2/19

What is AI?

Algorithms and systems exibit human-like intelligence

What is ML:

A subset of AI. We want to compensate the lack of knowledge (with knowledge you write programs) by data. It can learn from available data and change according to the scenario changes.

DL:- Imitating human brain.

⇒ Machine learning is a set of algorithm

Machine learning types
- Supervised → classification, Regression.
- Unsupervised

Not Required

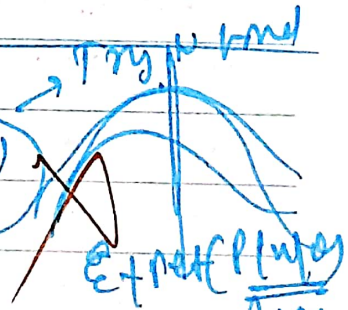
MAE

$P(\theta|x)$

Find the point where $P(\theta|x)$ is maximum

$$P(\theta|x) = \frac{P(x|\theta) \cdot P(\theta)}{P(x)}$$

$P(x|\theta)$



⇒ Much learning involves aquiring general concepts from specific training example.

or Approximating a boolean valued function from example.

⇒ Example

Target Concept:

- Day's on which my friend enjoys water sports:

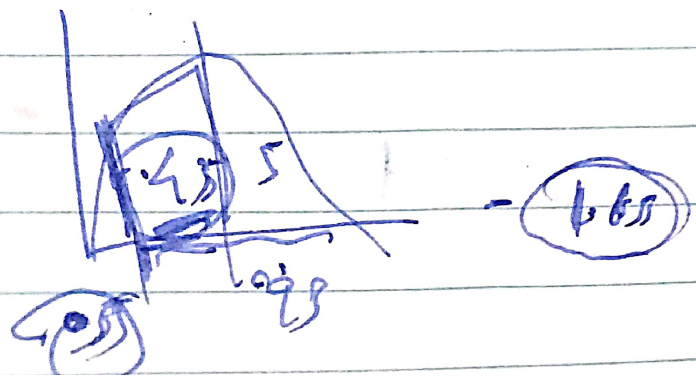
Example	Sky	Airtemp	Humidity	Wind	Water	Forecast	Enjoy
1	Sunny	Warm	~~Strong~~ Normal	Strong	Warm	Same	yes
2	Sunny	Warm	High	Strong	Warm	Same	yes
3	Rainy	cold	High	Strong	Warm	Change	~~No~~
4	Sunny	Warm	High	Strong	Cool	Change	yes

What hypothesis representation shall we provide on which each hypothesis is a conjunction of constraints on instance attributes.

∴ ? for any value Acceptable for attribute

∴ specify single value if this value is required.

∴ indicate ϕ that no value is acceptable.



⇒ If some instance x satisfies all the constraints of hypothesis h, then h classifies x as the example (h(x)=1)

⇒ To represent a hypothesis that your friend enjoys his favorite sports only on cold days with high humidity.

$$\langle ?, cold, High, ?, ?, ? \rangle$$

⇒ A Learning task can be described by the set of instances over which target function is defined.

Notations

→ Items over which the concept is defined is called set of instances.

⇒ Concept or the function to be learned is called target function.

$$X \rightarrow \{0, 1\}$$

Given:

→ Instances X: possible days defined by attributes.
→ Sky, AirTemp, Humidity, Wind, Water, Forecast.

Sunny, cloudy, rainy; Warm, cold; Normal, Humid; Strong, Weak; Warm, cool; Same, change

→ Hypothesis H: Each hypothesis is described by a conjunction of constraints on at attributes

→ Target concept: $\boxed{c: X \rightarrow \{0,1\}}$

→ Training examples D: positive & negative examples

Determine:

A hypothesis h in H such that h(x) = c(x) for all x ∈ X

transpose

⇒ Learner is trained by set of examples along with target concept value

$c(x) = 1$ are the instances → Members of target function

$c(x) = 0$ -ve examples → Non-Members of Target function

$\langle x, c(x) \rangle$ training example.

⇒ D is set of all available training example

⇒ H is set of all possible hypothesis determined by the representation of human designer's choice of representation.

⇒ Each hypothesis $h \in H$ represents a boolean-valued function defined over X; $h: X \to \{0,1\}$. Goal of the learner is to find a hypothesis h such that $h(x) = c(x)$ for all $x \in X$.

Inductive Learning Hypothesis

⇒ Entire set of instance of X are not always available. except the limited training data.

⇒ Lacking any further information we assume the best hypothesis regarding unseen instances is the hypothesis that best fits the observed data.

⇒ Defn Any hypothesis found to approximate the target function well over a sufficiently large set of training examples will also approximate the target function well over other unobserved examples.

Scanned with OKEN Scanner

Concept of learning as Search

⇒ Concept of learning is searching through a large space of hypothesis.

⇒ There are $3 \times 2 \times 2 \times 2 \times 2 \times 2 = 96$ distinct instances and

$5 \times 4 \times 4 \times 4 \times 4 \times 4 = 5120$ syntatically distinct hypothesis

(2 symbols added are ϕ and ?)

⇒ Where ϕ represents every instance as negative, even at any value. Hence actual No of hypothesis

$= (4 \times 3 \times 3 \times 3 \times 3 \times 3) + \underbrace{1}_{\phi} = 973$

⇒ This hypothesis space is finite.

General to specific ordering of Hypothesis

$h_1 = \langle sunny, ?, ?, strong, ?, ? \rangle$ — More specific

$h_2 = \langle sunny, ?, ?, ?, ?, ? \rangle$ — More general

→ preferred

Defn Let h_j and h_k be boolean-valued functions defined over X. Then h_j is more-general-than-or-equal-to h_k ($h_j \geq_g h_k$) iff

$(\forall x \in X)[h_k(x) = 1] \rightarrow (h_j(x) = 1)]$

⇓

this is a partial ordering

⇒ Based on this the total Hypothesis class can be organized by partial ordering.

Find-S: Finding a Maximally Specific hypothesis

1. Initialize h to most-specific hypothesis in H
2. for each +ve training instance x
   - For each attribute constraint a_i in h
     If the constraint a_i is satisfied by x
     then do nothing
     Else replace a_i in h with next more general constraint that is satisfied by x.
3. output h

Example

~~Take~~ Find the most specific hypothesis (as partially ordered)

$h_0 \leftarrow \langle \phi, \phi, \phi, \phi, \phi, \phi \rangle$

$x_1 = \langle$ Sunny, warm, Normal, Strong, warm, same $\rangle$ +
$x_2 = \langle$ Sunny, Warm, High, strong, warm, same $\rangle$ +
$x_3 = \langle$ Rainy, cold, High, strong, warm, change $\rangle$ -
$x_4 = \langle$ Sunny, warm, High, strong, cool, change $\rangle$ +

$x_1 \Rightarrow$ From the first example x_1, it is clear that we need to replace the constraints to more general values

$h' \leftarrow \langle$ sunny, warm, Normal, strong, warm, same $\rangle$

$x_2 \Rightarrow$ In 2nd training set Again a contradiction in 3rd attribute make h' to be more general

$h'' \leftarrow \langle$ sunny, warm, ?, strong, warm, same $\rangle$

$x_3 \Rightarrow$ -ve Example is neglected as long as we assume the target function c is inside H and training data are correct.

⇒ The 2nd example forces one more change

$h''' \leftarrow \langle$ sunny, warm, ?, strong, ? $\rangle$

Several issues in Find-S algorithm

⇒ Has the learner converged to the correct target concept? There can be many other consistent hypothesis?

⇒ Among all the consistent hypothesis Find-S will get the most specified one.

⇒ There is always a chance that training data can have some noise

Version Space and Candidate-Elimination Algorithm

⇒ Output of the algorithm is the set of all hypotheses consistent with training examples.

Defn: A hypothesis h is consistent with a set of training examples D iff $h(x) = c(x)$ for each $\langle x, c(x) \rangle$ in D

$$\text{Consistent}(h, D) \equiv (\forall \langle x, c(x) \rangle \in D)\ h(x) = c(x)$$

⇒ Set of all hypotheses consistent with observed training example is known as <u>version space</u>.

$$VS_{H,D} \equiv \{ h \in H \mid \text{Consist}(h, D) \}$$

→ The List-then-Eliminate algorithm :- simply lists all hypotheses in H and eliminates any hypothesis found inconsistent with training examples. Version space shrinks with more examples. If insufficient training examples donot lead to a single hypothesis, set of all hypotheses are displayed.

→ Can be applied to all finite spaces.
Need exhaustive search for even trivial hypotheses.

Candidate Elimination Algorithm

→ Initialize G to the set of maximally general hypothesis in H
→ Initialize S to the set of maximally specific hypothesis in H
→ For each training example d do
  → If d is +ve
    → remove from G any hypothesis inconsistent with d
    → For each hypothesis s in S that is not " " d
    → Remove s from S
    → Add to S all minimal generalization h of s such that
      → h is consistent with d & some member of G is more general than h.
    → Remove from S any hypothesis that is more general than any hypothesis in S.
  → If d is -ve
    → Remove from S any hypothesis inconsistent with d
    → For each hypothesis g in G that is not consistent with d
      → Remove g from G
      → Add to G all minimal specialization h of g such that
        → h is consistent with d & some member of S are more specific than h.
    → Remove from G any hypothesis that is less general than other hypothesis in G.

x_1 = <Sunny, Warm, Normal, Strong, Warm, Same>
x_2 = <Sunny, Warm, High, Strong, Warm, Same>
x_3 = <Rainy, Cold, High, Strong, Warm, Change>
x_4 = <Sunny, Warm, High, Strong, Cool, Change>

+ve example generalizes specific ~~c~~ hypothesis
-ve example specifies generalized

⇒ We start by most specialized hypothesis boundary S_0 and most generalized boundary G_0

S_0 = <ϕ, ϕ, ϕ, ϕ, ϕ, ϕ>
S_1 = <Sunny, Warm, Normal, Strong, Warm, Same>
$S_2 = S_3$ = <Sunny, Warm, ?, Strong, Warm, Same>
S_4 = <Sunny, Warm, ?, Strong, Warm, ?>

G_4 = <Sunny, ?, ?, ?, ?, ?> <?, Warm, ?, ?, ?, ?>
G_3 = <Sunny, ?, ?, ?, ?, ?> <?, Warm, ?, ?, ?, ?> <~~?, ?, ?, ?, ?~~, Same>

Eliminated at G_4

$G_0 = G_1 = G_2$ = <?, ?, ?, ?, ?, ?>

because the +ve example of h_4 is inconsistent

→ x_4 is the ad

⇒ The Candidate-Elimination algorithm converge to target concept if (i) no error in traing data
(ii) there ~~are~~ some hypothesis in H that correctly describes target function.

→ Target concept is exactly learned when S and G boundaries converge to a single hypothesis.

→ For G if an specific attribute of negative example, then replace with a specific values of +ve example.

Inductive Bias

⇒ A Baised Hypothesis Space

Take the example

Example	Sky	Air temp	Humidity	Wind	Water	Forecast	Enjoy
1	Sunny	Warm	Normal	Strong	Cool	change	Yes
2	Cloudy	Warm	Normal	Strong	Cool	change	Yes
3	Rainy	Warm	Normal	Strong	Cool	change	No

Let the target function be "Sky = Sunny or Sky = Cloudy"

S1: <Sunny, Warm, Normal, Strong, Cool, change>
<Cloudy, Warm, Normal, Strong, Cool, change>

S2: <?, Warm, Normal, Strong, Cool, change>

→ However this is overely general which will also make 3rd example true (wrongly)

The problem is; we have a baised learner to include only conjuctive hypothesis, not disjunctive.

⇒ We require more expressive hypothesis space.

