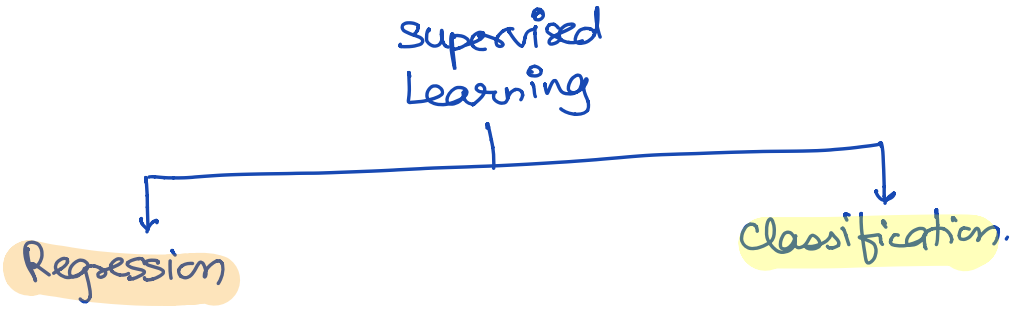
Agenda :

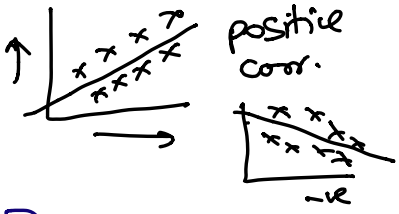
Support Vector Machines

Linearity Behavior in a dataset



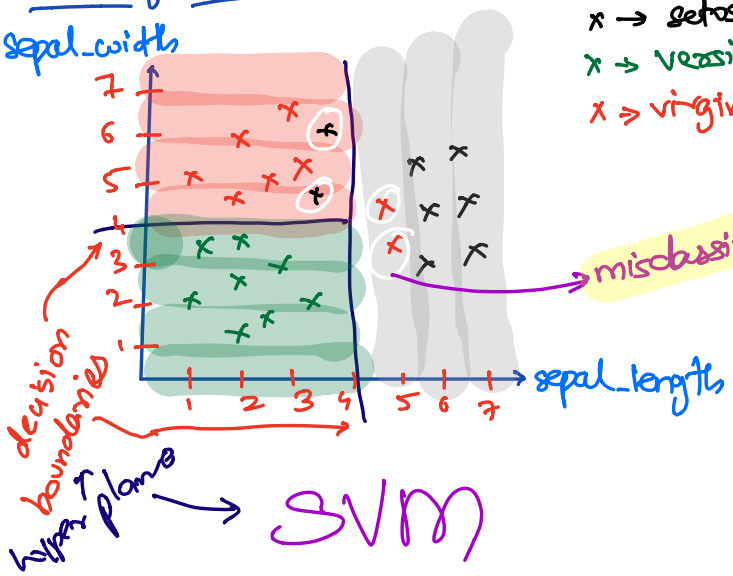
Is your dataset linearly seperable or not

If there exists any kind of linear relationship in your dataset (Regression) → corr()



Classification Problem (iris.csv)

x → setosa (0)
x → versicolor (1)
x → virginica (2)



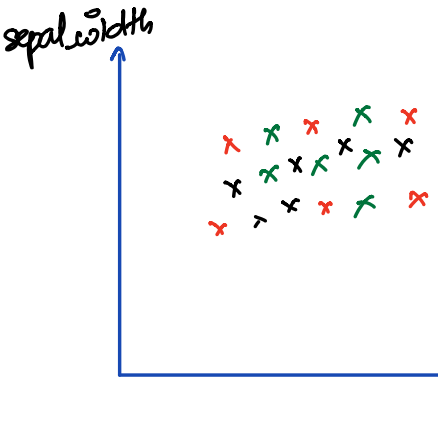
Is this data linearly seperable or not ?

SVM

x <= 4 & y <= 4 → versicolor.

x > = 4 → setosa

x <= 4 & y > 4 → virginica



K-NN

Is this data linearly separable?

No-linear data (expert to deal with (KNN)

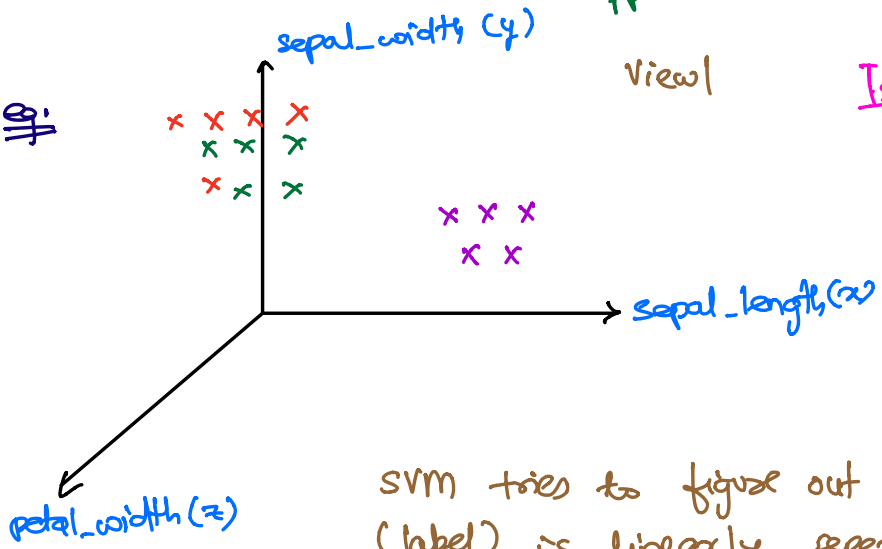
What is Support vector machine?

