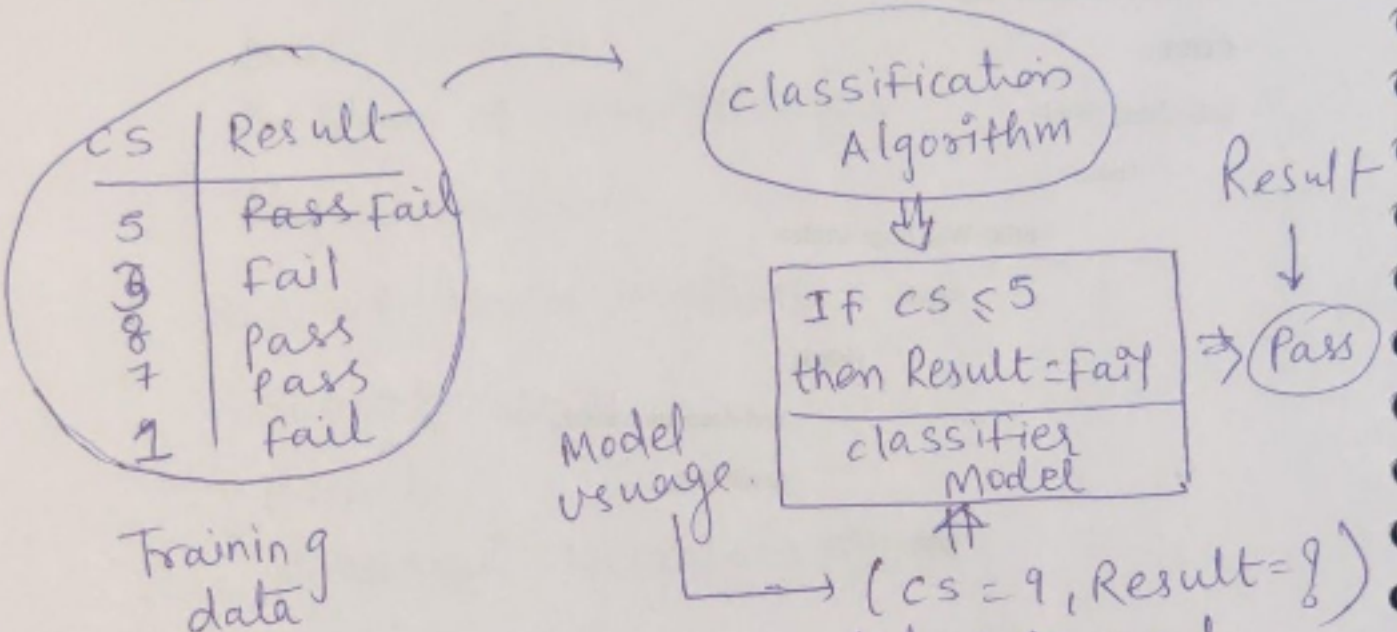
ML - Unit-3.

* classification

1) Model construction
2) Model usage



- After analyzing training dataset and applying classification algorithm, we can make classifier model

↓ that is

Model construction

→ So, after Model construction we can give new dataset i.e. Test data like

CS = 9 ⇒ Result = ?

And if it is given true value according to the Model then our model is O.K.

For example:

Online shopping

where we bought something then doing a review.

You | Friend.

↳ bought chips.

Stars Review

You: chips were excellent (5*.) → (+ve)

Review.

Friend: chips were Unhealthy. (2*) → (-ve)

So, from review we can make model,

review → (M) → +ve (1) } → -ve (0) }

↓

Fit in these 2 classes

as +ve for class 1.

-ve for class 0.

- When we give any new review now then it will from those 2 classes.

