→ Decision Tree and Randomforest

→ Support Vector Machine

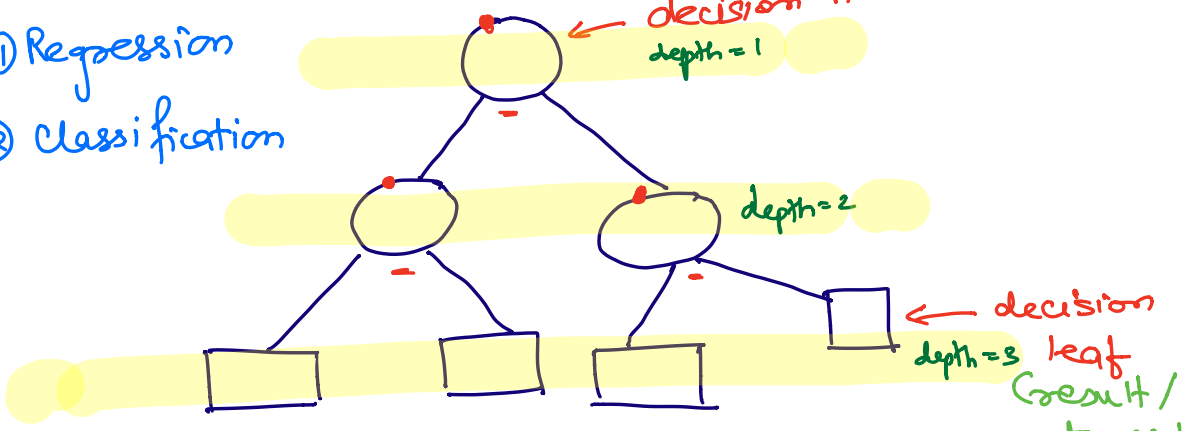
→ Feature Selection and Sampling Techniques

→ Unsupervised Learning (k-means algo).

Decision Tree: (CART - classifn & Regression Trees)

① Regression

② Classification



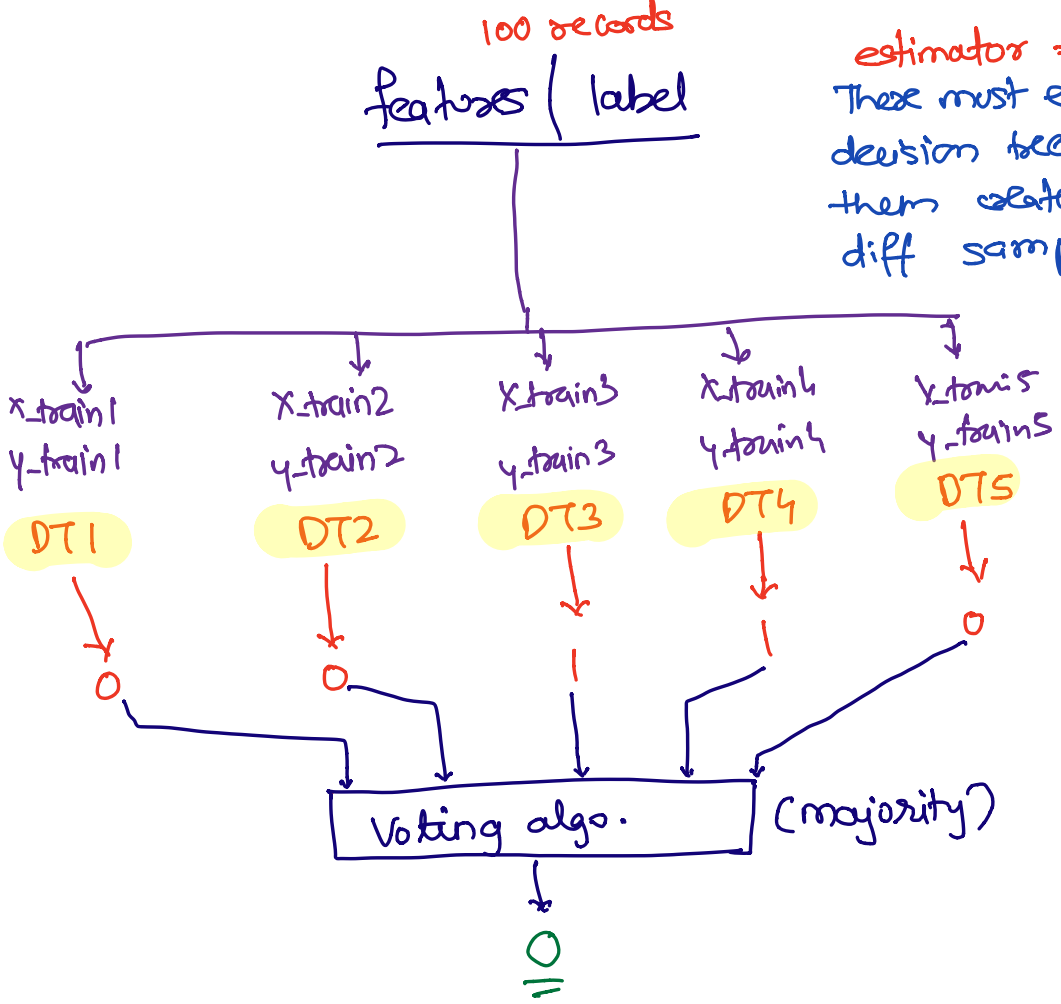
max_depth: ? .... Hit and try.
(max depth v/s accuracy)

