Unit-I

Machine Learning

Date 4/08/2022
Page

Intro. ML, Types, Linear Regression, classification & logistic regression, DT & RF, Naive Bayes & SVM, Applications of ML

Regression Analysis (RA)

A predictive modelling technique that investigate relationship b/w dependent (target) & independent variable (predictor).

Used for forecasting & finding relation b/w variables.

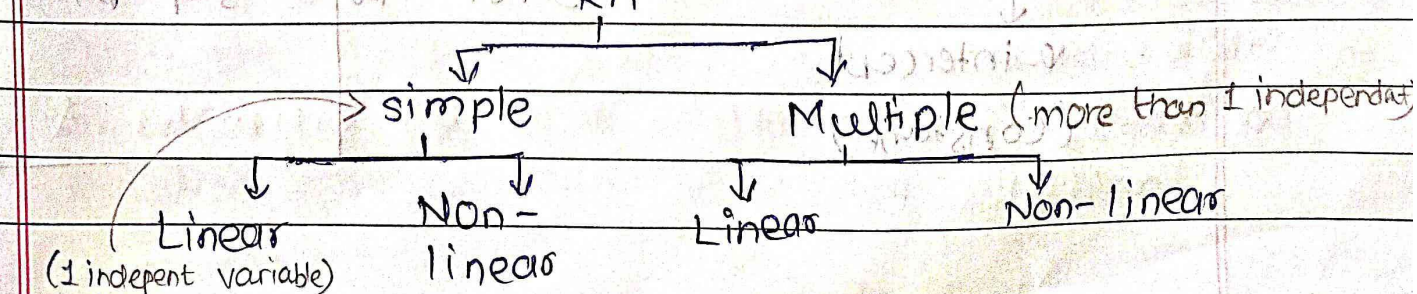
E.g. relationship b/w rash driving & no. of road accidents by driver

In RA, curve/ line is fitted to data pts., in such manner that differences b/w distances of data pts. from curve or line is minimized.

Benefits

Indicates significant relationship b/w dependent & independent variable.

Indicates strength of impact of multiple independent variables on dependent variables.



$\hat{y} = b_0 + b_1 x$

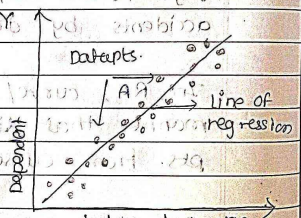
$b_1 = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2}$

Linear Regression - Continous

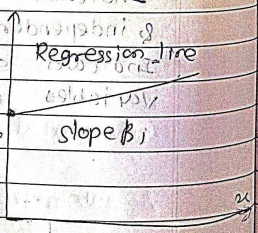
Most widely known modelling technique.
A linear approach to modelling the relationship blw a dependent variable & one or more independent variables using best fit straight line (known regression line)
In this technique dependent variable is continous
independent ——— " ——— or discrete
nature of regression line is linear.

Simple Linear Regression

Objective fit
To find best-line called regression line through the distribution which is nearest to most of pts.



Example
Predict height from age
House price
Sales in upcoming months



$Y = \beta_0 + \beta_1 x_1 + \epsilon$ → random error
↓ β_0
Y-intercept (constant)
↓ slope β_1

$Y = bx + a$

$a = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2}$

$b = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}$

Multiple Linear Regression

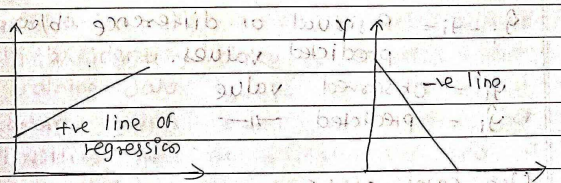
It is generalization of simple linear regression that has more than 1 independent variable, & 1 dependent variable.

$Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \epsilon$

Types of LR line

1) +ve Linear relationship
If dependent variable ↑ses on Y-axis & independent vari. ↑ses on X-axis.