⇓

Known as

⇓

Binary classification

Binary classification

Popular algorithms
- Logistic Regression
- K-Nearest Neighbors
- Decision Trees
- Support Vector Machine
- Naive Byes

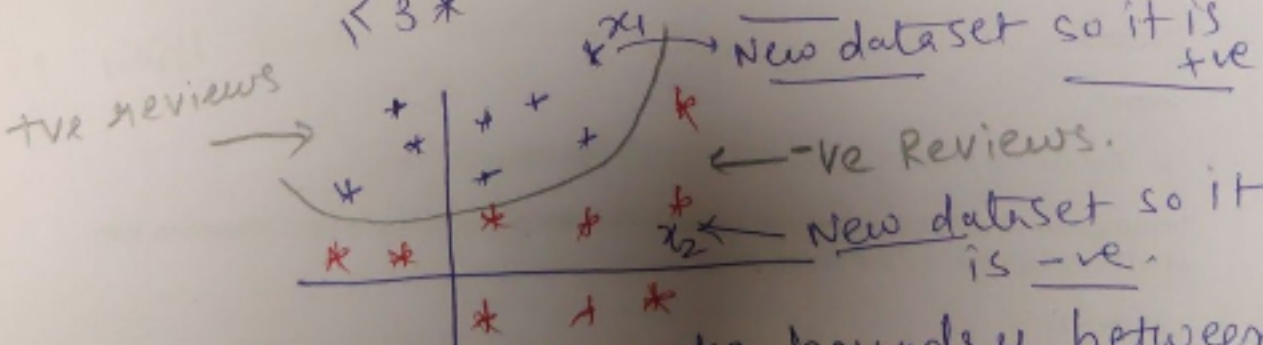
Model

You will be given a model to which you have to feed lots n lots of data (training datasets) like here we will give millions of reviews. (class-labels as +ve & -ve)

like $R_1 \rightarrow$ +ve
$R_2 \rightarrow$ +ve
$R_3 \rightarrow$ -ve
$R_4 \rightarrow$ -ve
$R_n \rightarrow$ -ve
} with star

→ And base on above datasets and star review system model will analyze all reviews and make classifier model.

as ≥3 * will be +ve review
<3 * " " -ve review.



→ After plotting it will make boundry between each class → so that new datasets can clearly differentiate betⁿ 2 classes.

Multiclass - classification

- Algorithms.
  - KNN
  - Decision Trees
  - Naive Bayes
  - Random forest
  - Gradient Boosting

→ When we have more than 2 classes for classes/category then it is multi-class classification

model

↑

lots n lots of data needs to be feeded for accurate result.

For ex: 1 picture is given with an animal and we need to find that it is from which category like cat, dog, monkey, cow etc.

↳ Multi-class

→ So millions n millions of picture to be feeded to model so that

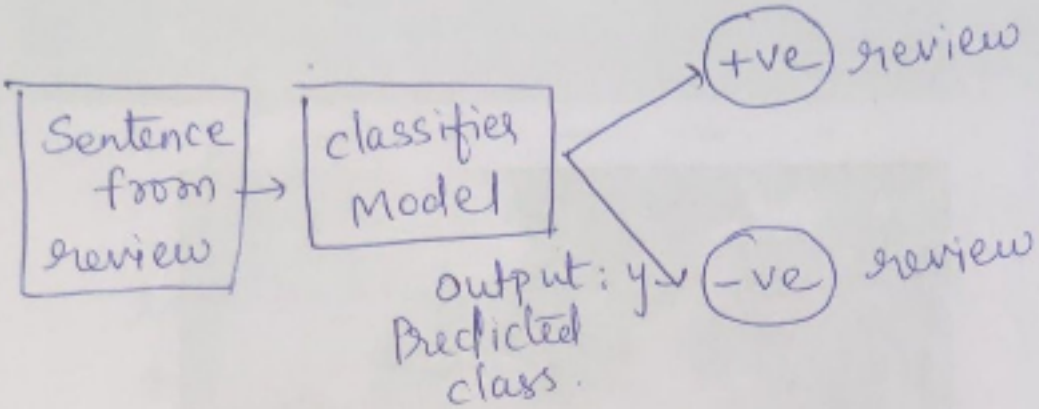
↳ with class-label

this model will draw a figure with boundry (like shown before in binary classification) and when new data comes in it will identify from boundry data that in which category it will fall.

more than 1 bandry.

classifier.

for review system.



A linear classifier.