Error function → Entropy (Gini) .... depth ∝ no of valid condn

RandomForest

→ Ensemble Learning
Create multiple models with random sample and get the best result based on majority vote.

Random forest is all about creating multiple Decision Trees with different samples.



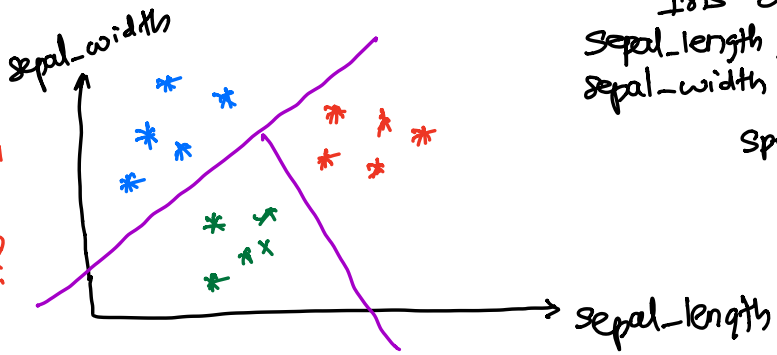
Support Vector Machine

- Works best when you have more than 2 feature columns. (multi-dimensional space)

- SVM → Regression, Classification } All supervised methods.

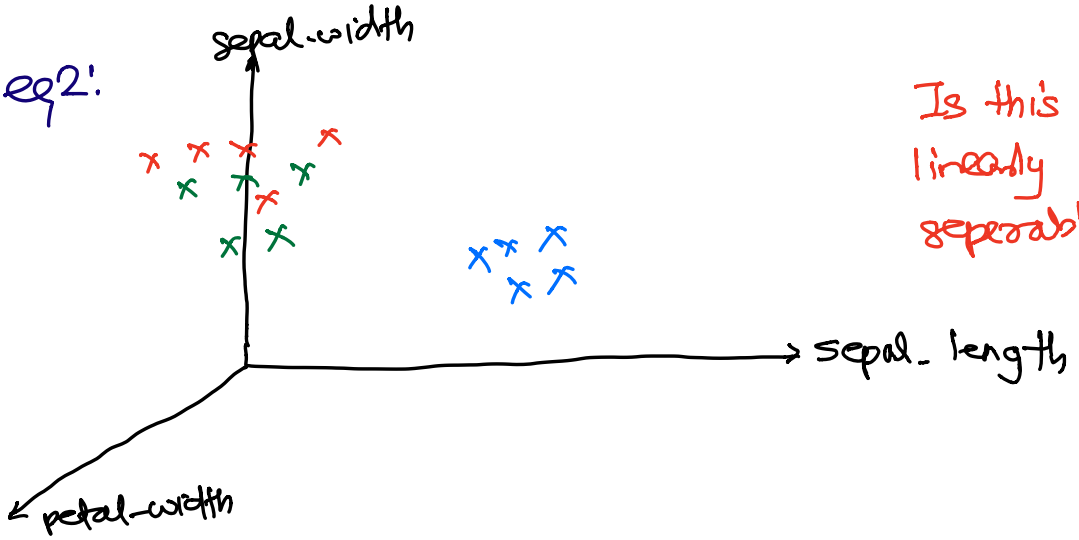
eg 1:

Is this data linearly seperable?