An unbaised Learner

⇒ One Hypothesis space should be large enough to represent all possible concept, which is power set of X i.e $2^{|X|}$ possible hypothesis, which is very high.

⇒ The conjuctive hypothesis space had approximally 973 hypothesis vs 2^{96} in power set.

⇒ Fundamental property of inductive inference that, the candide-elimination algorithm was able to generalize beyond the observed training is because of the inductive bais that the hypothesis is conjuction of attributes.

⇒ Inductive learning requires some form of prior assumption or inductive learning.

⇒ Given a learning algorithm L, arbitary set of training data $D_c = \{\langle x, c(x)\rangle\}$

⇒ After training L is asked to classify a new instance x_i. Let $L(x_i; D_c)$ denote a classification of x_i based on seen examples of D_c.

→ We can define the inductive inference as follows $(D_c \wedge x) \succ L(x_i, D_c)$

Inductive Inference

However $D_c \wedge x \vdash L(x_i, D_c)$ (can't be claimed.

$$\boxed{\text{Howere } [B \wedge D_c \wedge x_i] \vdash L(x_i, D_c)}$$

B → inductive bais

$[B \wedge D_c \wedge x_i] \vdash L(x_i, D_c)$ → Deductive Inference

~~Above bais lead~~ to ~~further~~ r

Probability Distributions

Random Experiment

An Experiment whose outcome is ~~unknown~~ not known with certainty.

Sample space

Sample space is the universal set that consists of all possible outcome of an experiment. represented as S. Individual outcomes are known as ~~events~~. elementary events.

Example

① Sample space of coin toss $S = \{H, T\}$

② Marks scored by a student in a subject

$S = \{x \mid x \in N, 0 \le x \le 100\}$

Event (E)

Event (E) is a subset of a sample space and probability is usually calculated with to an event.

Random variable

⇒ A random variable is an function that maps every outcome in the sample space to a real number. It can be either discrete or continuous.

→ If a random variable can assume only a finite or countably infinite set of values it is called discrete random variable.

Ex - 1 - Credit ratings AAA, AA, A, BBB

2 - Number of orders received

3. Fraud.

⇒ If a random variable can take infinite set of values, it is called Continuous Random variable.

Ex 1. Market share of an company
2. Percentage of attrition in an Company.
3. Time to failure in a system.
4. Time taken to complete an order placed in an e-commerce platform.

⇒ Discrete random variables are described by Probability mass functions (PMF) and cummulative distribution function (CDF)

⇒ PMF is the prob. that a random variable ~~x~~ takes a specific value k.

Ex Prob. of no. of fraudent transcation is 10

Written as $P(X=10)$

⇒ CDF is prob. that random variable x takes atmost k value written as

$P(X \leq 10) \rightarrow$ No. of transcation is between 0 to 10.

⇒ Continuous random variables are described by Prob. a density function (PDF) & CDF. (cummulative distribution function)

PDF is the prob. that a continuous random variable takes values in a small neighbourhood of x given as

$$f(x) = \lim_{\delta \to 0} [x \leq X \leq x+\delta x]$$

→ random variable

CDF of continous random variable is

i.e prob that $X \le a$

$$F(a) = \int_{-\infty}^{a} f(x)\,dx$$

Binomial distribution

⇒ It's discrete prob. distribution and has several applications in many business contexts. A random variable X said to follow a binomial distribution when

1- It can have two outcomes

2- Objective is to find prob. of getting n sucess in n trails.

3- Prob. of sucess is p and the prob. of failure is 1-p

4- The prob p is constant and does not change betn the trails.

PMF (Prob. of No. of sucess is exactly n out of n trails) is given by $PMF(n) = P(X=n) = \binom{n}{n} * p^n * (1-p)^{n-n}$

$$CDF(n) = P(X \le n) = \sum_{k=0}^{n} \binom{n}{k} * p^k * (1-p)^{n-k}$$

swpy. stat. binom

Example of binomial distribution:

(Q) A fashion company sells its cloths online out of which 10% are returned. On a specific day 20 items were purchased

(a) find ~~no of~~ prob that exactly 5 will return items

(b) Prob. that a maximum 5 will return item.

(c) Prob. that more than 5 will return items.

(d) Average no. of customers who are likely to return the items and the variance and the standard deviation of the number of returns.

$\boxed{\Rightarrow \text{Average of binomial distribution} = n \times p \quad \text{variance of distribution} = n \times p \times (1-p)}$

Ans (a) $n = 5$, $n = 20$, $p = 0.1$

$PMF(n=5) = \binom{20}{5} \times (0.5)^5 \times (.9)^{15}$

Poisson Distribution (Discrete prob. distribution)

⇒ Number of events that may occur over a period of time / space.

ex No. of orders in a e-commerce portal

No. of cash withdrawal in ATM

No. of errors in a book

PMF is given as

$$P(X=k) = \frac{e^{-\lambda} \lambda^k}{k!}$$

λ = rate of occurance of event per unit of measurement

↓

Generally mean value.

Example of Poisson Distribution

⇒ No. of calls arriving at a call center follows Poisson distribution at 10 call per hour.

1. Calculate the prob. that the no. of calls will be maximum 5.

As No. of event = 5

Second parameter : Average no. of events = 10

cdf = 0.0608

(02) Calculate prob. that no. of calls over a 3-hour period exceed 30

mean (λ) = 10

mean over 3 hours = 3 × 10 = 30

$\lambda t = 30$

1 - cdf(30, 30) = 0.4516

Exponential Distribution

⇒ Single parameter continuous distribution ~~Used for model~~ generally used for modelling time-to-failure of electronic components

⇒ This distribution represents a process in which events occur continuously and independently at a constant rate.

Prob. density function given by:

$$\boxed{f(x) = \lambda e^{-\lambda x}, \quad x \geq 0}$$

λ is scale parameter & represents the rate of occurence of the event. $\frac{1}{\lambda}$ is Mean of exponetial distribution.

Example

Time-to failure of an avoinic system follows an exponential distribution with a mean time betn failure (MTBF) of 1000 hours. calculate

1. Prob. that system will fail before 1000 hours.
2. Prob. that it will not fail up to 2000 hours.
3. The time by which 10% of the system will fail

Mean as 910 like.

$\left(\lambda = \frac{1}{1000}\right)$

1. calculate the system will ~~to~~ fail before 1000 hours.

Ans Cummulative prob. up to 1000 for exponetial distribution will give the prob. that the system will fail before 1000 hours.

$$\int_0^{1000} \left(\frac{1}{1000}\right) * e^{\left(-\frac{1}{1000}\right) * x} = 0.632$$

(2) ~~Prob~~ Calculate the prob. that it will not fail upto 2000 hours:

$1 - \text{cdf}(2000, \lambda = \frac{1}{1000}, \text{scale} = 1000) = 0.135$

(3) Calculate the time by 10% of system will fail.

Edit :- Let x be the no of time ~~after which~~ within which 10% system will fail.

$\lambda = \frac{1}{1000}$

$$\int_0^x \frac{1}{1000} \times e^{-\frac{x}{1000}} = 0.1$$

$$\Rightarrow \frac{1000}{1000} - \frac{1000}{1000} e^{-\frac{x}{1000}} = 0.1$$

$$\Rightarrow e^{-\frac{x}{1000}} = 0.9$$

$$\Rightarrow \frac{1}{e^{\frac{x}{1000}}} = 0.9$$

$$\Rightarrow 0.9 \times e^{\frac{x}{1000}} = 1$$

$$\Rightarrow e^{\frac{x}{1000}} = \frac{1}{0.9} = \frac{10}{9}$$

$$\Rightarrow \log_e\left(\frac{x}{1000}\right) = \log_e(10/9)$$

$$\Rightarrow \frac{x}{1000} = 0.1053503$$

$$\Rightarrow \boxed{x = 105.3}$$

Normal Distribution (Gaussian Distribution)

It's one of the most popular continuous distributions. It's observed across many naturally occurring measures such as age, salary, sales volume, birth weight, height etc. It's popularly known as bell curve.

It has two parameters
mean of distribution μ and variance $= \sigma^2$

$$\bar{x} = \mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

~~Confidence Interval~~
~~⇒ Find out expected range~~

$$\text{pdf (Prob density function)} = f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

μ - mean or expectation of distribution (median & mode)

σ^2 = variance

$x \in \mathbb{R}$

$$\text{CDF} = \Phi\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{2}\left[1 + e\right.$$

Example

→ An investor want to understand the risks and returns associated with various stocks before investing in them. He is evaluating two stocks: BEML, GLAXO. It's observed that there is an upward trend in close price of GLAXO during 2010-2017 period. However BEML had a downward trend during 2013-2015, followed by an upward trend during 2014 and then again a price downward from mid of 2015 to 2015.

1. What is the expected daily rate of return on these stocks?

2. Which stocks have higher risk or volatility as far as daily daily returns are concerned.

3. What is the expected range of return by 95% confidence interval?

4. Which stock has higher prob. of making daily return of 2% or more?

5. Which stock has higher prob. of making loss (risk) of 2% or more.

$$\text{gain} = \frac{\text{close price}_t - \text{close price}_{t-1}}{\text{close price}_{t-1}}$$

What are the properties of Normal distribution

→ the mean, median and mode are exactly the same.
→ Distribution is symmetric about the mean, half fall below the mean and half above mean.
→ Distribution described by mean (μ) & standard deviation. Mean is location parameter & S.D is scale parameter.

⇒ Mean is peak of the curve, S.D stretches or squeezes the curve. Small S.D narrows the curve while large S.D widens the curve

Empirical rule

68-95-99.7 rule.

⇒ Around 68% of values are within 1 S.D from mean.
⇒ Around 95% values are within 2 S.D.
⇒ Around 99.7% of values are within 3 SD from mean.

$$\sigma(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

μ = mean
σ = S.D
σ^2 = variance.

⇒ Standard normal distribution (Z-distribution) is a normal distribution with mean 0 and S.D is 1.

⇒ Every normal distribution can be converted to standard normal distribution. Z scores tell you how many S.Ds away from the mean each value lies.

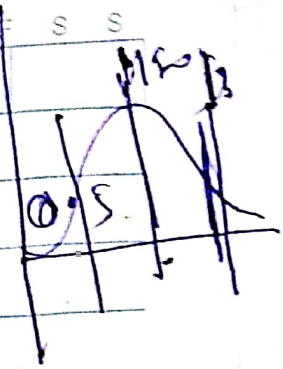
Scanned with OKEN Scanner

$z = \frac{x - \mu}{\sigma}$

$x \rightarrow$ observed value

$\mu \rightarrow$ hypothesis mean

$\sigma \rightarrow$ S.D

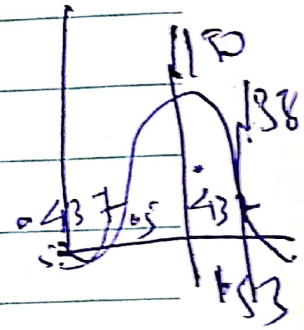


⇒ Find the prob. using z-distribution.

Each z-score is associated with a probability

Ex Find prob. of a SAT score in your sample exceeding 1380, with mean 1150 and S.D

$$\boxed{z = \frac{x - \mu}{\sigma}} = \frac{1380 - 1150}{150} = 1.53$$



corresponding probability = .937

$1 - .937 =$

Central Limit Theorem (CLT)

⇒ Let $S_1, S_2 \ldots S_k$ be samples of size n drawn from an independent and identically distributed population with mean μ and S.D σ.

Let $x_1, x_2 \ldots x_k$ are sample mean

According to CLT, distribution of $x_1, x_2 \ldots x_k$ follows a normal distribution with mean μ.

and S.D $= \sigma / \sqrt{n}$

⇒ CLT is key to hypothesis testing which primarily deals with sampling distribution

Hypothesis Test

⇒ Hypothesis is a claim and objective of hypothesis testing is either to reject or retain null hypothesis (current belief).