- will use training data to learn a weight or co-efficient for each word.

word	co-efficient
good	1.0
great	1.5
awesome	2.7
bad	-1.0
terrible	-2.1
awful	-3.3
restaurant, the, we, etc..	0.0

by analyzing above word's co-efficients when new word or sentence given we can know that score of statement.

ex: Sushi was great.
the food was awesome
but the service was terrible.

score(x_i) = 1.2 + 1.7 - 2.1 = 0.8 > 0
⇒ +ve review.

So, it is called a linear classifier, because output is weighted sum of inputs.

So,

word	co-efficients
:	:

Simple Linear classifier

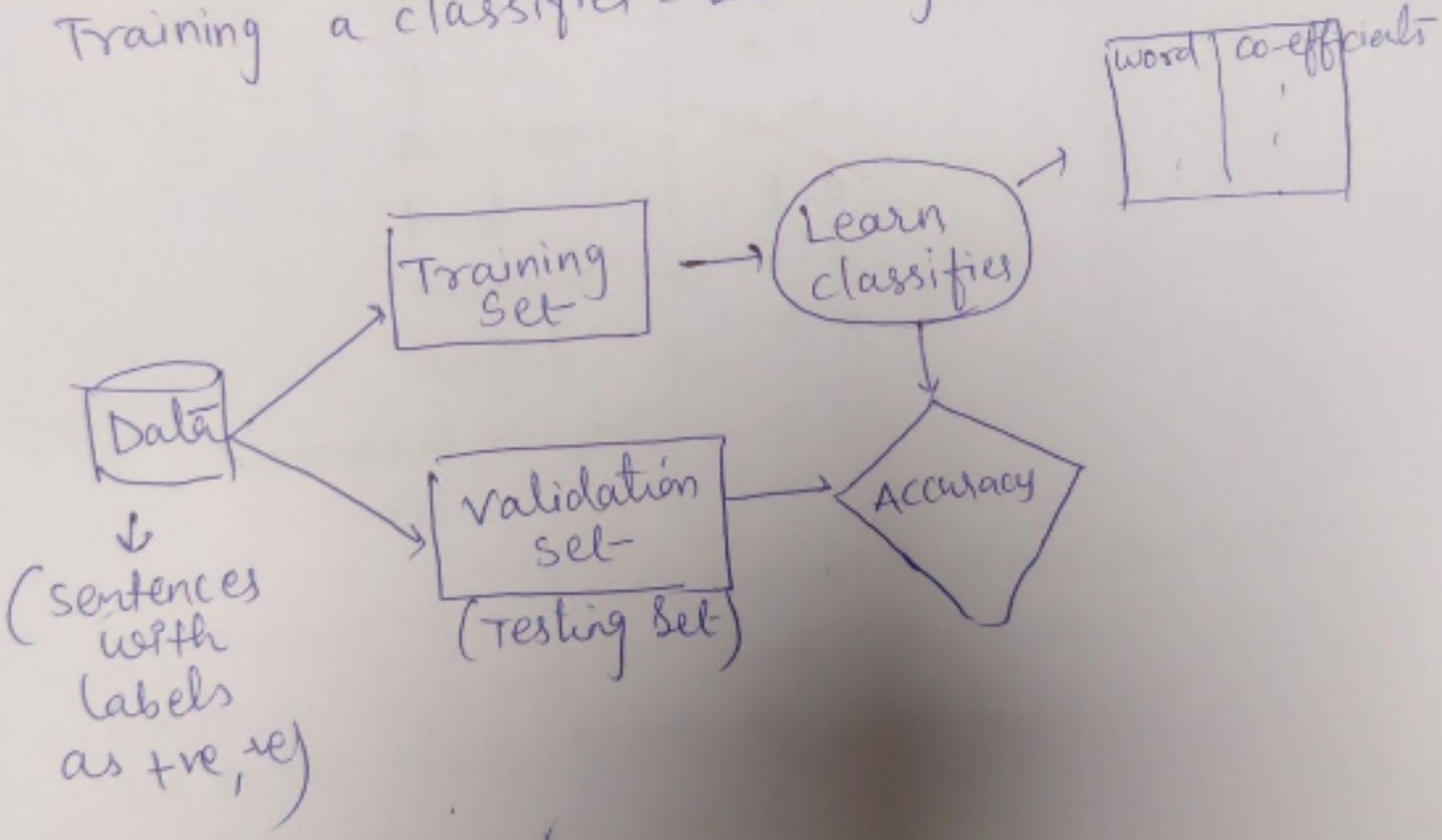
Score(x) = weighted count of words in Sentence

