Lecture: 1, Machine Learning

Date – 1-8-2024, P23CS0010

② What is machine learning?

→ ML is a discipline of artificial intelligence (AI) that provides machines ability to automatically learn from data and past experiences.

- Any process by which system improves its performance by experience.
- ML is study of algorithms that :—
  - → improves performance P
  - → at some task T
  - → with experience E.

→ A well defined learning task is given by $< P, T, E >$

eg:- we get suggestion of next sentence's word while writing mail or while searching something on Google.

⇒ ML algorithms adaptively improves performance with an increase in number of samples during 'learning' process.

- Why Machine Learning is used ?

We'll learn with the help of some examples :—

Example 1 :- Let's see, mathematic digit 2 can be written in various way by human as:

→ handwritten digit 2 :

goal :- identifying different digits from handwritten text.

- All 2 are in different shape, there is so much variation in data.

⇒ Shape alone is not sufficient to identify digits.

Example 2 :- Let's take example of a sentence :—

original sentence :- I will watch football Match.

understood sentence :- I will watch football Match.

- We need to eliminate observer based decision.

- Because of so many rules and variation in data we need to learn from experience in ML.

Now let's see some examples of machine learning based day to day tasks :-

- Self-driving cars and driver assistance features, automatic stopping, improve overall vehicle safety.
- A bank's fraud detection service automatically flag suspicious transactions.
- Recommendation engines suggest products, songs, television show such as Netflix, Amazon or Spotify.
- Speech recoginition software that allow us to convert voice memos to text.

- **Applications of ML :-**

(i) Image Recognition
(ii) Automatic language Translation
(iii) Medical Diagnosis
(iv) Stock Market trading
(v) Online Fraud Detection
(vi) Email spam/malware filtering
(vii) Self driving cars
(viii) Product Recommendations
(ix) Traffic Prediction
(x) Virtual Personal Assistance

Lecture-2

Date : 06-Aug-2024

- **Why ML is used ?**
- ML is used for Automation, Speed, Observer independence.

→ Observer Independence is done through decision support systems.

eg1:- For disease Prediction different doctor diagnosis differently.

eg2:- Astrophysics uses ML to calibrate simulation parameters based on observational data.

eg3:- Stock market prediction.

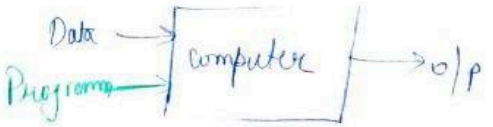
Note:- Final decision will be made by human.
↓
ML systems supports in making decisions.

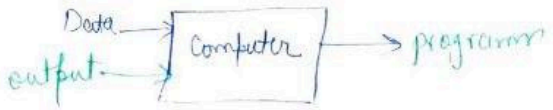
- **Definition :-**
  ML is field of study that gives computer ability To learn without being explicitly programmed.

- Rule based program fails because of variation in data.

Inductive learning :- It is learning from experience, i.e :- refining our observation based on experience.

Traditional Programming :-

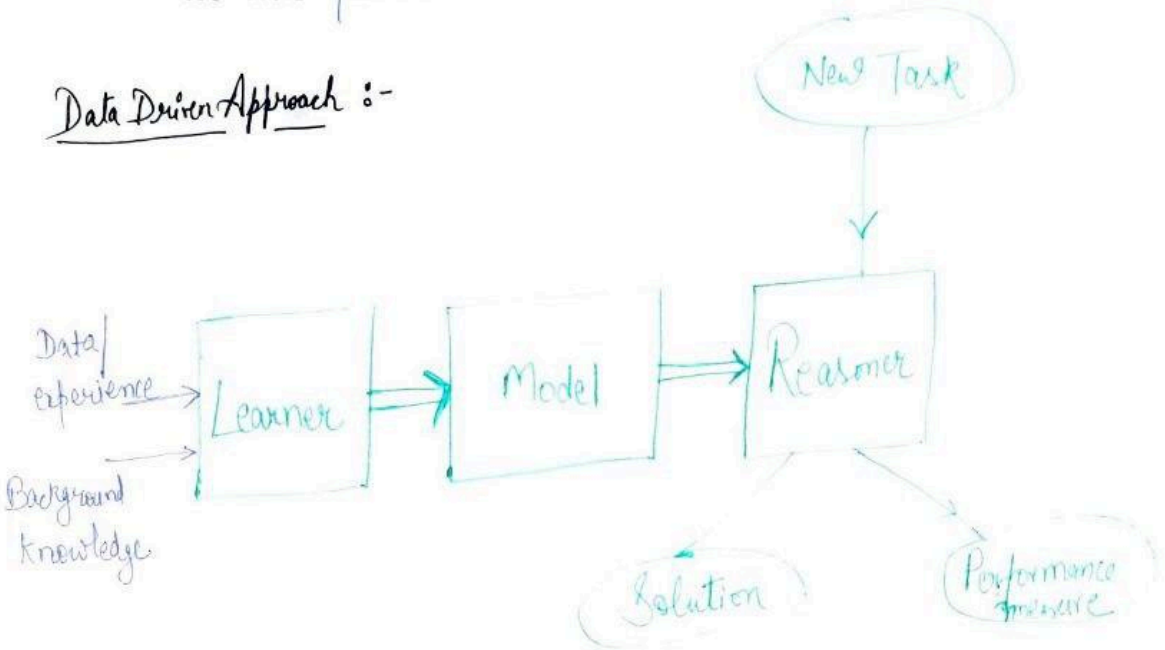




- In programming language :- behaviour of program don't depend on data.

In ML :- How computer comes up with program depends on type of data we have provided to learn.

Data Driven Approach :-



Steps of creating a learner :-

(i.) Choose the training data / experience
   → extract features from data.
   → different method may require different feature extraction.

(ii.) Choose target function (for classification between cat, dog some fun^{n} give different set of values of cat, dog).

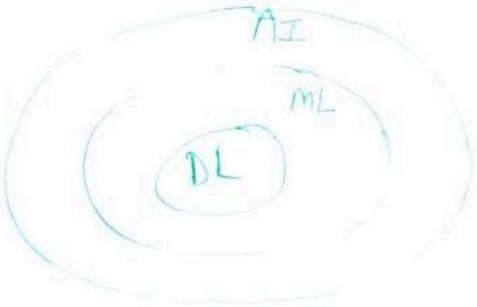
(iii) Choose how to represent target function.

   eg:- we want target function to Linear or circular or else?

(iv.) Choose a learning algorithm to infer target function.

- How to represent target function :—
  - → may be complex (but richer representation better result)
  - → difficult to learn

- Set of target function : Hypothesis.



History of ML :-

(i) ML is used to known as Pattern Recognition (PR).

(ii) 1950 : Samuel's checker playing algorithm.

(iii) 1990 : support vector machine

(iv) After 1990 :- Adaboost got tremendeous success (ensemble learning)
Bioengineering especially.

(v) Popularity of this field in recent time :- new software, new hardware,
↓ GPU
neural n/w
→ cloud enabled = Availability of big data.

(vi) Alexnet changed Innovation = Deep learning came into existence.

eg :- Superresolution problem after ML Boom.

- Explainability : Some algo just work we don't know why.
- Generalizability
- Computational Requirements
- Data requirements

Applications of ML :-

- Recommendation System
- Computer vision
- Medical data Analysis
- Robotics, Biometrics
- Speech recognition
- Autocomplete.
- Security, banking, finance, physics.

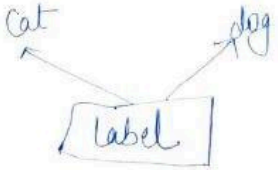
Different types of ML :-

(i) Supervised (ii) Unsupervised (iii) Reinforcement learning

Types of M/c learning :-

(i) Supervised learning :-

Input :- Image + Label = Training

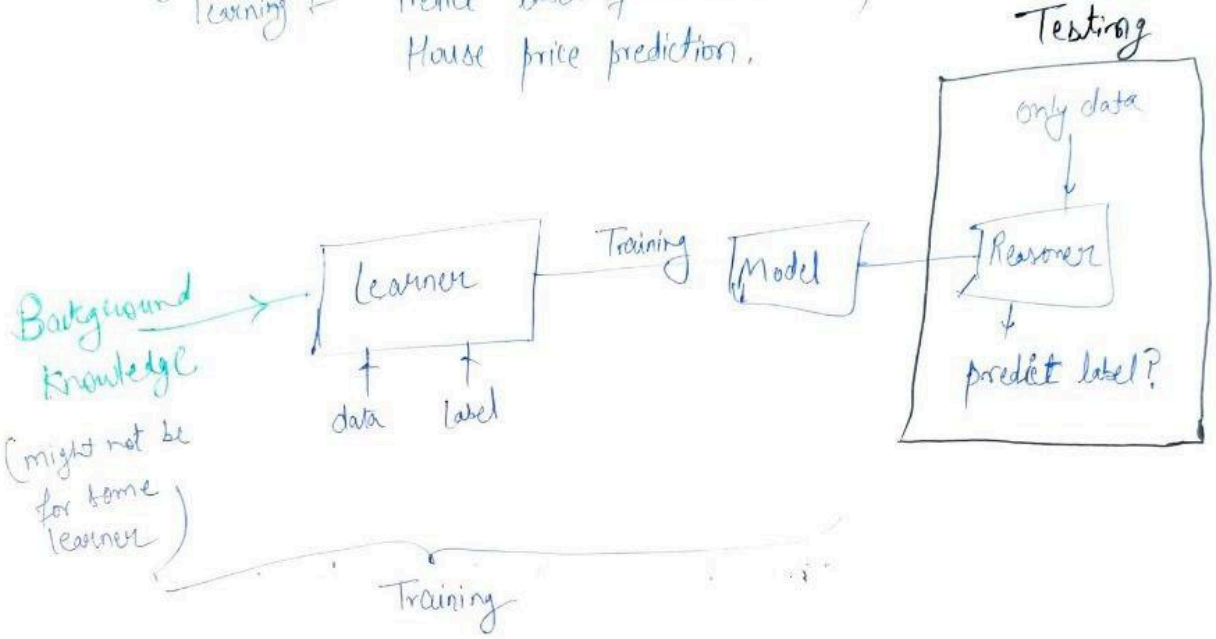


→ decide learning algorithm

- Asking Label of image — cat or dog?

→ It is kind of ml where algorithm is trained with Label data.

Task of Supervised learning :- Predict label of unlabeled data,
House price prediction.



Validation :- (i) Train model
(ii) with some amount of data validate

Validation is internal testing after training.

eg:- Training :- when company manufactures car.
Validation :- Based on internal checking deciding which thing to fit, tune, optimize change.

Hypothesis tuning / optimization.

Testing :- when customer drives the car / selling of car.

- During training Background knowledge might not be there for some learner.

- For some we may have seperate Training, Testing.

eg:- finding similar viruses based on genetic pattern recommendation system based on customer's previous choices.

→ Making cohesive group of data inputs points.

- we can define performance measure but need not to have training data set.

eg:- . Clustering . Dimensionality reduction . Association rule mining.
. Generative model . unsupervised feature extraction.
↑
extract feature from image.

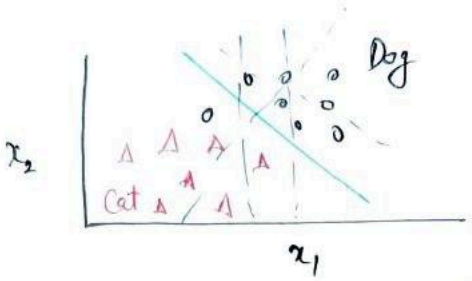
- **Dimensionality Reduction:-** All features might not be meaningful, finding salient features in data.
↓
Data compression.

- most of generative model is unsupervised.
- Discriminative model :- creates new image.

Feature Space:- The space defined by features is called feature space.
or
It is an Euclidian space defined by features.
↑
- Any data/sample is a point in the feature space.

need not to be always euclidian → we assume feature space to be euclidian space

Hypothesis :-

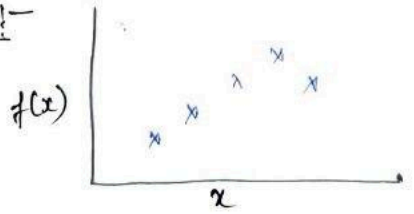


(fⁿ that approximates target function.)

→ How many such hypothesis are possible ?
: Let's assume that I have N data points.
· Each data points can be classified in one of the two classes.
· So possible class combination for N data points is 2^N.

→ 1 hypothesis for each class combination so, 2^N hypothesis are possible.
It is very very huge.

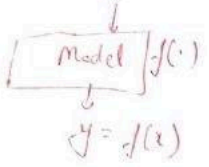
Inductive Bias :-



· A regression problem with data samples x and regression value f(x)

Inductive learning :- (for supervised learning) → necessary coz we have some amount of data not entire data.

· Given data points X and label Y.
or we may consider data X and a function f(x)



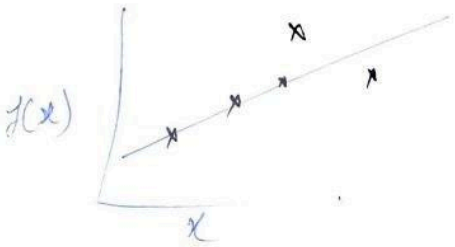
→ f(x) : Categorical for classification
→ f(x) : Real no. for Regression
→ f(x) : Probability of X for probability estimation

· learning from experience.

How to obtain features ?

- Features may be obtained from raw data
  eg: as texture, SIFT feature etc
- may be obtained through measurement at the time of data collection.
  eg: height, weight etc. of different animals.
- Feature Engineering
- Many NN methods can directly deal with raw data without explicitly requiring features at input.

Hypothesis :



- Regression problem with data samples x and regression value f(x)
  Which curves fits best ?

Decision Boundary :- The boundary by which decision before and after changes.

→ The (set of all legal hypothesis) which help to solve ML problem is known as Hypothesis Space.

Q:- How many hypothesis can exist?

Sol^{n}:- Let's assume that we have 'N' data points.

If each data points can be classified into one of 2 classes.

→ So possible class combination for 'N' data points is 2^{N}.

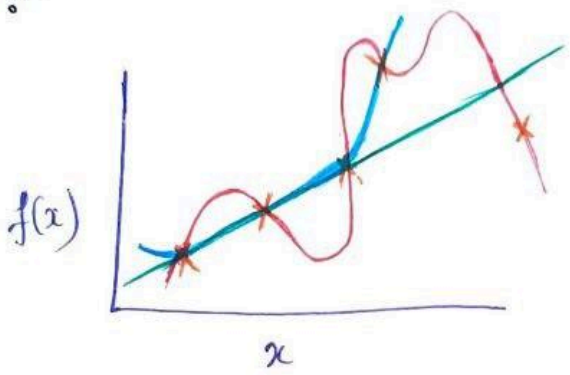
↑

not possible as could be very large.

- We have to make some restrictive expression.

Date: 13-08-2024 Lecture-4

Hypothesis :-



Let's consider regression problem with data samples x and regression value $f(x)$.

Q:- which curve fits best ?

Note:- our decision of fitting the curve depends on our assumption about problem.

> Our assumption is :- Inductive bias
> (based on my Prior)
> Inductive bias reduces search space

→ Inductive bias reduces 2^n assumptions to finite no of hypothesis.

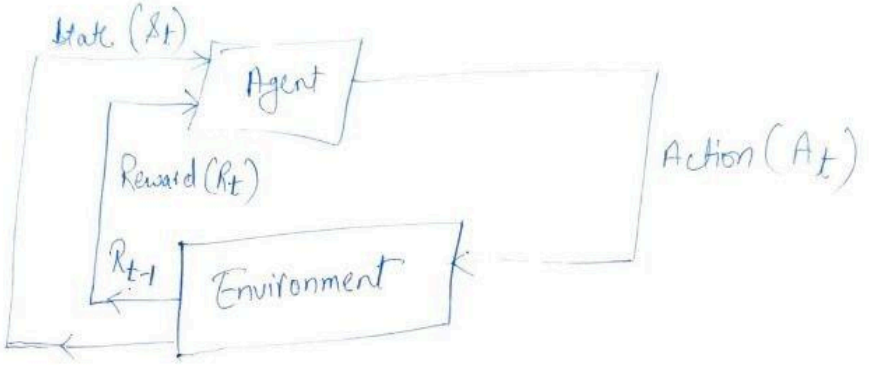
▲ No Free Lunch theorem :-

In essence, unless you don't make assumption about problem, you can't make solution.

> There is issue in ML which is Generalizability.

Generalized one hypothesis that not only fits traing data, also unseen data.

Reinforcement learning :-



What do we learn ?

To control an agent (bicycle) from sequence of good decisions

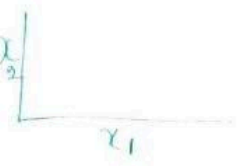
How do we learn it ?

From experiences through reward from the environment.

Inductive learning :- (for Supervised learning)

- Given data X and label Y
  - → we may consider data X and a function $f(x)$
- Return a function h that approximates f :-
  - → we don't know f
  - → But we assume that there is an f.
  - → h is called a hypothesis.

Feature Space :-



x_1 : Height
x_2 : Weight

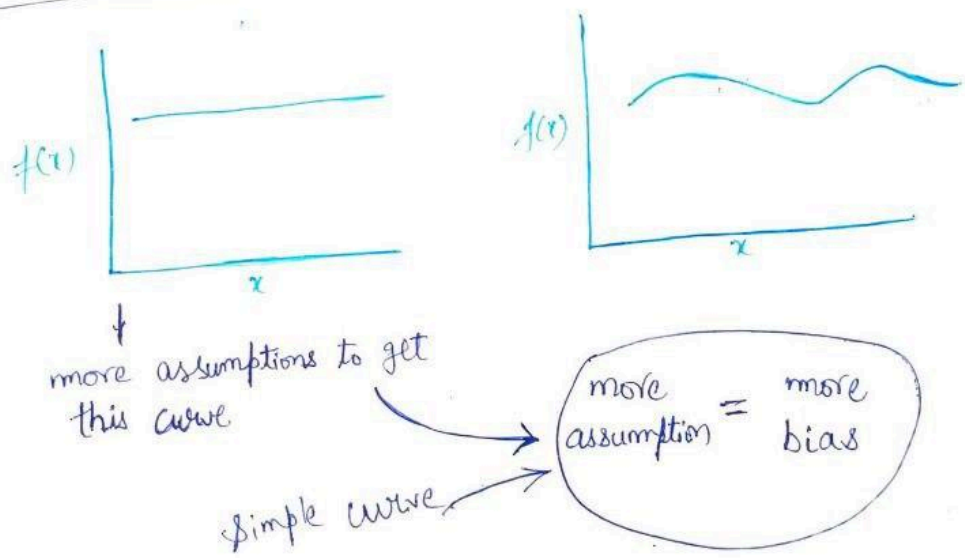
- we want generalized hypothesis.

▲ **Ockham's razor** : As training data may fit to multiple curves,

→ Prefer simplest hypothesis consistent with data.

- This is just an assumption and doesnot always work. (not universal truth)

eg:- Nature follows simplest rule
i.e:- mixing of gas in Room.



eg:- In order to drive on straight line
we consider some assumptions like good wheather, road, good vehicle, good condⁿ, expert rider etc.

- Realistic curve → less assumption → more complex curve

Types of Inductive Bias :-

(a) Restriction / Absolute bias
(b) Relative / Preference bias

(a) Restriction bias / Absolute bias :-

- representational power of algorithm.

i.e:- Set of hypothesis that algo will consider.

eg:- my algo will consider only straight line.

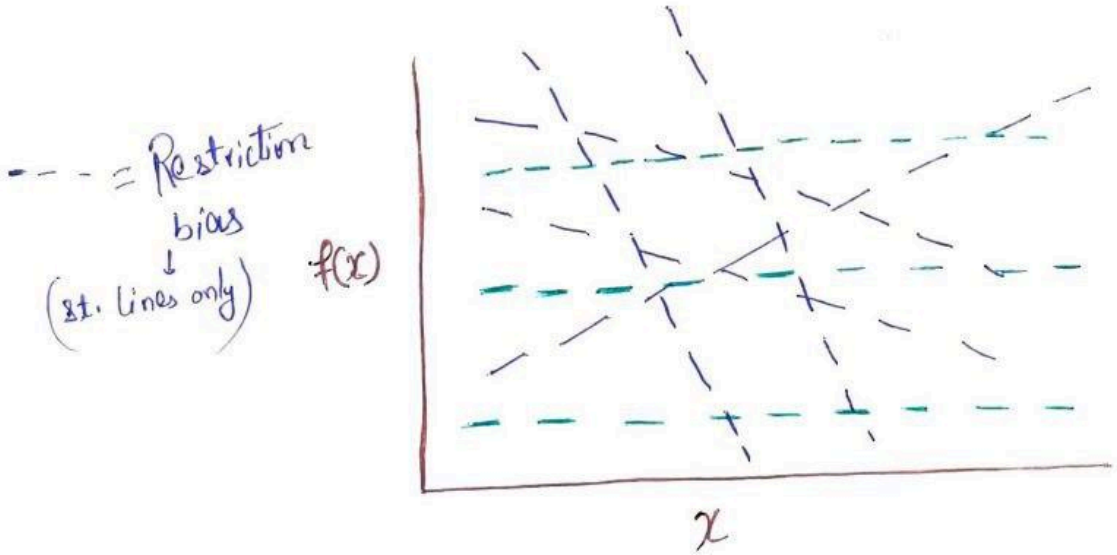
- even of this assumption many straight lines possible

⇓

(b) Preference bias / Relative bias :

- what type of representation my algo perfers.

eg:- my algo may preper straight lines parallel to x-axis.



(- - - - = Relative bias = st line || to x-axis only)

Inductive learning :

- Let's take general fun^n from training examples.

(a) Construct h that best approxes target fun^n f.

$$h \longrightarrow f$$

(b) h is consistent if it agrees with all training examples.

(c) h is generalized well if correctly predicts label for new examples

(d) learning $\longrightarrow$ refining hypothesis space.

> Inductive learning is an ill-posed problem.

- **Ill-posed Problem:** unless we see all possible examples related to problem, inductive learning algo can't find unique soln.

i.e:- one hypothesis can't be generalized to all examples.

> it is not practical to have all examples

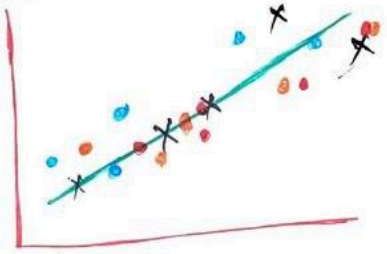
* Assumption in Inductive learning :

- A hypothesis h approxes f well in sufficiently large set of examples, will also approximate target fun^n well over unobserved examples.

Tasks in ML :- (a) come up with good hypothesis space
(b) find algo that works well in that hypothesis space
(c) generalizability in future data points
(d) Analyse confidence in result.
(e) Analyse computational tractability

[NN earlier wasn't computationally tractable.]

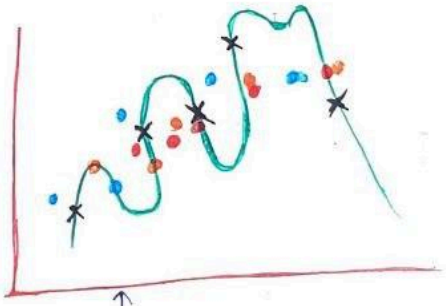
For Test data :-



↑
It has more error

* Similar performance on training, test data
- Underfitting didn't fit well to data = consistent performance

[almost same, similar type of error ↑ to new data]



↑
It has less error as curve-fit to all points.

- It has different performance on training data, test data.
- Overfitting : outstanding training data

[diff. performance every time → inconsistent performance to testing data.]

	T_1	Testing data 2	T_3
eg:- $e_{Green(st.)}$	34	2.9	3.1
$e_{red(curve)}$	17	5.7	10.9

↓

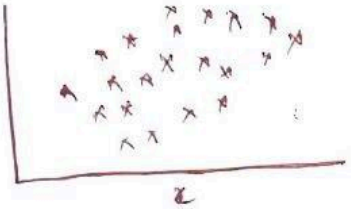
more error to curve which is more complicated.