2) -ve Linear relationship.
If dependent variable ↓ses Y-axis & independent vari. ↑ses on X-axis.

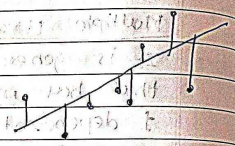


LSE Method

- common method used for fitting regression line.
- Calculates best fit line for observed data by minimizing sum of squares of vertical deviations from each data pt. to line.

Note: LSE method that builds models & RMSE is metric that evaluate model's performance

$$R^2 = \frac{\sum (\hat{y} - \bar{y})^2}{\sum (y - \bar{y})^2}$$



If R^2 close to 1, → good fit
R^2 close 0 → no relationship

RMSE

Checks whether model is good enough to predict in future.

RMSE tells measure of dispersion of predicted values from actual values.

$$RMSE_{rems} = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n}}$$

$\hat{y}_i - y_i$ = Residual or difference b/w actual & predicted values.
y_i = observed value
$\hat{y}_i$ = predicted —11—

Use cases of LR

1) Sales of product, pricing, performance, risk parameters.
2) Generating insights on consumer behavior, profitability & other buisness factors.
3) Evaluation of trends, making estimates & forecasts.
4) Determining marketing effectiveness, pricing & promotions on sales of product.
5) Assessment of risk in financial services & insurance domain
6) Studying engine performance from test data in automobiles
7) Calculating causal relationship b/w parameters in biological sm.
8) Conducting market research studies & customer survey results analysis
9) Astronomical data analysis
10) Predicting house prices with rise in sizes of houses

Classification

Classification algo is supervised ML technique used to identify category of new observations on training data basis.
Program learns from given dataset or observations & classifies new observation into no. of classes, such as Yes or No, 0 or 1, spam or not spam, cat or dog.
o/p variable is category, not value such 'G or B'.
A discrete o/p function (y) is mapped to i/p variable (x). $y = f(x)$

Algo. which implements classification on dataset is called classifier.

2 types

1) Binary classifier
If classification problem has only 2 possible outcomes

2) Multi class classifier
More than 2 outcomes
Eg: classification of types of crops

Learners in classification problem

1) Lazy learners: Firstly stores training dataset & wait until receives test dataset.
The classification is done on basis of most related data stored in training dataset.
Takes less time in training but more time in prediction.
Ex:- K-NN.

2) Eager Learners: They develop classification model based on training dataset by reciving test data.
More time learning, less time prediction.
Ex: DT, NB, ANN.

Types of classification problems

Linear models	Non-linear model
logistic regression	K-NN
SVM	Kernel SVM
	NB
	DT, RF

Evaluating model
Log loss or cross-entropy
confusion matrix
AUR - ROC curve.

1) Log loss / cross entropy
for evaluating performance of classifier, whose o/p is probability value b/w 0 & 1
for good binary classification, value should be near to 0
The lower log loss represents higher accuracy of model

2) Confusion Matrix / error matrix:

gives table / matrix as o/p & describes the performance.

gives predictions in summarized form, which has total no. of correct predictions & incorrect predictions.

	Actual +ve	Actual -ve
predicted +ve	TP	FP
predicted -ve	FN	TN

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

$$\text{Recall (R)} = \frac{TP}{TP+FN}$$ → measures model's ability to detect +ve samples

Sensitivity / Precision (P) $= \frac{TP}{TP+FP}$ → it is +ve predictive value that reflects how reliable model is in classifying samples as +ve.

$$\text{F-1 score} = \frac{2 \times (P \times R)}{(P+R)}$$

Use case of classification Algo:

Email spam detection
Speech recognition
Identification of cancer tumor cells
Drugs classification
Biometric identification

AUC-ROC

It is graph that shows performance of the classification model at diff. thresholds.