⇒ Hypothesis testing has two complementary statements
(i) Null hypothesis → current belief.
(ii) Alternate hypothesis

⇒ Hypothesis test → parametric (About population parameters of distribution of mean, proportion, S.D etc)
→ non-parametric
↓
About other characteristic
(independence of events or data following certain distributions)

* Precision:- Range of values within which the result is ~~exact~~ acceptable, e.g. 3%, 4% precision.

⇒ Confidence level:- 95% confidence limit. Mean out of 100 95 times the result is within ~~precision~~ acceptable range.

⇒ Significance level:- Likelihood the answer falls outside the range 95% confidence limit mean 5% significance level.

---

t-test, f-test, Chi-square test, Anova

$Z = \frac{|x-\mu|}{\sigma} = \left(\frac{1}{5}\right)$

Type-I error (α)

↳ ye H_0 is true but rejected

Type-II error (β)

H_0 is ~~accepted when it was~~ false but accepted.

Two tailed & one tailed test

⇒ Two tailed test rejects the null hypothesis if sample mean is significantly higher or lower than hypothesized value of mean of population. Hence two rejection regions.

Eg

Accept: $|Z| \leq 1.96$
Reject: $|Z| > 1.96$ } to 5% significance, 95% confidence

⇒ One tailed test when we test population mean is either lower or higher than hypothesized mean.

A: $Z > -1.645$
$Z < -1.645$ } 5% significance

Prob

$Z = \frac{\bar{x} - \mu}{\sigma_p/\sqrt{n}}$ ⇒ $Z \times (\sigma_p/\sqrt{n}) = \bar{x} - \mu$

⇒ $\bar{x} \leq \mu + 1.28 (\sigma_p/\sqrt{n})$

to .90

.951

Hypothesis Testing Mean

1- Population normal, infinite, large sample size or (~~small variance~~ small but population variance is known. H_a may be one-side or two-sided.

You observed some value and compare the observed average with hypothesis. check it lies in same side

$$z = \frac{\bar{x} - \mu_{H_0}}{\sigma_p / \sqrt{n}}$$

Used z-test

→ Calculate z

it correspond to the z ~~the prob. of values~~ is greater th

Take that probability and place it on graph to check it it is in accepted or rejected region.

(2) Population Normal, finite, Sample size is large or small but variance is unknown. H_a may be one-side/two-sided, use z-test

$$z = \frac{\bar{x} - \mu_{H_0}}{(\sigma_p / \sqrt{n}) * \sqrt{(N-n)/(N-1)}}$$

(3) population normal, infinite, small sample size with variance unknown. H_a may be one/two sided. use t-test

$$t = \frac{\bar{x} - \mu_{H_0}}{\sigma_s / \sqrt{n}}$$ with d.f = n-1

$$\sigma_s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{(n-1)}}$$

Scanned with OKEN Scanner

(5) Population may not be normal if sample size is large enough.

⇒ If variance of population not known then sample variance.

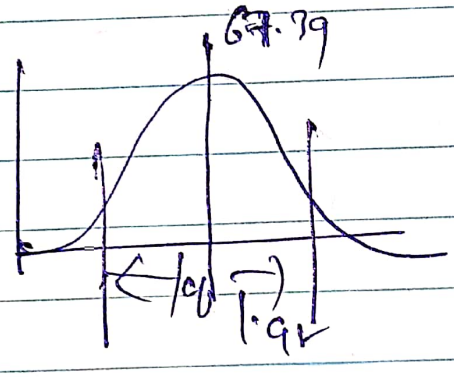
(Q) A sample of 400 male students with mean height 67.47 inches. Can it be reasonably regarded as a sample taken from a population with mean height 67.39 inches with S.D = 1.3 inches. Test at 5% level of significance.

Ans $H_0: \mu = H_0 = 67.39$

$H_A: \mu \neq 67.39$

$$z = \frac{\bar{x} - \mu_{H_0}}{\sigma_p/\sqrt{n}} = \frac{67.47 - 67.39}{1.3/\sqrt{400}} = 1.231$$

5% ~~Confidence~~ significance means 95%



Divide by $\frac{.95}{2} = .4775$

Check for

1.96

S. Table

Accepted

(Q) There are 20 industrial units of same size. All of them facing excessive labour turnover problem. The past records show that the mean of the distribution of annual turnover is 320 employees with S.D 75. A sample of 5 units taken at random and to have 300 employee turnover. Is it consistent with population mean at 5% ~~level~~ significance level

$H_0: \mu_{H_0} = 320$

$H_a: \mu_{H_0} \neq 320$

$\bar{x} = 300$, $\sigma_p = 75$, $n = 5$, $N = 20$

$$z = \frac{300 - 320}{75/\sqrt{5} * \left(\sqrt{(20-5)/(20-1)}\right)} = \frac{20}{(33.54) \cdot 88} = 0.67$$

$$\frac{\bar{x} - \mu_{H_0}}{\sigma_p/\sqrt{n} * \left(\sqrt{(N-n)/(N-1)}\right)}$$

(finite population)

$R: |z| > 1.96$

Accept

(Q) Mean of certain production is known to be 50 with S.D = 2.5. The production manager ~~[illegible]~~ Ch see product more than 50 not less. He takes 12 items with mean 48.5. What inference the manager would take about product with 5% significance.

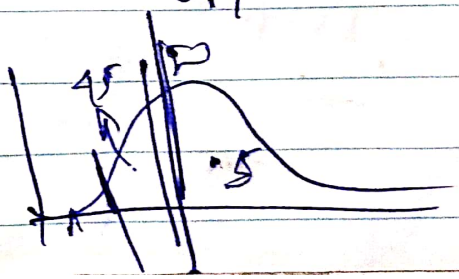
$H_0: \mu_{H_0} = 50$

$H_a: \mu_{H_0} < 50$

$$z = \frac{\bar{x} - \mu_{H_0}}{\sigma_p/\sqrt{n}} = -2.0784$$

$z_{5\%} \approx -1.645$

Hence rejected.



(Q) The specimen copper wire drawn from large lot have following breaking strength in kg weight:

578, 572, 570, 568, 572, 578, 570, 572, 596, 544

Whether it can be assumed that breaking strength is 578 kg. 5% level of significance.

(population normal, infinite, sample small, variance of population unknown)

$H_0: \mu = \mu_0 = 578$ kg

$H_a: \mu \neq \mu_0$

Since $n = 10$ and population s.d unknown we go for t-test.

$$t = \frac{\bar{x} - \mu_0}{\sigma_s / \sqrt{n}} = -1.488$$

(Sample size is small)

Degree of freedom $= n-1 = 10-1 = 9$

(two tail test)

$H_0: \mu = \mu_0 = 578$ kg

$H_a: \mu \neq \mu_0$

$$t = \frac{\bar{x} - \mu_0}{\sigma_s / \sqrt{n}} = \frac{572 - 578}{12.72/\sqrt{10}} = -1.4888$$

$\bar{x} = 572$

$\sigma_s = 12.72$

At 9 d.f. and .05 level of significance $= 2.262$

Accepted

8-1-24

2-tailed t-test

Sample of sales in similar shops in two towns are taken for a new product with following results.

Town	Mean sales	Variance	Size sample
A	57	5.3	5
B	61	4.8	7

Is there any evidence of difference in sales in the two towns? Use 5% level of significance for testing this difference betn the mean sales.

Soln:- $H_0: \mu_1 = \mu_2$
$H_a: \mu_1 \neq \mu_2$

$\bar{x}_1 = 57$, $\sigma_1^2 = 5.3$ $n_1 = 5$
$\bar{x}_2 = 71$ $\sigma_1^2 = 4.8$ $n_2 = 7$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(n_1-1)\sigma_{s_1}^2 + (n_2-1)\sigma_{s_2}^2}{n_1+n_2-2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

d.f = 5+7-2 = 10

$= -3.5053$

10 d.f, 5% significance, 2-tailed t-test = 2.228

$|t| > 2.228$

$t = -3.053$

Reject $H_a: \mu_1 \neq \mu_2$

Select

Paired t-test

Comparing two related samples

$D_i = X_i - Y_i$, $\bar{D} = \frac{\sum D_i}{n}$

X_i → before test, Y_i → After

Variance of D

$$(\sigma_{diff})^2 = \frac{\sum D_i^2 - (\bar{D})^2 \cdot n}{n-1}$$

$$t = \frac{\bar{D} - 0}{\sigma_{diff} / \sqrt{n}}$$ → target difference is zero with $(n-1)$ d.f

Example Memory capacity of 9 students tested before & after training. State at 5% level of significance whether the training was effective?

No. Student	1	2	3	4	5	6	7	8	9
Before	10	15	9	3	7	12	16	17	9
After	12	17	8	5	6	11	18	20	3

Ans $H_0: \mu_1 = \mu_2$ i.e $H_0: \bar{D} = 0$

$H_a: \mu_1 < \mu_2$

One tailed test

Check only lower right tailed test.

$\sum D = -7$ $\sum D_i^2 = 29$

$n = 9$
d.f = 8

$\bar{D} = -7/9 = -0.778$

$\sigma_{diff} = \sqrt{\frac{\Sigma D_i^2 - (\bar{D})^2 \cdot n}{n-1}} = \sqrt{\frac{29 - (0.778)^2 \cdot 9}{9-1}}$

$= 1.715$

left-right tailed test

$t = \frac{-0.778 - 0}{1.715/\sqrt{9}} = -1.361$

~~t-test value & d.f 5%. = -1.860~~

t-test value at 8 d.f 5% = -1.86 tabulated

t test value is -1.361 (Accept region) tabulated

Hence H_0 is accepted which means no statistical difference betn before and after experiment.

Chi-square Test

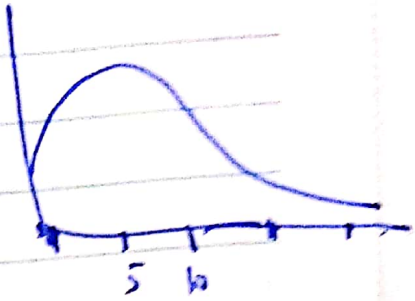
$$\chi^2 = \frac{\sigma_s^2}{\sigma_p^2}(n-1) \{ \text{or d.f} \}$$

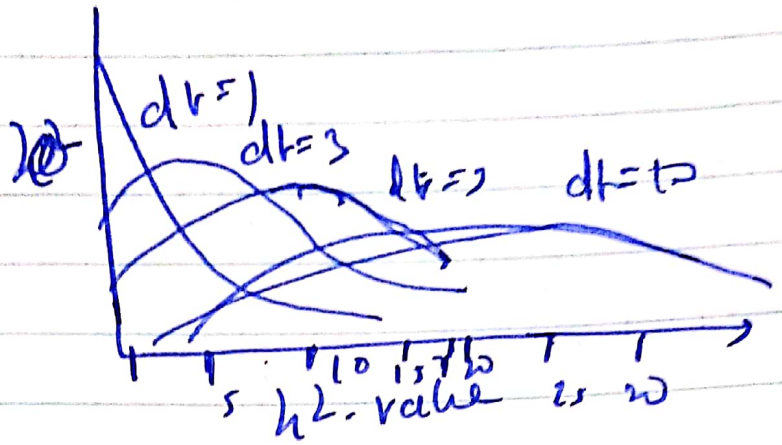
⇒ Used for comparing a variance to theoritical variance

⇒ It's a non parametric test, used to determine if categorical data shows dependency or two classifications are independent.

⇒ Used to test (i) goodness of fit (ii) significance of association betn two attributes (iii) test homogeneity or significance of population variance.

⇒ It's not symmetrical and all values are +ve.





⇒ Smaller d.f, more is the skewness of the curve.

