SRFP Internship

Prof :- Anil Kumar Ghosh.

Meetings fixed
Every Monday → 7 pm to 8 pm

Introduction Meet :- 13/5/2021

→ we will be working ~~w~~ with the following kind of problems
  ↳ honest evaluation metric of a candidate when the interview of each candidate is taken ~~from~~ ~~a~~ by a combination of members of the panel
  ↳ Ranking the teams based on some 1 v 1 matches (all teams dont play against each other)
  ↳ Estimating the life of a bulb when we can't logistically ~~check~~ wait for the complete life of the bulbs.

→ The methods used for such problems will be:-
  ↳ Inference → Studying about the population based on a ~~[illegible]~~ Sample

ESTIMATION
→ No prior idea of the population
→ Form idea based on sample

HYPOTHESIS TESTING
→ Have a prior idea / statement
→ test the validity of statement using samples.

The best way out is always through.
- Robert Frost

→ METHODS OF ESTIMATION

Simplest methods → Method of Moments
→ Method of Maximum Likelihood
Read these first

→ EM algorithm
→ Generalization of EM ie MM (Minorization Maximization or Majorization Minimization) Algo
Read Next

Casella Barger
Chapter 7 → Point Estimation (Page 337)

→ Sampling is done from a population that is described by pdf or pmf $f(x, \theta)$ where $\theta \to$ knowledge of the entire population

→ θ is the point estimator. Sometimes some function $\tau(\theta)$ may also be of our interest.

Def :- A point estimator is any function $W(X_1, \dots, X_n)$ of a sample; that is, any statistic is a point estimation

Note:- 1. There need not be correspondence b/w estimator & the parameters it is to estimate

2. In general, the range of the statistic $W(X_1, \dots, X_n)$ must coincide with that of the parameter, but that is not always the case

→ Estimate vs Estimator → function of the sample
↳ realized value of the estimator

Method of Moments

→ simple & almost everytime yields some sort of estimate
→ Estimators yielded by this method might have to be improved upon

Method:- • Let $X_1, \ldots X_n$ be a sample from a population with pdf or pmf $f(x|\theta_1, \ldots \theta_k)$
• Take first k raw moments of the sample and equate it to first k raw moments of the population.
• Solve the simultaneous equations

$$m_1 = \frac{1}{n}\sum_{i=1}^{n} x_i^1 , \quad \mu_1' = EX^1$$
$$m_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k , \quad \mu_k' = EX^k$$

→ functions of $(\theta_1 \ldots \theta_k)$

$$\Downarrow$$

$$m_1 = \mu_1'(\theta_1 \ldots \theta_k)$$
$$m_2 = \mu_2'(\theta_1 \ldots \theta_k)$$
$$m_k = \mu_k'(\theta_1 \ldots \theta_k)$$

→ Solve to obtain $\theta_1, \theta_2 \ldots \theta_k$

Note:- In theory, the moments of the distribution of any statistic can be matched to those of any distribution, but in practice, it is best to use similar distribution

Normal method of moments → $X_1 \ldots X_n \overset{iid}{\sim} N(\mu, \sigma^2)$

→ $N(\mu, \sigma^2) \Rightarrow \theta_1 = \mu, \ \theta_2 = \sigma^2$

Now

$m_1 = \bar{X}$ $\qquad \mu_1' = \mu$

$m_2 = \frac{1}{n} \sum X_i^2$ $\qquad \mu_2' = (\mu)^2 + \sigma^2$

Solving for μ & σ^2

$$\mu = \bar{X}, \quad \sigma^2 = \frac{1}{n} \sum X_i^2 - \bar{X}^2 = \frac{1}{n} \sum (X_i - \bar{X})^2$$

Binomial Method of Moments → $X_1 \ldots X_n \overset{iid}{\sim} bin(k, p)$

$$P(X_i = x \mid k, p) = \binom{k}{x} p^x (1-p)^{k-x}, \quad x = 0, 1 \ldots k$$

→ p & k are unknowns → we want point estimators p & k

$$\bar{X} = kp$$
$$\frac{1}{n} \sum X_i^2 = kp(1-p) + k^2 p^2$$

$$\Rightarrow k = \frac{\bar{X}^2}{\bar{X} - \frac{1}{n} \sum (X_i - \bar{X})^2} \quad \Bigg| \quad p = \frac{\bar{X}}{k}$$

#Note :- here we can get -ve values for k & p which should not be the case

→ case where range of the estimator does not coincide with the range of the parameter to be estimated.

measures the goodness of fit of a statistical model

Method of Maximum Likelihood (most popular technique)

Let $x_1, \dots x_n$ iid pdf or pmf $= f(x|\theta_1, \dots \theta_k)$

Likelihood function $L(\theta|x) = L(\theta_1 \dots \theta_k | x_1 \dots x_n)$
$$= \prod_{i=1}^{n} f(x_i|\theta_1, \dots \theta_k)$$

Notation :- Maximum Likelihood ~~parameter~~ estimator (MLE) of the parameter θ based on a sample X is $\hat{\theta}(x)$
($\hat{\theta}(x)$ is the parameter value at which $L(\theta|x)$ attains its maximum as a function of θ, with x held fixed)

→ Range of MLE = Range of the parameter

→ Drawbacks → finding the global maximum & varifying it is tough sometimes
→ Numerical sensitivity can affect the finding of maxima

→ Possible candidates of MLE are the values that suffice
$$\frac{\partial}{\partial \theta_i} L(\theta|x) = 0 \quad , i = 1, \dots, k$$

→ Note :- these points might just be local maxima / minima, or global minima, or inflection pts. Also extremes have to be checked for maxima manually.

Eg:- Normal likelihood $\rightarrow x_1, \ldots x_n \overset{iid}{\sim} N(\theta, 1)$

$$L(\theta|x) = \prod_{i=1}^{n} \frac{1}{(2\pi)^{1/2}} e^{-(1/2)(x_i-\theta)^2}$$

$$= \frac{1}{(2\pi)^{n/2}} e^{(-1/2)\sum_{i=1}^{n}(x_i-\theta)^2}$$

$$\frac{d}{d\theta}(L(\theta|x)) = 0 \implies \sum_{i=1}^{n}(x_i-\theta) = 0$$

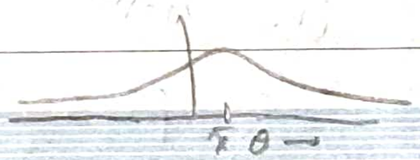
$\longrightarrow$ Solution $\Rightarrow \hat{\theta} = \bar{x}$ (candidate)

Checking for global maxima.

We know that at $\pm\infty$ boundaries $L(\theta|x) \longrightarrow 0$

Also, $\hat{\theta} = \bar{x}$ is the only candidate solution inside the boundaries

So the possible graph is



as $L(\theta|x) > 0$ at $\bar{x} = \hat{\theta}$

Hence $\hat{\theta} = \bar{x}$ is the maxima

Log likelihood :- In many cases it is easier to work with $\boxed{\log L(\theta|x)}$ as log function strictly ↑ on $(0,\infty)$ and the extrema of $\log L(\theta|x)$ and $L(\theta|x)$ coincide

Eg: Bernoulli MLE → $X_1, \dots X_n \overset{iid}{\sim}$ Bernoulli(p)

$$L(p|x) = \prod_{i=1}^{n} p^{x_i} (1-p)^{1-x_i} = p^y (1-p)^{n-y}$$

where $y = \sum x_i$

Now

$$\log(L(p|x)) = y \log p + (n-y)\log(1-p)$$

If $y \in [0, p]$ → $\frac{d}{dp}\log(L(p|x)) = 0 \Rightarrow \hat{p} = \frac{y}{n}$

$$\therefore \boxed{\hat{p} = \frac{\sum x_i}{n}}$$

Note :- MLE ~~mg~~ might be valid only for a range of the parameter values.