To visualize performance of multi-class classification, use AUC-ROC

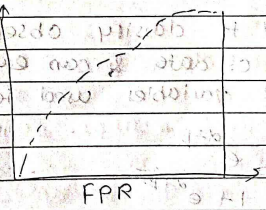
ROC curve plotted with TPR & FPR

Recall ← TPR → Y-axis → $\frac{TP}{TP+FN}$

FPR → X-axis $\quad FPR = \frac{FP}{FP+TN}$

ROC → probability curve

AUC → degree of separability.



In logistic, logit transformation is applied on odds – prob. of success / prob. failure. (log odds)

LR estimates probability of event occuring based on data

Logistic Regression – categorical

for predicting categorical dependent variable using given set of independent variable.

o/p categorical or discrete (Y or N, 0 or 1)

But instead of giving exact value as 0 or 1, it gives probabilistic value b/w 0 & 1.

In this regression, instead of fitting regression line, we fit an "S" shaped logistic function which predicts 2 max. values (0 & 1)

The curve from logistic function indicates the likelihood of something such as whether cells are cancerous or not, whether a mouse is obese or not based on its weight etc.

It is ML algo. bcoz it has ability to provide probabilities & classify new data using continous & discrete datasets.

It can used to classify observations using diff. types of data & can easily determine most effective variables used for classification

dependent variable

$$P(Y|x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$

$$\text{Logit}(y) \quad G(x) = \frac{1}{1+e^{-y}}$$ Sigmoid

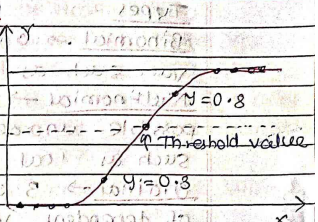
LR doesn't require linear relationship b/w dependent & independent variables. It can handle various types of relationships bcoz it applies a non-linear log transformation to predicted.

It requires large sample size bcoz max. likelihood estimates are less powerful at low sample size than ordinary least square.

Logistic function (Sigmoid)

Mathematical funct used to map the predicted values to probabilities.

It maps any real value into another value within range of 0 & 1, which can't go beyond this limit.



We use concept of threshold value.

Assumption

Logistic regression eqn can obtained from LR eqn. Mathematical steps to get eqn given as

Eqn of straight line (independent)

estimated via MLE $y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_n x_n$

In logistic, y can be b/w 0 & 1 only, so for this let's divide above eqn by (1-y)

$$\frac{y}{1-y}: \begin{cases} 0 & \text{for } y=0 \\ \infty & \text{for } y=1 \end{cases}$$

But need range b/w $-\infty$ to $+\infty$, take log.

$$\log\left[\frac{y}{1-y}\right] = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_n x_n$$

MLE test diff values of β thro. multiple iteration to optimize for best fit.

Types

Binomial → only 2 possible types of dependent var. such as 0 or 1, Pass or fail

Multinomial → there can be 3 or more possible unordered types of dependent var. such as 'cat', 'dog' or 'sheep'.

Ordinal → 3 or more possible ordered types of dependent variables such as 'low', 'medium' or 'high'. Ranges 1 to 5

Use

1) Trying to figure out who will win election
2) Whether student will pass or fail exam
3) Whether customer will come back
4) Whether email is spam
5) A credit card company want to know, whether transaction amount & credit score impact prob. of given transaction being fraudulent

Advantages

1) Makes no assumption about distribution of data in feature space.
2) Easily extended to multiple classes
3) Quick to train
4) Very fast at classifying unknown records
5) Good accuracy for many simple datasets

Decision Tree → Non-linear classification both

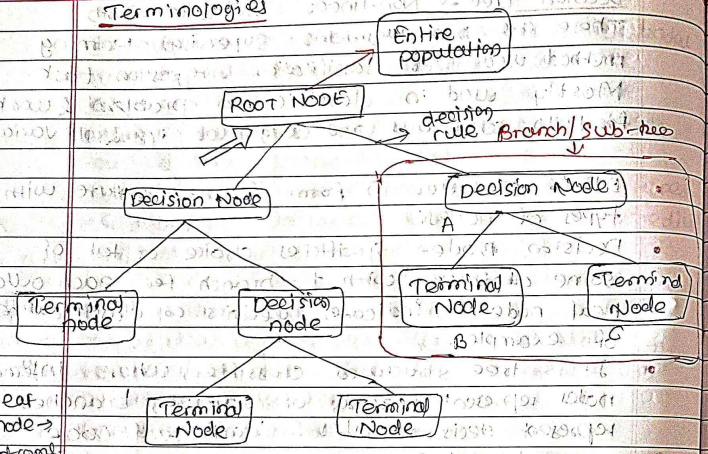
These are non-parametric supervised learning method used for classification & regression task.