Green Curve	Red curve
• High bias, underfitting • but low variance • Consistent but poorer performance	• Low bias, overfitting • High variance, inconsistent performance across datasets.

→ Bias = Avg. prediction of our model − correct value we want to predict.

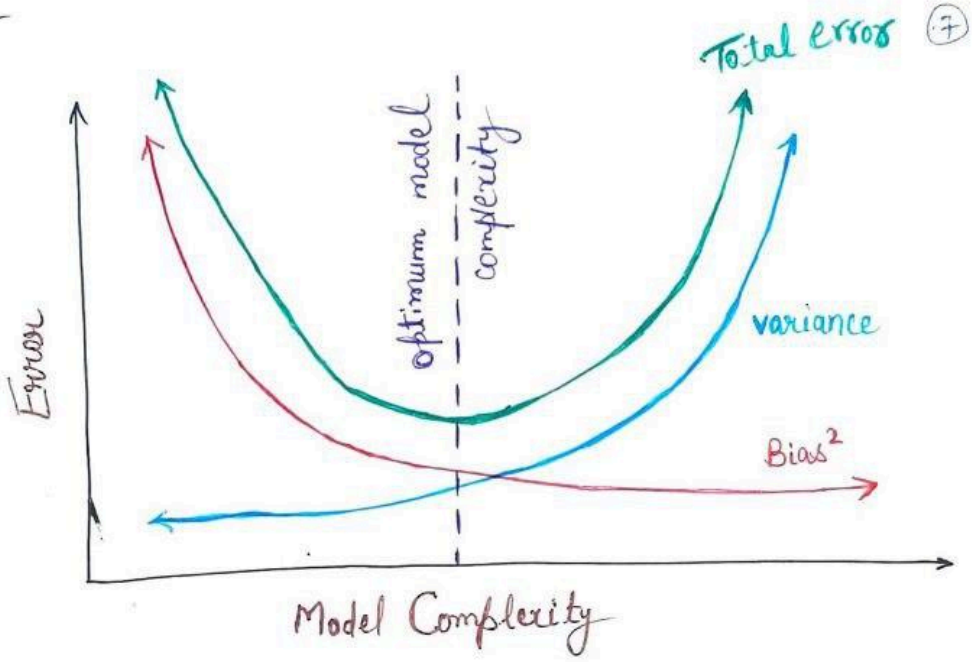
- Model with high bias pays very less attention to training data and oversimplifies the model.
  ⇓
  It leads to high error on training, test data.
- Variance is variability of model prediction across different sets of data.

If we have a lot of training data of sufficient diversity.



→ Training data of sufficient diversity
↓
prevents overfitting

Errors :-



- ↑↑ model Complexity ⟹ ↓↓ Bias
  ↑↑ variance

- ↓↓ model complexity ⟹ ↑↑ Bias
  ↓↓ variance

- we need optimum model complexity ⟹ ↓↓ Bias
  ↓↓ variance

→ Keep lot of training data
→ with sufficient diversity.

Date - 14/08/2024

Probability

- **Uncertainty :** Real world is complex.

  eg:- there is a moment of particle, we can't find out absolute moment. ↑ Real world is uncertain.

eg:- Tossing of a coin = certain

[Based on many factors like physics formula, how much force gonna be on tossing of coin, calculate rotation of coin, properties of atmosphere etc ⟹ If I have too many parameters = I can predict right.

but it is impossible to have microscopic details.

- In practice we often don't model exact system
  - → we choose some abstractions.
  - → result in the probabilistic nature of our experiments.

  Use probability to model the system.

Experiment

Deterministic

- outcome is known

→ If i throw object, it will fall.

Random

- outcome is not known beforhand

→ Rolling a dice.

Sample Space :- Set of possible outcomes.

(a) **Finite**

eg :- Dice

$S = \{1, 2, 3, 4, 5\}$

(b) **Infinite**

eg:- Tomorrow's temperature

Amount of rainfall at Jodhpur in July

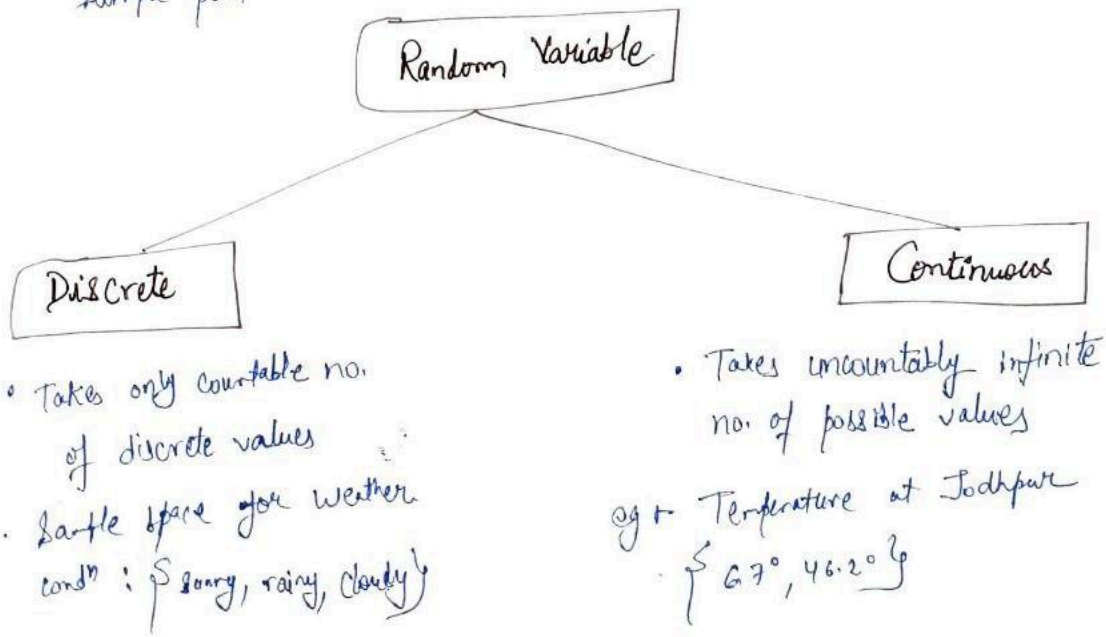
Event :- Any subset of the sample space.

eg:- if while rolling dice outcome is a no. < 3

Event (A) = $\{1, 2\}$

Random Variable :- it is numerical description of outcome of a random experiment.

→ A fun that assigns numerical values (real or Boolean) to each sample point.

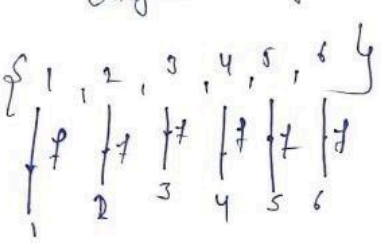


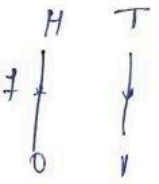
- R.V usually indicated by Capital letter. (eg: X)

- Outcome of Dice is not random variable. A fun 'f' assigns value to event.

Dice : {1, 2, 3, 4, 5, 6} Coin : {0, 1}

categorical weight





Probability of an Event :—

- A = event
- Probability of event A

$$P(A) = \frac{\text{No. of elements in set A}}{\text{No of elements in sample space S.}}$$

$$P(A) = \frac{\text{No of favourable outcome}}{\text{Total no of outcome}}$$

$P(\text{outcome} < 3) \Rightarrow$ event set = {1, 2}

$$P(\text{outcome} < 3) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3}$$

$P(\text{outcome odd})$ event set = {1, 3, 5}

$$= \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6} = \frac{1}{2}$$

Probability :- Sample space is not equally likely in realistic way.

Probability defⁿ applies if
- Sample space is equally likely
- It should be Countable.

Frequentist Approach of Probability

- If a experiment is done n times out of it event A occurs $n(A)$ no. of times.
- Relative frequency of A is $f_r(A) = \frac{n(A)}{n}$.

for true probability outcome experiment should be done 'as no. of times.

$P(A) = f_r(A)$ only if experiment done nearly '∞' times.

Probability of A is $P(A) = \lim_{n \to \infty} f_r(A) = \lim_{n \to \infty} \frac{n(A)}{n}$

- Probability measure probability to an event.

or

$P(\cdot)$ Probability function assigns probability to an event.

$P(A)$ = chance that event A occurs.

Probability Properties

- $P(A)$ = the no. i.e :- $\left(0 \leq P(A) \leq 1\right)$
- if ϕ = null event, no outcome from an experiment, $P(\phi) = 0$
  It is impossible that experiment has no outcome.
- if S = sample space
  $P(S) = 1$
  probability that something happens is always 1.
- If $A_1, A_2 - - - A_n$ are countable sequence of disjoint events
  $P(A_1 \cup A_2 \cup - - - A_n) = P(A_1) + P(A_2) - - - + P(A_n)$

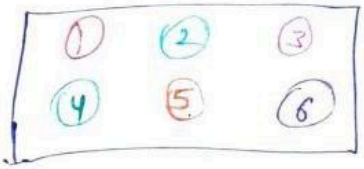
eg:- mutually exclusive o/p

$P(2 \cup 4 \cup 5) = P(2) + P(4) + P(5)$

getting either 2, 4, 5 $= \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6} = \frac{1}{2}$

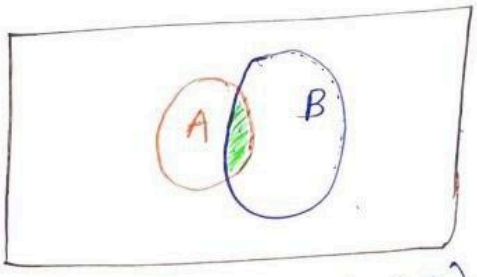
- **Disjoint / Mutually exclusive :-** Two events are disjoint if they can't occur simultaneously.

eg:- Probability of getting either 2 or 4 or 5.
Rolling a dice.



Venn diagram

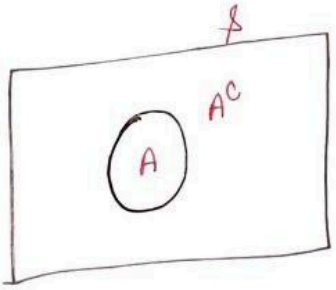
Probability + Addition Rule



$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Revisiting Random Variable

- X is rv with k possible values $x_1, x_2 \dots x_k$
- $P(X = x_j) = p_j$
- Then $\sum_{j=1}^{k} p_j = 1$



$$\Rightarrow P(A \cup A^c) = P(S) = 1$$

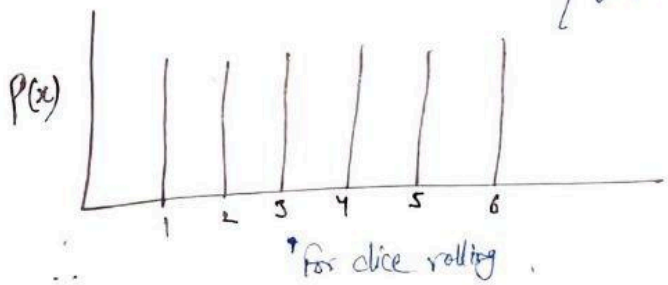
$$P(A) = 1 - P(A^c)$$

- Probability distribution : A list containing $p_1, p_2, \dots p_k$

$\left\{\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}\right\}$

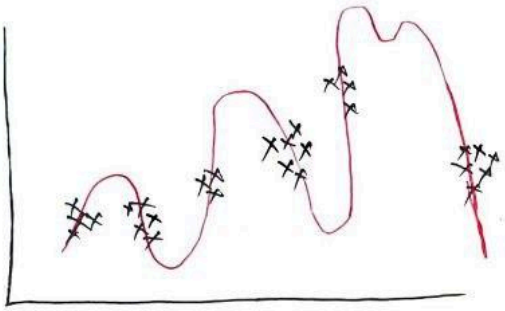
↓

Often represented by Histogram



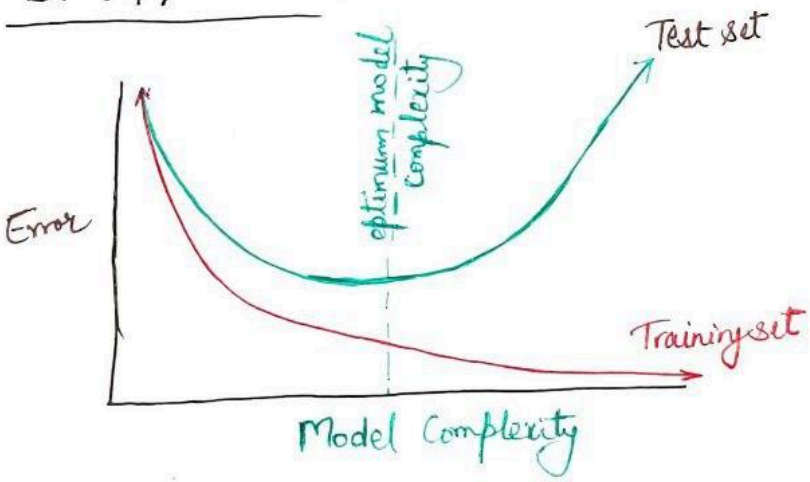
∴ for dice rolling.

Date : 16-08-2024 Lecture-6



- Now, we have a lot of training data but of similar types.

Training Vs. Test set Error :

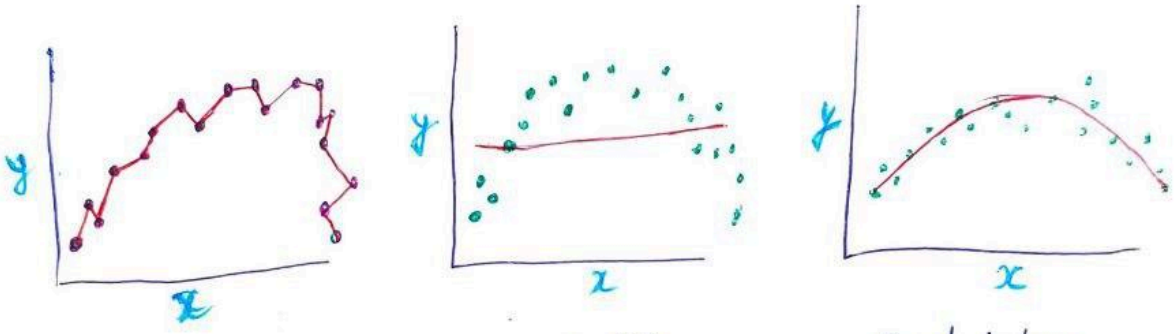


- For every line we can quantitatively calculate bias. Variance can't be calculated around multiple error.
- Qualitatively we can't say bias, variance.
- There is trade-off b/w bias and variance, based on my experience i need to choose reasonable limit.

Ideally we want low bias, low variance.

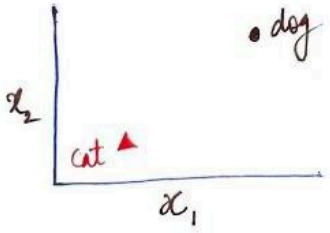
Bias - Variance Tradeoff

- We want ↓↓ bias, variance.
- we need to find sweet spot b/w simple and complex model.

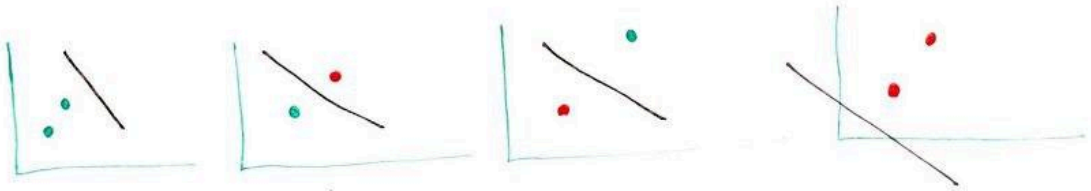


overfitting underfitting Good balance

No. of Data points that a hypothesis can classify correctly :-



x_1 : Height
x_2 : Weight



→ Hypothesis space of straight lines is expressive enough to classify two points in 2D plane.

Note

Atleast 1 configuration of 2 points st. line which can classify correctly.

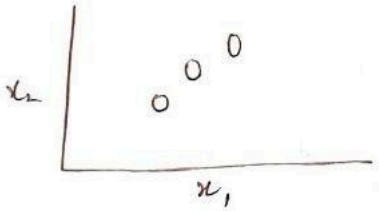
- Hypothesis space of st. line can shatter two points in a 2D plane.

Shattering :- A set of N points is said to be shattered by a hypothesis space H if :-

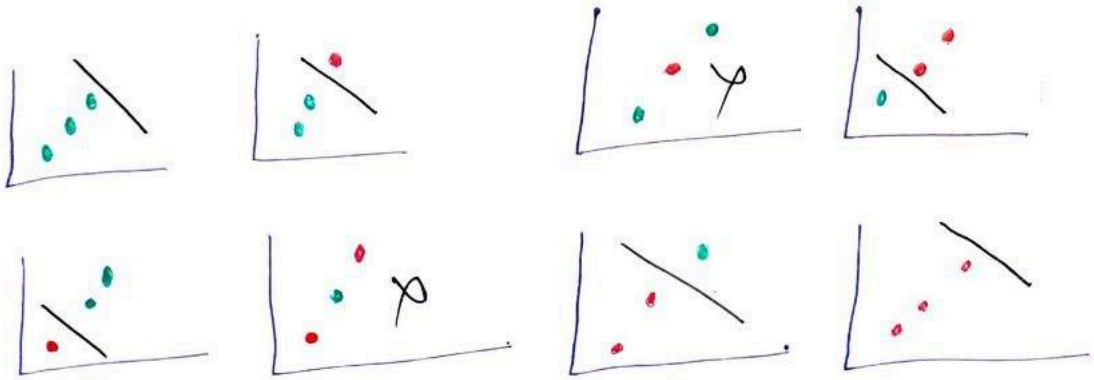
↓

there are hypotheses in H that can seperate +ve and -ve examples in all 2^N possible ways for atleast one configuration of data points.

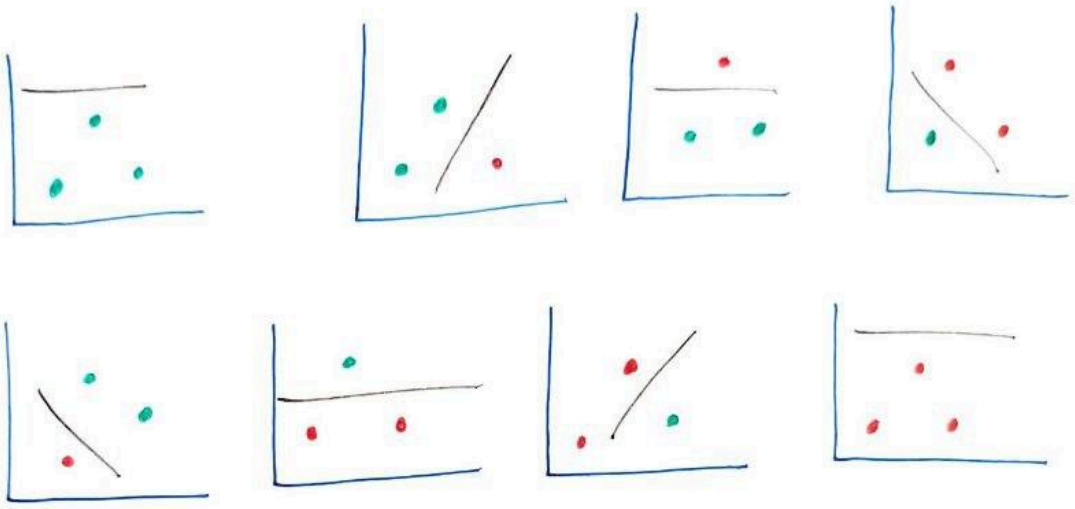
Three Points :-



Can st. line classify all possible combination of classes for 3 points?

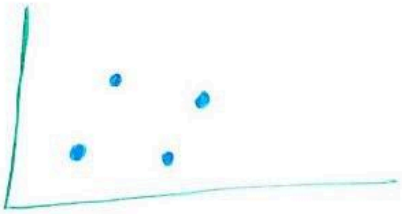


If 3 points are not Co-linear → this is one of possible configuration :- ①



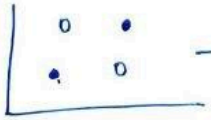
Note :- one classification by 1 configuration is okay.

4 Points :-



Can st line shatter any 4 points in 2D plane : No

Note :- Maximum no. of points that can be shattered by st. line in 2D plane is 3.

eg :-  → can be seperated by st. line.

Vapnik – Chervonenkis (VC-dimension):

VC dimension of hypothesis space consisting of st. lines in 2D plane : 3.

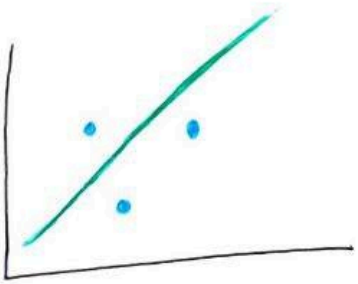
> VC-dimension characterizes the expressive power or capacity of a hypothesis space.

- Maximum # of points that can be shattered by a hypothesis space is called VC dimension of the Hypothesis space.

eg:- If classificatn model can shatter D points atmost by changing its parameters. VC dimension of classification : D.

> ↑↑ expressive hypotheses = VC dimension ↑↑

VC dimension:-

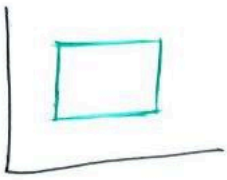


- Consider a classification model that creates a decision boundary:-

$$y = mx + c$$

where m, c = parameters of model

By changing m, c of classification model we'll get diff hypothesis space.



: VC-dimension of hypothesis space consisting of axis-aligned rectangle.

- Lets see classification model with VC-dimension D
- N = no. of training data / samples
- D = VC dimension

Then :

$$P\left(\text{test error} < \text{training error} + \sqrt{\frac{1}{N}\left[D\left(\log\left(\frac{2N}{D}\right)+1\right) - \log\left(\frac{\eta}{4}\right)\right]}\right) = 1-\eta$$

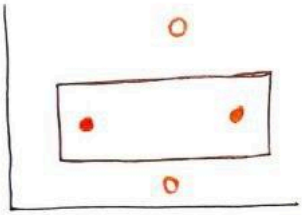
$$0 \leq \eta \leq 1$$

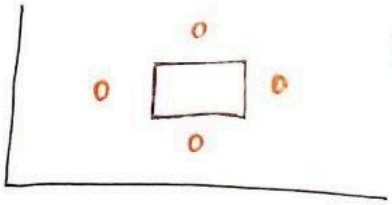
This is applicable when $D << N$.

No. of training sample is way higher than VC-dimension.

If VC-dimension is high, model is more expressive.

- More complex → more chance of overfitting → ↑↑ test error.





VC = 4

We want same error for Testing, training.

η is ↓ Bound to train, test error

$$P\left(\text{test error} < \text{training error} + \sqrt{\frac{1}{N}\left[D\left(\log\left(\frac{2N}{D}\right)+1\right) - \log\left(\frac{\eta}{4}\right)\right]}\right) = 1-\eta$$

Case 1:
If $\eta \downarrow\downarrow\downarrow$ $\quad 1-\eta = 1-0 = 1 \Rightarrow$ Probability $= 1$

Probability of test error < training error $= 1$ $\quad$ Confidence ↑↑

η should be minimum.

Case 2:

if VC dimension is ↑↑ ⇒ test error becomes ↑↑
. can't bound test error

if VC dimension ↑↑ = model is expressive
↓
more complex

↳ more complex means → more chance of overfitting.

So, higher test error.

Challenges in Data for ML :-

- Good ML models can't be designed without good quality data.

> Domain Knowledge is needed to handle such data.

- Various problems with most practical data :-
  - → Noise
  - → Duplicate data
  - → Missing data
  - → Inconsistent data etc.