Theorem (Invariance property of MLEs)

If $\hat{\theta}$ is the MLE of θ, then for any function $\tau(\theta)$, the MLE of $\tau(\theta)$ is $\tau(\hat{\theta})$

Catchup Meeting 1 :- (17th May 2021)

tasks
- → Reread about the distributions, Method of Maximum likelihood & Method of Moments throughly.
- → Read about Method of least squares
- → Solve the fish catch recatch problem

Discussed
- → We discussed the library books estimation problem
- → parametric vs non-parametric (construction of freq-dist histogram to prob-" graph)
- → Method of Maximum likelihood & Method of Moments was used to find parameter estimates for several distributions.
- → Meaning of likelihood was understood.

Probability

Revision

→ mgf → $M_X(t) = E(e^{tx}) \Rightarrow \frac{\partial}{\partial t^r} M_X(t)\Big|_{t=0} = E(X^r) = \mu'_r$

→ Markov Inequality → $P(X > t) \leq \frac{E(X)}{t}$

→ Chebyshev's " → $P(|X-\mu| > \varepsilon) \leq \frac{E[(X-\mu)^2]}{\varepsilon^2}$

→ Discrete Uniform → $P(X = x_i) = 1/k$, $E(X) = (k+1)/2$, $V(X) = \frac{(k+1)(k-1)}{2}$

→ $U[0,1]$ → $P([a,b]) = b - a$, $\forall (0 \leq a \leq b \leq 1)$

Life is like riding a bicycle. To keep your balance you must keep moving.

– Albert Einstein

mgf $= 1/N \sum_i^N e^{it}$

$E(X+Y) = E(X) + E(Y)$
$E(X.Y) = E(X).E(Y)$] → independence
$V(X+Y) = V(X) + V(Y)$]
$E(aX) = aE(X)$
$V(aX) = a^2 V(X)$

→ Bernoulli(p) → $f(x) = \begin{cases} p^x (1-p)^{1-x}, & x \in \{0,1\} \\ 0, & \text{otherwise} \end{cases}$

$E(X) = p$
$V(X) = p(1-p)$
↳ max at $p = 0.5$

$F(x) = \begin{cases} 0, & x < 0 \\ 1-p, & x \in [0,1) \\ 1, & x \geq 1 \end{cases}$

mgf $= (1-p) + pe^t$

→ Binomial(n,p) → $f(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x}, & x \in \{0,1,\dots n\} \\ 0, & \text{otherwise} \end{cases}$

$E(X) = np$
$V(X) = np(1-p)$

mgf $= [(1-p) + pe^t]^n$

→ Geometric (p) → $f(x) = \begin{cases} (1-p)^x p, & x \in \{0,1,\dots \infty\} \\ 0, & \text{otherwise} \end{cases}$

$E(X) = \frac{(1-p)}{p} = \frac{q}{p}$

$V(X) = \frac{(1-p)}{p^2} = \frac{q}{p^2}$

mgf $= \frac{p}{1-(1-p)e^t}$

→ Def^n :- → No. of failures before first success

OR

$f(x) = \begin{cases} (1-p)^{x-1} p, & x \in \{1,2 \dots \infty\} \\ 0, & \text{otherwise} \end{cases}$

$E(X) = 1/p$
$V(X) = (1-p)/p^2$

mgf $= \frac{pe^t}{1-(1-p)e^t}$

→ Def^n :- No. of trials for first success

→ negative Binomial(r,p) → $f(x) = \begin{cases} \binom{x+r-1}{x} q^x p^r, & x \in \{0,1 \dots\} \\ 0, & \text{otherwise} \end{cases}$

$E(X) = (rq)/p$
$V(X) = (rq)/p^2$

mgf $= \left(\frac{p}{1-(1-p)e^t}\right)^r$

→ Def^n : No. of failures before r^{th} success

→ Hypergeometric(n_1, n_2, r) → $f(x) = \frac{\binom{n_1}{x}\binom{n_2}{y}}{\binom{n_1+n_2}{x+y}}$, $x \in \{0,1 \dots \min(n_1, x+y)\}$

$E(X) = r \frac{n_1}{n_1+n_2}$

$V(X) = r\left(\frac{n_1}{n}\right)\left(1-\frac{n_1}{n}\right)\left(1-\frac{r-1}{n-1}\right)$

here $y \Rightarrow$ no. of n_2
$x \Rightarrow$ no. of n_1
$r = x+y$
$n = n_1 + n_2$

$\sum_{i=1}^{n}$ pois(1) = pois(n) (additive)

→ limiting case of binomial(n,p), $n\to\infty$, $p\to 0$ & $np=\lambda$

8 Poisson(λ) → $f(x) = \begin{cases} e^{-\lambda} \cdot \dfrac{\lambda^x}{x!}, & x \in \{0,1,\dots\}, \lambda>0 \\ 0, & \text{otherwise} \end{cases}$

9 $E(X)=\lambda$
$V(X)=\lambda$

No. of things in an interval

mgf $= e^{\lambda(e^t-1)}$

10

— CONTINUOUS —

11

→ Uniform(a,b) → $f(x) = \begin{cases} 1/b-a & \text{if } x \in [a,b) \\ 0, & \text{otherwise} \end{cases}$

12 $E(X) = (a+b)/2$

1 $V(X) = (b-a)^2/12$

$F(x) = \begin{cases} 0, & x<a \\ x-a/b-a, & x\in[a,b] \\ 1, & x\geq b \end{cases}$

mgf $= \dfrac{e^{bt}-e^{at}}{(b-a)t}$

$Y \sim U(0,1)$
$X = a+(b-a)Y \sim U(a,b)$

2

→ Exponential(λ) → $\lambda>0$ → $f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \in (0,\infty) \\ 0, & \text{otherwise} \end{cases}$

3 $E(X) = 1/\lambda$

4 $V(X) = 1/\lambda^2$

$F(x) = \begin{cases} 0, & x<0 \\ 1-e^{-\lambda y}, & x\geq 0 \end{cases}$

mgf $= \dfrac{\lambda}{\lambda-t}$ when $t<\lambda$

5 • $X\sim \exp\lambda$ if $Y\sim U(0,1)$ & $X = (-1/\lambda)\ln(1-Y)$, $\lambda>0$
• waiting time / interarrival time of similar events

6 • Memoryless property • continuous form of Geometric
$X\sim\exp(\lambda)$, $Z=[X] \sim geo(1-e^{-\lambda})$

→ Gamma(n,λ) → $f(x) = \begin{cases} \dfrac{e^{-\lambda x} \cdot x^{(n-1)} \cdot \lambda^n}{\Gamma n}, & y>0, \lambda,\alpha>0 \\ 0, & \text{otherwise} \end{cases}$

$E(X) = n/\lambda$

$V(X) = n/\lambda^2$

mgf $= \left(\dfrac{\lambda}{\lambda-t}\right)^n$, $t<\lambda$

→ Def^n: Arrival time of n^{th} event (sum of n exp(λ))

• Additive property
$Z_1\sim$ Gamma(n_1,λ), $Z_2\sim$ Gamma(n_2,λ)
$Z_1+Z_2 \sim$ Gamma(n_1+n_2,λ)