⇒ We calculate χ^2-value based on formula and compare with tabulated value (based on d.f & level of significance). We may accept ~~or reject~~ if calculated χ^2 value is less than tabulated value

We reject null hypothesis if ~~tabulated~~ calculated value χ^2 is greater than or equal to tabulated value.

(Q) See the next table say that variance of distribution of weight of all students from which above sample of 10 students was drawn is equal to 90 lbs. Test at 5% and 1 percent significance.

S.N	1	2	3	4	5	6	7	8	9	10	
weight (kg)	35	40	45	53	47	43	55	48	52	44	$\Sigma x_i = 470$
variance	81	49	9	36	0	16	64	1	25	9	$\Sigma(x_i-\bar{x})^2 = 290$

H_0: variance $= 20$

H_a: variance $\neq 20$

$\bar{x} = \frac{\Sigma x_i}{n} = \frac{470}{10} = 47$ kg , $\Sigma(x_i-\bar{x})^2 = 290$

$\sigma_s = \sqrt{\frac{290}{9}} = \sqrt{31.11}$, $\sigma_s^2 = 31.11$

$\chi^2 = \frac{\sigma_s^2}{\sigma_p^2} \times (n-1) = \frac{31.11}{20} \times 9$

$= 13.999$

At 5% and 9 d.f χ^2-tabulated value $= 16.919$

χ^2 $13.999 < 16.919$

χ^2_{cal} $\chi^2_{tabulated}$

(Accepted)

At 1% and d.f 9 χ^2-tabulated value $= 21.666$

$13.999 < 21.666$

(Accepted)

⇒ Chi-squre as a non-parametric test is used for (i) test of goodness of fit
(ii) test of independence

How assumed distribution (Binomial/Poisson) fit to observed data

Whether or not → two attributes are related.

→ Here H_0: is two attributes are independent.

$$\boxed{\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}}$$

value of χ^2

⇒ O_{ij} = observed frequency of the cell in i^{th} row and j^{th} column

⇒ E_{ij} = Expected frequency of the cell i^{th} row & j^{th} column

⇒ degree of freedom :- 95 there are 10 ~~frequency~~ frequency classes and one independent constraint then $10-1=9$ df is.

⇒ In case of table df = $(C-1)(R-1)$
column row.

⇒ Expected ~~frequency~~ frequency of cell

$$= \frac{\text{Row total} \times \text{Column total of cell}}{\text{Grand total}}$$

Scanned with OKEN Scanner

(Q) A die is thrown 132 times with following given. Is die unbiased

No turned up	1	2	3	4	5	6	
Freq	16	27	21	14	29	25	
Expected freq	22	22	22	22	22	22	
$(O_i - E_i)^2 / E_i$	36/22	4/22	9/22	64/22	49/22	36/22	$\left(\sum \frac{(O_i - E_i)^2}{E_i}\right) = 9$

Ans

H_0: die is unbiased.

Expected value $= \frac{132}{6} = 22$

$\chi^2 = 9$

d.f $= 6-1 = 5$

At 5 d.f and 5% significance tabulated value $= 11.070 < 9$

χ^2-square

H_0 is Accepted.

(Q) Given table is data during smallpox outbreak. Check the effectiveness of vaccine.

	Attacked	Not attacked	Total
Vaccinated	(54) 31	469 (446)	500
Not vaccinated	(162) 185	1315 (1338)	1500
	216	1784	2000

$\chi^2 = 9.7962 + 13.2649 + 1.18 + 0.3958 = 24.6389$

$\chi^2 = \frac{(31-54)^2}{54} + \frac{(185-162)^2}{162} + \frac{(469-446)^2}{446} + \frac{(1315-1338)^2}{1338}$

d.f $= (2-1)\cdot(2-1) = 1$

H_0: Vaccine not effective in smallpox outbreak → from table values?

$= 14.6392$

(Q) 8 coins are tossed 256 times

No of heads	0	1	2	3	4	5	6	7	8
Freq	2	6	30	52	67	56	32	10	1

is the coin unbaised

Ans

→ 0 heads expected value $= {}^8C_0\left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^8 \times 256 = 1$

→ 1 " " " $= {}^8C_1\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^7 \times 256 = 8$

→ 2 " " " $= {}^8C_2\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^6 \times 256 = 28$

→ 3 " " " $= {}^8C_3 \times \left(\frac{1}{2}\right)^3 \times \left(\frac{1}{2}\right)^5 \times 256 = 56$

→ 4 " " " $= {}^8C_4 \times \left(\frac{1}{2}\right)^4 \times \left(\frac{1}{2}\right)^4 \times 256 = 70$

→ 5 " " " $= {}^8C_5 \times \left(\frac{1}{2}\right)^5 \times \left(\frac{1}{2}\right)^3 \times 256 = 56$

→ 6 " " " $= {}^8C_6 \times \left(\frac{1}{2}\right)^6\left(\frac{1}{2}\right)^2 \times 256 = 28$

→ 7 " " " $= {}^8C_7 \times \left(\frac{1}{2}\right)^7\left(\frac{1}{2}\right)^1 \times 256 = 8$

→ 8 " " " $= {}^8C_8 \times \left(\frac{1}{2}\right)^8\left(\frac{1}{2}\right)^0 \times 256 = 1$

calculate χ^2

d.f $= 9 - 1 = 8$

S.L. | even or not

ANOVA (Analysis of variance & co-variance)

⇒ It is used when examining the significance of more than two sample means at the same time (For 2 samples we have z-test or t-test)

⇒ One can draw inferences about whether the samples have been drawn from population having the same mean.

Ex :- Compare several varities of seeds
→ Gasoline mileage of automobiles.
→ Smoking habit of 5 groups of university students.

⇒ If we take one factor and investigate the difference amongst its various categories having numerous possible values, we say one-way ANOVA
⇒ If we investigate two-factor at a time, we call two-way ANOVA.

⇒ Basic Principle of ANOVA → Multiple samples

Examining amount of variance within given sample relative to the amount of variation betn the sample.

⇒ We assume the absence of many factors that might affect our conclusion concerning the factor to be studied by controlling other factors.

F-test, $F = \frac{\text{Estimate of population variance based on betn samples variance}}{\text{Estimate of population variance based on within sample variance}}$

Scanned with OKEN Scanner

⇒ The F value to be compared with F-limit for given degree of freedom. If F value we calculated exceeds the F-limit value, we may say there are significant difference betn the sample means.

ANOVA Technique one way

For k samples

(i) Obtain the mean of each sample $\bar{x}_1, \bar{x}_2 \ldots \bar{x}_k$

(ii) Mean of sample means $= \frac{\bar{x}_1 + \bar{x}_2 \ldots \bar{x}_k}{k}$

(iii) SS betn $= n_1(\bar{x}_1 - \bar{\bar{x}})^2 + n_2(\bar{x}_2 - \bar{\bar{x}})^2 \ldots n_k(\bar{x}_k - \bar{\bar{x}})$

Sum of square of variance

(iv) MS betn $= \frac{\text{SS betn}}{(k-1)} \rightarrow$ d.f betn samples

Mean square

(v) SS within $= \sum(x_{1i} - \bar{x}_1)^2 + \sum(x_{2i} - \bar{x}_2)^2 \ldots \sum(x_{ki} - \bar{x}_k)$

(vi) MS within $= \frac{\text{SS within}}{n-k} \rightarrow$ d.f within sample $n = n_1 + n_2 + n_k$

(vii) SS of total variance $= \sum(x_{ij} - \bar{\bar{x}})^2$ (ith sample jth data) { $n-1$ d.f

SS total = SS betn + SS within

$(n-1) = (k-1) + (n-k) \rightarrow$ ~~MS~~ ANOVA-1 relationship

(viii) F-ratio $= \frac{\text{MS betn}}{\text{MS within}}$

(iii) check table with $k-1$ d.f and $n-k$ d.f

One-way ANOVA Short cut

$n = n_1 + n_2 \cdots n_k$

k samples

(i) Take total of all individual item $T = \Sigma x_{ij}$

(ii) correction factor $= (T)^2/n$

(iii) Total SS $= \Sigma x_{ij}^2 - \frac{(T)^2}{n}$

(iv) SS betn $= \boxed{\frac{\Sigma (T_j)^2}{n_j} - \frac{(T)^2}{n}}$ $j = 1, 2, 3$

Total value of each sample

(v) SS within = Total SS − SS betn

$= \Sigma x_{ij}^2 - \frac{(T)^2}{n} - \frac{\Sigma (T_j)^2}{n_j} + \frac{(T)^2}{n}$

$\boxed{= \Sigma x_{ij}^2 - \frac{\Sigma (T_j)^2}{n_j}}$

(Q) Plot land per acre production data

	variety of wheat		
	A	B	C
1	6	5	5
2	7	5	4
3	3	3	3
4	8	7	4

3 variety of wheat grown at 4 plots and state if the varity ~~of wheat~~ difference in wheat is significant. Check at 5% level of significance

We have 3 varieties, sample size $k=3$

$\bar{x}_1 = \frac{6+7+3+8}{4} = 6$ $\quad n_1 = 4$

$\bar{x}_2 = \frac{5+5+3+7}{4} = 5$ $\quad n_2 = 4$

$\bar{x}_3 = \frac{5+4+3+4}{4} = 4$ $\quad n_3 = 4$

$n = 12$, $k = 3$

$\bar{\bar{x}} = \frac{6+5+4}{k} = \frac{6+5+3}{3} = 5$

SS betn $= n_1(\bar{x}_1 - \bar{\bar{x}})^2 + n_2(\bar{x}_2 - \bar{\bar{x}})^2 + n_3(\bar{x}_3 - \bar{\bar{x}})^2$

$= 4(6-5)^2 + 4(5-5)^2 + 4(4-5)^2$

$= 8$

SS within $= \sum(x_{1i} - \bar{x}_1)^2 + \sum(x_{2i} - \bar{x}_2)^2 + \sum(x_{3i} - \bar{x}_3)^2$

$= 24$

MS betn $= \frac{8}{k-1} = \frac{8}{3-1} = 4$

MS within $= \frac{24}{12-3} = \frac{24}{9} = 2.67$

F ratio $= \frac{4}{\frac{24}{9}} = \frac{4 \times 9}{24} = 1.5$

At 5% level of significance F-limit =

d.f $= 9, 2$ $= 4.26$ - Hence within limit and we accept H_0,

i.e the difference is insignificant and 3 varieties are similar.

TWO-Way Anova → Two factors

Example of two way Anova

The agricultural o/p may be classified on the basis of different varities of seeds and also on the basis of different varity of fertilizers used.

Two ways of Two-way ANOVA

~~(i) ANOVA two way design without repeated values; As there is no repeated values we cannot calculate~~

Two-Way ANOVA

No repeated values

(i) Take total value of individual items (T)

(ii) Correction factor $= (T)^2/n$

(iii) ~~Find square~~ Calculate total sum of ~~varia~~ squares or variance or total SS $= \sum x_{ij}^2 - \frac{(T)^2}{n}$

$\sum \frac{T_j^2}{n_j}$

(iv) Take total of different columns, obtain the square of each column total and divide by no. of items in the column. Take the total of result obtained and ~~$T_j = \sum x_{ij}$ where i = 1, 2 ...~~ subtract correction factor from this total to obtain the sum of squares of deviation for variance betn ~~rows~~ column (or SS betn ~~rows~~) (columns)

$\sum \frac{(T_j)^2}{n_j}$

(v) Similarly calculate SS betn ~~columns~~ rows.

(vi) SS for residual or (error variance)

= Total SS − (SS betn columns + SS betn rows)

(vii) d.f for total variance = c.r − 1

d.f variance betn columns = c − 1

" " " " rows = r − 1

Residual variance = (c−1)(r−1)

Scanned with OKEN Scanner

⇒ MS residual or residual variance provide basis for F-ratio

→ F-ratio concerning column and row calculated with MS-within, MS-treatment, use residual-error as denominator.

