Gives estimation of what variables are important in the classification.
(5) Generates an internal unbiased estimate of the generalization error as the forest building progresses.
(6) It has effective method for estimating missing data & maintains accuracy when a large proportion of data are missing.
(7) It has methods for balancing errors in class pop^n unbalanced data sets.
(8) Generated forest can be saved for future use on other data.
(9) It offers an experimental method for detecting variable interactions.

Disadvantages

(1) It has been observed to overfit for some datasets with noisy classification regression tasks.
(2) For data including categorical variables with diff. no. of levels, random forests are biased in favor of those attributes with more levels.

Random Forest estimating the test error

→ while growing forest, estimate test error from training samples.
→ For each tree grown (33% to 36%) samples are not selected in Bootstrap Called out of Bootstrap (OOB) samples.
→ Using OOB samples as input to the corresponding tree, predictions are made as if they were novel test samples.
→ Through book keeping majority vote, average regg., is computed for all OOB samples from all trees.

estimating the Importance of each predictor:-

→ Denoted by $\hat{e}$ the OOB estimate of the loss when using original training set, D.

→ For each predictor x_p where $p \in \{1, \ldots, k\}$

- Randomly permute p^{th} predictor to generate a new set of samples $D' = \{(y_1, x'_1), \ldots, (y_N, x'_N)\}$
- Compute OOB estimate $\hat{e}_k$ of prediction error with the new samples.

→ A measure of Importance of predictor x_p is $\hat{e}_k - \hat{e}$, the increase in error due to random perturbation of p^{th} predictor.