→ Beta (α_1, α_2) → $f(x) = \begin{cases} \dfrac{x^{(\alpha_1-1)}(1-x)^{(\alpha_2-1)}}{B(\alpha_1,\alpha_2)}, & x \in [0,1] \\ 0, & \text{otherwise} \end{cases}$

$E(X) = \alpha_1/(\alpha_1+\alpha_2)$

$V(X) = (\alpha_1\alpha_2)/(\alpha_1+\alpha_2)^2(\alpha_1+\alpha_2+1)$

$Z_1 \sim \text{Gamma}(\alpha_1, \lambda)$

$Z_2 \sim$ " (α_2, λ)

$X = \dfrac{Z_1}{Z_1+Z_2} \sim \text{Beta}(\alpha_1, \alpha_2)$

→ Normal (μ, σ^2) → $f(x) = \left\{ \dfrac{e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}}{\sigma\sqrt{2\pi}} \right.$ $\mu \in \mathbb{R}$, $\sigma \in \mathbb{R}^+$

$E(X) = \mu$

$V(X) = \sigma^2$

mgf $= e^{\mu t + \sigma^2 t^2/2}$

$Y \sim N(0,1)$

$X = \mu + \sigma Y \sim N(\mu, \sigma^2)$

• Height - Weight Distribution

→ Chi-Square - $\chi^2_n \equiv \text{Gamma}(n/2, 1/2)$

$Z_1, Z_2, \dots Z_n \sim \text{iid } N(0,1)$

$\chi^2_n = \sum_{i=1}^{n} Z_i^2 \sim \text{Gamma}(n/2, 1/2)$

Theorem

$h(y) = f(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right|$

old pdf — old variable in terms of new

→ Cauchy (μ, σ) $f(x) = \dfrac{1}{\sigma\pi} \dfrac{1}{1+\left(\frac{x-\mu}{\sigma}\right)^2}$, $x \in \mathbb{R}$

No Expectation, variance or mgf as $\int_{-\infty}^{\infty} x f(x)\, dx$ diverges

→ Cauchy $(0,1)$

$F(y) = \frac{1}{2} + \frac{1}{\pi} \tan^{-1} y$

$X_1, X_2 \sim \text{iid } N(0, \sigma^2)$

$\dfrac{X_1}{X_2}, \dfrac{X_1}{|X_2|}, t_1 \sim C(0,1)$

It is the mark of an educated mind to be able to entertain a thought without accepting it.

– Aristotle

Catch-Recatch problem (MY SOLUTION) (Wrong)

P.S:- Assume there are N fish in the pond. We have tagged M fish. M is known. These tagged fishes are left in the pond again. As N is unknown we need to find it.
Next day we catch L many fishes out of which K are tagged.
Write an equation for N in terms of N, L & K it estimate N.

⇒ The pond is vast and fishes are free to roam about. Hence assuming that each fish has equal probability to get caught is valid.

Tagged fishes have $\frac{M}{N}$ probability to get caught.

Now, as the fishing is random, the proportion of tagged fishes in the caught set should be equal to the proportion of tagged fishes in the whole pond

i.e $\frac{K}{L} \simeq \frac{M}{N}$

$$\boxed{N \simeq \frac{ML}{K}}$$

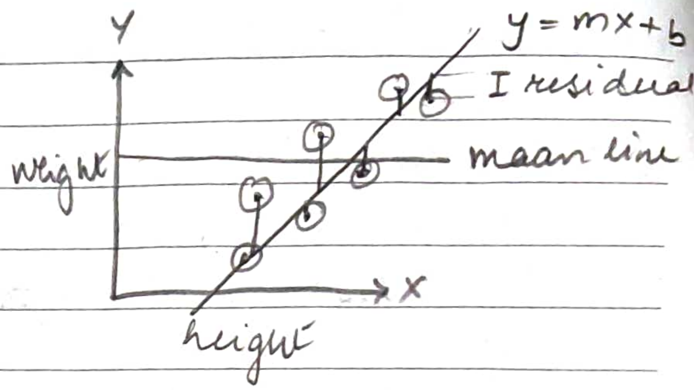
where N → Total fishes
M → Total tagged fishes
L → No of recaught fishes
K → " " " tagged fishes

OCTOBER						
M	T	W	T	F	S	S
				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	31

Dasshera (India); Mother's Day (Malawi)

Week 41

8 <u>Method of least squares</u>



9 m, b → unknowns

10 y = mx+b → regression line

11 Residual :- difference between actual pt & point predicted by a regression line

12

1 $r_1 = \cancel{y_1} \, y_1' - y_1$ → y_1 → actual y for some $x = x_1$
y_1' → predicted y for the same $x = x_1$

2 Best fit line → minimizes the sum of the squares of the residuals

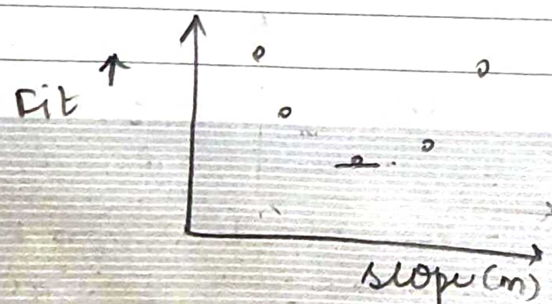
3 (squares are used as the residuals might be -ve as well)

4

5

Fit = Sum of squared residuals = $\Sigma (r_i)^2$ → has to be minimized

6



considering just the slope

best fit → where the derivative is '0'

Now

Variance of fit = $\dfrac{\text{Sum of squared residuals}}{\text{Total No of pts}}$ = Avg sum of squares

$$Var(fit) = \frac{\Sigma r_i^2}{N} = \frac{\Sigma (y_I - y_i')^2}{N}$$

→ In least squares method, we try to minimize the variance of the fit line

→ R^2 tells us how much variation in weight can be explained by height.

$$R^2 = \frac{Var(\overbrace{mean}^{\text{variance abt mean line}}) - Var(fit)}{Var(mean)}$$

$= \dfrac{\text{Variation of weight explained by height}}{\text{" " " without taking " into accant}}$

Larger the R^2, better is the fit

→ $F = \dfrac{\text{Variation of weigh explained by height}}{\text{" " " not " " "}}$

$$F = \frac{\text{var(mean)} - \text{var(fit)} / (p_{fit} - p_{mean})}{\text{var(fit)} / (n - p_{fit})}$$

→ explains variation reduced due to extra parameter (m here)

explains variation remaining

p_{fit} = No. of parameters in fit line (2 here → m & b)

p_{mean} = No. of parameters in mean line (1 here → y intercept)

n = no. of pts

F is a really large number if the fit is good

Note: we use $(n - p_{fit})$ and not just n as we need more data when there are more parameters to estimate & this part of equation takes that into account.