Mostly used in classification problem & works for both continuous and categorical i/p & o/p variable

- DT is classifier in form of tree structure with 2 types of nodes:
- Decision node → Specifies choice or test of some attribute with 1 branch for each outcome
- Leaf node → Indicates decision or classification of example.
- It is tree structured classifier, where internal nodes represent features of dataset, branches represent decision rules & each leaf node represent outcome.
- The decisions or tests are performed on basis of features of dataset.
- It is graphical representation for getting all the possible solns to problems / decision based on given conditions.
- A DT simply asks question & based on the answers (Y/N), it further splits the tree into subtrees.

internal nodes represent → features of data

Terminologies



leaf node → opt-comes

Pruning: Removing sub-node of decision node

DT Algorithms

ID3 → (Iterative Dichotomizer 3)
C4.5 → (successor of ID3)
CART → (classification & regression tree)
CHAID → (chi-square automatic interaction detection performs multi-level splits when computing classification trees)
MARS → multivariate adaptive regression splines

0 to 1

Attribute selection measures

While implementing DT, main issue arises that how to select best attribute for root node & for sub nodes.
So, to solve such problems there is technique which is called Attribute selection measure or ASM.
By this measurement, we can easily select the best attribute for nodes of tree.
2 techniques for ASM:
1) Info. gain
2) Gini index

Info. Gain

It is measurement of changes in entropy after segmentation of dataset based on attribute.
It calculate how much info. feature provides us about class.
According to value of info. gain, we split node & build DT.
DT algo. always tries to maximize value of info. gain & node/attribute having highest info. gain is split first.

Info. Gain = Entropy(S) − [(weighted avg) * Entropy (each feature)]

info. theory metric that measures impurity / uncertainity in observations. It determines how DT chooses to split data

Entropy: It is metric to measure impurity in given attribute. It specifies randomness in data.

Entropy(S) = $-P(yes)\log_2 P(yes) - P(no)\log_2 P(no)$

↓

Total no. of samples

to construct DT

← **Gini Index**:- It is measure of impurity or purity used while creating DT in CART. An attribute with low Gini index preferred compared to high Gini.

It is found that determine how well DT was split.

It only creates binary splits & CART uses Gini index to create

Gini Index $= 1 - \sum_j p_j^2$

Major factors considered to improve DT performance.

- **Choosing splitting attribute**: By examining data in training set & getting i/p from domain experts.
- **Ordering splitting attributes**: Order in which attributes are choosen is important.
- **Tree structure**: A balanced tree with fewest levels is desirable.

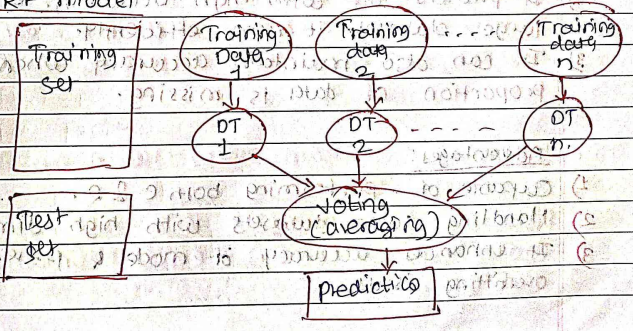
- **Stopping criteria**: The creation of tree stops when training data are perfectly classified. In addn, stopping earlier may be performed to prevent overfitting.
- **Training data**: If training dataset is
  - too small, tree may not work properly with more
  - too large, tree may overfit.
- **Pruning**: Once tree is constructed, pruning removing unwanted branches phase might remove redundant comparisons or remove subtrees to achieve better performance
  - ~~cost~~ cost complexity pruning
  - reduced error pruning.
- **Advantages**:
  - Simple to understand as it follows same process that human follows while making any decision in real life.
  - Useful for solving decision-related problems
  - It helps to think about all possible outcomes for problem.
  - There is less requirement for data cleaning compared to other algo.