If Score(x) > 0 : +ve

Else : -ve

Sentence from review → Input x :

Training a classifier = Learning the co-efficients



Multi-Label classification

⇓

Refers to those classification tasks that have two or more class labels, where one or more class labels may be predicted for each example.

ex: photo classification

⇓

where more than one type of class label objects are there

ex: bicycle
apple
person etc.

Where In binary & Multi-class classification, single class label is predicted for each example.

→ Specialized versions of standard classification algorithms can be used,

- Multi-label Decision Trees
- Multi-label Random Forests
- Multi-label Gradient Boosting

Or

Use a seperate classification algo to Predict the labels for each class.

Imbalanced classification

⇓

Refers to classification tasks where the no. of examples in each class is unequally distributed.

ex: where the majority of examples in the training dataset belong to the normal class and a minority of examples belong to the abnormal class.

like Fraud detection
Outlier detection
Medical diagnostic tests

Classification Evaluation Metrices

⇓

There are different ways to find out how effective our model is., one of them is Confusion Matrix.

Confusion Matrix

↳ easiest way to measure the performance of a classification problem where the output can be of 2 or more type of classes.

↳ Table with 2 dimensions
viz. "Actual" vs. "Predicted"

Predicted \ Actual	1	0
1	True Positive (TP)	False Positives (FP)
0	False Negative (FN)	True Negative (TN)

TP: When both Actual class & predicted class of data point is 1.

TN: When both Actual class & Predicted class of data point is 0.

FP: When actual class data point is 0 & Predicted class data Point is 1

FN: When actual class data point is 1 & Predicted class data Point is 0.

code: Sklearn.metrices import confusion_matrix

Logistic Regression ⇒ used for solving classification problems.

⇓

Used for predicting the categorical dependent variable using a given set of independent variables.

⇓

It predicts the output of a categorical dependent variable

⇓

Therefore the outcome must be categorical or discrete value.

⇓

Can be either Yes or No, 0 or 1, true or False, etc.

⇓

Instead of giving exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

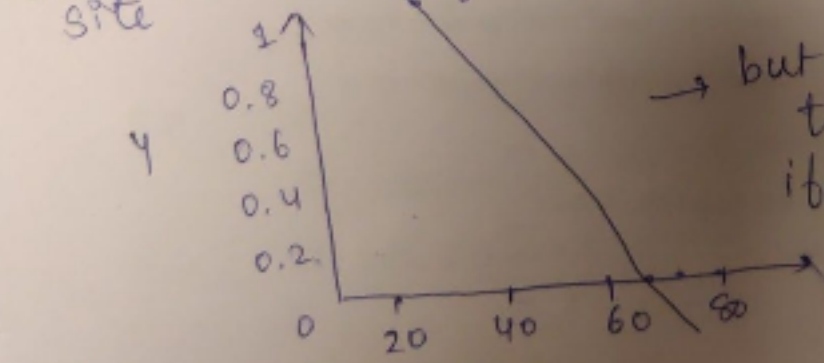
ex:

Time (x)	clicked on Ad (y)
68.95	No
80.23	No
69.45	No
14.15	No
50.0	Yes
55.5	Yes

→ categorical data

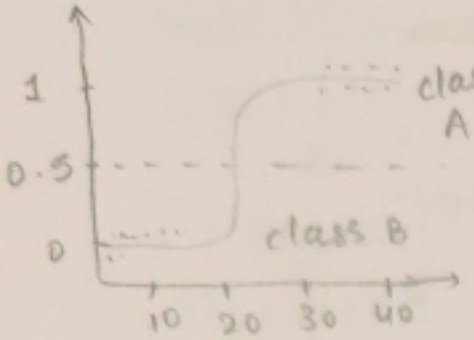
user active on any site

→ click on Ad or Not



→ but not helping to predict if we are using Linear Regression

So, If you have categorical value then use logistic regression.