- Data Cleaning is done to :-
  - → Reduce noise
  - → handle inconsistent data
  - → handle duplicate data
  - → handle missing data
  - → May require domain knowledge to deal with these problems.

Date : 20|08|2024 Lecture - 7 P23CS0010 ①

Challenges in Data for ML

(i) Missing data eg :-

Student id
10012
10013
10024
10098

(ii) Inconsistent data

Classification
Junior
Senior
Sophomore
Sr
Senior
Jr

ID
2102
2214
AUC8 → id starting with letter.
2308

(iii) Noisy data

Age
21
200
24
23
-2.8
30

Quality of Model = Quality of data

↓

performance of model depend on data

How to clean data, fill data?

:Q: Pattern can only identified correctly by correct data.

⊛ While coming up with a hypothesis :—

Step 1 :- Clean data
Step 2 :- Normalization
Step 3 :- Coming up to a best hypothesis
step 4 :- Error Minimization

Step 1 :- Clean Data :- Data cleaning involves :

- Reducing noise
- Handling duplicate data
- Handling inconsistent data
- Handling missing values

} • Domain Knowledge is required to deal with these problems.

eg :- If a person has (hemoglobin = 30) ← incorrect data

↑ need domain knowledge to fill / deal with data.

Dealing with cleaning of data

(i) Dealing with missing values :

- Data deletion : results in loss of data
- Imputation : predict data, through informed guess, by using regression.

(ii) Dealing with noise :

(a) Binning :- Divide total data into bins.

Data → 

- replace value of noise data based on some characteristic of bin (mean, avg, min, max).

(b) Regression : based on other parameters → predict

(c) Outlier analysis : treat them seperately or eliminate those.

Scale features :- suppose I want to find abnormalities :—

eg :- avg Hb : 13500 mg/dL
avg Creatine : 0.7 mg/dL
— Avg values - normal level

suppose if a person has :-

Hb : 12500 mg/dL
Creatine : 1.7 mg/dL

Hb : 13500 - 12500 = 1000
Creatine : 1.7 - 0.7 = 1

based on huge change in Hb compared to Creatinine

↓

we might conclude person has Hb disorder

→ This wrong conclusion may happen because:

↓

normal values are different range.

- We want to bring different features at same range.

- If we don't normalise

↓

features with lower value range may loose their meaning when combined response is created.

- Large spread in feature value may mislead the ML model.!

[may mislead ML model
lower feature may loose their meaning when collectively taken.

↓

Normalization is solution

↓

→ data Standardization
→ Min-max normalization

(i) Data Standardization : Z-Score normalization :

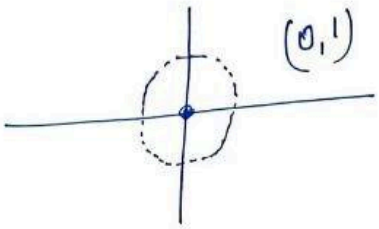
- Std^n at input :- create feature value with mean = 0, variance = 1 / unit
- m data points.
- feature values for n data points : $f_1, f_2 \cdots f_n$
- Standardized feature value :

$$f_i' = \frac{f_i - \mu_f}{\sigma_f}$$

μ_f = mean of $f_1, f_2 \cdots f_n$

σ_f = std of $f_1, f_2 \cdots f_n$

- Std. value will have (mean 0, unit variance)



(ii) Min-Max Normalization :-

- n data points
- feature values for n data points : $f_1, f_2 \cdots f_n$
- normalized feature value :

$$f_i' = \frac{f_i - \min(f_1, f_2 \cdots f_n)}{\max(f_1, f_2 \cdots f_n) - \min(f_1 \cdots f_n)}$$

$$f_i' = \frac{f_i - \min(f_1, f_2 \cdots f_n)}{\max(f_1, f_2 \cdots f_n) - \min(f_1 \cdots f_n)}$$

$\max(f_1, f_2 \cdots f_n) - \min(f_1 \cdots f_n)$ = denominator

$f_i - \min(f_1 \cdots f_n)$ = numerator

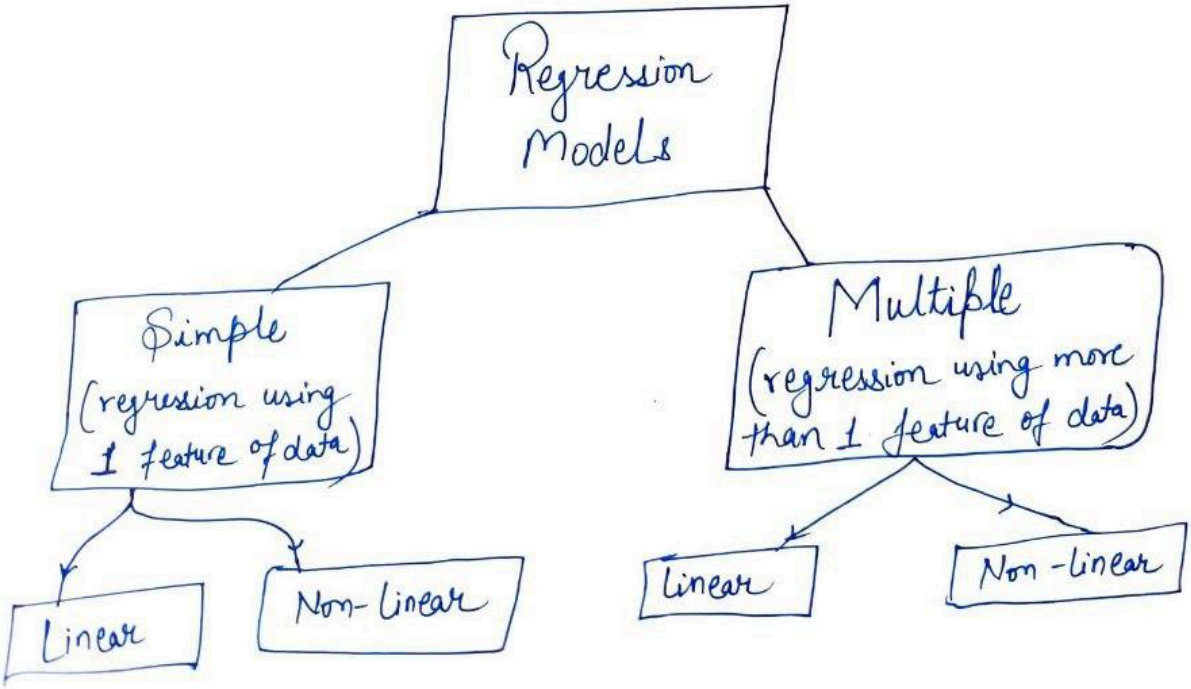
deno. $\geq$ numerator

Q :- Given H_b values, find min-max normalization :-

H_b
11.7
13.4
14.5
12.2

$$f_1' = \frac{11.7 - 11.7}{14.5 - 11.7} = 0$$

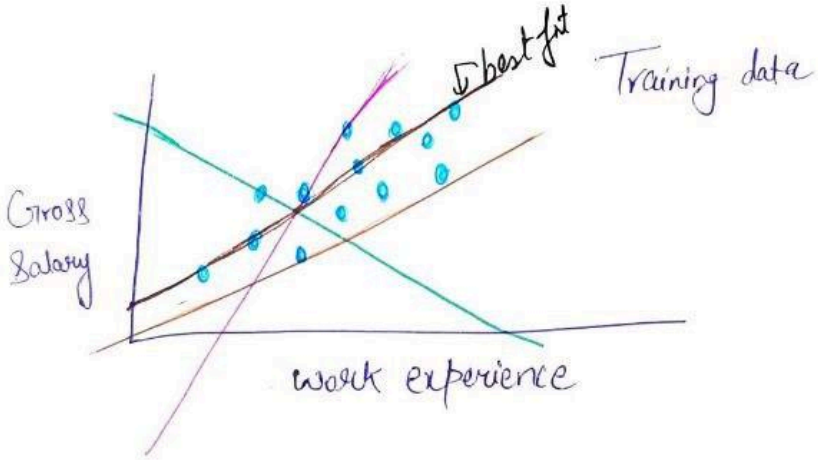
$$f_2' = \frac{13.4 - 11.7}{14.5 - 11.7} = 0.6$$



Regression :- eg: predicting price of house.

- ML introduces Inductive Bias : (a) restrictive (b) Reference

eg:- restrictive inductive bias is used for St. line.



- Straight line best fits on given data.

St. line ⇒ explain more points in good way.

Suitable line ⟶ finding Suitable value of (m, c)

- eqⁿ of St. line

$$y = mx + c$$

parameter of eqⁿ / regression model

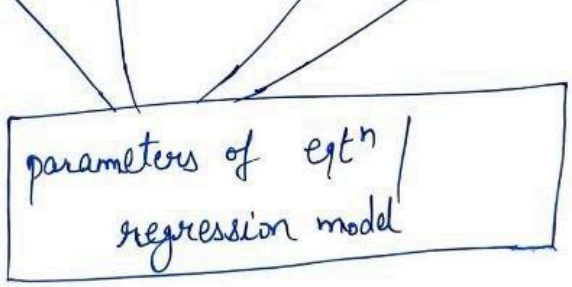
- different values of (m, C) → different straight lines
- Different st. lines, which to prefer?

  ↓ use Error

  [Brown line has least total error]
- Find st. line (m, C parameter) that minimizes total error in training data.

For multiple linear Regression :-

$$y = m_1 x_1 + m_2 x_2 \ldots\ldots m_p x_p + C$$

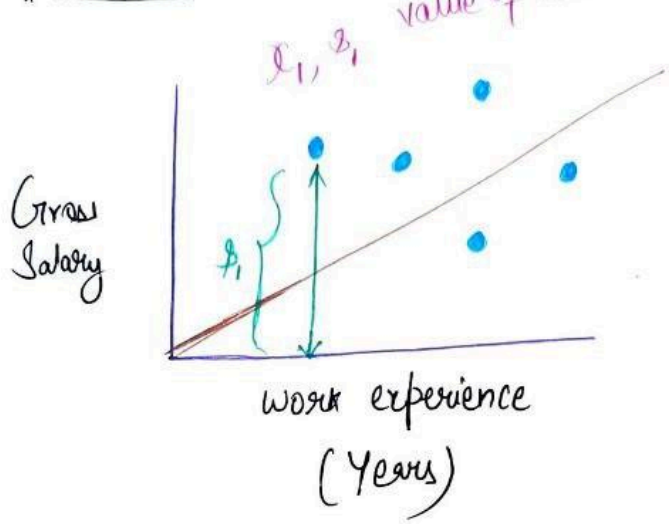


→ find st. line ($m_1, m_2 \ldots m_p$ and C) that minimizes total error in training data.

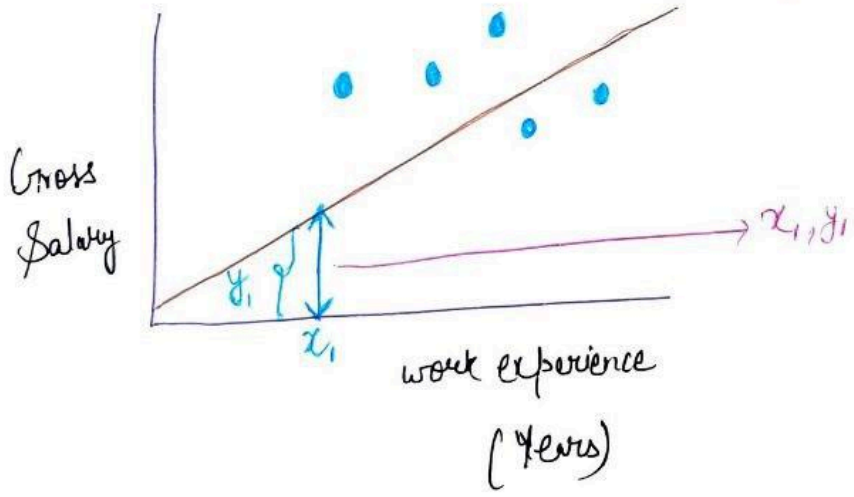
eg:- y = credit score depend on multiple factor :

x_1 = how many loan
x_2 = how many bank account
x_3 = previous installment history
x_4 = #. of experience in bank account
etc.

Use of Error :-



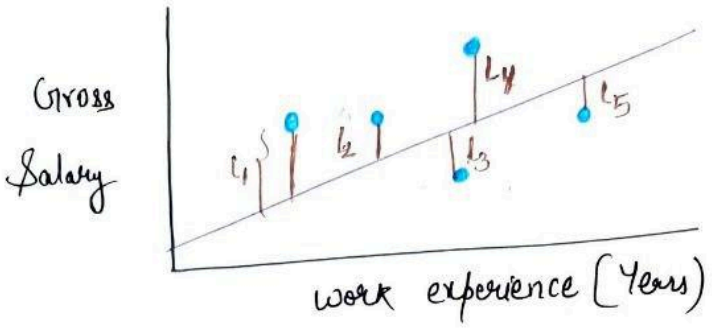
$$Y_1 = mx_1 + c$$



Error for this data point w.r.t prediction

$$e_1 = (y_1 - \hat{y}_1)^2 + (x_1 - x_1)^2$$

$$= (y_1 - \hat{y}_1)^2$$



$$E = e_1 + e_2 - \cdots + e_N$$

- We can calculate error for training data point and sum them upto get total error.

$$E = (y_1 - s_1)^2 + (y_2 - s_2)^2 \cdots\cdots + (y_N - s_N)^2$$

$$E = \sum_{n=1}^{N} (y_n - s_n)^2$$

- Goal :- To find a hypothesis such that error E is minimized

$$J(m,c) = \frac{1}{2} \sum_{n=1}^{N} (y_n - s_n)^2$$

[find m, c such that above is minimized]

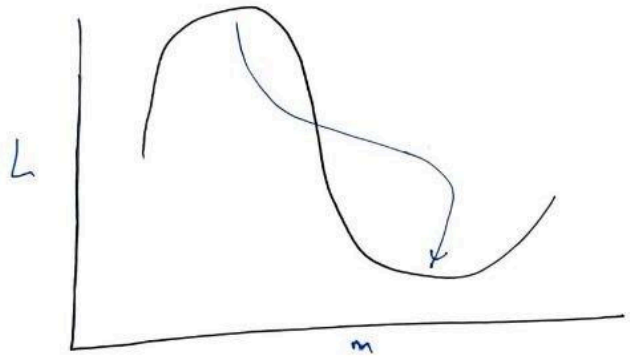
- minimize squared error = Least-Square regression.

$$J(m,c) = \frac{1}{2} \sum_{N=1}^{N} \left(mx_n + c - y_n \right)^2$$

↓

minimize square regression

How to minimize $J(m,c)$:—

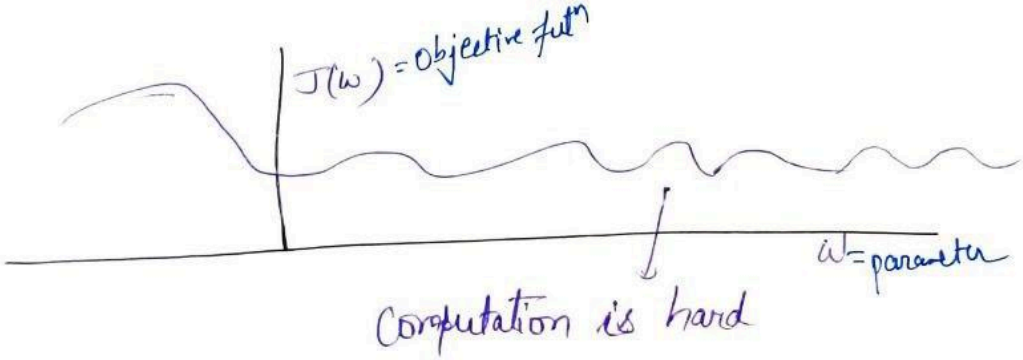


Goal :- of training is to get m, c

↑

to minimize objective function

- goal is to reach global minima
- Double derivative finds min, max values.



Computation is hard

Lecture -8

Revisiting Random Variable

- X is a discreate rv with K possible values $x_1, x_2 \dots x_K$.

  eg:- X contain $\{0, 1\}$ while tossing coin.

  $$P(X = x_j) = p_j \qquad \text{Then} \sum_{i=1}^{K} p_i = 1$$

- **Probability Distribution** : list containing $p_1, p_2 \dots p_K$
  - ↳ often represented as histogram.

Cond^n :- Sample Space — finite, Equally likely

- If probability isn't equally likely : **frequentist approach**
  - ↑ do experiment 'os' no. of times.

(r.v discrete)
- **Skewed** o/p : if experiment is less no. of times.
- **Equally Probable** :- if experiment done many no. of times.

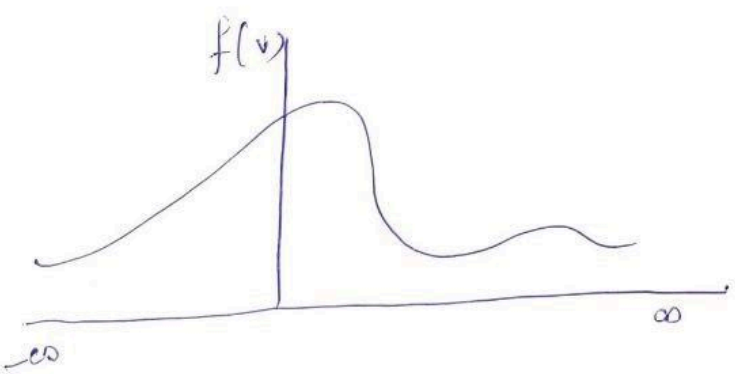
(b) If S.S is continuous :- For probability density function ↑ probability range is needed.

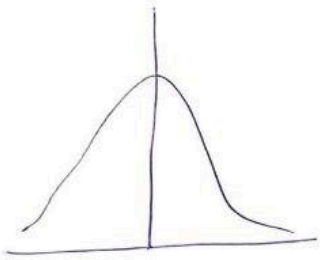
$$P(a < X \leq b) = \int_a^b f(v)\,dv$$

- If X is continuous r.v :-
  probability of value of X are represented by a curve $f(v)$.
  - → $f(v) \geq 0$ at all points.
  - → Area under curve = 1.

$$\int_{-\infty}^{\infty} f(v)\,dv = 1$$

Probability range is needed for continuous data.





gaussian, normal distribution

- In ML we want to learn distribution of data.

Conditional Probability :-

In a box, Samsung phones = 40, MI phones = 20

out of these 10 samsung, 2 MI phone are defective / not working

You pick up a phone and find probability of (phone isn't working) :- ?

i/p of Samsung Company ?

Given :- picked up phone is not working.

Goal :- if so probability of its being Samsung ?

$$P\left(\frac{SP}{NW}\right) = \frac{\#\text{ Samsung phone that are not working}}{\#\text{ phones that aren't working}}$$

$$= \frac{10}{12}$$ multiplying, ÷ by 40, 60

$$= \frac{\frac{10}{40} * \frac{40}{60}}{\frac{12}{60}}$$

$\frac{10}{40}$: Given Samsung phone that is NW

$\frac{40}{60}$: Probability of SP in box

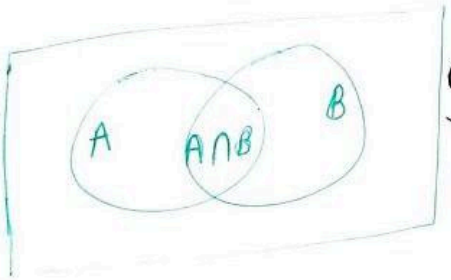
$\frac{12}{60}$: prob. of NW phones

$$P\left(\frac{SP}{NW}\right) = \frac{P(NW/SP)\ P(SP)}{P(NW)}$$

$$P\left(\frac{SP}{NW}\right) = \frac{P(SP \cap NW)}{P(NW)}$$

Conditional probability ⊢

$$P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{P\left(\frac{B}{A}\right) \cdot P(A)}{P(B)}$$ → Prior probability



Baye's Theorem

Conditional / Posterior probability

$$P\left(\frac{A}{B}\right) = \frac{P\left(\frac{B}{A}\right) \cdot P(A)}{P(B)}$$

→ prior (indicate my belief about occurance of A)

↓ evidence

that probability that B occurs

likelihood

$$P(\text{Cause}/\text{Effect}) = \frac{P(\text{Effect}/\text{Cause})\, P(\text{Cause})}{P(\text{Effect})}$$

- If a system with causes C_1, C_2 and e_1, e_2

Q:- If we want to find probability of diff. causes given effect e_2?

$$P(C_1/e_2) = \frac{P(e_2/C_1)\, P(C_1)}{P(e_2)}$$

$$P(C_2/e_2) = \frac{P(e_2/C_2)\, P(C_2)}{P(e_2)}$$

Since e_2 may be caused either by C_1, C_2

$$\boxed{P(C_1/e_2) + P(C_2/e_2) = 1}$$

$$\boxed{\text{Posterior} = \frac{\text{Prior} * \text{Likelihood}}{\text{Evidence}}}$$

Similarly $P(C_1/e_1) = \frac{P(e_1/C_1)\, P(C_1)}{P(e_1)}$ $\quad P(C_2/e_1) = \frac{P(e_1/C_2) * P(C_2)}{P(e_1)}$

- Evidence are fever

also $P\left(\frac{C_1}{e_1}\right) + P\left(\frac{C_2}{e_1}\right) = 1$

Posterior ∝ likelihood * prior

eg: given i get cough/fever (A) — A doesn't depend on B.
There is power cut (B)

P(fever) = only depends on A.

Q:- If among two person what is probability of having birthday on same month ?

$P(A) = \frac{1}{12}$ $\quad P(B) = \frac{1}{12}$

$P(A).P(B) = \frac{1}{12} * \frac{1}{12}$

If A,B born in Jan = $\frac{1}{144}$

in Feb = $\frac{1}{144}$ - - -

$P(x) = P(\text{Jan}) + P(\text{Feb}) - - - + P(\text{Dec})$

$= \frac{1}{144} + \cdots \frac{1}{144} = \frac{12}{144} = \frac{1}{12}$

- Since e_1 may be caused by C_1 or C_2.

$$P(C_1/e_1) + P(C_2/e_1) = 1$$

- For both $P\left(\frac{C_1}{e_1}\right)$ and $P\left(\frac{C_2}{e_1}\right)$ denominator is same :-

Let's consider $\frac{1}{P(e_1)} = \alpha$

Then :-

$$P(C_1/e_1) = P\left(\frac{e_1/C_1)\, P(C_1)}{P(e_1)}\right) = \alpha\, P\left(\frac{e_1}{C_1}\right) P(C_1)$$

$$\therefore P(C_2/e_1) = \alpha\, P\left(\frac{e_1}{C_2}\right) \cdot P(C_2)$$

Posterior $\propto$ (likelihood $\times$ prior)

Independent Events :- Two events A, B are independent if :-

- $P\left(\frac{A}{B}\right) = P(A)$ (A doesn't depend on B)

$$P\left(\frac{A}{B}\right) = \frac{P(A \cap B)}{P(B)} = P(A)$$

- $P\left(\frac{B}{A}\right) = P(B)$

$$P(A \cap B) = P(A) \cdot P(B)$$

Lecture - 9 (22-08-2024)

How to reach Global Minima : 2D ex. :-

Important point to note :-

- if superscript : Sample
- if subscript : variables

eg :- x^1, x^2 1st person, 2nd person .... so on.

$x_1, x_2 \ldots$: $y = mx + c$ = independent data points.