- Disadvantages
1) The DT contains lots of layers, which makes it complex.
2) It may have an overfitting issue, which can resolved using RF algorithm.
3) For more class labels, computational complexity of decision tree may rise.

Use
- Building knowledge management platforms for customer service that improve 1st call resolution, average handling time & customer satisfaction rates.
- In finance, forecasting future outcomes & assigning probabilities to those outcomes.
- Binomial option pricing predictions &
- real option analysis.
- Customer's willingness to purchase given product in given setting, online & offline both
- Product planning, for example, Gerber products, Inc. used decision trees to decide whether to continue planning PVC for manufacturing toys or not.
- General business decision-making
- Loan approval.

Random forest - Non linear classificaton, categorical
Used for both classification & regression.
Based on concept of ensemble learning, which is process of combining multiple classifiers to solve complex problem & improve performance.
RF is classifier that contains no. of decision trees on various subsets of given dataset & take average (majority voting) to improve predictive accuracy of that dataset.
Instead of relying on 1 DT, RF takes predicts from each tree, & based on majority votes of predictions, it predicts final o/p.
The greater no. of trees in forest leads to higher accuracy & prevents overfitting
condn when model completely fits training data but fails to generalize unseen testing data.

RF model



Assumptions

- Since RF combines multiple trees to predict class of dataset, it is possible that some DT may predict correct o/p, while others may not. But together, all the trees predicts correct o/p.

So below 2 assumptions

- There should be some actual values in feature variable of dataset so that the classifier can predict accurate results rather than guessed result.
- The predictions from each tree must have very low correlations.

Need for RF

1) Takes less training time compared to other algo.
2) It predicts o/p with high accuracy, even for large dataset it runs efficiently.
3) It can also maintain accuracy when large proportion of data is missing.

Advantages:

1) Capable of performing both C & R.
2) Handling large datasets with high dimensionality
3) It enhances accuracy of model & prevents overfitting issue.

Disadvantage

Although RF used for both C & R, it is not more suitable for Regression tasks.

Uses

1) Banking: uses this algo. for identification of loan risks.
2) Medicine: Disease trends & risks of disease can identified.
3) Land use: We can identify areas of similar land use
4) Marketing: marketing trends can identified.

Naive Bayes:

- Based on Bayes theorem & used for solving classification problems
- Mainely used in text classification that includes high-dimensional training dataset.
- It is simple & most effective classification algo. which helps in building fast ML models that can make quick predictions.
- It is probabilistic classifier, means predicts on the basis of prob. of object.

- Naive: It is called Naive bcoz it assume that occurence of certain feature is independent of occurance of other features, such as if fruit is identified on bases of color, shape & taste, then red, spherical & sweet fruit is recognized as apple. ∴ each feature individually contribute to identifying that it is apple w/o depending on each other.

- Bayes: called Bayes bcoz based on principle of Bayes theorem.

Fundamental assumption is that each feature makes an independent, equal contribution to outcome.

binary & multi-class

Bayes Theorem / Bayes Rule / Bayes law

Used to determine prob. of hypothesis with prior knowledge. It depends on conditional prob.

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

- $P(A|B)$: posterior prob. — prob. of hypothesis A on observed event B
- $P(B|A)$: likelihood prob. — Prob. of evidence given that prob. of hypothesis true.
- $P(A)$: prior prob. (prob. of hypothesis before observing evidence)
- $P(B)$: Marginal prob. (prob. of evidence)

Types of Bayes Model:

1) Gaussian:- This model assumes that features follow a normal distribution. This means if predictors take continous values instead of discrete, then model assumes that these values are sampled from Gaussian distribution.