Example

Set up an Analysis of variance table for the following two-way design results:

Per Acre Production Data of wheat

Varieties of seeds / Varieties of fertilizers	A	B	C	
N	6	5	5	16
P	7	5	4	16
Y	3	3	3	9
Z	8	7	4	19
	24	20	16	

→ $T = 60$, $n = 12$ correction factor $= \frac{(T)^2}{n} = \frac{60 \times 60}{12} = 300$.

→ Total SS $= (6)^2 + (7)^2 + 3^2 + 8^2 + 5^2 + 5^2 + 3^2 + 7^2 + 5^2 + 4^2 + 3^2 + 4^2 - 300$

$= 332 - 300 = 32$

→ SS betn columns $= \left[\frac{24 \times 24}{4} + \frac{20 \times 20}{4} + \frac{16 \times 16}{4}\right] - 300$

$= 8$

→ SS betn rows $= \left[\frac{16 \times 16}{3} + \frac{16 \times 16}{3} + \frac{9 \times 9}{3} \times \frac{19 \times 19}{3}\right] - 300$

$= 18$

→ SS residual error $=$ Total SS $-$ (SS betn columns + SS betn rows)

$= 32 - (8 + 18) = 6$

$V_1 = 1, V_2 = 2$

Source of variation	SS	d.f	MS	F-ratio	5% limit for table
Betn columns	8	3-1=2	8/2=4	4/1=4	F(2,1)=5.4
Betn rows	18	4-1=3	18/3=6	6/1=6	F(3,6)=4.76
Residual/error	6	(3-1)×(4-1)=6	6/6=1		
	32	3×4-1=11			

Note
y residual MS
error MS + calculate
F-ratio

betn rows
not within
limit

Seed dont
have effect

Fertilizer
has an effect

Linear Regression

We have training set $x = \{x^1 \dots x^m\}$
with $y = \{y^1 \dots y^m\}$ o/p

training data $\rightarrow$ [ML Algo] $\rightarrow$ hypothesis.

⇒ In regression the o/p is a real value in comparision with classification where values are catagorical.

Example

Try to predict the c.g.p student from his/her reading per day.
So hypothesis can be written as

$h(x) = \theta_0 + \theta_1 x$ — linear function / Affine function.

x → days per hour.

⇒ If more than one attributes, let's say d attributes

$$h(x) = \sum_{i=0}^{d} \theta_i x_i$$

where x_0 is dummy feature $= 1$

$$\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_d \end{bmatrix}, \quad x = \begin{bmatrix} x_0 \\ x_1 \\ x_d \end{bmatrix}$$

the vectors are developed column vectors.

θ = parameter weight, m = no. of training examples
d = dimensions, x = i/p vector. y = o/p or target vector.
(x^i, y^i) - one training example.

→ with dummy features, we have d+1 dimensions.

⇒ Now the target is to choose θ, so that $h_\theta(x) \approx y$ for training examples.

Mean of hypothesis expected value

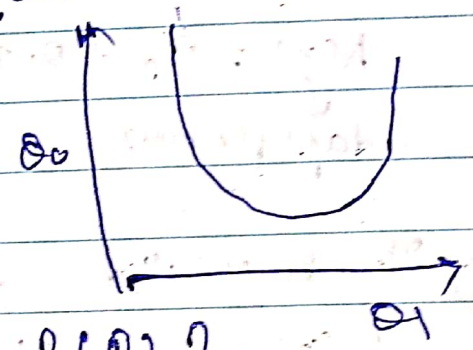
⇒ Now we define error $E(\theta)$ w.r.t the training data w.r.t θ as

$$E(\theta) = \frac{1}{2} \sum_{i=1}^{m} \left(h_\theta(x^i) - y^i\right)^2$$

known as squared error.

Minimize w.r.t to θ

⇒ the Error function $E(\theta)$ is bowl shaped with a single minimum value, and no local minima.



How to find θ to Minimize $E(\theta)$?

Gradient descent Algorithm

→ start with $\theta^0 = \langle 0, 0, 0, 0 \ldots \rangle$

→ update $\theta_j = \theta_j - \alpha \left(\frac{\partial E(\theta)}{\partial \theta_j} \right)$

$\alpha \left(\frac{\partial E(\theta)}{\partial \theta_j} \right)$

jth component of θ

α is learning component set to 0.01

$$\frac{\partial E(\theta)}{\partial \theta_j} = \frac{\partial}{\partial \theta_j} \cdot \frac{1}{2}\left[(h_\theta(x) - y)^2\right]$$

$$= \frac{1}{2} \times \frac{\partial [h_\theta(x) - y]^2}{\partial [h_\theta(x) - y]} \times \frac{\partial [h_\theta(x) - y]}{\partial \theta}$$

$$= [h_\theta(x) - y] \cdot \frac{\partial (\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 \cdots - y)}{\partial \theta_j}$$

$$= \underbrace{[h_\theta(x) - y]}_{\text{loss}} \cdot x_j$$

$\hookrightarrow$ jth component of x

$$\boxed{\theta_j = \theta_j - \alpha (h_\theta(x) - y) \cdot x_j}$$

$$\hat{\theta}_1 = \frac{\sum_{i=1}^{m} (x^i - \bar{x})(y^i - \bar{y})}{\sum_{i=1}^{m} (x^i - \bar{x})^2}$$

$$\hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x}$$

This only works with only one variable or attribute ordinary least square error.

For Batch Gradient descent

$$\theta_j = \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^i) \cdot x_j^i$$

$\rightarrow$ jth component or ith training set x^i

This method is done for $j = 0, 1 \ldots d$

Repeat till converge.

Disadvantage of Batch Gradient descent -

(i) For larger data set - take time to converge (slowly converges)

Alternative Stochastic Gradient descent

For larger data set -

For all training examples

$$\theta_j = \theta_j - \alpha (h_\theta(x^i) - y^i) \cdot x_j^i$$

$\Rightarrow$ Generally fluctuating θ value

$\Rightarrow$ Gives Approximate values.

Scanned with OKEN Scanner

Why least square Error?

Let $y^i = \theta^T x^i + \epsilon^i$ → error term or random noise

$\boxed{Y = X\theta + \epsilon}$ → For particular or unmeasured value attributes

For whole

$\theta^T = [\theta_0, \theta_1, \theta_2 \cdots \theta_d]$

$x_i = [x_0^i + x_1^i + x_2^i \cdots x_d^i]$

Assume ϵ^i is normally distributed $N(0, \sigma^2)$

$$P(\epsilon^i) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{\left[-\frac{\epsilon^{i2}}{2\sigma^2}\right]}$$

↓ prob. density function

$f(x)$ in Normal distribution $\frac{1}{\sqrt{2\pi}\,\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$

Here ϵ^i are indentiperdly & identically distributed iid

Now

$$P(y^i \mid x^i; \theta) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{\left[\frac{-(y_i - \theta^T x_i)^2}{2\sigma^2}\right]}$$

prob of y^i given x^i and θ

↳ we assume variance is due to Noise.

Given x^i, θ y^i is Normally distributed because of added noise which is following standard normal distribution. Otherwise x^i, θ would always be $y^i = \theta^T x^i$

Here $\underbrace{y^i | x^i ; \theta}_{\text{random variable}}$ Normally distributed $\sim N(\theta^T x^i, \sigma^2)$

$\boxed{L(\theta)}$ $= P(y | x, \theta)$

likelihood of θ — All examples

we need to maximize $L(\theta)$

$$= \prod_{i=1}^{m} P(y^i | x^i ; \theta)$$

Take log both sides.

$$l(\theta) = \log \prod_{i=1}^{m} P(y^i | x^i, \theta)$$

$$= \sum_{i=1}^{m} \log P(y^i | x^i ; \theta)$$

$$= \sum_{i=1}^{m} \log \left[\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{[y^i - \theta^T x^i]^2}{2\sigma^2}} \right]$$

$$= \sum_{i=1}^{m} \log \frac{1}{\sqrt{2\pi}\sigma} + \sum_{i=1}^{m} -\frac{[y^i - \theta^T x^i]^2}{2\sigma^2}$$

cnst

To maximize $l(\theta)$ we need to maximize

$\sum_{i=1}^{m} -(y^i - \theta^T x^i)^2$ which means

minimize $\frac{1}{2} \sum_{i=1}^{m} (y^i - \theta^T x^i)^2$ which is same as least squared error

$J(\theta)$

Matrix differentiation

$\Delta_x(x^T y) = \nabla(y^T x) = y$ — (1)

$\nabla_x(x^T x) = 2x$ — (2)

$\nabla_x(x^T A y) = Ay$ — (3)

$\nabla_x(y^T A x) = A^T y$ — (4)

$\nabla(x^T A x) = (A + A^T)x \,/\, 2Ax$ — (5)

$\Rightarrow 2Ax$ symmetric

Gradient of $b(x)$ w.r.t x if x is a vector $(x_1, x_2, \ldots x_d)$

$\nabla_x b(x) = 0$

Rough

$\nabla_x b(x) = \begin{pmatrix} \frac{\partial b(x)}{\partial x_1} \\ \frac{\partial b(x)}{\partial x_2} \\ \frac{\partial b(x)}{\partial x_d} \end{pmatrix}$

scalar, 1st vector

$d \times m$

$1 \times m$

$m \times d$

$\text{let } (y^T x)$

$a = a^T x^T y$

$x^T y$

$\frac{\partial (x^T y)^T a}{\partial a} = x^T y$

$= 2x^T y$

Least square Linear equations

Solving with Normal form

$$\underset{m\times d}{X}\ \underset{d\times 1}{\theta} = \underset{1\times d}{Y}$$

$$\left[\underset{1\times 1}{\theta^T x}\right]$$

For individual examples

$$\left[y^i = \theta^T x^i + e^i\right]$$

where $X = \begin{bmatrix} 1 & x_1 & x_2 & x_3 & \cdots & x_d \\ 1 & & & & & \\ \vdots & & & & & \\ 1 & & & & & \\ 1 & x_2 & \cdots & & & x_d \end{bmatrix}_{m\times d}$ ($x_0 = 1$) $* \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_{d-1} \end{bmatrix}_{d\times 1} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}_{m\times 1}$

To uniquely identify θ, we need $m \geq d$. If X is square & non-singular we can solve by $\theta = X^{-1} Y$ $\{m == d\}$

$\Rightarrow$ Due to noise $m == d$ examples may not be exact tolerant. So we need $m > d$. Assume X is non sing.

$$X\theta + e = y \qquad \text{where } e = \begin{bmatrix} e^1 \\ \vdots \\ e^m \end{bmatrix}_{m\times 1}$$

$\Rightarrow$ We want to search for $\hat{\theta}$ to minimize

$$E(\theta) = \sum_{i=1}^{m} (y^i - \theta^T x^i)^2 = e^T e = (Y - X\theta)^T (Y - X\theta)$$

$\theta = \hat{\theta}$ minimum, when $\frac{\partial E(\theta)}{\partial \theta} = 0$

$$\Rightarrow \frac{\partial (Y - X\theta)^T (Y - X\theta)}{\partial \theta} = \frac{\partial (Y^T - \theta^T X^T)(Y - X\theta)}{\partial \theta}$$

$$= \frac{\partial (Y^T Y - Y^T X\theta - \theta^T X^T Y + \theta^T X^T X \theta)}{\partial \theta}$$

$$= \underbrace{\frac{\partial Y^T Y}{\partial \theta}}_{0} - \underbrace{\frac{\partial Y^T X\theta}{\partial \theta}}_{-X^T Y} - \underbrace{\frac{\partial \theta^T X^T Y}{\partial \theta}}_{-X^T Y} + \frac{\partial \theta^T (X^T X)\theta}{\partial \theta}$$

rules 3 rules 2 rules

symmetric $(m \times m)$

$= X^T X$

$$\Rightarrow -2X^T Y + \underline{2X^T X \theta} = 0$$

$$\Rightarrow X^T X \hat{\theta} = X^T Y$$

If $X^T X$ is non-singular, then $\hat{\theta}$ can be solved uniquely

$$\hat{\theta} = (X^T X)^{-1} X^T Y$$

$$E_2(\hat{\theta}) = (Y - X\hat{\theta})^T (Y - X\hat{\theta})$$

$$= Y^T Y - Y^T X (X^T X) X^T Y$$

$\Rightarrow$ If $X^T X$ is singular then LSE is not unique. ~~But we assume X^T to be non-sing~~.