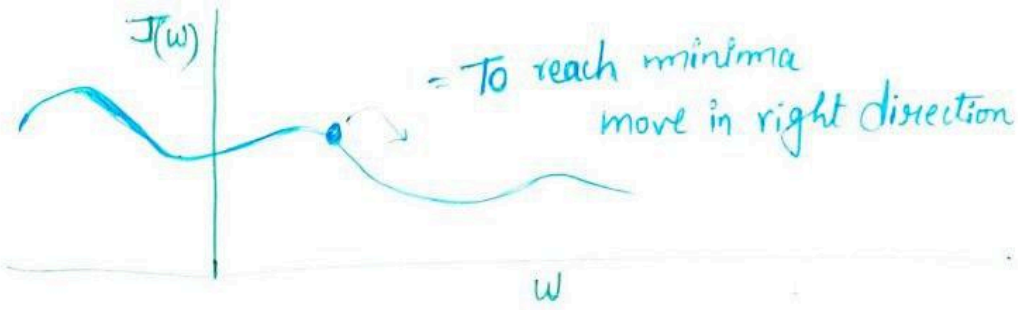
- Minimization can be done in/by many approaches
  - best is Gradient Descent
    - minimizing $(m, c) \Rightarrow$ changing $(m, c) ==$ diff. st. lines

$$J(m,c) = \frac{1}{2} \sum_{n=1}^{N} \left(mx^n + c - s^n \right)^2$$

minimize $J(m, c)$

- if multiple minimizⁿ of linear regression :-

$$y = m_1 x_1 + m_2 x_2 - \ldots - + m_p x_p + c$$



if objective fun^n in right side = error decreases.

- moving in right side :-

$\therefore$ [increasing value of w
reduces value of L(w)]

$$\frac{\partial J(w)}{\partial w} < 0$$

It concludes as :- $J(w)\downarrow$ — [$\partial J(w)\ \downarrow\downarrow = -ve$; $\partial w\ \uparrow\uparrow = +ve$]

$\left(\frac{\partial J(w)\ -ve}{\partial w\ +ve}\right)$ overall its -ve

moving towards right $\longrightarrow$ need new w value.

$$w_{new} = w_{old} + \eta(\Delta w) \rightarrow +\ (-)*(-ve)$$

$$w_{new} = w_{old} + \eta(\Delta w)$$

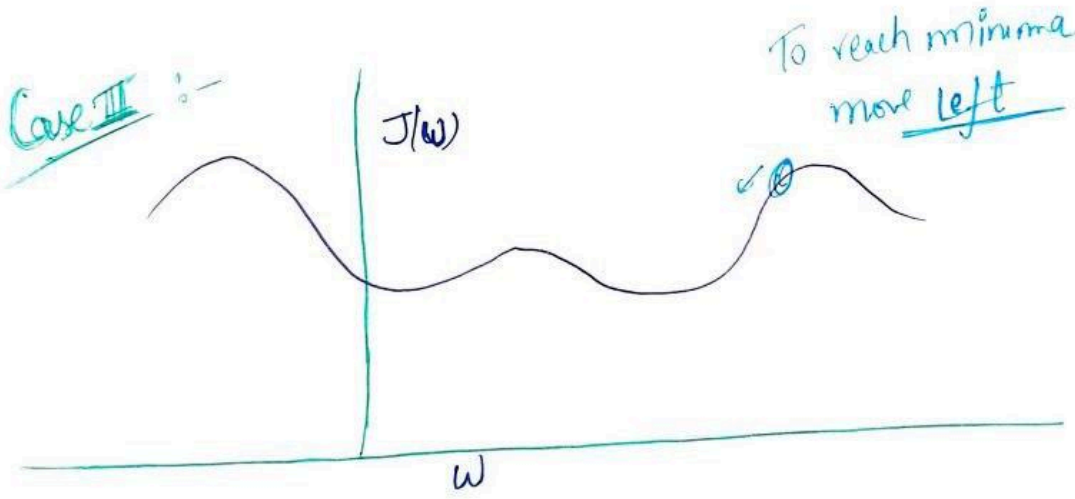
$\frac{\partial J(w)}{\partial w}$ = -ve $\Rightarrow$ to make it +ve, multiply by -ve value.

$$w_{new} = w_{old} - \eta \left[\frac{\partial J(w)}{\partial w}\right] \rightarrow -ve$$

overall +ve

η = +ve constant

Case III :-



To reach minima move left

To decrease w :-

$$w \leftarrow w - \eta \frac{\partial J(w)}{\partial (w)}$$

$\left(\frac{\partial J(w) \rightarrow -ve}{\partial w \rightarrow -ve}\right)$ +ve

- After multiple checking of η values, testing on :-

$\eta = 0.1, 0.2 \ldots$ we conclude at particular value of η, we are getting less error, best result.

η = learning rate, generally = 0.01

How to decide η : ? Test performance of model on validation data.

How to know minima is reached: Training error will reduce.

eg: I may fix epoch 1000 times and i assume minima will be reached and return particular value.

x^n = n^{th} data point

x_p = p^{th} # of variable

$x_1, x_2 \ldots x_p$ = P # of parameters

- w_t is updating at every step.

Gradient Descent :-

(i) Initialize parameter w

(ii) loop until convergence

(a) Compute gradient $\frac{\partial J(w)}{\partial w}$

(b) update parameters $w \leftarrow w - \eta \frac{\partial J(w)}{\partial w}$

(iii) Return parameters

η = hyperparameter whose value should be set using validation performance.

for many parameters :-

(i) Initialize parameter $m = \{m_1, m_2 \dots m_p\}$

(ii) loop until convergence

(a) Compute gradient $\frac{\partial J(m)}{\partial m}$

(b) update parameters $\theta_i \leftarrow \theta_i - \eta \frac{\partial J(m)}{\partial m_i}$

(iii) return parameters.

$$J(m) = \frac{1}{2} \sum_{n=1}^{N} \left(y^n - s^n\right)^2$$

find expression for update :-

$$\theta_i \leftarrow \theta_i - \eta \frac{\partial J(m)}{\partial (m_i)}$$

$$m_i \leftarrow m_i - \eta \sum_{n=1}^{N} \left(y^n - s^n\right) x_i^n$$

Gradient descent for all training samples :→ Batch G.D

Challenges :- As it computes gradient over all data, computing gradient over millions of training data is
- → computationally hard
- → need huge storage

In Batch G.D :- N training samples (data points)
↑
we compute total objective fun^n considering all N training samples.

$$J(m,c) = \frac{1}{2} \sum_{n=1}^{N} \left(mx_n + c - s_n\right)^2$$

$$\frac{\partial J(w)}{\partial w} = \frac{1}{2} \sum_{n=1}^{N} \frac{\partial}{\partial w} \left(mx_n + c - s_n\right)^2$$

→ for each update : it will take long time if we have huge training samples

→ so we need to compute and store N gradients for each update.

Batch G.D → Stochastic Gradient Descent

- In Batch G.D with N samples :-

$$J(m,c) = \sum_{N=1}^{N} \frac{1}{2}\left(mx_n + c - y_n\right)^2 = \sum_{n=1}^{N} J_n(m,c)$$

Cond^ :- If I calculate objective fun for 1 sample and do parameter update.
Do same process for 2nd points. [For Batch G.D : need to wait till n data points]

Stochastic G.D :- each update would require only one gradient computation [less time, less storage for gradient values]

eg :- Going from jodhpur to mehrangarh fort
↓
multiple paths are there
↓
reach at particular point → again discuss → decide.

[1 particular step = will take significant time,
Collective opinion (lots of Noise)
going by discussion based decision has as diff person has diff opinions.

SGD :-

1. Initialize parameters
2. loop until convergence
   (i) Randomly shuffle samples in training dataset
   (ii) For training sample $n = 1, 2, \ldots N$
      (a) Compute gradient $\frac{\partial J_n(w)}{\partial w}$
      (b) Update parameters $w \leftarrow w - \eta \frac{\partial J_n(w)}{\partial w}$
3. Return parameters.

Challenges in Stochastic and Batch GD :-

- BGD :- not data efficient when data is very similar.
- BGD :- high computational time for each step.

SGD :- not computationally efficient

- noisy results since at a time, one data point is considered.

Solution :- Minibatch GD.

Minibatch GD :- neither to
- single points
- nor to whole group.

It is more likely to be correct., less noisy-, fast moment.

Minibatch GD (i) Initialize parameters

(ii) Loop until convergence

(a) Randomly shuffle samples in training dataset

(b) for training dataset, create b minibatch of size k

(c) for each minibatches $n = 1, 2 \cdots b$

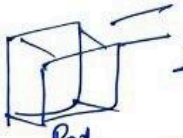
- Compute gradient $\frac{\partial J_n(w)}{\partial w}$
- Update parameters $w \leftarrow w - \eta \frac{\partial J(w)}{\partial w}$

(iii) Return Parameters

How to initialize ? Can be all zeros ?
Can be random ?

. Performance depends on initialization.

Old ex. in New form :- not working phones



Tell whether phone is Samsung or MI ?

Find $P\left(\frac{SP}{NW}\right)$ and $P\left(\frac{MIP}{NW}\right)$?

If $P\left(\frac{SP}{NW}\right) > P\left(\frac{MIP}{NW}\right) \Rightarrow$ i had picked Samsung phone.

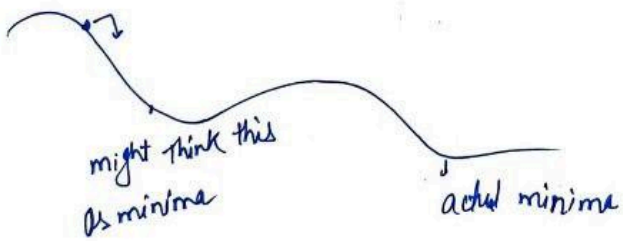
else MI

$$P\left(\frac{MI}{NW}\right) = \frac{\#\ MI\ phones\ That\ are\ not\ working}{\#\ phones\ that\ are\ not\ working} = \frac{2}{12}$$

$$P\left(\frac{SP}{NW}\right) = \frac{10}{12}$$

$P\left(\frac{SP}{NW}\right) > P\left(\frac{MI}{NW}\right) \Rightarrow$ concludes that i picked up Samsung phone

Q:- Eventual minima is not effected by linear regression?



In linear regression :- gradient is constant because relationship is a straight line. This means it always leads directly to same minimum — point where error is smallest regardless of where we start.

In Curves :- gradient changes as we move along curve, which lead to different paths and potentially different minima depending on where we start.

Lecture – 10 (P23CS0010) Date: 27/08/2024

Baye's Decision theory :

Joint Probability Distribution :

Consider 2 random variables (X, Y) :

x : Corresponding to weather $\{$ sunny, rainy, cloudy $\}$

$P(x)$: $\{0.6, 0.1, 0.3\}$

Y : Corresponding to power cut $\{$ power cut, no power cut $\}$

$P(Y) = \{0.15, 0.85\}$

> A joint probability distribution of x and Y (pairs considered)
> ↓
> probability distribution on all possible pairs of x's.

∴ $3 \times 2 = 6$ possible cases

- $\sum$ of all possible outcome $= 1$

3×2 matrix of values

- Joint probability is necessary in daily life.

  eg: AC temp. depend on wheather, our mood etc.

Chain Rule :-

$$P(AB) = P(A \cap B) = P(A/B)\,P(B)$$

If $A_1, A_2 \ldots A_n$ are n events, then

$$P(A_n \cap A_{n-1} \ldots \cap A_1) = P\left(\frac{A_n}{A_{n-1} \ldots A_1}\right) P(A_{n-1} \cap \ldots A_1)$$

Similarly :-

$$P\left(\frac{A_n}{A_{n-1} \cap \ldots A_1}\right) P\left(\frac{A_{n-1}}{A_{n-2} \cap \ldots A_1}\right) P(A_{n-2} \cap \ldots A_1)$$

eg :- .... so on

If 4 boxes :-

$$P(A_4 \cap A_3 \cap A_2 \cap A_1) =$$

$$P\left(\frac{A_4}{A_3 \cap A_2 \cap A_1}\right) P\left(\frac{A_3}{A_2 \cap A_1}\right) P\left(\frac{A_2}{A_1}\right) P(A_1)$$

	fever		¬ fever	
	Cough	¬ Cough	Cough	¬ Cough
Covid	0.21	0.10	0.11	0.08
¬ Covid	0.11	0.07	0.09	0.23

$P(\text{fever})$ = Cough - Covid possible pairs { all exhaustive Combination of variable

$P(\text{fever}) = 0.21 + 0.10 + 0.11 + 0.07 = 0.49$

$P(f \vee C) = P(f) \vee P(C) \vee P(f \wedge C)$

$= 0.21 + 0.11 + 0.10 + 0.07 + 0.11 + 0.08 = 0.68$

$$P(\neg \text{Covid} \mid \neg \text{fever}) = \frac{P(\neg \text{Covid} \cap \neg \text{fever})}{P(\neg \text{fever})} = \frac{0.09 + 0.23}{(0.11 + 0.09 + 0.08 + 0.23)}$$

$= 0.627$

Conditional probability :- tells joint prediction

↓

if know whole distribution ; can answer any query

if $P\left(\frac{SP}{NW}\right) = P\left(\frac{MI}{NW}\right) \Rightarrow$ can't make any decision

eg2:- Sample space corresponding to X is $S_x = \{s, r, c\}$

" " " " Y . $S_y = \{pc, npc\}$

Sample space corresponding to joint distribution is:-

$$S_J = \{(s, pc), (s, npc), (r, pc), (r, npc), (c, pc), (c, npc)\}$$

eg3:- if $P\left(\frac{SP}{NW}\right) > P\left(\frac{MI}{NW}\right)$: phone picked up is samsung.
(10/12) (2/12)

if $P\left(\frac{MI}{NW}\right) > P\left(\frac{SP}{NW}\right)$: picked up phone is MI

eg4:- Let's say 2 class classification problem with C_1, C_2

C_1 = cat , C_2 = tiger

If i have a data point x as feature that i need to classify to one of these 2 classes

x : weight of animal

evidence : it tells proportion of animal there in zoo.

Prior : wt

likelihood : animal wid particular weight is x, yz etc

Using Bayes' decision rule, I will find out :-

$$P(Cl_1/x) \quad \text{and} \quad P(Cl_2/x)$$

$$P\left(\frac{Cl_1}{x}\right) > P\left(\frac{Cl_2}{x}\right) \Rightarrow Cl_1$$

$$P\left(\frac{Cl_1}{x}\right) = \frac{P\left(\frac{x}{Cl_1}\right) * P(Cl_1)}{P(x)} \qquad P\left(\frac{Cl_2}{x}\right) = \frac{P(x/Cl_2)\,P(Cl_2)}{P(x)}$$

Let's use Bayes theorem :-

Class conditional probability

$$P(Cl_1/x) = \frac{P(x/Cl_1)\,P(Cl_1)}{P(x)} \qquad P\left(\frac{Cl_2}{x}\right) = \frac{P(x/Cl_2)\,P(Cl_2)}{P(x)}$$

Prior probability (my belief about existence of particular class)

Total no of animal in 200 → fixed

Particular ratio of animal in 200

probability depend on ↑ weight i picted up of particular animal. $P\left(\frac{x}{Cl_i}\right)$

$$P(Cl_1|x) > P(Cl_2|x) \Rightarrow Cl_1$$

After expending :—

$$P\left(\frac{x}{Cl_1}\right) P(Cl_1) > P\left(\frac{x}{Cl_2}\right) P(Cl_2) \Rightarrow Cl_1$$

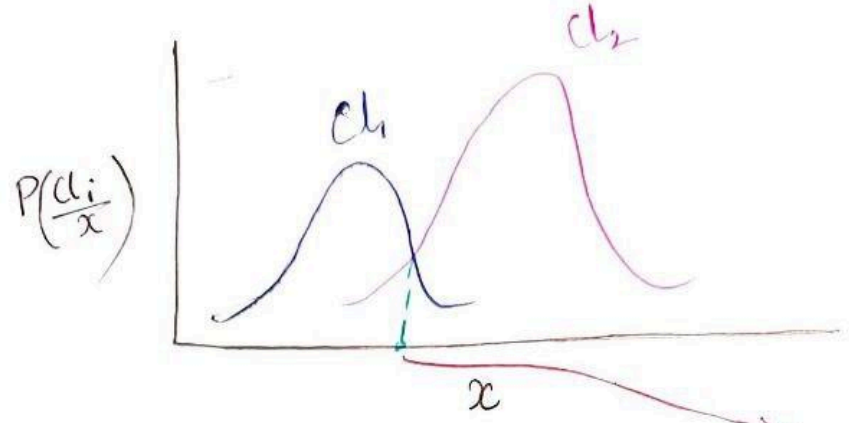
$$P\left(\frac{Cl}{x}\right) = \text{Posterior} = \frac{\text{Prior} \times \text{Likelihood}}{\text{Evidence}}$$

$P\left(\frac{x}{Cl_i}\right)$ = Changes : Class Conditional probability.

[$P(x)$ = fixed , common in both classes.
Before coming to zoo inside, already presence of no. of animal in it fixed]

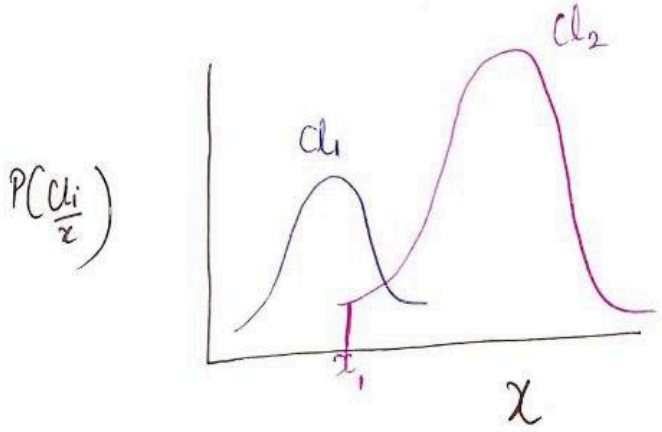
So Posterior depends on :—

Posterior $\propto$ likelihood $\times$ prior



Decision Boundary

$$P\left(\frac{Cl_1}{x}\right) = P\left(\frac{Cl_2}{x}\right)$$



Error in decision :

$$\min\left\{P\left(\frac{Cl_1}{x}\right),\ P\left(\frac{Cl_2}{x}\right)\right\}$$

i.e $P\left(\frac{Cl_1}{x}\right) = 0.6$ $\quad P\left(\frac{Cl_2}{x}\right) = 0.4$

↑ error of this much

Class Cl_1 correct probability $P\left(\frac{Cl_1}{x}\right)$

Correct decision is obtained only when we consider minimum error probability.

[if I take decision = C_2
Error probability $P\left(\frac{C_1}{x}\right)$

min error probab = $P\left(\frac{C_2}{x}\right)$

- Probability 1 ⟶ Consider error of other class i.e C_2
- Probability 2 ⟶ error,

check : $\{error_1, error_2\}$ min

↓

that is class.

Loss Function

Lecture-11

Date : 29/8/2024

- Consider there are K no. of classes :-
  - → $Cl_1, Cl_2 \dots Cl_K$
  - States of nature
- Consider that there are 'q' no. of actions :-
  - $\alpha_1, \alpha_2 \dots \alpha_q$
  - → Action can be assigning one class to the data
  - → Action can also be assigning no class when there is tie.
- Loss function :-

  $\lambda\left(\frac{\alpha_i}{Cl_j}\right)$: Loss incurred for taking action α_i when state of nature is Cl_j

- We consider a data point x to be d-dimensional feature vector.
- Generic error in terms of Loss.

  Action :- assigning class label to any class

  State of Nature : $Cl_1, Cl_2 \dots Cl_K$ [K Classes]

- State of nature = all classes
- If classes are 'equal' : might not take action

~ Given state of nature,
Actual class label Ck_2.

For error = 'λ' function

eg: State of nature { Cat, dog }

Data point $\in$ to Dog and we say 'cat'
Acc. to this two state of nature loss λ
if told correct classification, loss = 0

- Earlier with example of goo, Animal weight, 1 feature = 1D was considered
now & D-dimension' is considered.

eg: weight, color, height, eyes ... etc = D-dimensional.

Expected loss :-

$X \rightarrow x$

$x_1, x_2 \ldots x_N$

Expected value :- $\sum_{i=1}^{N} x_i P(x_i)$

x_i = don't know which class it belongs to

α_i = Action taken , K = # of classes.

Cl_j = True state.

Posterior probability = $P\left(\frac{Cl_j}{x}\right)$

- Expected loss for taking action α_i when we observe data x

$$\sum_{j=1}^{K} \left(\begin{array}{l}\text{Loss of being in } C_i \\ \text{considering class } j \\ \text{as true}\end{array}\right) * \text{Posterior probability}$$

Expected loss / Risk function / Conditional Risk :-

$$R\left(\frac{\alpha_i}{x}\right) = \sum_{j=1}^{K} \lambda\left(\frac{\alpha_i}{Cl_j}\right) P\left(\frac{Cl_j}{x}\right)$$

↑

expected loss for taking action α_i , when we observe data x.

$\Rightarrow \lambda\left(\frac{\alpha_i}{Cl_1}\right) \cdot P\left(\frac{Cl_1}{x}\right) + \lambda\left(\frac{\alpha_i}{Cl_2}\right) P\left(\frac{Cl_2}{x}\right) \ldots\ldots$

$\boxed{\lambda_{ij} = \lambda\left(\frac{\alpha_i}{(Cl_j)}\right)}$

- We want to take an action which minimizes the risk.

Minimum Risk Classifier.

eg:- If state of Nature

$$X = \begin{Bmatrix} \text{Pneumonia} \\ \text{Viral} \\ \text{Cardiac} \\ \text{None} \end{Bmatrix}$$

If doctor predicts for pneumonia, but actually its cardiac.
If doctor starts medicine, risk of health, time.

Goal is = minimum error

$$\lambda\left(\frac{\alpha_i}{C_j}\right) = \lambda_{ij}$$

Classification Problem :- Assigning point to class, given risk (observation).

- Minimum Risk Classifier = do classification that minimizes risk.

λ_{ij} = loss of taking action α_i, given true class C_j

if $i == j$ = correct action
∴ loss should be minimum.

Two class Classification :-

$$R\left(\frac{\alpha_i}{x}\right) = \sum_{j=1}^{2} \lambda_{ij} \cdot P\left(\frac{C_j}{x}\right)$$

α_{11} = True class 1, Action α_1 → True, error should be small.

α_{21} :- action → class 2, actual class 1 → error

- There might be no decision too.

extra
- Action may not be related to classification
  eg:- Doctor might call ambulance not classified in any disease

- α_j = I may have more than K no. of actions
  K = # of class
  It may be $k \geq K$ for not taking action.

$$R(\alpha_1|x) = \sum_{j=1}^{2} \lambda_{1j} P\left(\frac{C_j}{x}\right) = \lambda_{11} P\left(\frac{C_1}{x}\right) + \lambda_{12} P\left(\frac{C_2}{x}\right)$$

True

$$R(\alpha_2|x) = \sum_{j=1}^{2} \lambda_{2j} \cdot P\left(\frac{C_j}{x}\right) = \lambda_{21} P\left(\frac{C_1}{x}\right) + \lambda_{22} P\left(\frac{C_2}{x}\right)$$

If Risk of assigning x to class 1 $\downarrow$ b.
we want to assign to class 1.

- $R\left(\frac{\alpha_1}{x}\right) < R\left(\frac{\alpha_2}{x}\right)$

$\Rightarrow \left(\lambda_{11} P\left(\frac{C_1}{x}\right) + \lambda_{12} P\left(\frac{C_2}{x}\right)\right) < \left(\lambda_{21} P\left(\frac{C_1}{x}\right) + \lambda_{22} P\left(\frac{C_2}{x}\right)\right)$

$\Rightarrow (\lambda_{11} - \lambda_{21}) P\left(\frac{C_1}{x}\right) < (\lambda_{22} - \lambda_{12}) P\left(\frac{C_2}{x}\right)$

- $(\lambda_{21} - \lambda_{11}) P\left(\frac{C_1}{x}\right) > (\lambda_{12} - \lambda_{22}) P\left(\frac{C_2}{x}\right)$

From Baye's decision rule :-
we need to have following to assign class 1 to i/p data x :-

- $P\left(\frac{C_1}{x}\right) > P\left(\frac{C_2}{x}\right)$

If we want to assign x to class 1, we want :-

- $(\lambda_{21} - \lambda_{11}) P\left(\frac{C_1}{x}\right) > (\lambda_{12} - \lambda_{22}) P\left(\frac{C_2}{x}\right)$

weight

- $(\lambda_{11}$ or $\lambda_{22})$ = (is always lower as they are correctly classified)
- $(\lambda_{12}, \lambda_{21}) \Rightarrow$ is always higher as they are misclassified.

$$\lambda_1 \leq \lambda_2$$

$$(\lambda_{21} - \lambda_{11})\, P\left(\frac{C_1}{x}\right) > (\lambda_{12} - \lambda_{22})\, P\left(\frac{C_2}{x}\right)$$

(λ_{21}: wrong high value; λ_{11}: lower value; implies)

[λ_1 = true action error so might be zero
$\therefore \lambda_1 \leq \lambda_2$, or zero.]