2) Multinomial:- This classifier used when data is multinomial distributed. It is primarily used for document classification problems, it means particular document belongs to which category such as sports, politics, eduction etc. The classifier uses freqn of words for predictors.

3) Bernoulli: This classifier works similar to Multinomial classifier, but predictor variables are independent Boolean variables. Such as if particular word is present or not in document. This model is also famous for document classification tasks.

Advantages:
- It is one of fast & easy ML algo. to predict the class of datasets.
- Can used for binary as well as multi-class
- It performs well in multi-class predictions compared to other algo.
- Most popular for text classification problems.

Disadvantage
- It assumes that all features are independent or unrelated, so can't learn relationship b/w features.

Uses:
1) For classifying articles
2) For credit scoring
3) Medical data classification
4) In real time predictions bcoz NB is eager learner
5) In text classification such as spam filtering & sentiment analysis.

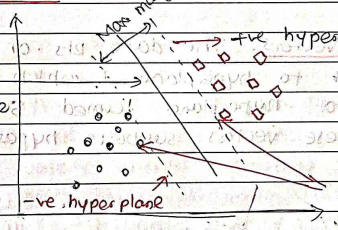
SVM - Linear classification - categorical.

It is 1 of most popular supervised learning algo., can used for both C & R.

The goal of SVM is to create best line or decision boundary that can segregate n-dimensional space into classes so that can easily put new data pt. in correct category in future. The best decision boundary - Hyperplane.

SVM chooses extreme pts/ vectors that help in creating hyperplane. These extreme cases called support vectors & algor. called SVM.

SVM model



Consider below dia. in which there are 2 diff. categories that are classified using decision boundary or hyperplane.

Hyperplane & SV in SVM algo.

Hyperplane: There can be multiple lines / decision boundaries to segregate classes in n-dimensional space, but we need to find out best decision boundary that helps to classify data pts.

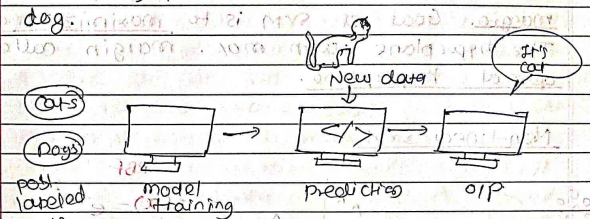
The dimensions of hyperplane depend on features present in dataset, which means if there are 2 features, hyperplane will be straight line.
3 features ——"—— 2 dimensional plane.

We always create hyperplane that has max. margin, which means max. dist. btw data pts.

Support Vectors: The data pts or vectors that are closest to hyperplane & which affect the position of hyperplane termed support vector. Since these vectors supports hyperplane called SV.

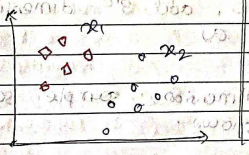
Example:
CAT & DOG.
Suppose we see strange cat that has some features of dogs, so if we want model that accurately identify whether it's dog or cat, so such model can be created by SVM algo.

We will 1st train our model with lots of images of cats & dogs so it can learn about diff. features & then test with strange creature.
So as SV creates decision boundary btw these 2 data (cat & dog) & chooses extreme cases (SV), it will see extreme case of cat & dog.



Types of SVM

Linear	Non-linear
Used for linearly separable data.	Non-linearly separable data.
Dataset can classified into 2 classes by straight line.	Dataset can't classified using straight line.



Linear SVM

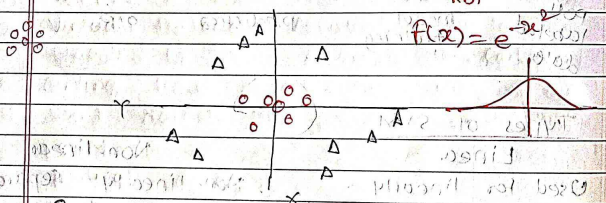
SVM algo. helps to find best line or decision boundary; this best boundary / region called hyperplane.