→ Appropriate b & m are found by some decent algorithms which update the weights by iteration

→ Cost function = $J(m,b) = \frac{fit}{2N} = \frac{1}{2N}\sum_{i=1}^{N}(\kappa_i)^2 = \frac{1}{2N}\sum_{i=1}^{N}(mx_i+b-y_i)^2$

→ Descent Algorithm (Gradient Descent)

Repeat

$$\left\{ \begin{aligned} m &:= m - \alpha \frac{\partial}{\partial m} J(m,b) \\ b &:= b - \alpha \frac{\partial}{\partial b} J(m,b) \end{aligned} \right\}$$

α = learning rate

$$\frac{\partial}{\partial b} J(m,b) = \frac{1}{N}\sum_{i=1}^{N} \kappa_i$$

$$\frac{\partial}{\partial m} J(m,b) = \frac{1}{N}\sum_{i=1}^{N} (\kappa_i)(x_i)$$

as $J(m,b) = \frac{1}{2N}\sum_{i=1}^{N}(mx_i+b-y_i)^2$

→ Method 2:-

- Take $\frac{\partial J}{\partial m} = 0$ & $\frac{\partial J}{\partial b} = 0$
- Solve for m & b

$$\frac{\partial J}{\partial b} = \frac{\cancel{2}}{\cancel{2}N} \sum_{i=1}^{N} (m x_i + b - y_i) = 0 \rightarrow ①$$

$$\frac{\partial J}{\partial m} = \frac{1}{N} \sum_{i=1}^{N} (m x_i + b_i - y_i) x_i = 0 \rightarrow ②$$

Solve ① & ② to get b & m

$$① \Rightarrow \sum_{i=1}^{N} (m x_i + b - y_i) = 0 \Rightarrow m \sum x_i + Nb - \sum y_i = 0$$

$$② \Rightarrow \sum_{i=1}^{N} (m x_i + b - y_i) x_i = 0 \Rightarrow m \sum x_i^2 + b \sum x_i - \sum y_i x_i = 0$$

∴ from ① $\boxed{b = \frac{\sum y_i}{N} - \frac{m}{N} \sum x_i = \bar{y} - m\bar{x}}$

Substituting b in ②

$$\sum_{i=1}^{N} (m x_i + \bar{y} - m\bar{x} - y_i) x_i = 0$$

$$\Rightarrow \sum x_i (\bar{y} - y_i + m(x_i - \bar{x})) = 0$$

$$\Rightarrow \sum x_i (\bar{y} - y_i) + m \sum x_i (x_i - \bar{x}) = 0$$

$$\boxed{m = \frac{\sum x_i (y_i - \bar{y})}{\sum x_i (x_i - \bar{x})} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}}$$

Basic Algorithm of method of least squares

→ fit a line $y_1 = a + bx_1 + cx_2 + \dots M x_{m-1}$

→ Calculate $r_i = y_1^i - y^i$ ($y_1^i \to$ calculated y, $y^i \to$ actual y)

→ Let fit $= \sum_{i=1}^{N} (r_i)^2$

Cost funⁿ $= J = \frac{1}{2N} \sum_{i=1}^{N} (r_i)^2 = \frac{1}{2N} \sum_{i=1}^{N} (a + bx_1^i + cx_2^i + \dots - y_i)$

$i \to$ pt. No.

x_j, $j \to$ variable No.

→ Set $\frac{\partial J}{\partial a} = 0$, $\frac{\partial J}{\partial b} = 0$, ... to get m eqⁿˢ

→ Solve those equations to get $a, b, \dots m$

$$\begin{bmatrix} \uparrow & \uparrow & x_2^1 & & \uparrow \\ x_0 & x_1 & x_2^2 & \cdots & x_n \\ \downarrow & \downarrow & & & \downarrow \end{bmatrix} \begin{bmatrix} a \\ b \\ \vdots \\ m \end{bmatrix} = \begin{bmatrix} y^1 \\ \vdots \\ y^n \end{bmatrix}$$ } equations we get

Catchup Meeting 2:-
24 May 2021

Tasks
→ ~~Try to code an iterative algorithm to find the minimum of $\sum |x_i m + c - y_i|$~~
→ ~~Try coding IRWLS algorithm~~
→ ~~We know that $\sum (x_i - a)^2$ is minimized at $\bar{x}$ &~~
~~$\sum |x_i - a|$ " " at median (X)~~
~~find the function $\sum \phi$ such that $\sum \phi(x_i, A)$ would be~~
minimized when $A = p^{th}$ quantile

Discussions

Solution of the fish problem (estimation of number of fishes in the pond) & book problem (estimating the no. of books in the library) were discussed

Regression was discussed and minimizing $\Sigma(x_i m+c-y_i)^2$ vs $\Sigma|x_i m+c-y_i|$ were discussed in detail.

Also minimizing $\Sigma(x_i m+c-y_i)^2$ is analogous to fining the mean/expectation while minimizing $\Sigma|x_i m+c-y_i|$ is analogous to finding the median.

Advantages / Disadvantages & methods to find both were discussed.

Algorithm to find m & c that minimize $\Sigma|y_i - mx_i - c|$

Wiki page: Least Absolute Deviations

Def^n:- Method of Least Absolute Deviations (LAD)

The method of fitting a curve that minimizes the absolute differences is known as Method of LAD

Here we are interested in finding parameters which result in minimum $S = \Sigma|y_i - f(x_i)|$

We do not have any analytical solving method. Hence we have to use some iterative method to find LAD parameters

Methods

Solution 1 (Iteratively Reweighted Least Squares)

We want to minimize $J = \sum_{i=1}^{n} |y_i - \theta^T x_i|$

We can write it in terms of least squares

$$J = \sum_{i=1}^{n} (y_i - \theta^T x_i)^2 \cdot \frac{1}{|y_i - \theta^T x_i|}$$

we know that, we have a solution for $J = \sum_{i=1}^{n} (y_i - \theta^T x_i)^2$ i.e

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

But this minimizes the least square only

To use this as an solution for $J = \sum_{i=1}^{n} (y_i - \theta^T x_i)^2 \cdot \frac{1}{|y_i - \theta^T x_i|} = \sum_{i=1}^{n} (y_i - \theta^T x_i)^2 \cdot w_i$

we need to use the above solution as an iterative form

$$[\theta]_{new} = (X^T W X)^{-1} (X^T W Y)$$

Algorithm:-

→ Randomly initialize θ

→ Calculate $\theta_{new} = (X^T W_{old} X)^{-1} (X^T W_{old} Y)$ → All the terms are explained on Oct 26 & Oct 28 pg at the end

→ Repeat until convergence

$\eta = \theta$

Now,

$\eta = \mu$ for Normal family → Linear Regression

$\eta = \text{logit}(\pi)$ " Binomial family → Logistic "

$$\frac{\partial L}{\partial B_j} = \left(\frac{y-\mu}{\text{var}(y)}\right)(x_{ij})\left(\frac{\partial \mu}{\partial \eta}\right)$$

$$\frac{\partial L}{\partial B_j} = \left(\frac{y-\mu}{a(\phi)}\right) x_{ij}$$

Iteratively Reweighted least squares (wikipedia)

→ IRWLS is used to solve optimization problems with objective functions of the form of p-norm.

→ p-norm :- (wikipedia) for some $p \geq 1$, p-norm on x is defined as

$$\|x\|_p = (|x_1|^p + |x_2|^p + \dots + |x_n|^p)^{1/p}$$

when $p \in \mathbb{R}$

when $p \to \infty$

$$\|x\|_\infty = \max\{|x_1|, |x_2|, |x_3|, \dots, |x_n|\}$$

→ IRWLS solves objective functions of the form $\min \sum_{i=1}^{n} |y_i - f_i(\beta)|^p$ to find β iteratively

→ In each iteration, β is updated as follows

$$\beta^{t+1} = \underset{\beta}{\overset{\text{that gives}}{\min}} \sum_{i=1}^{n} w_i^t \, |y_i - f_i(\beta)|^2$$

$$= (X^T W^{(t)} X)^{-1} (X^T W^{(t)} y)$$

($\because$ here $f_i(\beta) = X_i \beta$)

where

$W^{(t)}$ is the diagonal matrix of weights (which are usually initially set to 1)

$$w_i^{(0)} = 1$$

$$w_i^{(t)} = |y_i - X_i \beta^{(t)}|^{p-2}$$

<u>Note</u> :- when $p = 1$ (our case)

$$w_i^{(t)} = \frac{1}{|(y_i - X_i \beta^{(t)})|} \rightarrow \text{might reach inf}$$

to avoid division by '0', it is a common practice to use

$$w_i^{(t)} = \frac{1}{\max\{\delta, |y_i - X_i \beta^{(t)}|\}} \quad \text{where } 0 < \delta \ll 1$$