$\lambda_{21} - \lambda_{22}$
$\lambda_{12} - \lambda_{11}$
— Both are always positive

[λ_{21} = error $\simeq$ approximately equal to 1
λ_{11} = true " " " 0

$$(\overset{1}{\lambda_{21}} - \overset{0}{\lambda_{11}})\, P\left(\frac{C_1}{x}\right) > (\overset{1}{\lambda_{12}} - \overset{0}{\lambda_{22}})\, P\left(\frac{C_2}{x}\right)$$

$(1-0)$ $(1-0)$

$$P\left(\frac{C_1}{x}\right) > P\left(\frac{C_2}{x}\right)$$ — Bayes' Theorem.

P23CS0010

Lecture - 12

30-08-2024

Minimum Risk Classifier : For Two class Problem

$$(\lambda_{21} - \lambda_{11})\, P\left(\frac{C_1}{x}\right) > (\lambda_{12} - \lambda_{22})\, P\left(\frac{C_2}{x}\right)$$

weight

- From Bayes decision rule, to assign class 1 to Input data x :-

$$P\left(\frac{C_1}{x}\right) > P\left(\frac{C_2}{x}\right)$$

Here, to assign class 1 to i/p x :-

$$(\lambda_{21} - \lambda_{11})\, P\left(\frac{C_1}{x}\right) > [\lambda_{12} - \lambda_{22}]\, P\left(\frac{C_2}{x}\right)$$

All the losses (λ_{ij}s) are pre-defined depending on problem.

- $\lambda_{ii} = 0$; indicates correct decision

Minimum error risk classifier is more general version of Baye's theorem

It is difficult in term of loss, error, risk.

How to minimize error :-

$$\lambda_{ij} = \lambda(\alpha_i / Cl_j) = \begin{cases} 0 & \text{when } i = j \\ 1 & \text{when } i \neq j \end{cases} \quad \text{---(i)}$$

we also have :-

$$R(\alpha_i / x) = \sum_{j=1}^{k} \lambda\left(\frac{\alpha_i}{Cl_j}\right) P\left(\frac{Cl_j}{x}\right) \quad \text{---(ii)}$$

Combining (i) and (ii) :-

$$R\left(\frac{\alpha_i}{x}\right) = \sum_{i \neq j} P\left(\frac{Cl_j}{x}\right)$$

Calculation of risk :-

Let $i = 2$ and there are 3 classes Cl_1, Cl_2, Cl_3

So, $\sum_{j=1}^{3} P\left(\frac{Cl_j}{x}\right) = 1$

$i=2$ → risk = 0

$$\sum_{j=1}^{3} P\left(\frac{Cl_j}{x}\right) = P\left(\frac{Cl_1}{x}\right) + \underbrace{P\left(\frac{Cl_2}{x}\right)}_{0} + P\left(\frac{Cl_3}{x}\right)$$

So $\sum_{i \neq j} P\left(\frac{Cl_j}{x}\right) = P\left(\frac{Cl_1}{x}\right) + P\left(\frac{Cl_3}{x}\right) = \boxed{\sum_{i \neq j} P\left(\frac{Cl_j}{x}\right) = 1 - P\left(\frac{Cl_2}{x}\right)}$

$$= 1 - P\left(\frac{Cl_2}{x}\right)$$

- Every incorrect decision doesn't have same impact.
- Some incorrect decision are more incorrect.
  ↓
  ideally we don't assign same assignment to all incorrect.

eg: Tomorrow is my exam → fail → coz watched movie (incorrect decision) rather than preparing

Limit changes acc to level of incorrect condition.

$$\sum_{i \neq j} P\left(\frac{C_j}{x}\right)$$

here Let's assume all as (incorrect $\lambda = 1$)

$$R\left(\frac{\alpha_i}{x}\right) = \sum_{i \neq j} P\left(\frac{C_j}{x}\right) = 1 - P\left(\frac{C_i}{x}\right)$$

(Posterior probability)

I want to minimize α_i risk.

Maximize $P\left(\frac{C_i}{x}\right)$

> whichever class maximize Posterior probability take decision for that.

↑

Minimum error rate classifier

This is similar to Bayes decision rule :-

$$P\left(\frac{C_1}{x}\right) > P\left(\frac{C_2}{x}\right) \longrightarrow C_4$$

Posterior probability → Baye's decision rule.

[every incorrect → different weight, α_{ij} diff. for diff incorrect cond^n]

Conditional Probability Table :- assume, we can tell any probability answer from this.

Drawback :- Storage.

- for a system with many causes and effects
  we have to maintain a large set of values and operate on those

eg:- If 3 random variables
$2^3 = 8$ entries

↑↑ storage
- For D possible answer on N random variable $= D^N$.
- 3 answers of 10 question $= 3^{10}$

Independent Events

Two events A and B are said to be independent if

$$P\left(\frac{A}{B}\right) = P(A) \quad \text{——— (1)}$$

we already have

$$P\left(\frac{A}{B}\right) = P\left(\frac{A \cap B}{B}\right) = \frac{P(B|A)\,P(A)}{P(B)} \quad \text{——— (2)}$$

From (1), (2) ←

$$P\left(\frac{B}{A}\right) = P(B)$$

$$\boxed{P(A \cap B) = P(A)\,P(B)}$$

If I want to deal with fever, covid, cough and Internet speed

(Internet speed doesn't depend on other 3)

So if we want to find out Joint Distribution :— $P(AB) = P(A) \cdot P(B)$

$$P(\text{fever, cough, covid, I\_sp}) = \underbrace{P(\text{fever, cough, covid})}_{A} \cdot \underbrace{P(\text{I\_sp})}_{B}$$

(B: independent)

If sample space for Internet speed (I_sp) is ←

{slow, medium, fast, very fast}

		fever		¬fever	
		cough	¬cough	cough	¬cough
covid	slow medium fast very fast	a_1	a_2	.	.
¬covid		.	.	.	a_{32}

Total entries : 32 (31 parameters)

Now I know :-

$$P(\text{fever, cough, covid, I\_sp}) = P(\text{fever, cough, covid})\, P(\text{I\_sp})$$

$$\sum P_1 + \cdots + P_{32} = 1$$

$P = 32 - 1$

No. of parameters needed = 1 − one variable left = 31

P(internet) = 4 entries

slow	fast	medium	very fast

Table for I_sp :-

4 entries, 3 parameters (4 − 1 = 3)

Before entries required was :- 32 (31 parameters)

- After performing factorization :-

$$P(\text{f, cough, covid, I\_sp}) = P(\text{f, cough, covid})\, P(\text{I\_sp})$$

Total entries :- $8 + 4 = 12$ (7 + 3 = 10 parameters)

(2^3, I_sp)

Complete Independence is extremly powerful.
↓
we will not include such factors (useless for our modelling purpose)

- only by using notion of Independent events = Reduced Storage.

Complete independent is not feasible

Some factor completely or may not completly, partiall independent to others.

partially ← weather condith, food
↓
Partially dependent.

Conditional Independence ←

I want to calculate probability of fever given a person has covid, cough.

$$P\left(\frac{\text{fever}}{\text{covid, cough}}\right)$$

- Knowledge abaut cough doesn't give additional infarn to predict fever.

[fever is outcome of covid
fever is not outcome of cough

Given covid, fever is conditional independent of cough.

eg 2 :- $P\left(\frac{\text{flood}}{\text{water leakage, heavy rainfall}}\right)$

flood is conditional independent of water leakage

So we can say :-

$$P\left(\frac{\text{fever}}{\text{covid, cough}}\right) = P\left(\frac{\text{fever}}{\text{covid}}\right)$$

also, $P\left(\frac{\text{fever}}{\neg \text{covid, cough}}\right) = P\left(\frac{\text{fever}}{\neg \text{covid}}\right)$

. fever is conditional independent of cough, given covid.

To model, $P(\text{fever, covid, cough})$ = 8 entries, 7 parameters

$$P(f, \text{cough}, \text{covid}) = P\left(\frac{\text{fever}}{\text{covid, cough}}\right) P(\text{covid, cough})$$

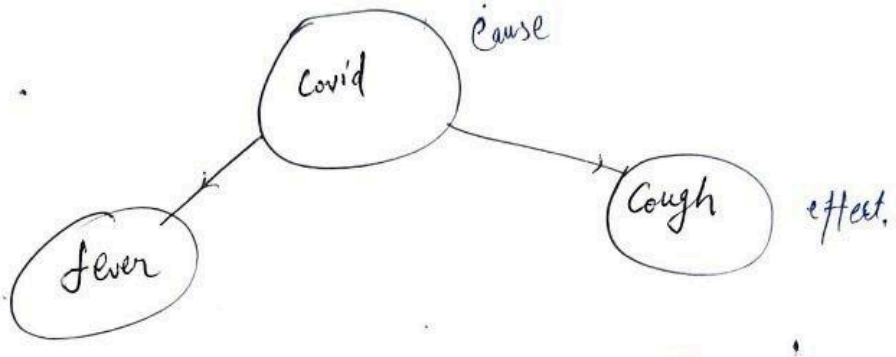
$$= P\left(\frac{f}{\text{covid}}\right) P(\text{covid, cough})$$

$$= P\left(\frac{\text{fever}}{\text{covid}}\right) P\left(\frac{\text{covid}}{\text{cough}}\right) P(\text{cough})$$

2 parameter, 2, 1

= 5 parameter.

Graphical Representation :—



one cause resulting in multiple independent effects.

Also known as Cause - Effect Graph.

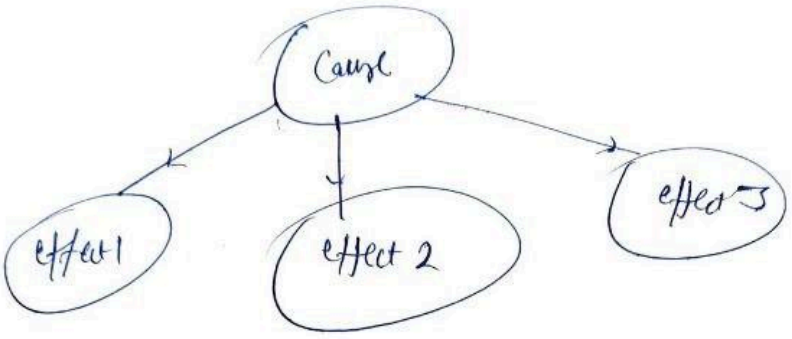
- Independent events don't interact with each other.

eg:- Genectic disease disorder
for blood, Kidney if Hemoglobin is low, kidney effect
not independent.

Genectic disease disorder
don't follow conditional independence.

eg2:- Heavy rainfall : cause [power cut, water cut]

for some causes [effects are independent
effects are dependent



Naïve Assumption [one cause resulting in multiple independent effects.

Navie assumption : effects are independent, in Cause-effect

Lecture - 13

5-Sept-2024

Problem 1:- At a certain university :—

4% men → over 6 feet tall.

1% women → over 6 feet tall.

Total students are in ratio = 3 : 2 in favor of women.

If a student is selected at random from all those 6 feet tall.

What is probability that student is woman?

Solution:-

Probability that student is male $P(M) = \frac{2}{5}$

" " " " female $P(W) = \frac{3}{5}$

Probability that a male that is over 6 feet tall ; $P\left(\frac{H}{M}\right) = 0.04$

" " " female " " - - - - - - ; $P\left(\frac{H}{W}\right) = 0.01$

Total probability of being over 6 feet tall:—

$$P(H) = P\left(\frac{H}{M}\right) \cdot P(M) + P\left(\frac{H}{W}\right) \cdot P(W)$$

$$P(H) = (0.04 * 0.4) + (0.01 * 0.6) = 0.016 + 0.006$$
$$= 0.022$$

Using Bayes' rule:—

$$P\left(\frac{W}{H}\right) = \frac{P(H/W) \cdot P(W)}{P(H)} = \frac{0.01 * 0.6}{0.022} \approx 0.273$$

Problem-2

Probability of certain disease = 0.01

Probability of test positive given person is infected with disease = 0.95

Probability of test positive given person is not infected with disease = 0.05

(a) Calculate probability of test positive

(b) using Baye's rule, calculate probability of being infected with disease given that test is +ve.

Soln:

$P(D) = 0.01$ $\quad P\left(\frac{T}{D}\right) = 0.95$

(a) $P\left(\frac{T}{\neg D}\right) = 0.95$

$P(T) = P\left(\frac{T}{D}\right) \cdot P(D) + P\left(\frac{T}{\neg D}\right) P(\neg D)$

$(0.95 * 0.01) + (0.05 * 0.99) =$

$0.0095 + 0.0495 = 0.059$

(b) $P\left(\frac{D}{T}\right) = \frac{P\left(\frac{T}{D}\right) \cdot P(D)}{P(T)} = \frac{0.95 * 0.01}{0.059} = 0.161$

Problem 3: A factory manufactures bolts using A, B, C machines. Proportion, Defect rates are :-

- Machine A : 25% proportion, Defect rate : 5%
- Machine B : 35% , Defect : 4%
- Machine C : 40% , Defect rate : 2%

A bolt is chosen at random, found defective.

Probability it came from :-

(a) Machine A
(b) Machine B
(c) Machine C

Soln: Total probability of Defect :-

$$P(D) = P\left(\frac{D}{A}\right) \cdot P(A) + P\left(\frac{D}{B}\right) \cdot P(B) + P\left(\frac{D}{C}\right) \cdot P(C)$$

$$P(D) = (0.05 * 0.25) + (0.04 * 0.35) + (0.02 * 0.40)$$

$$= 0.0345$$

using Bayes' rule

$$P\left(\frac{A}{D}\right) = \frac{P\left(\frac{D}{A}\right) \cdot P(A)}{P(D)} = \frac{0.05 * 0.25}{0.0345} = 0.3623$$

$$P\left(\frac{C}{D}\right) = \frac{P\left(\frac{D}{C}\right) * P(C)}{P(D)} = \frac{0.02 * 0.40}{0.0345} = 0.2319$$

$$P\left(\frac{B}{D}\right) = \frac{P\left(\frac{D}{B}\right) \cdot P(B)}{P(D)} = \frac{0.04 * 0.35}{0.0345}$$

Q:- A garage machine has box with :-

good springs : 5

faulty springs : 3

Machine picks from box = 2 springs.

Find probability of :- (a) 1st spring drawn is faulty

(b) 2nd " " "

Soln:- Probability of 1st spring drawn is faulty :-

$$P(\text{First faulty}) = \frac{\text{No of faulty spring}}{\text{Total no of spring}} = \frac{3}{8}$$

Total probability that 2nd spring is faulty :-

Case 1:- First spring is faulty :-

$$P\left(\frac{\text{2nd faulty}}{\text{First faulty}}\right) = \frac{2}{7}$$

Case 2:- first spring is not faulty :-

$$P\left(\frac{\text{second faulty}}{\text{first not faulty}}\right) = \frac{3}{7}$$

Overall probability :-

$$P(\text{second faulty}) = \frac{3}{8} * \frac{2}{7} + \frac{5}{8} * \frac{3}{7} = \frac{21}{56} = \frac{3}{8}$$

Q: Manufacturing plant produces three types of widgets :- (5

Type	Proportions	Defects rate
A	40%	5%
B	35%	4%
C	25%	3%

(a) what is probability that defective widget is of Type A ?

$$P(D) = P\left(\frac{D}{A}\right) \cdot P(A) + P\left(\frac{D}{B}\right) \cdot P(B) + P\left(\frac{D}{C}\right) \cdot P(C)$$

$$= (0.05 * 0.40) + (0.04 * 0.35) + (0.03 * 0.25)$$

$$= 0.0415$$

using Bayes rule :-

$$P\left(\frac{A}{D}\right) = \frac{P\left(\frac{D}{A}\right) \cdot P(A)}{P(D)} = \frac{0.05 * 0.40}{0.0415} = 0.4819$$

(b) Recalculate if defect rate of Type A increases to 10%.

New defect rate for Type A = 10%

Recalculate Total probability of Defect :-

$$P(D) = (0.10 * 0.40) + (0.04 * 0.35) + (0.03 * 0.25)$$

$$= 0.0615$$

Using Bayes' Rule :-

$$P\left(\frac{A}{D}\right) = \frac{0.10 * 0.40}{0.0615} = 0.6504$$

(c) Similarly recalculate if defect rate for Type B is reduced to 1%. (6)

Ans:-

x_1	x_2	t
0	0	0
0	1	1
1	0	1
1	1	0

$w_0 = -0.3$ $\quad \eta = 0.5$

$w_1 = w_2 = 0.5$

$-x_0 = 1$ Default

Calculate gradient :-

Soln :- Model setup :-

$$y_{pred} = \text{step}\left(w_0 x_0 + w_1 x_1 + w_2 x_2\right)$$

(i) :- for $x = (0,0)$

$$z = w_0 x_0 + w_1 x_1 + w_2 x_2$$

$$= (-0.3 * 1) + (0.5 * 0) + (0.5 * 0)$$

$$= -0.3$$

$$\Delta w_0 = \Delta w_0 + \eta(t - o)\, x_0$$

(t: true value, o: predicted)

$$= 0 + 0.5\left(0 - (-0.3)\right)0 = 1$$

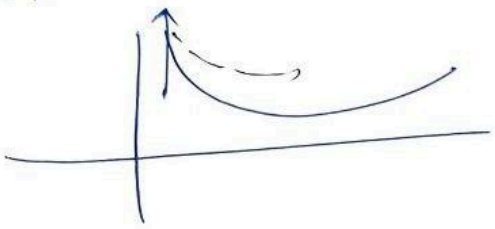
$$y_{pred} = \text{step}(-0.3) = 0$$

$$\Delta w_0 = -0.3 + 0.15$$

$$= 0.5$$

$$\sum w_i x_i = w_0 + w_1 x_1 + w_2 x_2$$
$$= -0.3 + 0.5(0) + 0.5(1)$$
$$= 0.2$$

only for 1st update it (error) spikes ↑ then decreases



$x_0 = 0, x_2 = 1$

$$\sum w_i x_i = w_0 x_0 + w_1 x_1 + w_2 x_2$$
$$= -0.3 * 1 + 0.5 * 0 + 0.5 * 1$$
$$= 0.2$$

$$\Delta w_0 = \Delta w_0 + \eta(t-o) x_0$$
$$= 0.15 + 0.5\,[1 - 0.2)\,1$$
$$= 0.55$$

$x_1 = 1 \quad x_2 = 0$

$$\Delta w_0 = \Delta w_0 + \eta(t-o) x_0$$
$$= 0.95 + 0.5\,(0 - 0.7) * 1 = 0.6$$

$$\sum w_i x_i = w_0 x_0 + w_1 x_1 + w_2 x_2$$
$$= 0.3 + 0.5(1) + 0.5(1)$$
$$= 0.7$$

$$w = w_1 + \Delta w_1 = -0.3 + 0.6 = 0.3$$

$$\text{loss} = \frac{1}{2}\left(\text{predicted} - \text{actual}\right)^2$$

$$\frac{\partial \text{loss}}{\partial w} = w_1 - \eta\left(t - o\right)x$$

$$\text{cost} = J = \sum \left(y - y_{pred}\right)^2$$

ML-14

Date : 9/09/2024 ①

Naive-Bayes Model

- Bayes's decision rule : $P(Cl_1/x) > P(\frac{Cl_2}{x}) = Cl_1$

Now let's use Bayes theorem :-

$$P(\frac{Cl_1}{x}) = \frac{P(x/Cl_1)P(Cl_1)}{P(x)}$$

Let's consider x to be d-dimensional feature vector $\{x_1, x_2 \dots x_d\}$

All features are conditionally independent of each other given the class

Now Let's use Bayes theorem :-

$$P(\frac{Cl_1}{x}) = \frac{P(x, Cl_1)}{P(x)}$$

$$= \frac{P(x_1, x_2 \dots x_d, Cl_1)}{P(x)}$$

$$= \frac{P(\frac{x_1}{Cl_1}) P(\frac{x_2}{Cl_1}) \cdots P(\frac{x_d}{Cl_1}) P(Cl_1)}{P(x)}$$

$P(\text{Effect 1, effect 2, } \dots \text{ Effect d, Cause})$

$$= P(\text{Cause}) \prod_{k=1}^{d} P(\frac{\text{Effect K}}{\text{Cause}})$$

Similarly :-

$$P\left(\frac{Cl_2}{x}\right) = \frac{P\left(\frac{x_1}{Cl_2}\right) \cdot P\left(\frac{x_2}{Cl_2}\right) \cdots P\left(\frac{x_d}{Cl_2}\right) P(Cl_2)}{P(x)}$$

- Bayes decision rule :-

$$P\left(\frac{Cl_1}{x}\right) > P\left(Cl_2/x\right) \rightarrow Cl_1$$

$$\frac{P\left(x_1/Cl_1\right) P\left(x_2/Cl_1\right) \cdots P\left(x_d/Cl_1\right) P(Cl_1)}{P(x)} > \frac{P\left(x_1/Cl_2\right) \cdot P\left(x_2/Cl_2\right) \cdots P\left(x_d/Cl_2\right) P(Cl_2)}{P(x)}$$

$\Rightarrow Cl_1$

also $P\left(x_1/Cl_1\right) P\left(\frac{x_2}{Cl_1}\right) \cdots P\left(\frac{x_d}{Cl_1}\right) > P\left(\frac{x_1}{Cl_2}\right) \cdot P\left(\frac{x_2}{Cl_2}\right) \cdots P\left(\frac{x_d}{Cl_2}\right) \cdot P(Cl_2)$

$\Rightarrow P(Cl_1)$

Since from denominator $P(x)$ is not there, we'll not get exact answer $\rightarrow$ Probable answer.

eg:- For a classification problem of Tiger, Lion classification.

Class $\longrightarrow$ cause

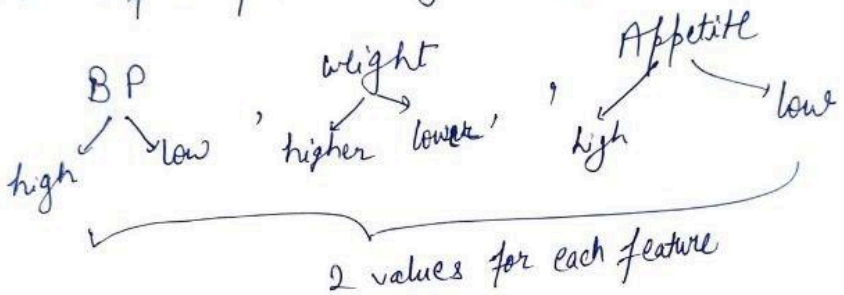
features $\longrightarrow$ effect

As features are conditionally independent given class.

eg 1:- $\frac{\text{Body color}}{\text{Lion}}$. $\frac{\text{Size of Lion}}{\text{Lion}}$. Lion

↖ ↗ Conditionally Independent

eg 2 :- If a person may or may not have Anemia. if either



given Anemia, features ~~BP~~ BP, ~~Appetite~~, weights are conditionally independent.

→ It follows chain rule.

→ $P_i \Rightarrow \Pi \rightarrow$ It is for product

* → Here feature values are discrete.

If feature value is Continuous ⟶ It becomes complex.

→ can't construct Table values of continous values (∞ value infinite)

- What if conditional probability becomes 0

  Completely out of context as whole output is zero,

→ Let's assign some small value to this case

↓

else output might end up = 0

Example of Naive Bayes :-

Sample no	Appetite	Weight	BP	Class
1	low	Normal	low	No Anemia
2	low	low	low	Anemia
3	normal	low	low	Anemia
4	low	low	Normal	No Anemia
5	normal	low	Normal	Anemia
6	normal	Normal	low	Anemia
7	normal	Normal	Normal	No Anemia

If we see a new sample who has normal appetite, normal weight, and Low BP. find if person has Anemia?