SVM find closer pt. of lines from both classes. These pts. SV.

The distance b/w vectors & hyperplane called margin. Goal of SVM is to maximize margin.

The hyperplane with max. margin called optimal hyperplane.

Non-linear SVM



So to separate these pts. we need to add 1 more dimension. for linear data, used 2 dimen. x & y.

For non-linear data, add 3rd dimension z.

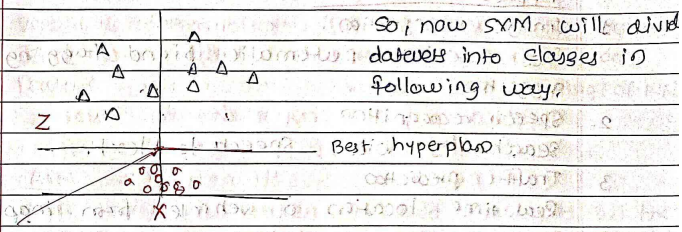
It can calculated as

$z = x^2 + y^2$

By adding 3rd dimension sample space will become as shown.

Decision rule $\vec{w}.\vec{u} \geq c$ $\vec{w}.\vec{u} - c \geq 0$, $\vec{w}.\vec{u} + b \geq 0$

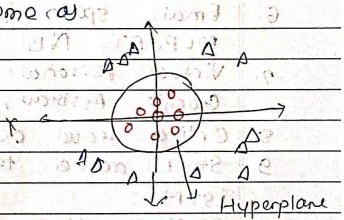
for any $y_i = \begin{cases} +1 & \text{if } \vec{w}.\vec{x_i} + b > 0 \\ -1 & \text{if } \vec{w}.\vec{x_i} + b < 0 \end{cases}$



So, now SVM will divide datasets into classes in following way.

∵ we are in 3-d space, hence it is looking like ⊥ to x-axis. If we convert it in 2d space with $z = 1$, it become as

∴ we get circumference of radius 1 in case of non-linear data.



Use

1) Face Detection
2) Image classification
3) Text categorization

$\vec{w}\vec{x} = \vec{w}.\vec{x} = |w||x|\cos\theta = 0$

$\theta = 90$

vectors ⊥ to hyperplane,

$w^T x + w_0 = 0$ / $w^T x = 0$

Optimization function.

$\vec{w}.\vec{x} + b \geq 1$. $d = \frac{(\vec{x_2} - \vec{x_1}).\vec{w}}{\|w\|}$

$\vec{w}.\vec{x} + b \leq -1$ $= \frac{\vec{x_2}\vec{w} - \vec{x_1}\vec{w}}{\|w\|}$

$y_i(\vec{w}.\vec{x} + b) \geq 1$. → constraint eqn

$y_i(\vec{w}.\vec{x} + b) = 1$ → for SV.

Applications of ML

1. Image recognition.
   face detection, Automatic friend tagging suggestion
2. Speech recognition
   Search by voice, Speech to text.
3. Traffic prediction
   Real time location of vehicles from map.
4. Product recommendations.
   Amazon, Netflix
5. Self-driving cars.
6. Email spam & malware filtering
   MLP, DT, NB classifier used.
7. Virtual personal Assistant
   Google Assistant, Alexa, Siri, Cortana
8. Online fraud detection.
9. Stock market trading
   LSTM
10. Medical diagnosis
11. Automatic language translation.

DT

(1) Suffer from overfitting if it's allowed to grow w/o any control.
(2) Single DT is faster in computation
(3) When dataset with features is taken as i/p by DT it will formulate some set of rules for prediction.

RF

(1) RF are created from subset of data & final o/p based on avg./majority & overfitting handled
(2) Slower compare to DT.
(3) It randomly selects observation builds DT & avg. result taken No use of any set of rules.