Example

Q)

Force (Newton)	Length of spring (inches)
1.1	1.5
1.9	2.1
3.2	2.5
4.4	3.3
5.9	4.1
7.4	4.6
9.2	5.0

take $l = \theta_0 + \theta_1 n$. find θ_0 and θ_1

$$X = \begin{bmatrix} 1 & 1.1 \\ 1 & 1.9 \\ 1 & 3.2 \\ 1 & 4.4 \\ 1 & 5.9 \\ 1 & 7.4 \\ 1 & 9.2 \end{bmatrix} * \underbrace{\begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix}}_{\theta} + \begin{bmatrix} e^1 \\ e^2 \\ \vdots \\ e^m = 7 \end{bmatrix} = \begin{bmatrix} 1.5 \\ 2.1 \\ 2.5 \\ 3.3 \\ 4.1 \\ 4.6 \\ 5.0 \end{bmatrix}$$

Scanned with OKEN Scanner

$$\boxed{\hat{\theta} = (X^T X)^{-1} X^T Y} = \begin{bmatrix} 1.2 \\ 0.44 \end{bmatrix}$$

If the target has q O/ps

$$\underset{m \times d}{X} \; \underset{d \times q}{\theta} + \underset{m \times q}{E} = \underset{(m \times q)}{Y}$$

$$E(\theta) = \sum_{i=1}^{m} \sum_{j=1}^{q} e_{ij}^2 \longrightarrow \text{error for } i^{th} \text{ examples } j^{th} \text{ output}$$

$$= \sum_{j=1}^{q} \underbrace{\left(\sum_{i=1}^{m} e_{ij}^2 \right)}_{\text{squared error of } j^{th} \text{ column or error for } j^{th} \text{ attributes}}$$

Hence $\min_{\theta} \sum_{j=1}^{q} \left(\sum_{i=1}^{m} e_{ij} \right)^2 = \sum_{j=1}^{q} \left(\min_{\theta_j} \sum_{i=1}^{m} e_{ij} \right)^2$

$\Rightarrow$ θ_j is the jth column of θ. Now we can calculate $\theta_j = \hat{\theta}_j$ which minimizes $\sum_{i=1}^{m} e_{ij}^2$

Each column responsible for one output

Now its a LSE subproblem

$$X\theta_j + e_j = Y_j \rightarrow j^{th} \text{ column vector}$$

Each θ_j can be uniquely calculated.

$$\underset{d \times 1}{\hat{\theta}_j} = (X^T X)^{-1} X^T Y_j$$

$$\underset{d \times q}{\hat{\theta}} = (X^T X)^{-1} X^T \underset{q \times m}{Y}$$

Geometric Interpretation of LSE

because of any metric

⇒ $X\theta$ is a liner combination basis vector of d-dimensional space.

$\langle a_1 \cdots a_d \rangle$

$$X\theta = \underbrace{\begin{bmatrix} a_1 & \cdots & a_d \end{bmatrix}}_{1\times d} \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_d \end{bmatrix}_{d\times 1} = \theta_1 [a_1] + \cdots \theta_d [a_d]$$

a_1 → $m\times 1$, a_d → $m\times 1$

each a_i is a vector of size $m\times 1$

$$a_i = \begin{bmatrix} x_1^i \\ x_2^i \\ \vdots \\ x_m^i \end{bmatrix}_{m\times 1}$$

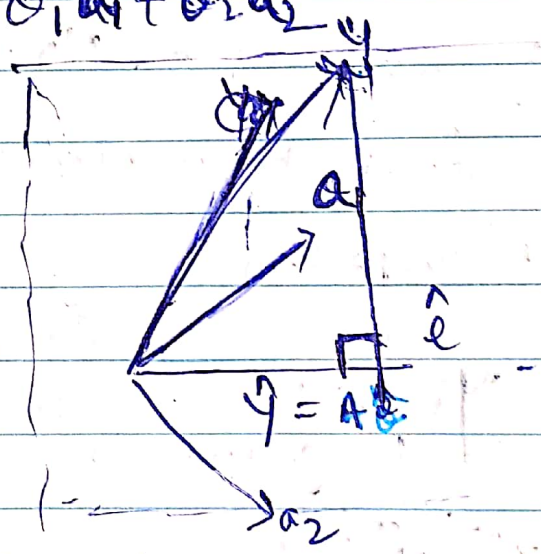
~~$X\theta$ is projection~~ $X\theta$ is a liner combination of basis vector $[a_1 \cdots a_d]$ is an m-dimensional space.

$X\theta$ is approximately y in

$X\theta$ is projection of y onto the space spanned by $\langle a_1 \cdots a_d \rangle$

$A = \{a_1, \ldots\}$

eg $X\theta = \theta_1 a_1 + \theta_2 a_2$



We can say projection of y on to space of a_1, a_2 the Min^m possible error $\perp$ to $\hat{y}$

$e = y - A\hat{\theta}$ is $\perp r$ to both a_1 & a_2 which is called principle of orthogonality.

$$\begin{cases} a_1^T (y - A\hat{\theta}) = 0 \\ a_2^T (y - A\hat{\theta}) = 0 \end{cases}$$

$$\text{Proj}_A(y) = \underline{A\hat{\theta}} = \{A = P\}$$

projection y to space spanned by A

predicted

$\hat{y} = A\hat{\theta}$ orthogonal to error $\hat{e}$

From Thakkra for Notes

Not Required

Simple linear Regression

$$y = \underbrace{\overset{\theta_0}{a} + \overset{\theta_1}{b}x}_{\text{predicted value}} + e$$

For points are (y_1, x_1) $(y_2$

$(y^1, x^1), (y^2, x^2) \ldots (y^m, x^m)$

$$y^{(i)} = \overset{\theta_0}{a} + \overset{\theta_1}{b}x^{(i)} + e^i$$

Sum of squares of error

$$= E = \sum_{i=1}^{m} e^{i^2} = \sum_{i=1}^{m} (y^i - \underbrace{\theta_0 - \theta_1 x^i}_{\text{predicted value}})^2$$

$$\Rightarrow \frac{\partial E}{\partial \theta_0} = \frac{\partial}{\partial \theta_0} \sum_{i=1}^{m} (y^i - \theta_0 - \theta_1 x^i)^2 = 0$$

for min E

$$\frac{\partial}{\partial \theta_0} \sum_{i=1}^{m}$$

$$= -2 \sum_{i=1}^{m} (y^i - \theta_0 - \theta_1 x^i) = 0 \quad (1)$$

similarly

$$\frac{\partial E}{\partial \theta_1} = -2 \sum_{i=1}^{m} (y^i - \theta_0 - \theta_1 x^i) \cdot x^i = 0 \quad (2)$$

Scanned with OKEN Scanner

Logistic Regression (Generally used for classification problem)

We want our $h_\theta(x)$ to betn $[0, 1]$

We are going to choose

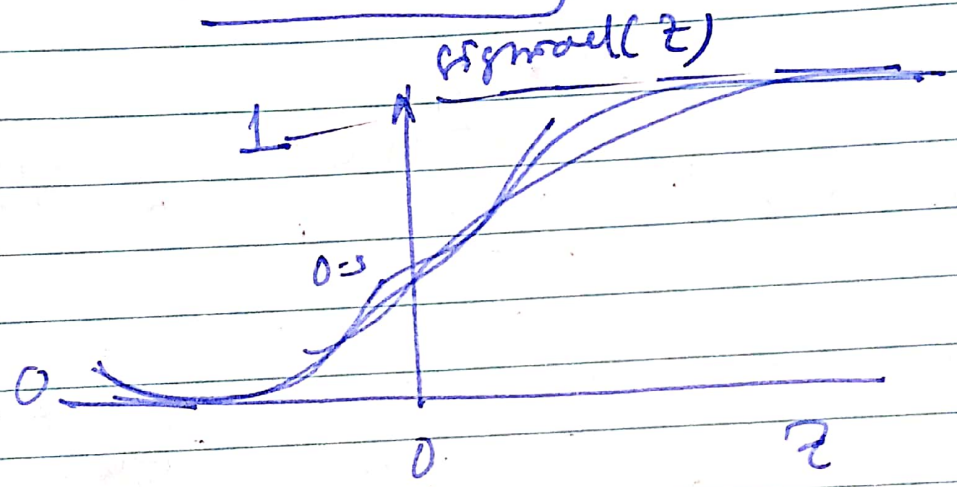
$$h_\theta(x) = \text{sigmoid}(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}} \rightarrow z$$

logistic

$$= \text{sigmoid}(z) = \frac{1}{1 + e^{-z}}$$

Where

$$\boxed{z = \theta^T x}$$



This value is 0.5 at $z = 0$. It's gives value betn 0 to 1. We want to give the value zero for one type of instance and 1 to other type.

Let's discuss only for binary classification.

$$P(y=1 \mid x, \theta) = h_\theta(x)$$ → prob. of belong to $y=1$

$$P(y=0 \mid x, \theta) = 1 - h_\theta(x)$$

Now combining both

$$P(y \mid x, \theta) = h_\theta(x)^{y} * (1 - h_\theta(x))^{1-y}$$

If $y=1$ then

$$P(y \mid x, \theta) = h_\theta(x)$$

If $y=0$

$$P(y \mid x, \theta) = [1 - h_\theta(x)]^{1-y}$$

Our aim is to maximize the likelihood (or prob) of $P(y \mid x; \theta)$ w.r.t θ.

$$L(\theta) = \cancel{\text{argmax}}_{\theta} \; \text{argmax}_{\theta} P(y \mid x; \theta)$$

likelihood of θ

$$= \prod_{i=1}^{m} P(y^i \mid x^i, \theta)$$

$$= \prod_{i=1}^{m} (h_\theta(x^i))^{y^i} (1 - h_\theta(x^i))^{1-y^i}$$

Take log both sides

$$\text{argmax}_{\theta} \, l(\theta) = \sum_{i=1}^{m} y^i \log h_\theta(x^i) + (1-y^i) \cdot \log (1 - h_\theta(x^i))$$

Take θ to maximize $l(\theta)$ to maximize probability which is to take partial derivative w.r.t θ_j which is slope of the likelihood function

Gradient ascent algorithm

$$\theta_j = \theta_j + \gamma \frac{\partial l(\theta)}{\partial \theta_j}$$

$$\cancel{\frac{\partial l(\theta)}{\partial \theta_j} = \sum_{i=1}^{m} y^i \cdot \frac{\partial \log e^{h(x)}}{\partial \theta_j}}$$

$$\boxed{\frac{\partial l(\theta)}{\partial \theta_j} = \sum_{i=1}^{m} (y^i - h_\theta(x^i))\, x_j^i}$$