Soln:-

Step 1:- Calculate Prior :-

$$P(\text{Probability of having Anemia}) = \frac{4}{7} \qquad P(\overline{\text{Anemia}}) = \frac{3}{7}$$

Step 2:- Calculate $P\left(\frac{\text{feature}}{\text{Anemia}}\right)$, $P\left(\frac{\text{feature}}{\text{No Anemia}}\right)$ from training data.

		Anemia	Anemia
Appetite	low	1/4	2/3
	Normal	3/4	1/3

$P(\text{Appetite} = \frac{\text{low}}{\text{Anemia}})$

		Anemia	Anemia
BP	low	3/4	1/3
	Normal	1/4	2/3

		Anemia	Anemia
Weight	low	3/4	1/3
	Normal	1/4	2/3

$P(\text{anemia}) = \frac{4}{7}$ $\quad P(\text{no anemia}) = \frac{3}{7}$

Step 3:- Testing :-

Suppose, I observe a test data with normal appetite, normal weight, and low BP.

- To predict if person has Anemia :-

$$P\left(\text{appetite} = \frac{\text{normal}}{\text{anemia}}\right) P\left(\text{weight} = \frac{\text{normal}}{\text{anemia}}\right) P\left(BP = \frac{\text{low}}{\text{Anemia}}\right) \cdot P(\text{anemia}) \quad (6)$$

$$= \frac{3}{4} * \frac{1}{4} * \frac{3}{4} * \frac{4}{7} = 0.08$$

$$P\left(\frac{\text{no anemia}}{\pi_{\text{no anemia}}}\right) = P\left(\text{appetite} = \frac{\text{normal}}{\text{Anemia}}\right) \cdot P\left(\text{weight} = \frac{\text{normal}}{\text{Anemia}}\right) \cdot P\left(BP = \frac{\text{low}}{\text{Anemia}}\right) \cdot P(\text{no anemia})$$

$$= \frac{2}{3} * \frac{2}{3} * \frac{1}{3} * \frac{3}{7} = 0.06$$

$$P(C_1) \cdot P\left(\frac{x_1}{C_1}\right) \cdot P\left(\frac{x_2}{C_1}\right) \cdots P\left(\frac{x_d}{C_1}\right) > P\left(\frac{x_1}{C_2}\right) \cdot P\left(\frac{x_2}{C_2}\right) \cdots P\left(\frac{x_d}{C_2}\right) \cdot P(C_2)$$

$$\Rightarrow C_1$$

Conclusion :- person has anemia as per Naive Bayes' Classifier

i.e. $\pi_{\text{anemia}} > \pi_{\text{no anemia}}$

Training data :-

- Assume N training samples and class label pairs $(x^1, y^1), (x^2, y^2) \ldots (x^N, y^N)$.
- Each training sample x^i is a D dimensional feature vector $\{x_1^{(i)}, x_2^{(i)}, \ldots x_D^{(i)}\}$

  eg:- 3 is value for wt, appetite, BP
  x_1 x_2 x_3
- Assume that attribute x_k can take values :-

  $x_{K_1}, x_{K_2} \ldots x_{K_v}$

  eg:- BP [High, low] $K=2$ here for each d there is K.
- The values of attributes are discrete.
- We have total of C number of classes $C_1, C_2 \ldots C_C$.

Training Method

for $j = 1 : C$

  Calculate $P(C_j)$ from training data

for $d = 1 : D$

  for $q = 1 : V$

    Calculate $P(x_{dq} \mid C_j)$

Naive Bays' classifier: The algorithm

Inference for a new test sample $x^{(new)}$, feature vector is $\{x_1^{new}, x_2^{new} \ldots x_d^{new}\}$

for $j = 1 : C$

→ get $P(Cl_j)$ computed during training

→ get $P\left(\frac{x_1^{new}}{Cl_j}\right) \cdot P\left(\frac{x_2^{new}}{Cl_j}\right) \ldots P\left(\frac{x_d^{new}}{Cl_j}\right)$ computed during train

- Calculate

$$\Pi_j = P\left(\frac{x_1^{new}}{Cl_j}\right) P\left(\frac{x_2^{new}}{Cl_j}\right) \ldots P\left(\frac{x_d^{new}}{Cl_j}\right) \cdot P(Cl_j)$$

- Find out j for which Π_j is maximum.

$$j^* = \underset{j}{\text{argmax}}\ \Pi_j$$

- Assign class label j^* to test data.

- Dependence of features in Algo is complex problem.
- "Zero problem" completly discard a particular class.

Life without Naive Bayes'

- Consider x to be a d-dimensional feature vector

$$\{x_1, x_2 \cdots\cdots x_d\}$$

- If all features are not conditionally independent of each other given the class

↓

w/o Naive Bayes = make solution very very complex.

> **Inductive Bias:** making assumption of Condition Independent makes life easier.

- Bayes theorem would be :-

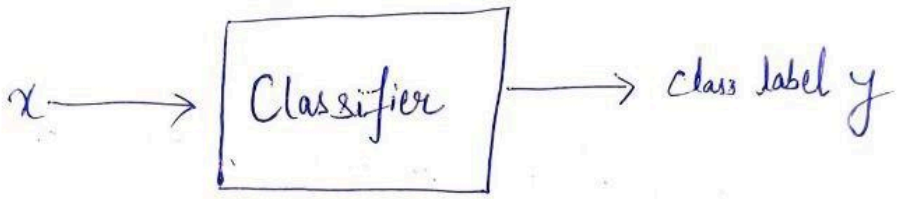
$$P\left(\frac{Cl_1}{x}\right) = \frac{P(x, Cl_1)}{P(x)}$$

$$= \frac{P(x_1, x_2 \cdots\cdots x_d, Cl_1)}{P(x)}$$

Computing these conditional probabilities would be extreme difficult.

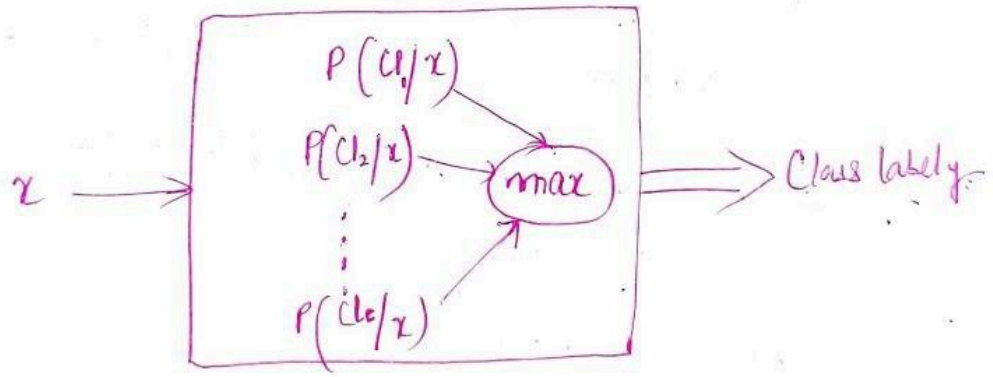
$$= \frac{P\left(\frac{x_1}{x_2, x_3 \cdots x_d, Cl_1}\right) \cdot P\left(\frac{x_2}{x_3, \cdots, Cl_1}\right) \cdots P\left(\frac{x_d}{Cl_1}\right) P(Cl_1)}{P(x)}$$

Decision rule : try to find probability which gives maximum value discriminant fun^n for x.

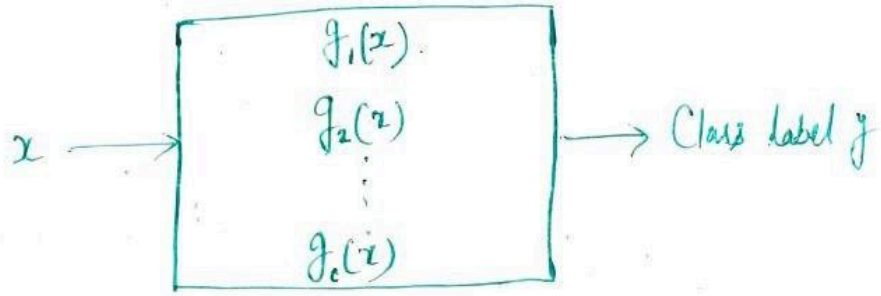


How does the classifier do it ?

Suppose I have C number of classes. I consider Baye's decision rule



$g_i(\cdot)$ is called discriminant function.



If $g_i(x) > g_j(x)$, $\forall j \neq i$ we assign class label i to the input data x.

Nature of Discriminant functions:-

For the minimum risk classifier:

We assign the class label corresponding to the action of minimum risk.

If the risk of action α_i is $R(\alpha_i/x)$.

- Minimum risk → in order to minimize risk = negative of risk.

$$g_i(x) = -R\left(\frac{\alpha_i}{x}\right)$$

- we assign class label corresponding to maximum posterior probability.

$$g_i(x) \propto P\left(\frac{C_i}{x}\right)$$

$$\text{discriminant fun}^n \propto \begin{pmatrix}\text{posterior}\\ \text{probability}\end{pmatrix} \uparrow\uparrow$$

- Choice of discriminant function is not unique.

eg:- $-R$, $h = \frac{1}{e^x}$, $R = \frac{1}{\text{Softmax}}$ .....

$-\left[R\left(\frac{\alpha_i}{x}\right)\right]^2$ = valid fun^n, also $-2\left[R\left(\frac{\alpha_i}{x}\right)\right]$ → also valid fun^n.

If $g_i(x)$ is a discriminant fun^n and $f(\cdot)$ is monotonically increasing function, $f(g_i(x))$ is also a valid discriminant fun^n.

If $g_i(x) > g_j(x)$, $\forall j \neq i$, we assign class label i to input data x.

$$g_i(x) \propto P\left(\frac{C_i}{x}\right)$$

$$\propto \frac{P(x/C_i) \cdot P(C_i)}{P(x)}$$

Since denomenator is common for every class i :—

we take

$$g_i(x) = P\left(\frac{x}{C_i}\right) \cdot P(C_i)$$

- fun^n monotonically increasing means :—
  with increasing value of argument, monotically increasing function gives higher values.

  if $x_1 \rightarrow x_2 \rightarrow x_3$

  $$f(x_1) < f(x_2) < f(x_3)$$

  monotonically increasing function don't have equality. (negative only have)

i.e $f(g_i(x)) \neq f(g_j(x))$

Can't be equal

we take

$$g_i(x) = P\left(\frac{x}{C_i}\right) P(C_i)$$

→ multiplication is difficult

↓

→ addition is better.

(log Likelihood)

we can also take :—

$$g_i(x) = \ln P\left(\frac{x}{C_i}\right) \cdot P(C_i)$$

$$= \ln P\left(\frac{x}{C_i}\right) + \ln P(C_i)$$

As log is monotonically increasing function.

- If we take a two-class problem, & discriminant fun^{n}s $g_1(x), g_2(x)$.

→ If $g_1(x) > g_2(x)$, we conclude that x belongs to class 1.

→ If $g_1(x) < g_2(x)$, we conclude that x belongs to class 2.

- $\boxed{g_1(x) = g_2(x)}$ ← Decision Boundary

$g(x) > 0$: class 1 $\quad \left[g(x) = \underset{\text{+ve}}{g_1(x)} - \underset{\text{-ve}}{g_2(x)}\right]$

$= +ve.$

$g(x) < 0$: class 2

$\left[g(x) = \underset{\text{-ve}}{g_1(x)} - \underset{\text{+ve}}{g_2(x)}\right]$

-ve value

As neumerator is good enough to do classificatn so removed denomenator $P(x)$:—

$$g(x) = P\left(\frac{C_i}{x}\right) = P\left(\frac{x}{C_i}\right) P(C_i)$$

Taking ln of prior, posterior to make it easy as addⁿ is easier than multiplication :—

$$g(x) \Rightarrow g_1(x) - g_2(x)$$

$$= \ln(g_1(x)) - \ln(g_2(x))$$

$$= \ln\left[P\left(\frac{x}{C_i}\right) \cdot P(C_i)\right] - \ln\left[P\left(\frac{x}{C_j}\right) P(C_j)\right]$$

$$= \ln\left[P\left(\frac{x}{C_1}\right) \cdot P(C_1)\right] - \ln\left[P\left(\frac{x}{C_2}\right) \cdot P(C_2)\right]$$

$$= \ln P\left(\frac{x}{C_1}\right) + \ln P(C_1) - \ln P\left(\frac{x}{C_2}\right) - \ln P(C_2)$$

$$= \underbrace{\ln\left[\frac{P(x/C_1)}{P(x/C_2)}\right]}_{\text{Posterior}} + \underbrace{\ln\left[\frac{P(C_1)}{P(C_2)}\right]}_{\text{Prior}}$$

here at outside we can't write as $\frac{\ln P(\,)}{\ln P(\,)}$ x , $\ln\left(\frac{P(\,)}{P(\,)}\right)$

$$g_i(x) = \ln P\left(\frac{x}{Cl_i}\right)$$

Talking about probability distribution :—

$P\left(\frac{x}{Cl_i}\right)$ is probability distribution which can take many forms.

- Assume that $P\left(\frac{x}{Cl_i}\right)$ follows Gaussian or Normal distribution in d-dimensions.

So,
$$P\left(\frac{x}{Cl_i}\right) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma_i|^{\frac{1}{2}}} \exp\left[-\frac{(x-\mu_i)^T \Sigma_i^{-1} (x-\mu_i)}{2}\right]$$

μ_i = Expected value of x (samples) given that x belongs to class Cl_i.

Σ_i = Covariance matrix computed from x (samples) given that x belongs to class Cl_i

d dimensional ⇒ d features

- Talking about probability distribution : Continuous terms.
  ↓
  value can be find out in range.

Expected value :- $\frac{1}{n}\sum x_i$

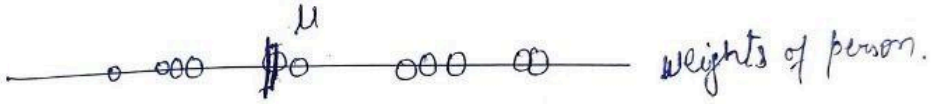
Variance :- Scattering measure from mean.

$$P\left(\frac{x}{C_i}\right) = \frac{1}{(2\pi)^{d/2}|\Sigma_i|^{\frac{1}{2}}} \exp\left[-\frac{(x-\mu_i)^T \Sigma_i^{-1} (x-\mu_j)}{2}\right]$$

μ_i, μ_j are two diff. mean, (may or may not)

Variance :-

$$\text{var}(x) = E\left[(x-\mu)^2\right]$$



Suppose, I want to measure mean weights

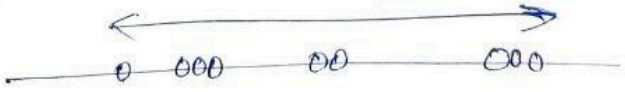
$$\text{Mean } \mu = E(x) = \int x \cdot p(x)\, dx$$

For N distinct (equally likely) data points $x_1, \dots x_N$

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Variance indicates Spread

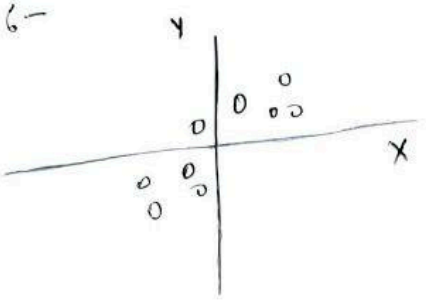
Variance



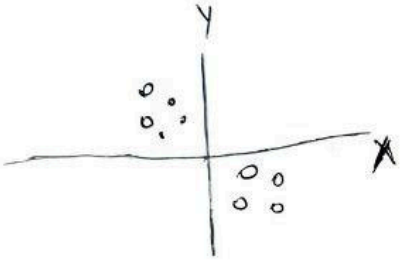
for N distinct (equally likely) data points $x_1, x_2 \dots x_N$

$$\text{Var}(X) = \frac{1}{N}\sum_{i=1}^{N} (x_i - \mu)^2$$

Co-variance :-



we observe that as X increases, Y also increases.



we observe that as X increases, Y decreases

If we look at x, y spread can't differentiate b/w x, y data individually.

→ we need to see x, y together

↓

join variance : Covariance.

$$Cov(X,Y) = E\left[(X-\mu)(X-\mu)^T\right]$$

μ_x, μ_y are different

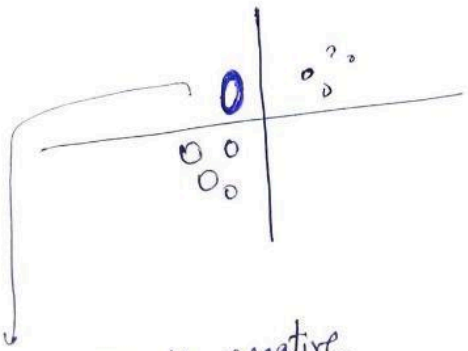
$$Cov(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu_x)(y_i-\mu_y)$$

pointwise product of differences from mean.

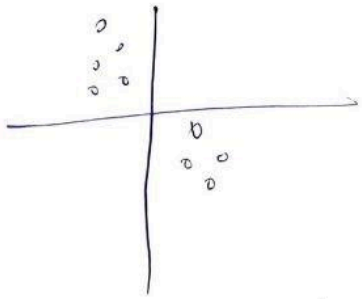
Shows a joint trend of different variables.

[(How much difference in x direction) . (How much it differs in Y direction)]

Tell variance of data along x, y direction = Correlated.



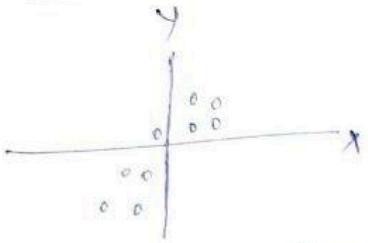
Covariance is negative

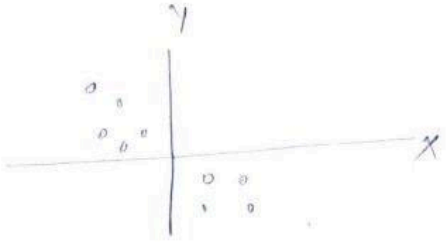


Covariance is +ve.

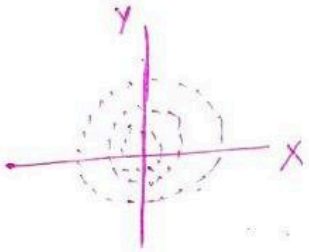
$$\Sigma = \begin{pmatrix} \text{Var}(X) & \text{Cov}(X,Y) \\ \text{Cov}(X,Y) & \text{Vari}(Y) \end{pmatrix}$$

Covariance Matrix:



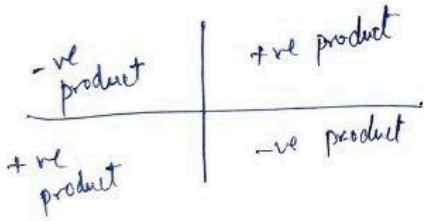


$$\Sigma = \begin{pmatrix} var(x) & Cov(x,y) \\ Cov(x,y) & var(y) \end{pmatrix}$$



$\Rightarrow$ Covariance = 0

"No trend/trade of data"



- Trade +ve = +ve covariance
- Trade -ve = -ve covariance

$\Sigma =$  variance off diagonal

off diagonal = co-variance

- Covariance is calculated for pair of random variable.
  ‼ only for pair
- Covariance Matrix of any no of variable.

$Cov(X, X) = Var(X)$

for same variable its covariance is another variable is X itself.

- Covariance for three random variables X_1, X_2, X_3 ⊢

$$\Sigma = \begin{pmatrix} Var(X_1) & Cov(X_1, X_2) & Cov(X_1, X_3) \\ Cov(X_2, X_1) & Var(X_2) & Cov(X_2, X_3) \\ Cov(X_3, X_1) & Cov(X_3, X_2) & Var(X_3) \end{pmatrix}$$

Probability Density and the Discriminant function

$$g_i(x) = \ln\left(P\left(\frac{x}{C_i}\right) \cdot P(C_i)\right)$$

$$= \ln P\left(\frac{x}{C_i}\right) + \ln P(C_i)$$

$P\left(\frac{x}{C_i}\right)$ = likelihood , $\ln P\left(\frac{x}{C_i}\right)$: log likelihood

$P(C_i)$ = Prior , $\ln P(C_i)$ = log prior

$P\left(\frac{x}{C_i}\right)$ x_i may be any of feature : weight, appetite, BP....

x = observed data

$P\left(\frac{x}{C_i}\right)$ = <u>likelihood of observed</u> data given a particular class.

given data for a particular class 'i'.

eg:- all data given for Anemia circumstances :-

Calculate $P\left(\frac{x_i}{Cl_i}\right)$ eg:- $P\left(\frac{\text{weight} = \text{low}}{\text{anemia}}\right)^{x}_{Cl_i}$

- Likelihood can be at any distribution.

→ we only have subset of data i.e:- Small sample size
So whole distribution can be at any Distribution.

We need to have inductive bias : need to assume any particular distribution of data.

eg:- For this small subset of data in d dimension i.e d no. of features given distribution is Normal distribution.

$$P\left(\frac{x}{Cl_i}\right) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma_i|^{\frac{1}{2}}} \exp\left[-\frac{(x-\mu_i)^T \Sigma_i^{-1} (x-\mu_i)}{2}\right]$$

✂ ignoring $P(x)$ as it was common for both class 'i,j'.

$|\Sigma_i|$ = determinant of covariance matrix

↳ corresponding to class i

eg:- only for anemia class patient $|\Sigma|$

μ_i = mean of class i = mean is feature vector.

$\mu_i = \frac{1}{N}\sum$ wt , $\frac{1}{N}\sum$ appetite , $\frac{1}{N}\sum$ BP ... d times

Vector quantity d dimensional.

Σ_i = d x d matrix , X = d dimension

→ In order to solve for i, need to take Log of that term.

$$g_i(x) = -\frac{(x-\mu)^T \Sigma_i^{-1} (x-\mu_i)}{2} - \frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln(\Sigma_i) + \ln P(C_i)$$

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1} (x-\mu_i) - \frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln(\Sigma_i) + \ln P(C_i)$$

independent of class label

So we may ignore this term in constructing discriminant function.

Parameters : μ_i, Σ_i

→ If μ_i, Σ_i are given, the Gaussian pdf can be uniquely identified.

μ_i, Σ_i can draw any distribution.

- $-\frac{d}{2}\ln(2\pi)$: it is constant ! doesn't depend on class label.

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i) - \frac{1}{2}\ln(\Sigma_i) + \ln P(C_i)$$

If we want to assign a particular class label as i

then $g_i(x) >> g_j(x)$

$-\frac{d}{2}\ln(2\pi)$ = Can ignore: Common term

- To compare, making computational easy, considering terms which are not constant taking i variability.

Normal Density and Discriminant Function

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i) - \frac{1}{2}\ln(\Sigma_i) + \ln P(C_i)$$

Assume :-

$$\Sigma_i = \sigma^2 I \quad \forall i$$

σ^2 : Constant

$I \Rightarrow$ only diagonal term 1

$\rightarrow$ off diagonal = 0

• Covariance b/w features = 0

↗

Not Correlated.

Similar to Naive Bayes : one feature doesn't depend on other, given class i.

$|\Sigma|_{d \times d}$: determinant of $\sigma^2 I$

$|\Sigma| = \sigma^{2d}$: determinant product of diagonal term d.

d no. of terms product diagonally.

In 3D:

$$I = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

→ only diagonal elements (variance terms) are non zero.

→ As off diagonal not dependent on each other, off diag. covariance : zero

→ It means features of data are statistically independent of each other, i.e no pair of features show any trend.