Iris dataset
sepal_length, petal_length
sepal_width, petal_width
species
→ Setosa
→ Versicolor
→ Virginica

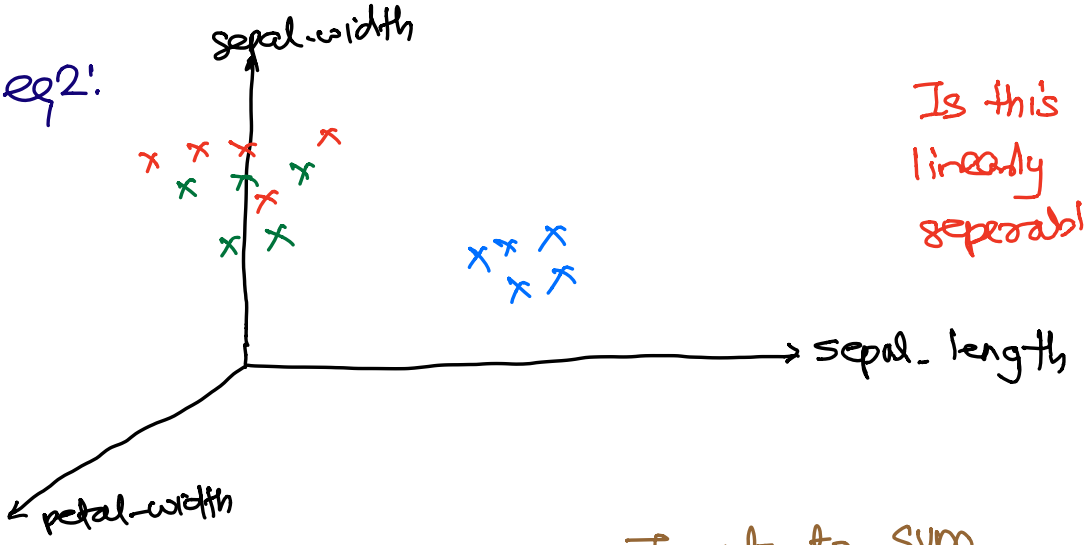
eg 2:



Is this data linearly seperable?

⇓ SVM

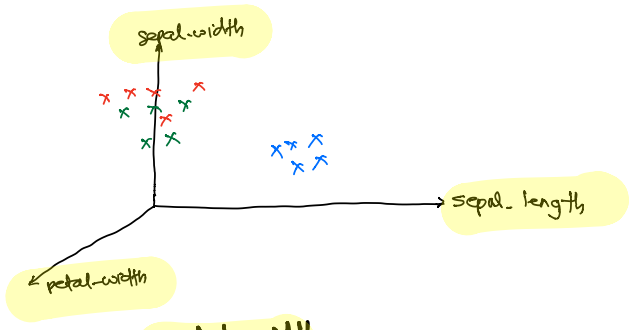
eg2:



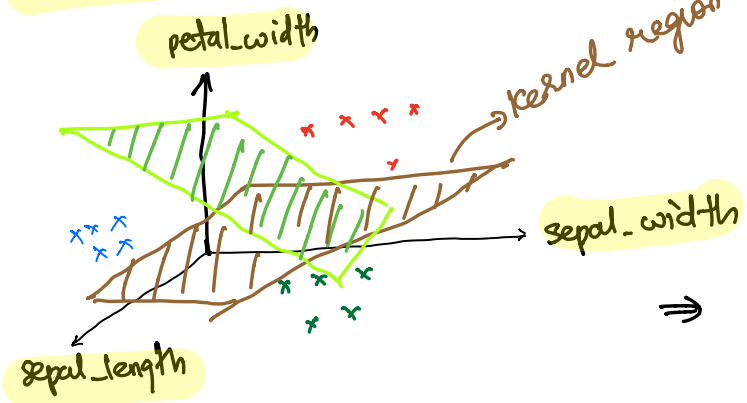
Is this data linearly seperable?

Input to svm

svm tries to figure out can I make this data linearly seperable by checking all possible dimensions?



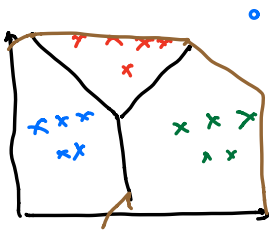
→ View 1

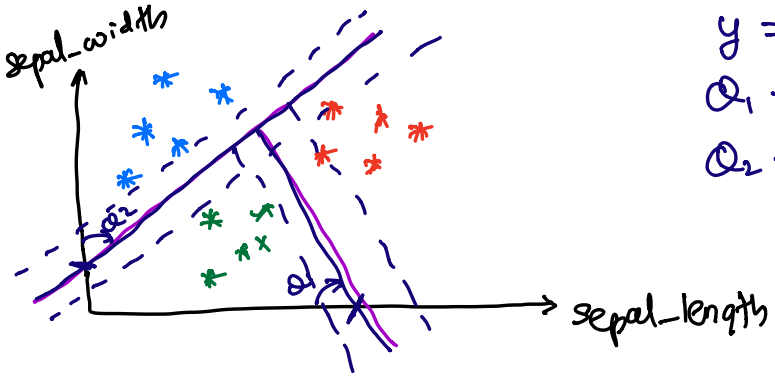


SVM

→ View 2

⇒





$x = ?$
$y = ?$
$\theta_1 = ?$
$\theta_2 = ?$

How to figure out if my data can be linearly seperated ?