Finding β that minimizes $\Sigma|y_i - x_i\beta|$

Intuition for

$$\beta^{t+1} = \beta \text{ that minimizes } \sum_{i=1}^{n} w_i^{(t)} (y_i - x_i\beta)^2$$

$$= (X^T W^{(t)} X)^{-1} (X^T W^{(t)} Y)$$

we know that

$\min \Sigma (x_i - A)^2$ is given by $A = \bar{x}$

Now, we have a frequency distribution as

x_1	f_1
x_2	f_2

and we need to find A such that we minimize $\Sigma f_i (x_i - A)^2$

A would be $\frac{\Sigma x_i f_i}{\Sigma f_i}$ i.e weighted mean of distribution

Similarly for the above problem, we would get the minimum of sum when

$$\beta = \frac{\sum_{i=1}^{n} \frac{w_i^{(t)}}{x_i^2} \left(\frac{y_i}{x_i}\right)}{\Sigma \frac{w_i^{(t)}}{(x_i)^2}}$$

when pts are just (y_i, x_i) i.e lie on a 2D plane

Another solution for this problem would be

$$\frac{\partial}{\partial \beta_j}\left(\sum_{i=1}^{n} w_i^{(t)}(y_i - x_i\beta)^2\right) = 0$$

i.e

$$\sum_{i=1}^{n} w_i^{(t)}(y_i - x_i\beta) = 0 \quad \forall\ \beta_j$$

$$W^{(t)}Y - W^{(t)}X\beta = 0$$

$$W^{(t)}X\beta = W^{(t)}Y$$

multiplying ~~of~~ by X^T to make $(X^T W^{(t)} X)$ invertible

$$X^T W^{(t)} X\beta = X^T W^{(t)} Y$$

$$\beta^{t+1} = \beta = (X^T W^{(t)} X)^{-1}(X^T W^{(t)} Y)$$

~~then~~ where $[X]_{m\times n}$, $[Y]_{m\times 1}$, $[\beta]_{n\times 1}$, $[W^{(t)}]_{m\times m}$,

$m \rightarrow$ no of instances

$(n-1) \rightarrow$ no of variables per instance

(1 extra variable ~~for~~ set as constant 1)

$$[W^{(t)}] = \begin{bmatrix} w_1^{(t)} & 0 & 0 & \cdots \\ 0 & w_2^{(t)} & 0 & \cdots \\ 0 & 0 & w_3^{(t)} & \cdots \\ \vdots & & & \ddots \\ \cdots & & & w_m^{(t)} \end{bmatrix}_{m\times m}$$

where $w_i^{(t)} = \dfrac{1}{\max\{\delta, |y_i - \beta^t x_i|\}}$

for i^{th} instance

[δ is is a very small +ve value]

Method 2

→ Step Wise Algorithm

8 **Maximum likelihood method to find LAD**

9 Lets consider linear LAD regression such that

10 $Y_i = ax_i + b + u_i$, where a is the slope of line
b is the y-intercept
11 & u_i is error at ith pt.

12 Objective function :- $F(a,b) = \sum_{i=1}^{n} |Y_i - ax_i - b| = \sum_{i=1}^{n} |u_i|$

1 Now,
Consider that u has laplace distribution, i.e

2 pdf of laplace dist. → $f(u|\mu,\beta) = \frac{1}{2\beta} \exp\left(-\frac{|u-\mu|}{\beta}\right)$, $\mu \rightarrow$ location parameter
$\beta > 0 \rightarrow$ Scale parameter

4 If $\mu = 0$, $\beta = 1$
$f(u|0,1) = \frac{1}{2} \exp(-|u|)$

5 $L(u|0,1) = \prod \frac{1}{2} e^{(-|u_i|)} = \left(\frac{1}{2}\right)^n e^{-\sum |u_i|}$

6 MLE for the above likelihood exists when $\sum |u_i|$ is minimum

$\therefore$ MLE exists at $\min(\sum |u_i|)$ ie $\min(\sum |Y_i - ax_i - b|)$

MLE Iterative Algorithm (Step wise Algorithm)

(1) let $a = a_0$, we need to find b such that we get $\min(\Sigma|u_i|)$

$$b = b_0 = \text{Median}\left[(y_i - a_0 x_i)\Big|_{i=1}^{n}\right]$$

(2) we have $b = b_0$, we need to find a such that we get $\min(\Sigma|u_i|)$ i.e $\min \sum_{i=1, x_i \neq 0} |x_i| \left|\frac{y_i - b_0}{x_i} - a\right|$

$$a = a_1 = \text{Median}\left(|x_i| \diamond \left(\frac{y_i - b_0}{x_i}\right)\Big|_{i=1}^{n}\right)$$

we repeat (1) & (2) iteratively until convergence.

> Note :- symbol $\diamond$ means $\left(\frac{y_i - b_0}{x_i}\right)$ replicated $|x_i|$ times

IRWLS Algorithm (can also be proved by MLE)

→ Initialize θ^0 randomly, [let $\theta^0 = \begin{bmatrix} b^0 \\ a_0 \end{bmatrix}$ here]

→ Let weight Matrix $[W^0]_{m \times m} = \begin{bmatrix} w_1^0 & 0 & 0 & 0 \\ 0 & w_2^0 & & \\ \cdots & & & w_m^0 \end{bmatrix}$ where $w_i^0 = \frac{1}{\max(\delta, |y_i - a_0 x_i - b_0|)}$

→ update $\theta^1 = (X^T W^0 X)^{-1}(X^T W^0 Y)$ where $[X]_{m \times n}$, $[Y] = m \times 1$

$(n-1) \rightarrow$ no. of features, (here $n = 2$)

[we add 1 dimentionless feature]

→ Repeat $\theta^{t+1} = (X^T W^t X)^{-1}(X^T W^t Y)$ until convergence

Quantiles

The quantile of order p or p^{th} quantile $(0<p<1)$ is a value of the variable which divides the whole frequency distribution into two parts such that p proportion of observations are less than it and $(1-p)$ proportion of observations are greater than it.

Now,

if $p = \tau \in (0,1)$, and q is the τ^{th} quantile
i.e for a sample $\{y_1, y_2 \dots y_n\}$, $P(y_i \leq q) =$

$\therefore$ we get minimum of $\displaystyle\sum_{i=1}^{n} \begin{cases} (\tau-1)(y_i - A) & \text{if } y_i < A \\ \tau\,(y_i - A) & \text{if } y_i \geq A \end{cases}$

When $A = q = \tau^{th}$ quantile

Special case:-

When $\tau = 0.5$, $A = q$ is the median which gives

$$\text{minimum of } \sum_{i=1}^{n} \begin{cases} -0.5\,(y_i - A), & y_i < A \\ 0.5\,(y_i - A), & y_i \geq A \end{cases}$$

$$\Downarrow$$

$$\text{minimum of } \sum_{i=1}^{n} |y_i - A|$$

Note:- If $m \to$ Median

$P(X \le m) \geq \frac{1}{2}$ & $P(X \geq m) \geq \frac{1}{2}$ $\to$ Discrete

$P(X \le m) = \frac{1}{2} = P(X \geq m) \longrightarrow$ continuous

Here we have a distribution that can be easily divided into 3:7, hence we take strict equality

Intuition

let us consider a sample $\{1, 2, 3, \ldots, 10\}$ ($y_1, y_2, \ldots, y_{10}$)

now if $p = 0.3$, we want to find q such that

$$P(y_i < q) = 0.3 \quad \& \quad P(y_i \geq q) = 0.7$$

→ We know that, $\sum |y_i - A|$ is minimized at the median.