$\sum_i \Rightarrow \sigma^2 I$

$|\Sigma_i| = \sigma^{2d}$ (assuming I to be a $d \times d$ identity matrix)

$\Sigma_i^{-1} = \frac{1}{\sigma^2} I$

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i) - \frac{1}{2}\ln(\sigma^2 I) + \ln P(C_i)$$

constant — independent of class label

→ So we may ignore this term in constructing the discriminant function.

$$\boxed{g_i(x) = -\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i) + \ln P(C_i)}$$

$$\Sigma^{-1} = \frac{1}{\sigma^2} I$$

$$g_i(x) = -\frac{\|x-\mu_i\|^2}{2\sigma^2} + \ln P(C_i)$$

$$= -\frac{1}{2\sigma^2}\left[x^T x - 2\mu_i^T x + \mu_i^T \mu_i\right] + \ln P(C_i)$$

$x^T x$ — independent of class label

$x^T x = \|x\|^2$

As x is input data, my observation independent of class.

When I calculated X didn't know which particular class does it belongs to.

X ← not dependent on i. = observation doesn't depend on class

. may ignore this term.

$$g_i(x) = -\frac{1}{2\sigma^2}\left[-2\mu_i^T x + \mu_i^T \mu_i\right] + \ln P(C_i)$$

Can't ignore $\mu_i^T x$ as one term is dependent on i.

$$\Rightarrow \frac{\mu_i^T x}{\sigma^2} + \ln P(C_i) - \frac{1}{2\sigma^2}\mu_i^T \mu_i$$

$$= g_i(x) \boxed{= w_i^T x + w_{i0}}$$

$$w_i^T = \frac{\mu_i^T}{\sigma^2}$$

$$w_{i0} = \ln P(C_i) = -\frac{1}{2\sigma^2}\mu_i^T \mu_i$$

→ Discriminant function is a linear function of input.

∴ This is called Linear Machine.

LDA assumption summary :-

(i) distribution is gaussian

(ii) features are equally weighted

(iii) features doesn't depend on each other.

Naive Baye's example –

$\Pi_{\text{no anemia}}$, Π_{anemia}

0.006 , 0.009

$\Pi_{\text{no anemia}} > \Pi_{\text{anemia}}$

Conclusion is that person doesn't have anemia as per Naive Baye's classifier.

Naive Baye's : Slightly Different Example –

Sample number	Appetite	Weight (kg)	BP	Class
1	low	81	low	No Anemia
2	low	49	low	Anemia
3	normal	47	low	Anemia
4	low	51	Normal	No anemia
5	normal	50	Normal	Anemia
6	normal	63	low	Anemia
7	normal	58	Normal	No Anemia

Step 1 – Calculate prior probability

$$P(A) = \frac{4}{7} \qquad P(\bar{A}) = \frac{3}{7}$$

Step 2: Calculate $P\left(\frac{\text{feature}}{\text{Anemia}}\right)$, $P\left(\frac{\text{feature}}{\text{No Anemia}}\right)$ from the training data.

		Anemia	No Anemia
Apetite	low	$\frac{3}{4}$	$\frac{1}{3}$
	Normal	$\frac{1}{4}$	$\frac{2}{3}$

$P\left(\text{Appetite} = \frac{\text{low}}{\text{Anemia}}\right)$, $P\left(A = \frac{\text{normal}}{A}\right)$, $P\left(A = \frac{\text{low}}{\bar{A}}\right)$, $P\left(\text{Appetite} = \frac{\text{normal}}{\text{no Anemia}}\right)$

All these are calculated

Similarly for BP all possible values are calculated.

		Anemia	No Anemia
BP	low	$\frac{3}{4}$	$\frac{1}{3}$
	Normal	$\frac{1}{4}$	$\frac{2}{3}$

All these are calculate. eg: $P\left(BP = \frac{\text{low}}{\text{Anemia}}\right)$ ...

What should I put in weight table entries ?

	Anemia	No Anemia
Weight	?	?

		Anemia	No Anemia
Weight	61	0	1
	49	1	0
	47	1	0
	51	0	1
	50	1	0
	63	1	0
	58	0	1

Naive Bayes fails here

Step 3 :- Suppose, I observe a test data with normal appetite, 56 kg weight, and low BP.
predict if person has anemia.

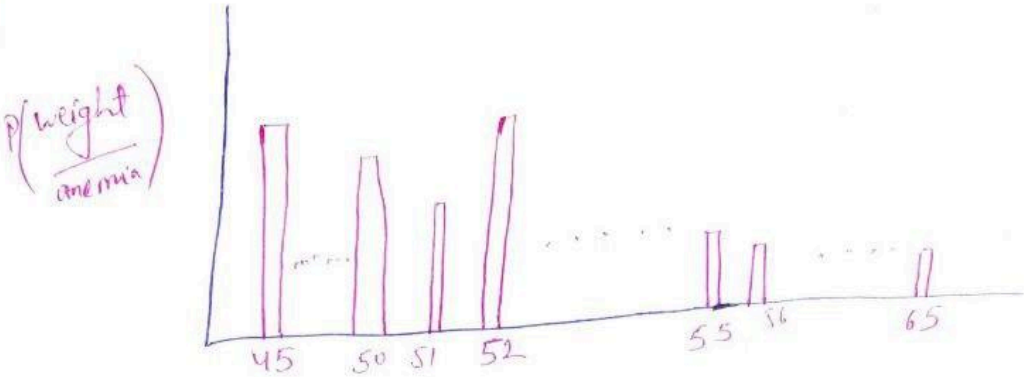
Let's first evaluate choice of Anemia:—

$$\pi_{anemia} = P\left(appetite = \frac{normal}{Anemia}\right) P\left(weight = \frac{56\,kg}{Anemia}\right) P\left(\frac{BP = low}{Anemia}\right) P(anemia)$$

$$\frac{1}{4} * ? * \frac{1}{4} * \frac{4}{7}$$

From my training data, I can't find $P(weight = 56\,kg / Anemia)$

If see, Table has no entry at $P\left(\frac{wt = 56\,kg}{anemia}\right)$.



Accuracy compromised if we consider range.

if we are given Probability / training data.
Goal :- to find distribution that maximaly fits data.

- See distance which lies nearby to which value.

eg:- given random variable value = weight,
also disease is Anemia.

Can calculate Likelihood.

How to find probability density function ?

Q :- Question is :- How to find probability density from training data so that we can estimate likelihood $P(\text{weight} = 56\,kg / \text{anemia})$?

$$P\left(\frac{C_i}{x}\right) = \frac{P(x/C_i)\,P(C_i)}{P(x)}$$

Maximum Likelihood Estimation :-

If I have c number of classes $Cl_1, Cl_2, \ldots Cl_c$ in my dataset

S_1 : Training samples of class Cl_1

S_2 : Training samples of class Cl_2

$\vdots$

S_c : Training sample of class Cl_c.

I have to find :- $P\left(\frac{x}{Cl_i}\right)$ such that $P\left(\frac{x}{Cl_i}\right)$ is maximized when I use training data of class Cl_i.

→ We assume that $P\left(\frac{x}{Cl_i}\right)$:- Known parametric form

- If $P\left(\frac{x}{Cl_i}\right)$ is Gaussian, it is completely specified by mean μ_i and covariance matrix Σ_i

parameter = μ, Σ of gaussian distn

If I can find out parameters Θ_i, I can find out $P(x|Cl_i)$

- Find θ_i which maximally fits data distribution.

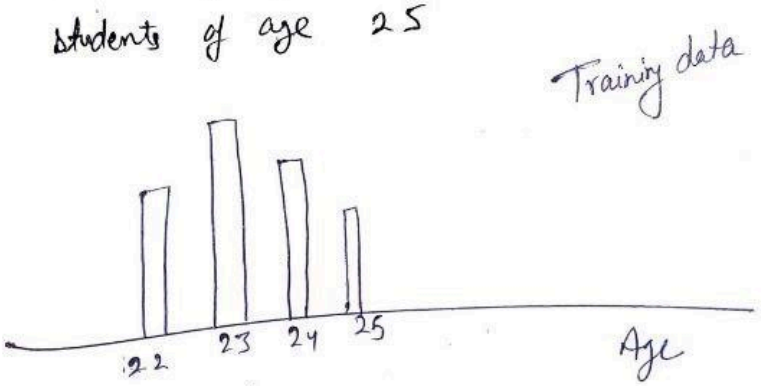
eg: In the classroom:

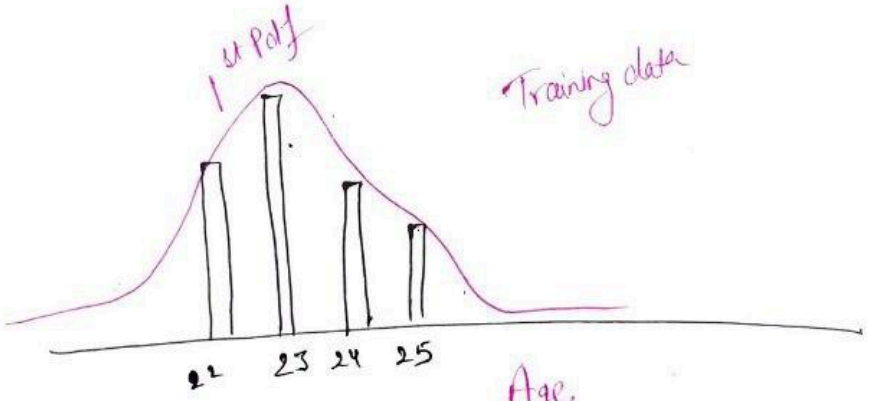
30 students of age 22

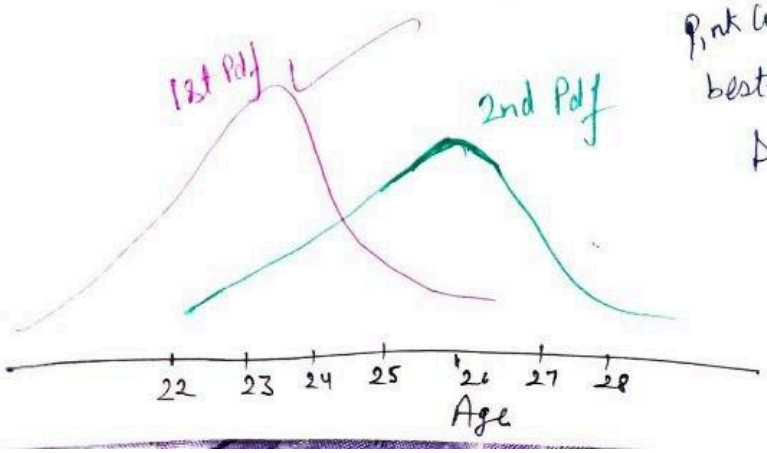
35 students of age 23

30 students of age 24

25 students of age 25







Pink cover (22, 23, 24, 25) best.

So Pink Pdf chosen.

$P\left(\frac{age=22}{A}\right) P\left(\frac{age=23}{A}\right) P\left(\frac{age=24}{S}\right) P\left(\frac{age=25}{S}\right)$ is maximum for which curve.

Likelihood values.

- We assume that $P\left(\frac{x}{C_i}\right)$ has a known parametric form.
- If $P\left(\frac{x}{C_i}\right)$ Gaussian, it is completely specified by mean μ_i and covariance matrix Σ_i (calle them parameters θ_i of distribution).

- Goal :- Find parameters θ_i,

  Given :- information of Training sample in sample i.e S_i

  such that $P\left(\frac{x}{C_i}\right)$ is maximized.

  → using training data of class C_i.

> We assume information in S_j doesn't affect θ_i if $i \neq j$

( )( )( )

Likelihood multiplication

Find curve = product of terms that maximizes

→ our goal is to use information from training sample in S_i

. To find parameters θ_i such that $P\left(\frac{x}{C_i}\right)$ is maximized when I use training data of class i.

→ We have to maximize $P\left(\frac{x}{C_i}, \theta_i\right) \forall i$

→ Since θ_i is parameters corresponding to C_i only, we can say we have to find maximize $P(x/\theta_i)$ $\forall i$.

Lecture -17

Date : 13 sept 2024

Our goal is to use information from training sample in S_i to find parameters θ_i such that $P\left(\frac{x}{C_i}\right)$ is maximized when we use training data of class C_i.

→ we have to maximize $P(x/C_i, \theta_i) \quad \forall i$

→ Since θ_i are parameters corresponding to C_i only

↓

we have to maximize $P\left(\frac{x}{\theta_i}\right) \forall_i$

- Since data from another class doesn't effect other class's distribution so considering whole data distribution will not effect θ_i of class i.

$$P\left(\frac{x}{\theta}\right)$$

→ In this, we assume information in S_j doesn't effect θ_i if $i \neq j$.

So even if we consider entire training dataset S instead of just S_i, parameter θ_i will only be influenced by S_i.

Likelihood is == probability distribution

(It can be bionomial, Log based ....)

→ A well defined distribution can write statistical in Composed form.

{ If 1 dimensional data = variance
  multiple dimensional " = Co-variance

θ_i = parameter of pdf of particular class i.

[In θ parameter is not necessarily two, it varies, diff. pdf has diff no. of parameters in θ]

eg:- for gaussian : μ, Σ

We can interchangably say $C_i \sim \theta_i$

→ Every training data is independent.

Maximum likelihood Estimation :-

If we have N training samples $x_1, x_2, \ldots x_N$

goal :- We want to maximize likelihood of each of training samples.

So we want to maximize :-

$$P\left(\frac{x_1}{\theta}\right), P\left(\frac{x_2}{\theta}\right), \ldots\ldots P\left(\frac{x_N}{\theta}\right)$$

We have to find θ that maximizes the product :-

$$P\left(\frac{x_1}{\theta}\right) P\left(\frac{x_2}{\theta}\right) \ldots P\left(\frac{x_N}{\theta}\right) = \prod_{K=1}^{N} P\left(\frac{x_K}{\theta}\right)$$

we have to find θ that maximizes product so :-

Step 1 :- Take log ln

Step 2 :- Equate ~~to~~ θ To maximize find gradient

Step 3 :- Equate gradients to zero

Step 4 :- Find value of 'θ'

It is not necessary that while find θ all times μ, Σ is calculated,

If given Σ we need find μ only.

Find θ that maximizes product :-

$$P\left(\frac{x_1}{\theta}\right) \cdot P\left(\frac{x_2}{\theta}\right) \cdots P\left(\frac{x_N}{\theta}\right) = \prod_{k=1}^{N} P\left(\frac{x_k}{\theta}\right)$$

Step 1:- It is equivalent to finding θ that maximizes :-

$$l(\theta) = \ln\left(\prod_{k=1}^{N} P\left(\frac{x_k}{\theta}\right)\right) = \sum_{k=1}^{N} \ln P\left(\frac{x_k}{\theta}\right)$$

Step 2:- To maximize, find gradient of $l(\theta)$ and equate to zero

$$\nabla_\theta\, l(\theta) = 0$$

If it gives min, max ~
else calculate double derivative.

→ If multiple max, choose maximum among all maxima.
Find highest maxima.

- If we have m no. of parameters $\theta_1, \theta_2, \ldots \theta_m$ we have to do :-

$$\begin{bmatrix} \frac{\partial l(\theta)}{\partial \theta_1} \\ \frac{\partial l(\theta)}{\partial \theta_2} \\ \vdots \\ \frac{\partial l(\theta)}{\partial \theta_m} \end{bmatrix} = 0$$

Let's assume pdf has Gaussian distribution with mean μ and covariance matrix Σ.

- Assume Σ is known, we have to find μ

$$P\left(\frac{x_k}{\mu}\right) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left[-\frac{(x-\mu)^T \Sigma^{-1} (x-\mu)}{2}\right]$$

→ Take $\mu = \theta$ parameter to be found

$$\ln P\left(\frac{x_k}{\theta}\right) = -\frac{1}{2}\ln(2\pi)^d |\Sigma| - \frac{1}{2}(x_k-\theta)^T \Sigma^{-1}(x_k-\theta)$$

x_k = Training data
feature vector containing d dimension

[Naive Bayes consider all dimension seperately]

→ we have to find θ_i that maximizes $P\left(\frac{x}{\theta_i}\right) \forall i$

We assume that information in S_j doesnot affect θ_i if

So even if we consider entire training dataset S instead of S_i, parameter θ_i will only be influenced by S_i

we can say that we have to find θ that maximizes $P\left(\frac{x}{\theta}\right)$

$$P\left(\frac{x_k}{\mu}\right) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{\frac{1}{2}}} \exp\left[-\frac{(x_k-\mu)^T \Sigma^{-1}(x_k-\mu)}{2}\right]$$

Take $\mu = \theta$

$$l_k(\theta) = \ln P\left(\frac{x_k}{\theta}\right) = -\frac{1}{2}\ln(2\pi)^d|\Sigma| - \frac{1}{2}(x-\theta)^T\Sigma^{-1}(x_k-\theta)$$

$$l(\theta) = \sum_{k=1}^{N} \ln P\left(\frac{x_k}{\theta}\right) = \sum_{k=1}^{N} l_k(\theta)$$

In can be shown as → take derivative ! 1st term independant of θ

$$\nabla_\theta l_k(\theta) = \Sigma^{-1}(x_k-\theta)$$

$$\nabla_\theta l(\theta) = \sum_{k=1}^{N} \nabla_\theta l_k(\theta) = \sum_{k=1}^{N} \Sigma^{-1}(x_k-\theta)$$

For MLE :- equate to zero →

$$\nabla_\theta l(\theta) = \sum_{k=1}^{N} \nabla_\theta l_k(\theta) = \sum_{k=1}^{N} \Sigma^{-1}(x_k-\theta) = 0$$

Since $\Sigma^{-1} \neq \theta$, $\sum_{k=1}^{N}(x_k-\theta) = 0$

$$\left(\sum_{k=1}^{N} x_k\right) - N\theta = 0$$

$$\theta = \frac{\sum_{k=1}^{N} x_k}{N}$$

= mean when all data are Equally likely.

eg 2:- Consider a Univariate Gaussian with unknown μ & σ^2.

Find values of μ, σ^2 for MLE :-

Univariate specified gaussian by 1 variable.

→ If 2 variable = Take partial derivative.

θ = for equally likely

$$P\left(\frac{x_k}{\mu, \sigma^2}\right) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left[-\frac{(x_k - \mu)^2}{2\sigma^2}\right]$$

Assuming $\theta_1 = \mu$, $\theta_2 = \sigma^2$:-

$$P\left(\frac{x_k}{\theta}\right) = \frac{1}{(2\pi\theta_2)^{\frac{1}{2}}} \exp\left[-\frac{(x_k - \theta_1)^2}{2\theta_2}\right]$$

$$l_k(\theta) = \ln p\left(\frac{x_k}{\theta}\right) = -\frac{1}{2}\ln(2\pi\theta_2) - \frac{1}{2\theta_2}(x_k - \theta_1)^2$$

$$\nabla_\theta l_k(\theta) = \begin{bmatrix} \frac{\partial l_k(\theta)}{\partial \theta_1} \\ \frac{\partial l_k(\theta)}{\partial \theta_2} \end{bmatrix} = \begin{bmatrix} \frac{1}{\theta_2}(x_k - \theta_1) \\ -\frac{1}{2\theta_2} + \frac{1}{2\theta_2^2}(x_k - \theta_1)^2 \end{bmatrix}$$

while equating to zero

$$\sum_{k=1}^{N} \nabla_\theta l_k(\theta) = 0$$

$$\sum_{k=1}^{N} \frac{1}{\theta_2}(x_k - \theta_1) = 0$$

$$\sum_{k=1}^{N} \left(-\frac{1}{2\theta_2} + \frac{1}{2\theta_2^2}(x_k - \theta_1)^2 \right) = 0$$

Assuming $\theta_2 \neq 0$

$$\theta_1 = \sum_{k=1}^{N} \frac{x_k}{N} \qquad \theta_2 = \sum_{k=1}^{N} \frac{(x_k - \theta_1)^2}{N}$$

This is estimate of μ This is estimate of σ^2

Q:- $P\left(\frac{x_k}{\theta}\right) = \begin{cases} \theta \exp(-\theta x_k) & \text{if } x_k \geq 0 \\ 0 & \text{otherwise} \end{cases}$

Sol^n:- Goal :- To find parameter θ that maximizes likelihood of observing given data.

given Probability distribution :—

$$P\left(\frac{x_k}{\theta}\right) = \begin{cases} \theta \exp(-\theta x_k) & \text{if } x_k > 0 \\ 0 & \text{otherwise} \end{cases}$$

Step 1:- for n independent data points $x_1, x_2 \cdots x_n$, function $L(\theta)$ is product of individual probability :—

$$L(\theta) = \prod_{k=1}^{n} P\left(\frac{x_k}{\theta}\right)$$

for given pdf, likelihood becomes :—

$$L(\theta) = \prod_{k=1}^{n} \theta \exp(-\theta x_k)$$

Since θ and x_k are independent :— Simplifying :—

$$L(\theta) = \theta^n \exp\left(-\theta \sum_{k=1}^{n} x_k\right)$$

Step 2 :- To simplify maximization process, take log of likelihood :-

$$\log L(\theta) = \log\left(\theta^n \exp\left(-\theta \sum_{k=1}^{n} x_k\right)\right)$$

Simplifying :-

$$\log L(\theta) = n \log \theta - \theta \sum_{k=1}^{n} x_k$$

Step 3 :- differentiate log. likelihood wrt θ, set it equal to 0, find maximum :-

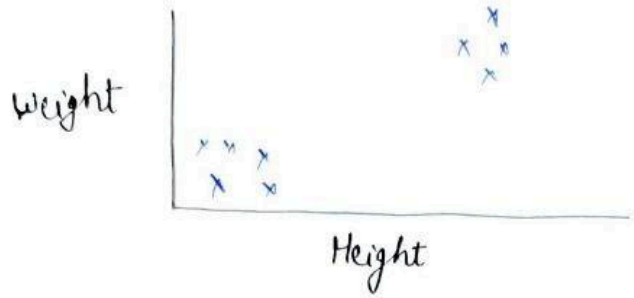
$$\frac{d}{d\theta} \log L(\theta) = \frac{n}{\theta} - \sum_{k=1}^{n} x_k = 0$$

Step 4 :- Solve for θ :-

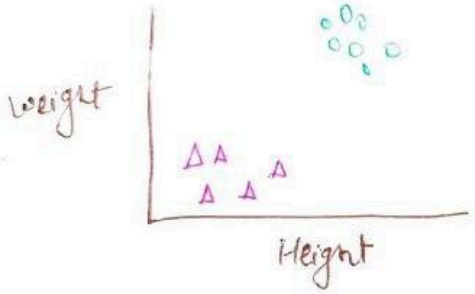
$$\theta = \frac{n}{\sum_{k=1}^{n} x_k}$$

It means that θ is reciprocal of average of observed data points x_k.

Lecture - 18



Some sample data of a few animals



In clustering, we create group of samples in such a way that sample in group are more similar to each other than sample in other groups

Let's 1st see **Notion of Distance** :-

There are multiple types of distances :-

Given : 2 points (x_1, y_1) (x_2, y_2) in 2D plane.

Goal : Is to find distance

Approaches are many :-

Euclidian distance :-

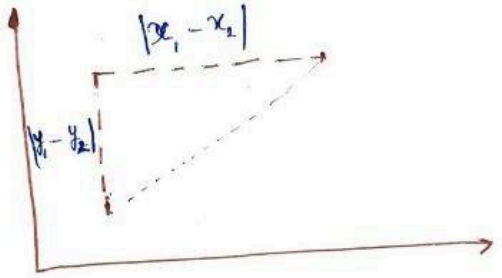
$$d_E = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

data point in 2 dimension feature.

When we move to higher dimension :-

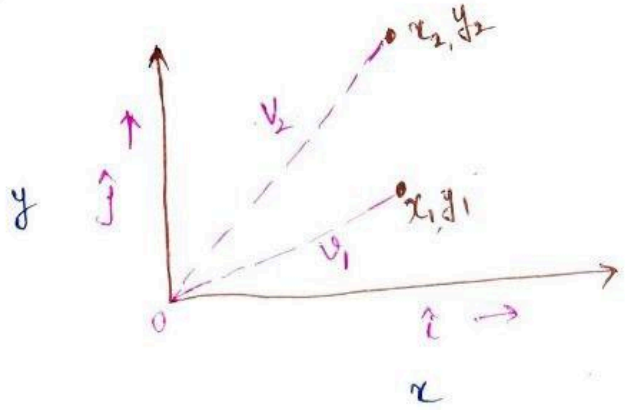
Euclidian distance tend to loose its significance.