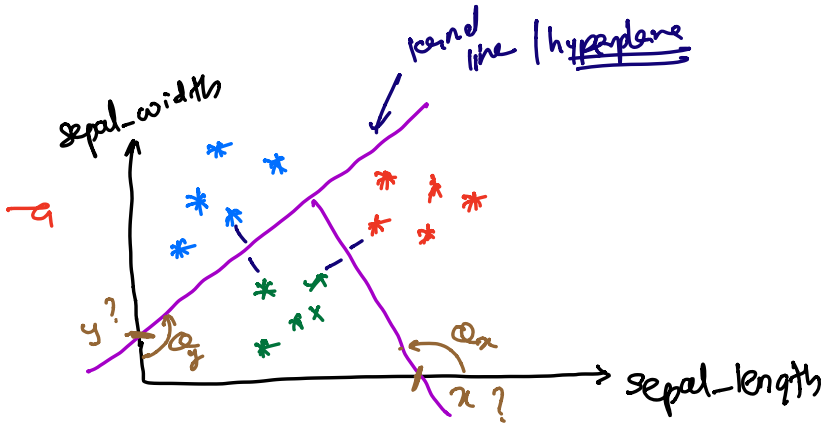
SVM

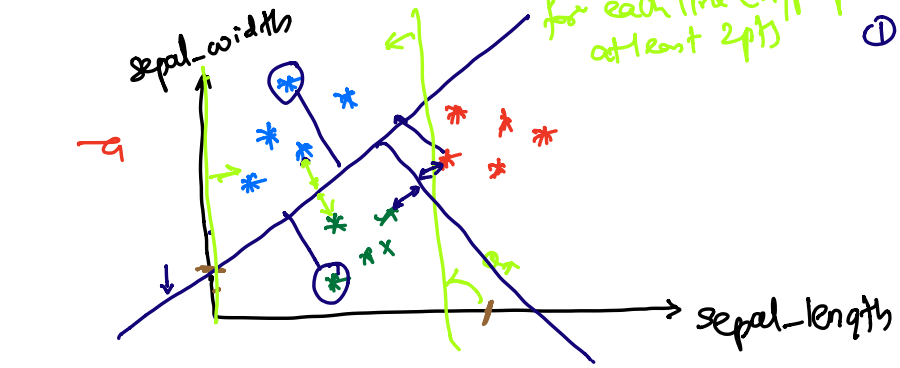
I will check all possible dimension to figure out if in any dimension, the data is linearly seperable or not ?

data is seperable
SVM will create the best model.

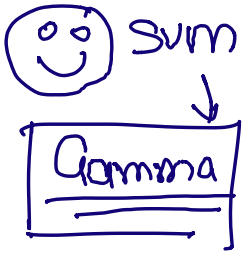
How to ensure my seperator line (hyperplane) is placed in correct & generalized location ?

$x \Rightarrow ?$ $y \Rightarrow ?$ $\theta_x \Rightarrow ?$ $\theta_y \Rightarrow ?$





gamma is high !



High → nearby pts from line

Low → far pts from line

for each line (hyperplane) atleast 2 pts

① create one line where $x=0$, $y=0$.
Increment x till I receive atleast one separation

② change α_y such that the distance both two label pts is same

Feature Selection and Sampling Techniques

Questions we are trying to address:

① Is it mandatory to use all columns in the given dataset as feature or is there a way to figure out which feature column is the best for the given algo?

② Is the sample I am using for model creation, the best sample? KFold, StratifiedShuffleSplit

③ What can be the maximum accuracy I can get from the given dataset? cross_val_score()

LinearRegression

Best features → corr()
→ OLS (Backward elimination.

what is the best accuracy I can get? → Cross validation for accuracy score. (This technique is applicable for all supervised learning algos)

Cross validation Technique for identifying best score for the given algorithm.

if you are using Kfold,

Split depends on Size

Feature Selection

* Based on Statistics and not Domain.

Iris → sepal_length, sepal_width, petal_length, petal_width.

Best / Worst

Statistics

① Correlation matrix } Regression

② OLS → Backward Elimination } Regression

③ RFE → Recursive Feature Extraction } Classifn Regression
R → LR, DT, RF, SVR
class → DT, RF

④ feature_importance } classification → DT, RF

⑤ PCA → Principal Component Analysis } Reg. Classfn
↓ Principal Component Analysis
↓ Doesn't require any model at all. } Principal Component algo. → Statistical algo
↳ Use me anywhere!