→ We also know that $q = 3.5$

→ If we transform the whole distribution such that, 3.4 becomes the median of the transformed sample set, we would achieve our goal

→ We also know that, $\sum f_i |y_i - A|$ is minimum when $A =$ weighted median of y_i ie median of $f_i y_i$

Consider a function f such that,

$\sum f |y_i - A|$ is minimum at $A = q = 3.5$ for the above sample

We know that 30% of samples are $< q$ & 70% of samples are $\geq q$

If we define f such that, we multiply 0.7 to samples $< q$ and $0.3 \geq q$, then, q would be the minimum

$$f = \begin{cases} 0.7, & y_i < A \\ 0.3, & y_i \geq A \end{cases} \mapsto \sum f|y_i - A| \text{ would be minimum at } A = q = 4$$

For easier understanding, let $f = \begin{cases} 0.7 \times 10 = 7, & y_i < A \\ 0.3 \times 10 = 3, & y_i \geq A \end{cases}$

When $A = q$, the distribution would be transformed as follows

frequency	y	cumulative frequency
7	1	7
7	2	14
7	3	21
3	4	24
⋮	⋮	⋮
3	10	42

We have 3.5 as the median of the transformed distribution.

Thus, if $q = p^{th}$ quantile then,

$$\sum f|y_i - A| \text{ is minimized at } A = q \text{ when } f = \begin{cases} 1-p, & y_i < A \\ p, & y_i \geq A \end{cases}$$

$\Downarrow$ can also be written as

$$\sum f(y_i - A), \text{ where } f = \begin{cases} p-1, & y_i < A \\ p, & y_i \geq A \end{cases}$$

Catchup Meeting 3:-
2nd June 2021

Tasks
- → Read EM, MM & NR algorithms
- → Code out the solution for ABO Problem

DISCUSSIONS:
- → Discussed formal proof of $\sum \rho|x_i - \theta|$ minimized at $\theta = p^{th}$ quantile when $\rho = \begin{cases} 1-p, & x_i < \theta \\ p, & x_i \geq \theta \end{cases}$
- → Discussed proof that LAD regression is MLE of its error when error ~ double exponential distribution
- → similarly discussed proof that least-squares regression is MLE of its error when error $\sim N(0, \sigma^2)$ distribution
- → Discussed convergence of the step-wise algorithm for LAD
- → Discussed MLE for probability p in the Urn problems where Urn has many white & red balls & p is the probability of picking a red ball. Discussed many such cases.
- → Discussed initial part of ABO problem.

\# Advised to Read MM & EM from the book Statistical computing (Compiled by Debashis Kundu & Ayendranath Basu)

• The chapter on EM (Expectation Maximization) Algorithm in this book is written by T. Krishnan

Problem of latent Variables for Maximum Likelihood

→ Problem arises on how to estimate joint probability Distribution of an Dataset from a sample

→ Probability Density estimation is basically the construction of an estimate based on observed data. It involves distribution function & the parameters of the function that best explains the joint probability of the observed data.

→ MLE :- In statistics, MLE is a method of estimating parameters of a probability distribution by maximizing the likelihood function in order to make the observed data most probable for the statistical model

→ Problem with MLE :- It assumes that data is completely observed and it does not mandate that the model will have access to all the data. It assumes that all the variables associated to the model are already present. But, in some cases, some relevent variables may stay hidden and cause inconsistancies. ~~This~~ These hidden variables are known as "LATENT VARIABLES"

Culture Day (Japan)

8 # In the presence of latent variables, a conventional MLE would not work as expected and
9 an approach to estimate parameters in presence of latent variables is <u>EM Algorithm</u>
10 (soft clustering)

<u>EXPECTATION-MAXIMIZATION (EM) Algorithm</u>

11
→ used to find the local maximum likelihood estimator
12 in case the data is missing/incomplete or in the presence of latent variables

1 <u>Algorithm</u>.

→ <u>Step 1</u>:- Consider a set of starting parameters in the incomplete data
2

→ <u>Step 2</u>:- (Expectation step) Estimates the values of the
3 missing data/parameters. It uses observed data to guess the value of the
4 missing data. It also uses the starting parameters.

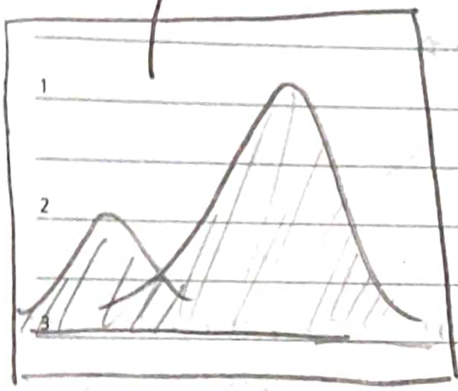
5
→ <u>Step 3</u>:- (Maximization step) This step generates
6 complete data after step 2 and updates the parameters set in step 1 using the estimate of values generated in step 2

→ <u>Step 4</u>:- Repeat step 2 & 3 until convergence
[Convergence in probability is based on intuition]
(Here convergence means the matching of the missing value with ~~the~~ its expected value)

Example

Gaussian Mixture Model → combination of probability distributions & requires estimation of mean & std. dev. parameters

Consider a case :-
- Datapts. are generated by 2 different processes such that each process has a normal (gaussean) dist.
- We dont know which dist. a datapt. belongs to as the data pts. are combined & dist. are similar & not disjoint.
- The processes used to generate datapts act as latent variables & influence the data



- Here EM algorithm helps us to guess if the datapt ~~datto~~ belongs to first curve or second
- It also allows us to estimate σ & μ of the curves, & hence we can dist. the pts as curve 1 or 2 acc. their density

now, • we need σ & μ of both the curves to say if pts are from curve 1 or 2
- we need to know the origin of the pts to estimate σ & μ of the curves

This is an infinite loop which can be converted into an iterative algorithm i.e EM.

EM algorithm in the mixed gaussian case

1 → Start ~~randomly~~ with $(\sigma_1, \mu_1), (\sigma_2, \mu_2)$ initialized randomly

2 → E step: Based on these σ's & μ's allot the pts as curve 1 or curve 2 (Acc to the probability)

3 → M step: Acc. to the ~~allo~~ allotments on previous step, recalculate $(\sigma_1', \mu_1'), (\sigma_2', \mu_2')$

4 → Check if $(\sigma_1', \mu_1'), (\sigma_2', \mu_2')$ are close enough to ~~initial~~ previous (σ_1, μ_1) & (σ_2, μ_2) respectively.
If Yes, stop
If No, repeat steps 2, 3 & 4

E-M algorithm is best explained with examples in Sir's class notes (Check notes of Meeting on 9th June 2021)

Algorithm

→ we find out the unknown data first $\rightarrow$ (n_{11}, n_{12} in examples) and the unknown parameters → (p in examples)

→ we compute likelihood / log likelihood based on the complete data

→ We find $E(\mathcal{L} \mid$ known data like $n_1 \dots n_2)$

1 → we set $p = p^0$ to some value

2 → E step: we find $E_{p^0}(\mathcal{L} \mid \text{known data})$ → we estimate unknown data using p^0 & known data