(ii) Absolute distance / Manhatan distance :-



$$d_m = |x_1 - x_2| + |y_1 - y_2|$$

(iii) Mostly for higher dimensions - we prefer cosine distance.



- Destance of x_1, y_1 and origin is v_1
- " $x_2 y_2$ - - - - = v_2

It indicates Angular distance.

$$V_1 = x_1 \hat{i} + y_1 \hat{j}$$

$$V_2 = x_2 \hat{i} + y_2 \hat{j}$$

$$\vec{V_1} \cdot \vec{V_2} = |\vec{V_1}| \cdot |\vec{V_2}| \cos\theta$$

$$\cos\theta = \frac{\vec{V_1} \cdot \vec{V_2}}{|\vec{V_1}| \cdot |\vec{V_2}|} = \frac{x_1 x_2 + y_1 y_2}{\sqrt{x_1^2 + y_1^2}\sqrt{x_2^2 + y_2^2}}$$

Cosine distance tells how much I need to rotate V_1 to reach V_2

→ It is indicator of direction mostly.

Clustering

Some examples of it :—

- Targeted Advertisement to group people of similar interest.
  - → Recommandation system

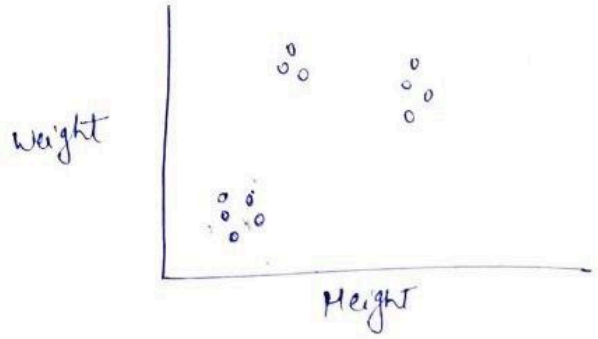
Clustering Problem :	don't assign label
Classification " :	It can assign Label

Clustering Techniques :—

- Centroid model : K mean, K mediod
- Graph based model : Spectral clustering

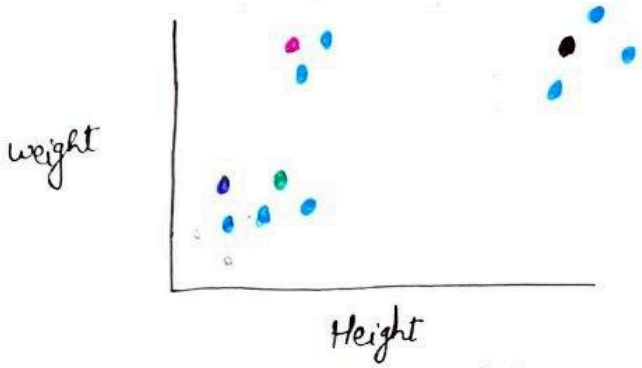
- Distribution model : Gaussian Mixture models
- Density models : DBSCAN
- Connectivity Based : Hierarchical clustering
- Neural models : Self-organizing map

(1) K-means :- Take data points :-

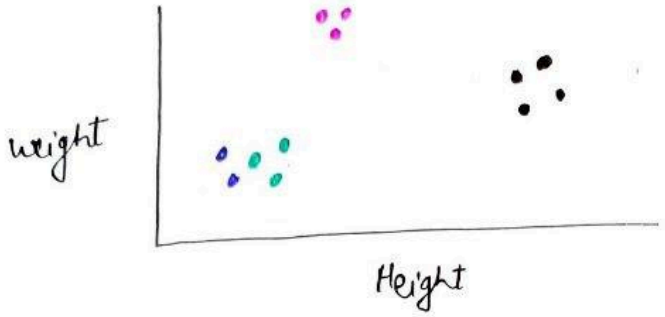


Let's say I want m clusters , i.e = m : 4

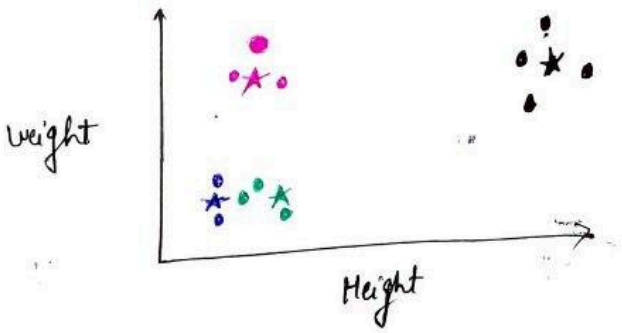
- Randomly initialize m=4 cluster centers.



- Assign each of other points to nearest cluster center.

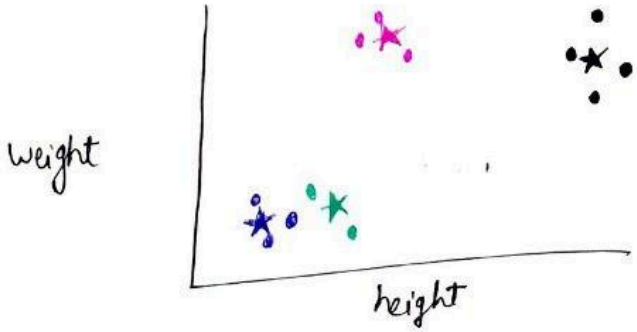


→ Recalculate to cluster mean.



→ Reassign points based on updated mean.

→ Keep in loop to mean calculation, reassign part until there is no change in assignment.



- Randomly assign points to 4 points centers.
- Each data point is d-dimensional data.

$$x^1 \rightarrow x_1^1, x_2^1 \ldots\ldots x_N^1$$
$$x_2^2 \rightarrow x_1^2, x_2^2 \ldots\ldots x_N^2$$
$$\vdots$$
$$x^N \rightarrow x_1^N, x_2^N \ldots\ldots x_d^N$$

Coordinate of centroid for cluster 1, Cl_1 :-

$$Cl_1 = \{C_1^1, C_2^1 \ldots\ldots C_d^1\}$$

$$C_1(1) = \frac{x_1^1 + x_1^2 \ldots\ldots x_1^p}{p}$$

If p points are there

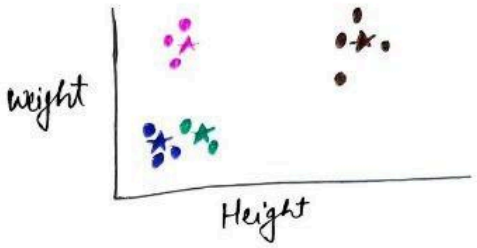
$x^1 \rightarrow$ no of data point = p data point

$x_1 \rightarrow$ dimension eg:- d dimension

- No gaurantee if there is convergence or not, infinite loop.
- We set some threshold when less than i points changes their cluster assignment.

- We can tune it to find K.
  - ↳ There is no full-proof method to find K (# of clusters)

Steps for clustering :-



(i) Let's say, I want m clusters

(ii) Let's take m = 4

(iii) Randomly initialize m = 4 cluster centers

(iv) Assign each of other points to nearest cluster center

(v) Recalculate cluster mean

(vi) Reassign points based on updated mean

(vii) Go to step 5, if there is any change in the assignment, Otherwise, stop.

Limitations :-

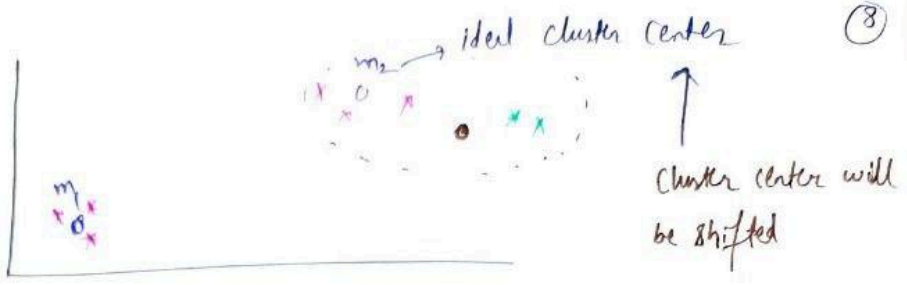
(i) Not gaurantee to converge

(ii) Difficult to find K

(iii) Sensitive to outliers

(iv) may be significantly affected by initialization

eg:-



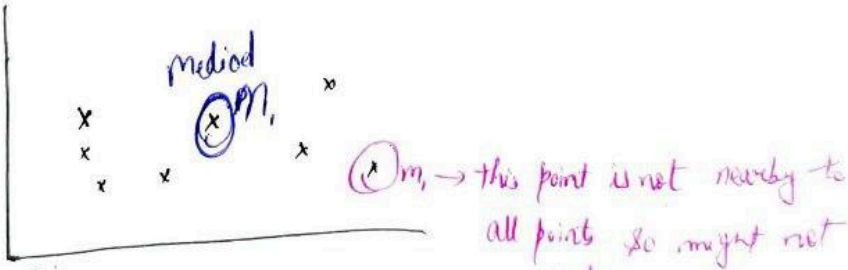
(i) 1, 2 data mean is far away from m_1, m_2 so mean shifted one of the way is to discard these points.

→ K mean significantly affected by initialization.

→ Mean may not lie on actual data point.

→ Medio lies on actual data point.

Mediod:-



m_1 → this point is not nearby to all point so might not mediod.

m_1 = mediod → it is nearby to all points (centre)

Mediod = minimum distance

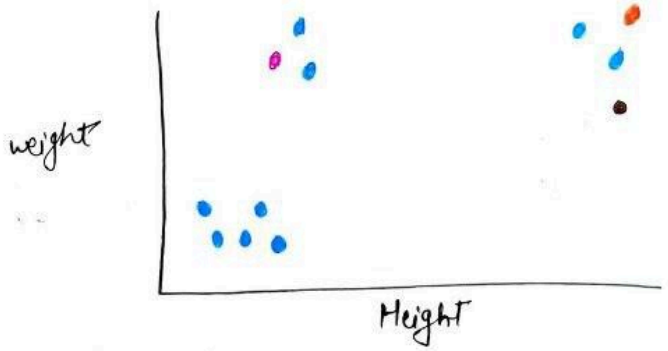
Mediod : We have to options :

(a) To assign points randomly to any point

(b) To assign point to same cluster.

Clustering Technique : K- mediod :-

- A mediod is a point in the cluster from which sum of distances to other data points is minimal.
- A Mediod is a point in cluster from which dissimilarities with all other points in cluster are minimal.
- Instead of Centroid as reference points in K-means algorithms, K-mediod algo takes a Mediod as reference point.



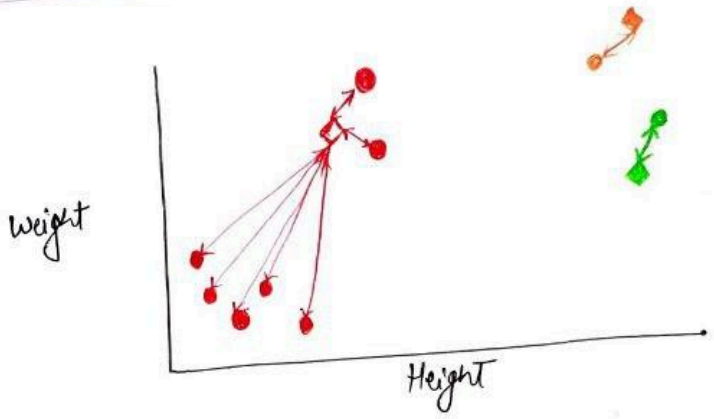
(i) Take the data points

(ii) Randomly assign mediods

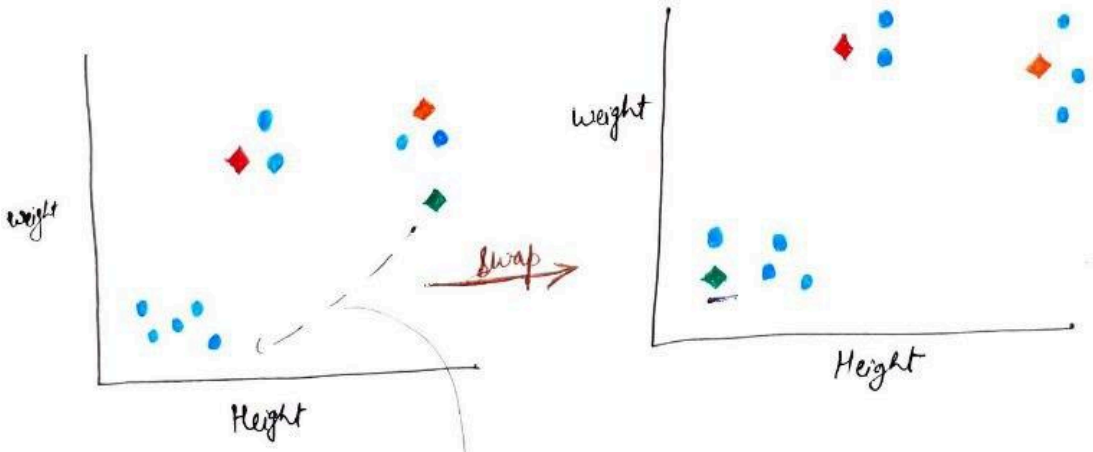
(iii) For each non-mediod data points, find nearest mediod and assign data point to the corresponding cluster.

(iv) Calculate distance of every data point from corresponding medoid. Sum of all these distances is called **Cost**. (10)

> Cost = total red distance + total orange distance + total green distance



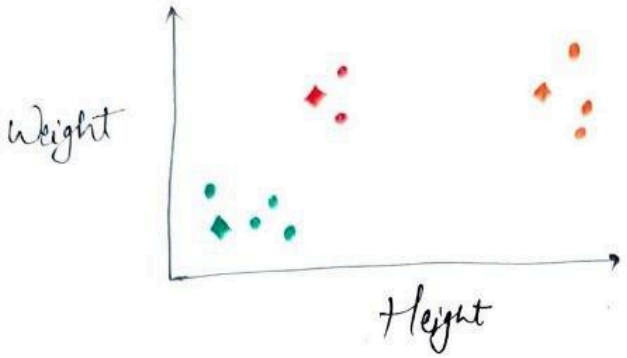
• Rhombus represents = medoids



old **New**

if restricted to same cluster green couldn't come at Bottom.

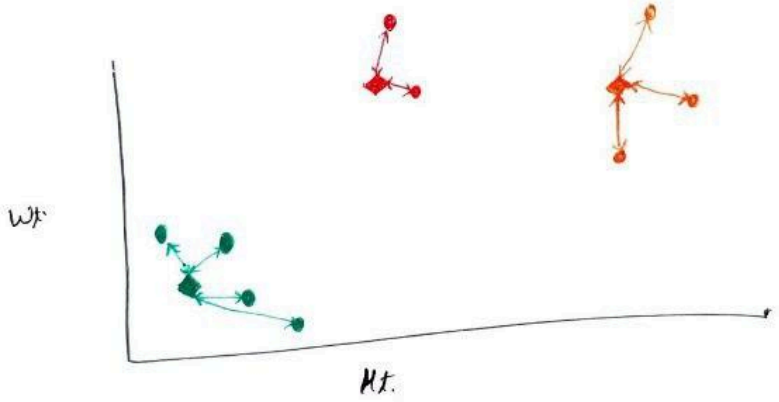
- With new Mediods, create the clusters using nearest mediods.



→ with the new Mediods, recalculate the cost.

- Calculate distance of every data point from corresponding Mediod. Sum of all distances is called Cost.

Cost = total red distance + total yellow distance + total black distance



If new cost > old cost → discard new mediods
↓
go back to old mediods.
↓
Algorithm converges.

else, keep new Mediod and redo random swapping.

Steps :-

(i) Select K random points from the dataset.

(ii) Assign data points to the cluster of nearest mediod.

(iii) Calculate total sum of distances of data points from their assigned mediod for each mediod.

(iv) Swap a non-mediod with mediod point and recalculate cost.

(v) Undo swap if recalculated cost with new mediod exceeds previous cost.

Until new cost > old — Terminate condition

↓

This applies when already tried for 3-4 times if always cost is greater than old one then stop.

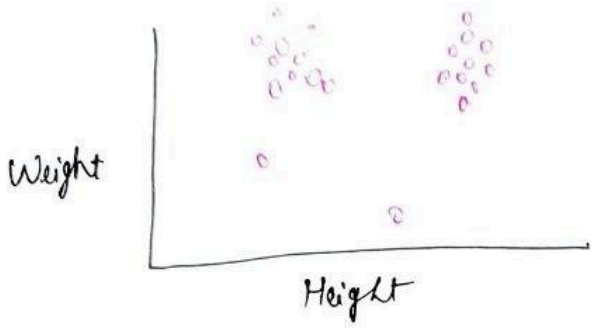
To get rid of K mean, K mediod initialization based method;—

Density Based Clustering.

Lecture - 19

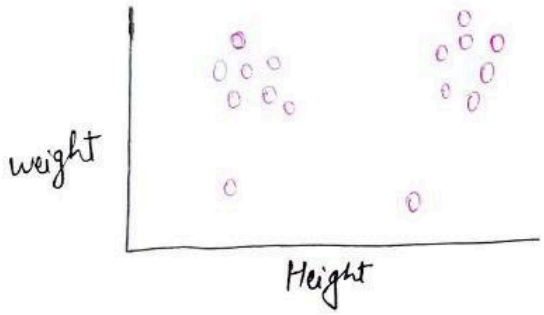
Date : 17/09/2024

Clustering Techniques : Use of Density :



which points should form cluster ?

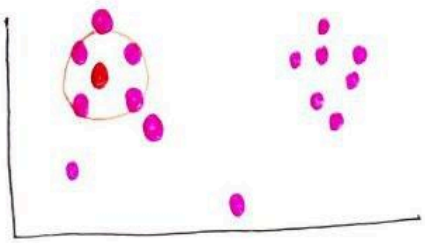
→ Take a disc of specific size (parameter)



take a disc of specific size (parameter)

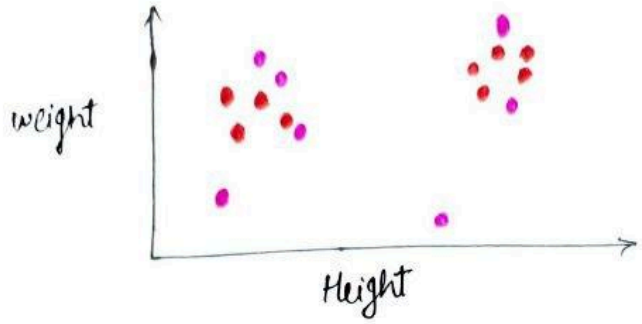
→ Put it across every point.

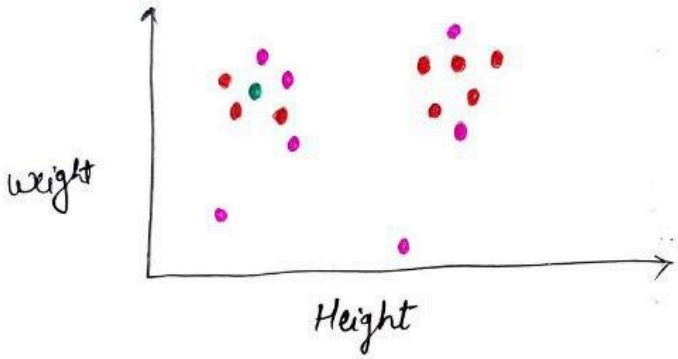
→ If there are m no. of points within disc when positioned on point x_i,
The point x_i is called as **Core Point.** m is parameter



• Put it across every point.
If there are m no. of points within the disc when positioned on point x_i,

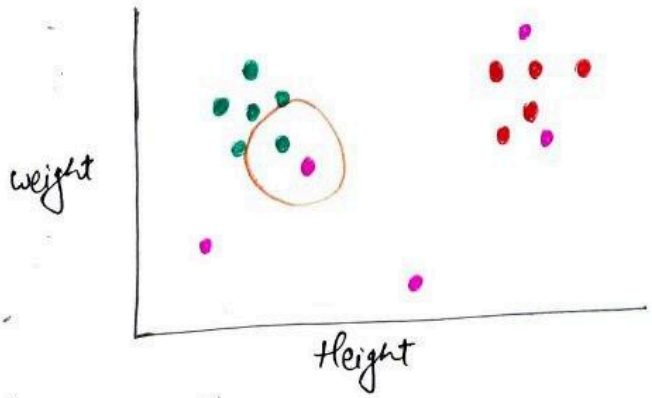
→ Point x_i will be called a core point. m- parameter.





Red : core points

- Assign a cluster label in a core point (green)



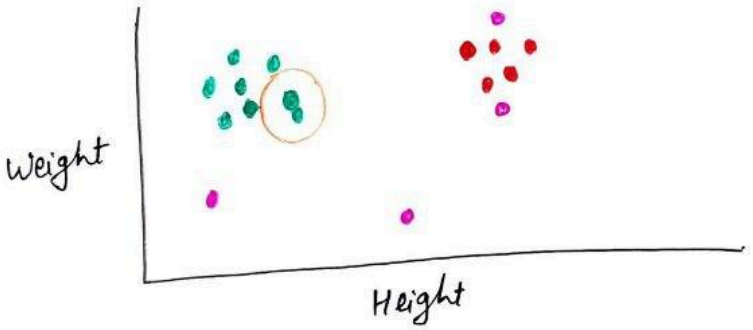
→ Assign a cluster label in a core label (green)

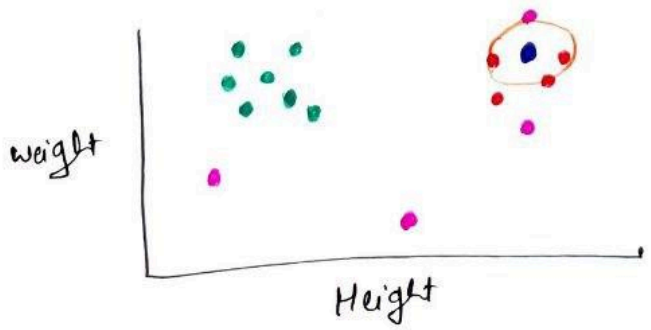
- Put every point within the disk inside the 1st cluster.

- Apply the dist from every point in the cluster.
- eg:- Customer has no idea, how many customer for delivery boy.
  So based on density based clustering.
- Based on density, I am making Top, bottom, right, left Clustering.

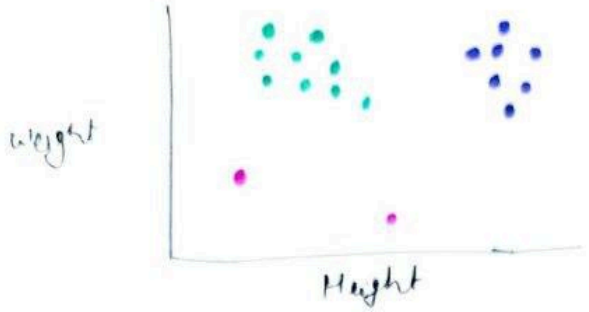
→ If no. of point inside disc is greater than outer, this point is core points.
i.e Core point = Crowded region", if m = some Threshold.

→ Apply disc from every point in cluster.





- Repeat it for core points



When we are done with all core points and left with only non-core points that can't be added to any clusters, we are done.

These non-core points is called outliers

- Its 2D, so considering circle, hypersphere else.
- Bottom two aren't core point as doesn't contain point greater than equal to m.

(i) Find core point
↓
based on putting circle disc m points, check if falling as core point, randomly choose any core, find its neighbour
↓
repeat.

(ii) Repeat with different radius values find which 'r' considers more point.

(iii) Choose intutive value of r

There are two parameters is m, r