$$y = \frac{1}{1+e^{-x}}$$

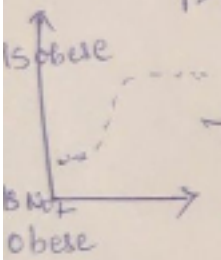
where, e = euler's constant
x = independent Val

⇓

Sigmoid function

It is trying to convert the independent Variable into expression of probability which ranges between 0 and 1 with respect to the dependent variable.

→ So, Logistic Regression predicts wheather Something is True or False, instead of predicting Something continuous like quantity.



⇓

That means that the curve tells you the Probability that mouse is obese or not

→ So, LR mostly used for Probality if the probability a mouse is obese is > 50% ⇒ then we'll classify it as obese. Otherwise "Not obese"

→ Not, calculate R^2 or least Square Method instead of that it uses maximum likelihood.

→ It is a predictive analysis algo & based on the concept of probability.

Cost Function

- Error Representation in ML

↓

Shows how our model is predicting compared to original given database

↓

Cost Function ↓ ⇒ Accuracy ↑

For Linear Regression,

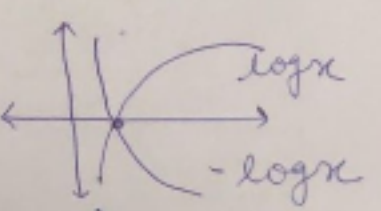
$$\text{cost} = \frac{1}{m} \sum_{i=0}^{m} |y_{pred} - y|$$

↑ don't work for this

For Logistic Regression,

$$\text{Cost} = \frac{1}{m} \sum_{i=1}^{m} [y \log(y_{pred}) + (1-y) \log(1 - y_{pred})]$$

$$y_{pred} = \sigma(w^T x + b)$$



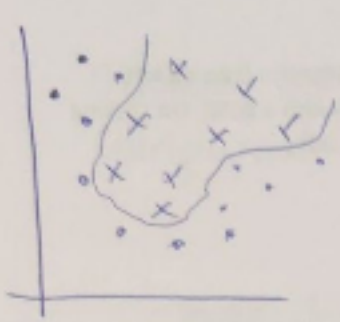
For 1 observation,

$y = 0$ | error $= -1 \times \log(1 - y_{pred})$

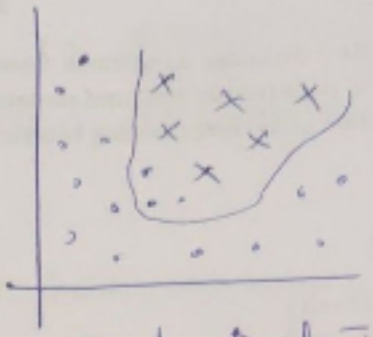
$y = 1$ | error $= -\log(y_{pred})$

When more than 1 → then average of all predictions in above equation

Overfitting



Train Dataset
↑ Train Accuracy



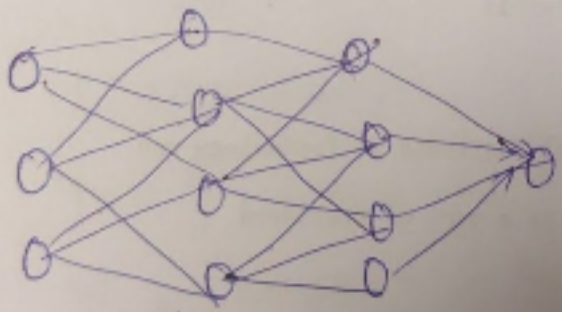
Test Dataset
↓ Test Accuracy

↘ Known as Overfitting ↙

- But we want Smooth Curve which fits all Training datasets as well Test dataset

↓ for that we are using **Regularization**

-



like Neural network → where more neurons and hidden layer may not get accurate Curve or predictions

→ So, eliminate few of them

↓ Known as **Regularization**