Creativity is intelligence having fun. – Albert Einstein

3 → M step: we maximize $E_{p^0}(\mathcal{L} \mid \text{known data})$ wrt p to get p^1

→ Repeat step 2 & 3 until convergence

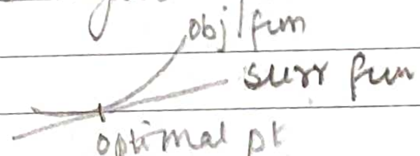
MM Algorithms

→ MM Algorithm is not an algorithm, but a prescription or principle for constructing optimization algorithms

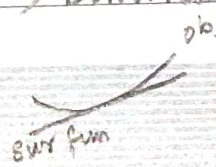
→ EM algorithm → spl case of MM algorithm

→ for minimization problems :- MM → majorize/minimize
" maximization " :- MM → minorize/maximize

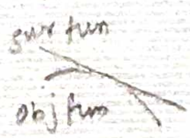
→ MM Algo works by creating a surrogate function that minorizes/majorizes the objective function. When surrogate function is optimized, the objective funⁿ is driven uphill/downhill as needed.
↳ Surrogate function is simpler than objective function
↳ Properties of surrogate funⁿ
1) Tangency



2) Domination requirement (Obj function always above surrogate function for minimization problem, opp. for maximization ")



maximization



minimization

Rationale for MM principle

↳ Can generate an algo that avoids matrix inversion
↳ Can separate parameters of a problem.
↳ ~~lin~~ can linearize the optimization problem
↳ can deal gracefully with equality/inequality constraints
↳ can Restore symmetry
↳ Can turn a ~~smo~~ non-smooth problem into a smooth problem

Majorization

1) A fun^n $g(\theta|\theta^n)$ is said to majorize the function $f(\theta)$ at θ^n provided

$f(\theta^n) = g(\theta^n|\theta^n)$ tangency at θ^n

$f(\theta) \leq g(\theta|\theta^n)$. domination

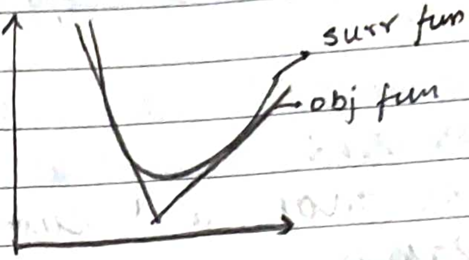
obj^{t} surrogate

The majorization relation ~~is~~ b/w fun^n is closed under sums, negative products & limits

2) A fun $g(\theta|\theta^n)$ is said to ~~mo~~ minorize function $f(\theta)$ at θ^n if $-g(\theta|\theta^n)$ majorizes at $-f(\theta)$

3) In minimization, we choose a majorizing fun^n $g(\theta|\theta^n)$ an minimize it to produce θ^{n+1}

<u>Example</u> (Quadratic Majorizer)



Quadratic function is the majorizer. We minimize it ~~it~~ to get the minimum of the objective fun^n

<u>Descent Property</u> → Makes MM very <u>stable</u>

→ $f(\theta^{n+1}) \le f(\theta^n) \xrightarrow[\text{when}]{\text{only}} g(\theta^{n+1}|\theta^n) = g(\theta^n|\theta^n)$ &
$f(\theta^{n+1}) = g(\theta^{n+1}|\theta^n)$

else,
$f(\theta^{n+1}) < f(\theta^n)$

↳ <u>Reason</u>

$f(\theta^{n+1}) \le g(\theta^{n+1}) \le g(\theta^n|\theta^n) = f(\theta^n)$

<u>Example</u>

Consider a sport league with m teams. Assign team i the skill level θ_i.

$$P(i \text{ beats } j) = \frac{\theta_i}{\theta_i + \theta_j}$$

Let b_{ij} = no. of times i beats j

$$L(\theta) = \prod_{(i,j)} \left(\frac{\theta_i}{\theta_i+\theta_j}\right)^{b_{ij}} \Rightarrow \log(L(\theta)) = \sum_{(i,j)} (b_{ij})\left[\log(\theta_i) - \log(\theta_i+\theta_j)\right]$$

$\Rightarrow \log(l(\theta)) = \sum_{(i,j)} b_{ij}\left[\log \theta_i - \log(\theta_i + \theta_j)\right]$ → maximize this to get θ

($\log(l(\theta))$ ↗ $f(\theta)$)

→ we need to separate parameters of $-\log(\theta_i + \theta_j)$ so that we can maximize the log likelihood easily

→ we use hyperplane property

$$h(y) \geq h(x) + \nabla h(x)^t (y-x)$$ of a convex fun $h(x)$

Let $h(x) = -\log(x)$

$$-\log(y) \geq -\log(x) - \frac{1}{x}(y-x)$$

→ Minorization $-\ln(y) \geq -\ln(x) - (y-x)x^{-1}$ produces

$$g(\theta|\theta^n) = \sum_{(i,j)} b_{ij}\left[\log \theta_i - \ln(\theta_i^n + \theta_j^n) - \frac{\theta_i + \theta_j}{\theta_i^n + \theta_j^n} + 1\right]$$

→ now, the optimal pt would be

$$\theta_i^{n+1} = \frac{\sum_{i \neq j} b_{ij}}{\sum_{i \neq j} (b_{ij} + b_{ji}) / (\theta_i^n + \theta_j^n)}$$

(by minimizing $\frac{d}{d\theta} g(\theta|\theta^n)$)

Catchup Meeting 4:- 9th June 2021

Task
→ Code out ABO problem using • EM, • NR & • fisher scoring
→ find Solution for Gaussean problem using E-M algo

Discussions
→ Found EM approach to Balls & Boxes problems and to ABO problem
→ Proved its convergence and convergence to the right value
→ Discussed and solved the gaussean problem

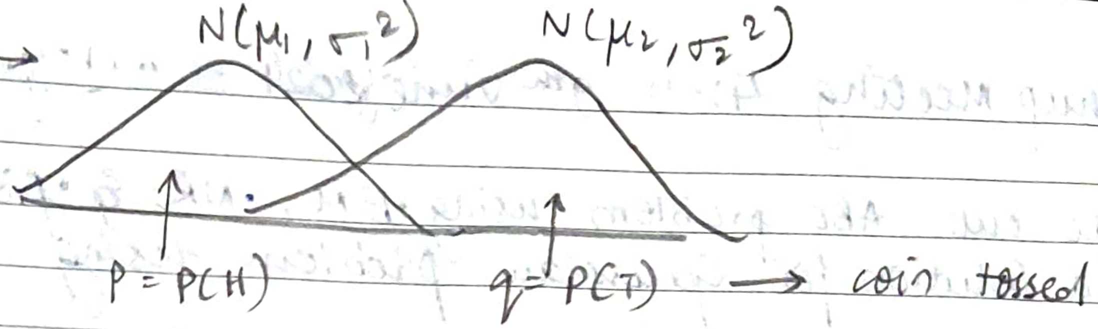
Q) We have two normal Distributions → $N(\underbrace{\mu_1, \sigma_1^2}), N(\underbrace{\mu_2, \sigma_2^2})$ unknown

We have a coin with $P(H) = p$ & $P(T) = 1-p = q$
If we get a head → we pick $N(\mu_1, \sigma_1^2)$ and choose a pt
" " " " tail → " " $N(\mu_2, \sigma_2^2)$ " " " "

We have n such observations → $x_1, x_2 \dots x_n$

We know → $x_1, x_2, \dots x_n$
We don't know → $\mu_1, \sigma_1, \mu_2, \sigma_2, p$
What is the MLE of p?