$\Rightarrow$ see the last page for proof

$$\theta_j = \theta_j + \alpha \sum_{i=1}^{m} y^i - h_\theta(x^i)\, x_j^i$$

$$\gamma \mid \gamma \frac{\partial}{\partial x} = -\frac{1}{2}$$

$z = \theta^T x = \theta_0 + \theta_1 x_1 + \theta_2 x_2$

Proof of logistic Regression

Prob. of seeing $y=1$

$$P(Y=1 \mid x, \theta) = h_\theta(x) = \frac{1}{1+e^{-\theta^T x}} = \frac{1}{1+e^{-z}}$$

$$P(Y=0 \mid x, \theta) = 1 - h_\theta(x)$$

Combining both

$$P(Y \mid x, \theta) = h_\theta(x)^{y} * (1-h_\theta(x))^{1-y}$$

For all examples

$$L(\theta) = P(Y \mid x, \theta) = \prod_{i=1}^{m} h_\theta(x^i)^{y^i} * (1-h_\theta(x^i))^{1-y^i}$$

Taking log both sides

$$l(\theta) = \sum_{i=1}^{m} y^i \cdot \log h_\theta(x^i) + (1-y^i) \log(1-\log h_\theta(x^i))$$

$$= \sum_{i=1}^{m} y^i \log \frac{1}{1+e^{-z}} + (1-y^i) \log\left(1-\frac{1}{1+e^{-z}}\right)$$

Maximize

$$\frac{\partial l_\theta}{\partial \theta_j} = \sum_{i=1}^{m} y^i \frac{\partial \log \frac{1}{1+e^{-z}}}{\partial \theta_j} + (1-y^i) \frac{\partial \log \frac{1}{1+e^{-z}}}{\partial \theta_j}$$

$$= \sum_{i=1}^{m} y^i \frac{\partial \log \frac{1}{1+e^{-z}}}{\partial 1+e^{-z}} * \frac{\partial 1+e^{-z}}{\partial z} * \frac{\partial z}{\partial \theta_j} + (1-y^i) \frac{\partial \log \frac{1}{1+e^{-z}}}{\partial 1+e^{-z}} * \frac{\partial (1+e^{-z})}{\partial z} * \frac{\partial z}{\partial \theta_j}$$

$$= \sum_{i=1}^{m} y^i x_j^i \frac{e^{-z}}{1+e^{-z}} + \frac{e^{z}}{1+e^{z}} x_j^i - x_j^i y^i * \frac{e^{z}}{1+e^{z}}$$

$$= \sum_{i=1}^{m} x_j^i y^i \left(\frac{e^{-z}}{1+e^{-z}} + \frac{e^{z}}{1+e^{z}} \right) - \frac{e^{z}}{1+e^{z}} x_j^i$$

$$= \sum_{i=1}^{m} y^i x_j^i - \frac{1}{1+e^{-z}} x_j^i$$

$$= \sum_{i=1}^{m} \left(y^i - \frac{1}{1+e^{-z}} \right) x_j^i$$

$$\boxed{= \sum_{i=1}^{m} \left[y^i - h_\theta(x^i) \right] x_j^i}$$

proved

Scanned with OKEN Scanner

Model diagnostics

⇒ It's important to validate the regression model to ensure its validity and goodness of fit before it can be used for practical applications. Following are used to validate the simple linear regression models.

1 - Coefficient of determination (R-squared)
2 - Hypothesis test for Regression Coefficient
3 - Analysis of variance for overall model validity (important for multiple linear regression)
4 - Residual Analysis to validate the regression model assumptions.
5 - Outlier Analysis, since the presence of outliers can significantly impact regression parameters.

(1) Coefficient of determination (R-squared)

Primary objective of regression is to explain the variation in y using knowledge of x. The R^2 measures the percentage of variation in y explained by the model $\beta_0 + \beta_1 x$

$$\underbrace{y_i}_{\text{variation in } y} = \underbrace{\beta_0 + \beta_1 x}_{\text{variation in } y \text{ explained by the model}} + \underbrace{\varepsilon_i}_{\text{variation of } y \text{ not explained by the model}}$$

Sum of square in variation y Total

$$\underbrace{\sum_{i=1}^{n}(y_i-\bar{y})^2}_{SST} = \underbrace{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}_{SSR} + \underbrace{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}_{SSE}$$

predicted

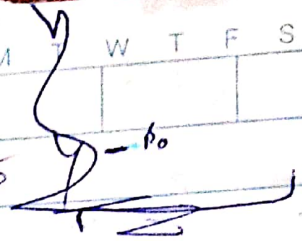
$$\boxed{\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i}$$

SSR → sum of squared variations explained by Regression

SSE → sum of squared variations ~~explained~~ ~~between actual~~

square of ~~coor~~ pearson coefficient

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$



lies betn 0 & 1

⇒ Higher co-relation coefficient is better fit

8) Hypothesis Test for Regression Co-efficient

$H_0: \beta_1 = 0$

$H_A: \beta_1 \neq 0$

For simple regression co-efficient (β_1) captures the existence of linear relⁿ betⁿ variable and feature. If $\beta_1 = 0$, we can conclude there is no statistically significant relⁿ betⁿ two variables.

$H_0: \beta_1 = 0$ $(\mu = 0)$

$H_A: \beta_1 \neq 0$

$$t_{\alpha/2, n-2} = \frac{\hat{\beta}_1 - 0}{se(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{\sqrt{\frac{(y_i - \hat{y}_i)}{n-2}} \, (x_i - \bar{x})}$$

standard error of coefficient

Standard error of estimate of regression coefficient is given

$$se(\hat{\beta}_1) = \frac{se}{\sum \sqrt{(x_i - \bar{x})^2}}$$

Where se is the standard error of the estimated value of y_i (and the residuals) are given by.

$$se = \sqrt{\frac{\sum (y_i - \hat{y}_i)}{n-2}}$$

⇒ Hypothesis test is a 2-tailed test - is a t-distribution with $n-2$ degrees of freedom. (T.

⇒ $Se(\hat{\beta}_1)$ is standard error of regression co-efficient $\hat{\beta}_1$.

Analysis of variance (ANOVA) in Regression Analysis

Null Hypothesis $H_0: \beta_1 = \beta_2 \ldots \beta_k = 0$

H_A: Not all regression coefficients are zero

Mean squared regression

$$F = \frac{MSR}{MSE} = \frac{SSR/k}{SSE/(n-k-1)}$$

for k degree

MSE → Mean squared error

F-test $(k, n-k-1)$:- Whether overall Regression model is statistically significant or not.

Note:- In Simple Linear Regression (SLR) p-value for t-test and F-test will be the same since the null hypothesis is same.

Residual Analysis

Residual error are diff^n betw^n the actual value of the outcome variable and predicted value $(y_i - \hat{y}_i)$

It checks if

1- Residuals are normally distributed.
2- Variance of residual is constant
3- The functional form of regression is correctly specified.
(4) No outliers.

Scanned with OKEN Scanner

Normal distribution of Residual

→ plot probability-probability plot (PP plot)

→ It compares the cumulative distribution function of two probability distributions against each other.

Ex

import matplotlib.pyplot as plt.

import seaborn as sn

mba-salary-resid = mba-salary-lm.resid

Probplot=

Hypothesis Test for the Regression Co-efficient

→ Regression coefficient (β_1) captures the existence of a linear relationship betn the outcome variable and the feature. If $\beta_1 = 0$ no statistically significant linear reln exists.

$H_0: \beta_1 = 0$

$H_A: \beta_1 \neq 0$

$t_{\alpha/2, n-2} = \frac{\hat{\beta}_1}{Se(\hat{\beta}_1)}$

$\hat{\beta}_1$ → estimated value of β_1

$Se(\hat{\beta}_1)$ → standard error of estimation

significance level

$$Se(\hat{\beta}_1) = \sqrt{\frac{n\hat{\sigma}^2}{n\sum x_i^2 - (\sum x_i)^2}}$$

$$Se(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{\sum_i (x_i - \bar{x})^2}}$$

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_i e_i^2$$

$$\boxed{e_i = y_i - (\beta_0 + \beta_1 x)}$$

Mean square error

Compare with n-2 degree of freedom.

Outlier Analysis

⇒ Presence of outliers can have significant influence on values of regression coefficients.

Z-score can be used to find outliers in data

$$Z = \frac{y_i - \bar{y}}{\sigma_y}$$

σ_y → ~~value of y~~ S.D of y

∴ Z score greater than 3 may be flagged as outlier.

Making Predictions and Measuring Accuracy

⇒ Done on test (validation) data.

Finding R^2 and RMSE

→ Mean square Error (MSE) $= \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$ → lower the better

→ Root Mean Square Error (RMSE) → lower the better

$$= \sqrt{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

⇒ Mean Absolute percentage Error (MAPE)

$$= \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i}$$

→ Less than 10% good model

10-20% A good model

More than 50% means bad model

Multi-collinearity and Handling Multi-collinearity

⇒ Few input variables may be highly correlated. Existance of a high correlation betn independent variables is called Multi-collinearity.

⇒ Multicolinearity can destabilize regression Model

Variance Inflation Factor (VIF)

~~$y_i = \beta_0 + \beta_1 x$~~

X_1, X_2 independent variables

$X_1 = \alpha_0 + \alpha_1 X_2$

Let R_{12} be R-squared value of this model

$$VIF = \frac{1}{1 - R_{12}^2}$$

VIF > 4 need to investigate ~~Multicolinearity~~ Multicollinearity.

Scanned with OKEN Scanner

Polynomial curve fitting

Let $y(x, w) = w_0 + w_1 x + w_2 x^2$

$= \theta_0 + \theta_1 x + w_2 x^2 \ldots = \sum_{j=0}^{M} \theta_j x^j$

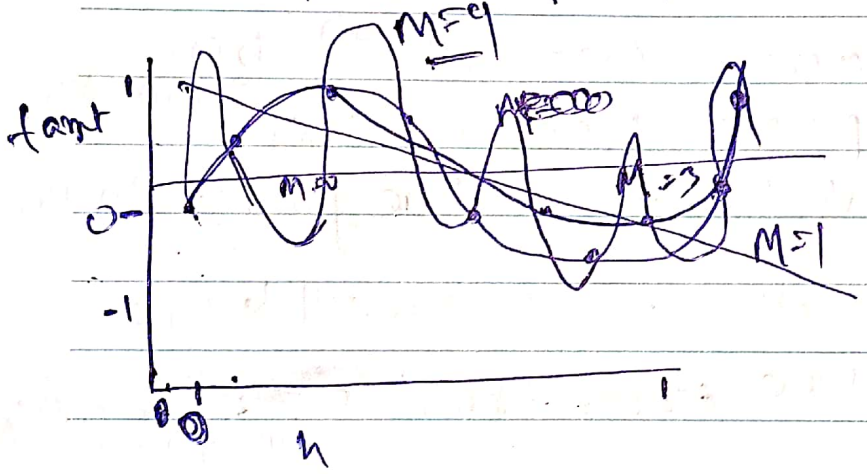
where M = order of polynomial x^j is raised to power j.

$E_n(\theta)$ error function $= \frac{1}{2} \sum_{n=1}^{N} [y(x_n, w) - t_n]^2$

~~Curve fitting~~

Let we have 10 training data point for a function $f(x) = \sin(2\pi x)$ + Add some Gaussian Noise

Take M=0 M=1, M=3, M=9 cases.



Let $E(\theta^*)$ be the Min^m error with a given θ^* values for a given model with order M.

How to compare which model is best?

→ M=0, M=1 have large error (underfit)
High training and testing error

→ M=3 is perfect fit

→ M=9 oscilates wildly even the $E(w^*)=0$ for training data also it matches exactly with all 10 points known as overfitting
⇒ It results in low training error but high testing error.