→ more than 2 features columns. (multi-dimensional space)

→ SVM → Regression → SVR (Support Vector Regressor) → linearity ↑ accuracy ↑

→ Classification → SVC (Support Vector classifier)

Eg:



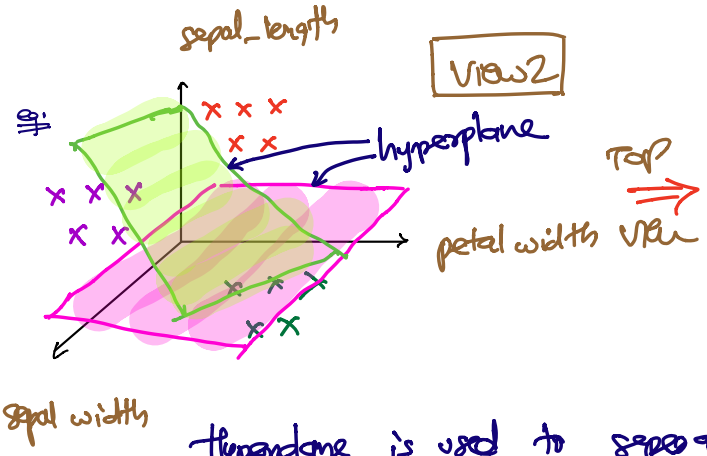
View!

Is this data linearly seperable or not?

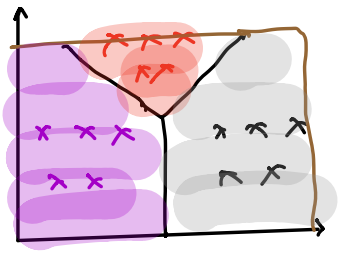
↑

answered by SVM

SVM tries to figure out whether the data (label) is linearly seperable or not by checking all possible dimensions (views)?



Engineering Drawing Top view.



Hyperplane is used to separate regions

SVM tries to place the hyperplane in such that misclassification of data is very less.

Some of questions SVM tries to answer is,

① How to figure out whether my data is linearly separable or not?

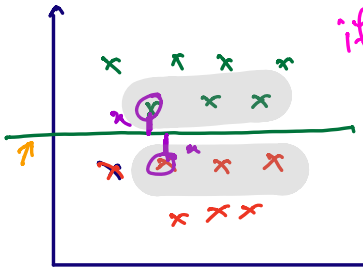
SVM → I will check all possible dimensions to figure out if in any dimension, the data is linearly separable or not.

② How to ensure my hyperplane is placed in the correct & generalized location?

SVM → He has a parameter called gamma (sklearn)

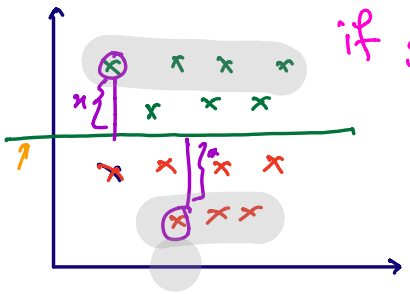
High (nearby pts)

Low (far pts)



decision of hyperplane placement is done using support vectors.

The points which is responsible to place the hyperplane at correct location is called SUPPORT VECTOR.



Sampling Techniques.

↑ how the data is picked and maintained in training & testing set.

Sampling Techniques helps us to understand whether the data has all possible combinations ∧ Population.

Questions we are trying to address.

→ accuracy } score.
→ generaliz^n }

① Is it mandatory to use all feature columns in the given dataset or can we remove features such that population & prediction both are not affected?

SelectByModel, RFE, ANOVA, OLS (regression)

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

→ KFold validation
→ StratifiedShuffleSplit.

③ What can be the maximum accuracy I can get from the given best sample ?

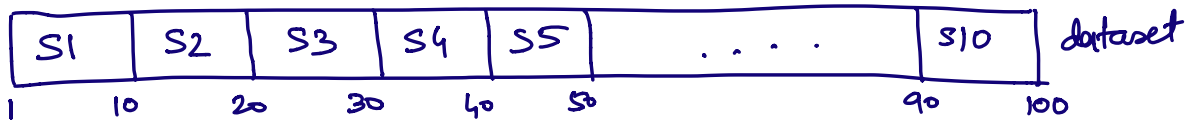
↳ cross validation score.

Cross validation

cv = 10 ⟶ dataset (100 records)

no of splits & no of iterations

(feature, label)



10 iterations :

1st iteration S1 → testing
remaining → traing
} → model → } score — ? score1

2nd iteration S2 → testing
remaining → traing
} → model → } score — ? score2

⋮

10th iteration S10 → testing
remaining → traing
} → model → } score — ? score10

Result :

[score1 , score2 , . . . , score10]