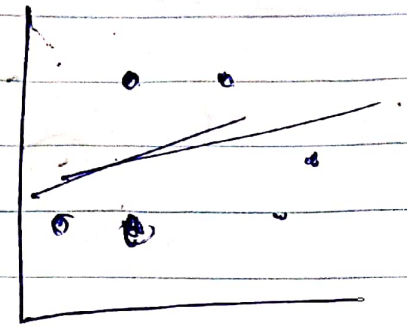
Aim of the Model selection
→ Make training error small ⇒ to choose best parameters
→ Make gap betn training and testing error less to (Reduce generalization gap)

⇒ Some times its more useful to use root mean square error (RMS) $E_{RMS} = \sqrt{2E(w^*)/N}$

This allows us to compare different data sets with a value diffn

⇒ Bias = $E(\hat{y}) - y$
↓
expect value of estimator $\hat{y}$

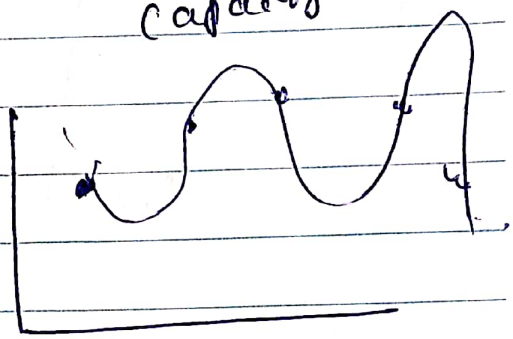
Variance :- Amount by which the performance of a predictive model changes when trained on different subsets of data.



Y o underfit

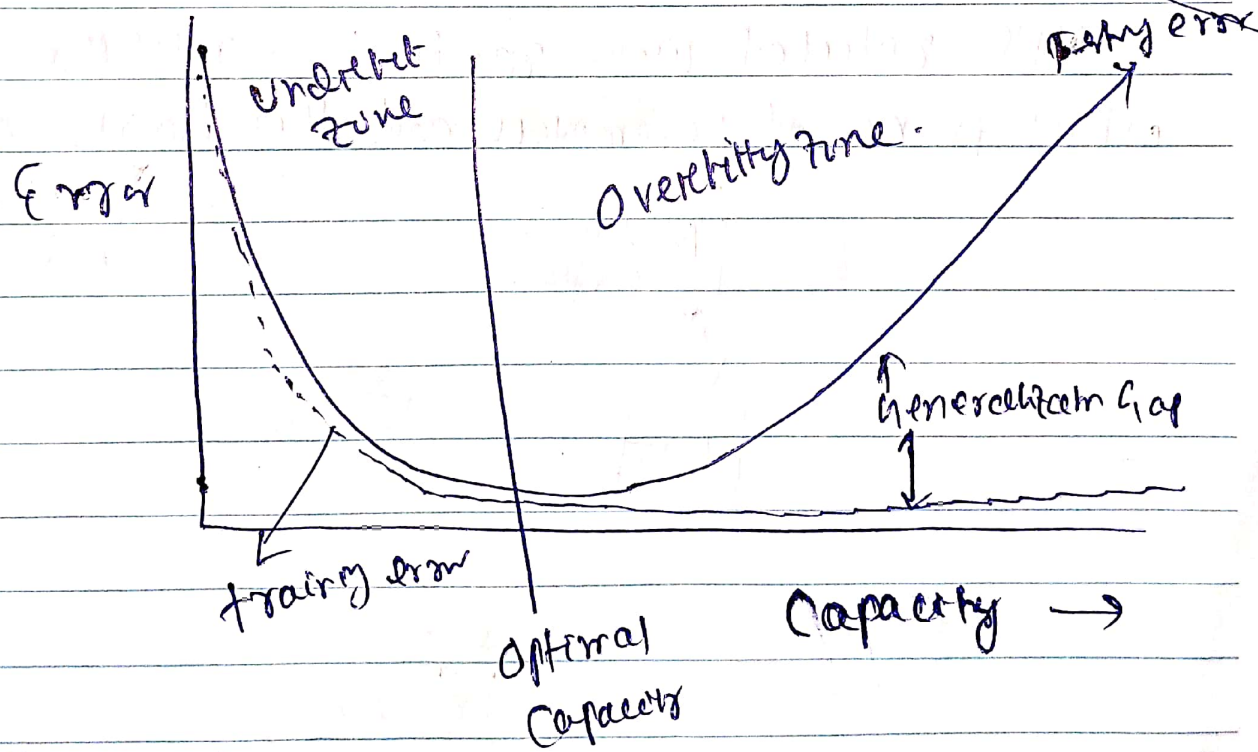
Appropriate capacity

* If your training set is large, less chance of overfit.



Overfit

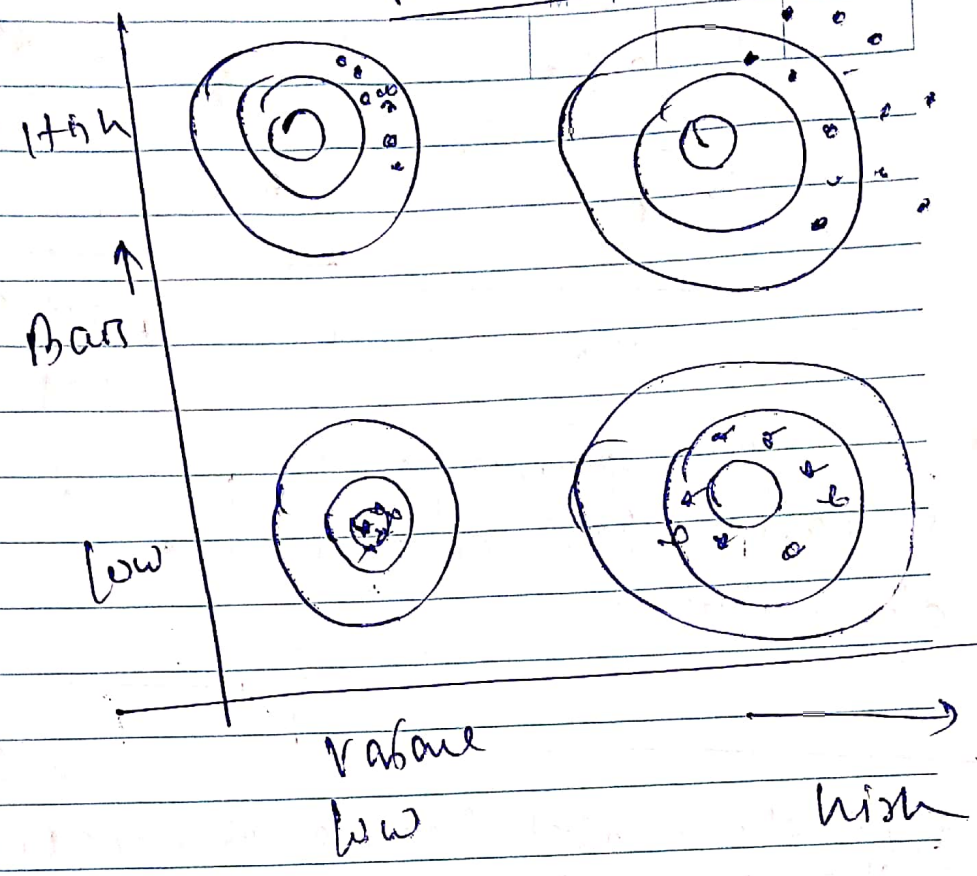
⇒ As the model become complex (higher degree), the associated coefficients or parameter values gets larger (Absolute values of the coefficients)



⇒ Bias-variance tradeoff is how to ~~deal with~~ give appropriate capacity to given model

Capacity :- Model's capacity is ability to fit a wide variety of functions.

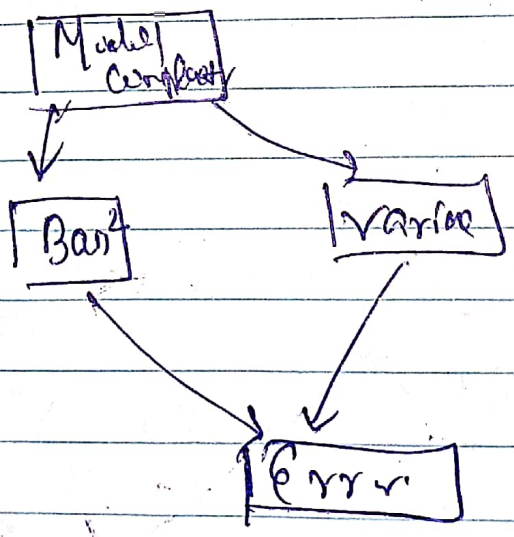
Part. Board example



Bais variance trade off

It's related with model's complexity which is no. of parameters and the bais functions.

Allows tin



Fitting error (training error)

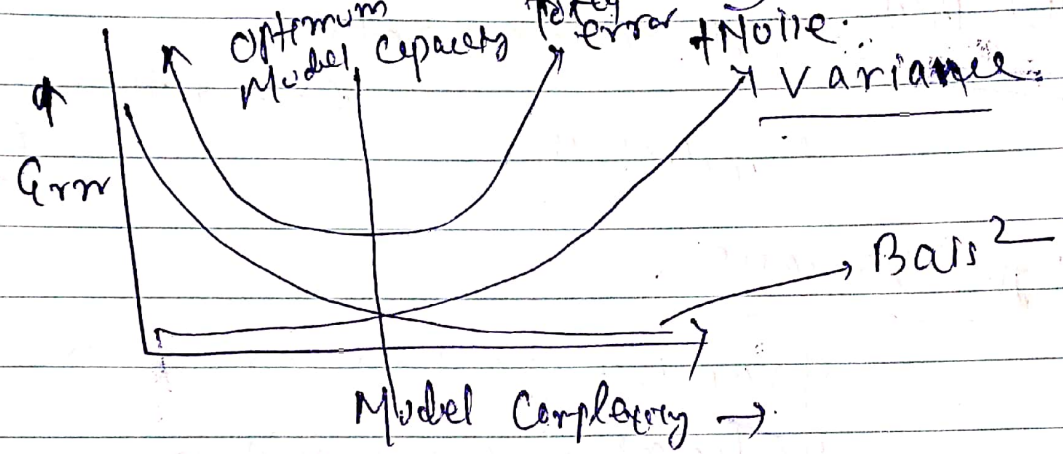
= Bais² + variance.

$$\text{Bais}^2 = E\left[(\hat{y} - y)^2\right]$$

$\hat{y}$ → Predicted value

y → ~~observed~~ current value the model should predict

$$\text{variance} = E\left[(\hat{y} - \bar{\hat{y}})^2\right]$$



Bais	Variance
(1) High Bais result in ~~over~~ underfitting	(1) Result in overfitting
(2) Insufficient feature	(2) Too may features.
(3) Simpler the model higher are the assumption high is Bais.	(3) Simple Model, low variance.
(4) High complexity model, low Bais	(4) Complex Model, high varace.

~~How to Measure Bias-Variance~~

How Bias is Measured

Divide data into 60% + 20% + 20%
training validation testing
or development set

To check if your model has high variance or Bias

Used to determine final accuracy

High Bias Both training + validation error is high (underfitting)

High Variance :-

Training has Low error with validation error is high (overfitting)

Method used to reduce overfit

→ Default Method is to take larger training data

→ Use regularization.

→ Involves Adding penalty term to error function to discourage the coefficient from reaching large values.

$$E_2(w) = \frac{1}{2}\sum_{i=1}^{m} [h(x_i, \theta) - y^i]^2$$

$$\|\theta\|^2 = \theta^T\theta = \theta_0^2 + \theta_1^2 + \dots \theta_d^2 \qquad + \frac{\lambda}{2}\|\theta\|^2$$

λ governs relative importance to regularization term compared with sum of square error.

θ^2 this particular regression is ridge regression.

L_1 norm is $\sum \|\theta\|$ → Absolute value.
→ Lasso

$\lambda \in [0,1]$ → Hyper parameter

$\log \lambda = 0$
→ strong regularization
(high bias Model)

$\log \lambda = -120$
→ Low bias model.

⇒ Cross validation also reduces variance.