Solution →



let us assume that $\mu_1 \leq \mu_2$

Points:- $x_1, x_2 \dots x_n$ ⟶ original Data

Complete Data:- $\begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \begin{pmatrix} x_2 \\ y_2 \end{pmatrix} \dots \dots \begin{pmatrix} x_n \\ y_n \end{pmatrix}$

where $y_i = \begin{cases} 0 & \text{if } x_i \in N(\mu_1, \sigma_1^2) \\ 1 & \text{if } x_i \in N(\mu_2, \sigma_2^2) \end{cases}$

But $y_1, y_2 \dots y_n$ are unknown. Hence we use E-M Algorithm

$$L(c) = \prod_{i=1}^{n} f(x_i, y_i) = \prod_{i=1}^{n} \underbrace{(1-p)^{y_i} p^{(1-y_i)}}_{\text{marginal density of } y_i} \phi(x_i)^{(1-y_i)}, \psi(x_i)^{y_i}$$

here

$$\phi(x_i) = [\text{pdf of } N(\mu_1, \sigma_1^2) \text{ at } x_i] = \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{1}{2}\frac{(x_i-\mu_1)^2}{\sigma_1^2}}$$

$$\psi(x_i) = [\text{pdf of } N(\mu_2, \sigma_2^2) \text{ at } x_i] = \frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{1}{2}\frac{(x_i-\mu_2)^2}{\sigma_2^2}}$$

$$\log(L(c)) = \sum_{i=1}^{n} y_i \log(1-p) + (1-y_i)\log p + (1-y_i)\log\phi(x_i) + y_i \log\psi(x_i)$$

here,

$$\log\phi(x_i) = \log\frac{1}{\sqrt{2\pi}} - \log\sigma_1 - \frac{1}{2}\frac{(x_i-\mu_1)^2}{\sigma_1^2}$$

$$\log\psi(x_i) = \log\frac{1}{\sqrt{2\pi}} - \log\sigma_2 - \frac{1}{2}\frac{(x_i-\mu_2)^2}{\sigma_2^2}$$

$$\log L(c) = \sum_{i=1}^{n} y_i \log(1-p) + (1-y_i)\log p - (1-y_i)\left[\log\sigma_1 + \frac{1}{2}\frac{(x_i-\mu_1)^2}{\sigma_1^2}\right] - y_i\left[\log\sigma_2 + \frac{1}{2}\frac{(x_i-\mu_2)^2}{\sigma_2^2}\right] + c_0$$

$\hookrightarrow$ terms independent of $p, \mu_1, \mu_2, \sigma_1$ & σ_2

$$E(\log L(c))\Big|_{x_1\ldots x_n} = \log(1-p)\sum_{i=1}^{n} E(y_i|x_1\ldots x_n) + \log p \sum_{i=1}^{n} 1 - E(y_i|x_1\ldots x_n)$$
$$- \sum_{i=1}^{n}\left[\log\sigma_1 + \frac{1}{2}\frac{(x_i-\mu_1)^2}{\sigma_1^2}\right](1 - E(y_i|x_1\ldots x_n))$$
$$- \sum_{i=1}^{n}\left[\log\sigma_2 + \frac{1}{2}\frac{(x_i-\mu_2)^2}{\sigma_2^2}\right] E(y_i|x_1\ldots x_n) + c_0$$

EM Algorithm

→ Start with $p = p_0 \in (0,1)$, $(\mu_1^0)(\mu_2^0) \in (\min_x x_i, \max_x x_i)$
Such that $\mu_1^0 < \mu_2^0$ & $\sigma_1^0, \sigma_2^0 > 0$

E Step →

$$E_{p_0,\mu_1^0,\mu_2^0,\sigma_1^0,\sigma_2^0}(\log L(c)|x_1\ldots x_n) = \log(1-p)\sum_{i=1}^{n} y_i^0 + \log p(n-\Sigma y_i^0)$$
$$- \log\sigma_1(n-\Sigma y_i^0) - \frac{1}{2\sigma_1^2}\Sigma(x_i-\mu_1)^2(1-y_i^0)$$
$$- \log\sigma_2(\Sigma y_i^0) - \frac{1}{2\sigma_2^2}\Sigma(x_i-\mu_2)^2(y_i^0)$$

where
$$y_i^0 = E(y_i|x_1\ldots x_n) = E(y_i|x_i) = 0\cdot p_i + 1(1-p_i)$$
$$= 1-p_i$$
↳ conditional probability of $N(\mu_2,\sigma_2^2)$

$$y_i^0 = E_{p_0,\mu_1^0,\mu_2^0,\sigma_1^0,\sigma_2^0}(y_i|x_i) = P(x_i \text{ belongs to } N(\mu_2^0,(\sigma_2^0)^2))$$

$$= \frac{P(x_i|N(\mu_2^0,(\sigma_2^0)^2))\,P(N(\mu_2^0,\sigma_2^0))}{P(x_i|N(\mu_2^0,\sigma_2^{0\,2}))\,P(N(\mu_2^0,\sigma_2^0)) + P(x_i|N(\mu_1^0,\sigma_1^{0\,2}))\,P(N(\mu_1^0,\sigma_1^0))}$$
(Bayes theorem

$$= \frac{\phi(x_i)|_{\mu_2^0,\sigma_2^0}(1-p_0)}{\phi(x_i)|_{\mu_2^0,\sigma_2^0}(1-p_0) + p_0\,\phi(x_i)|_{\mu_1^0,\sigma_1^0}}$$

$$(y_i)^0 = \frac{(1-p_0)\frac{1}{\sqrt{2\pi}\sigma_2^0} e^{-\frac{1}{2}\frac{(x_i-\mu_2^0)^2}{(\sigma_2^0)^2}}}{\text{Numerator} + p_0 \frac{1}{\sqrt{2\pi}\sigma_1^0} e^{-\frac{1}{2}\frac{(x_i-\mu_1^0)^2}{(\sigma_1^0)^2}}}$$

M step $\rightarrow$ Maximizing $E_{p_0, \mu_1^0, \mu_2^0, \sigma_1^0, \sigma_2^0}(\log L(\theta) | x_1, \dots x_n)$

gives

$$\boxed{p^1 = \frac{n - \sum y_i^0}{n}} \longrightarrow \frac{\partial E}{\partial p} = 0$$

Set $\frac{\partial E}{\partial \mu_1} = 0 \Rightarrow \sum(x_i - \mu_1)(1 - y_i^0) = 0$

$$\sum(x_i - \mu_1) = \sum(x_i - \mu_1) y_i^0$$

$$\sum x_i - n\mu_1 = \sum x_i y_i^0 - \mu_1 \sum y_i^0$$

$$\mu_1 = \frac{\sum x_i y_i^0 - \sum x_i}{(\sum y_i^0) - n}$$

$$\boxed{\mu_1^1 = \frac{\sum x_i y_i^0 - \sum x_i}{\sum y_i^0 - n}}$$ $\rightarrow$ Gives mean of pts belonging to curve 1 $\rightarrow$ intuitively right

Set $\frac{\partial E}{\partial \mu_2} = 0 \Rightarrow \sum(x_i - \mu_2) y_i^0 = 0$

$$\boxed{\mu_2^1 = \frac{\sum x_i y_i^0}{\sum y_i^0}}$$

8 ~~Set~~ $\cancel{\frac{\partial E}{\partial \sigma_1} = 0 \Rightarrow}$ ~~tough to do directly~~

9

~~We can we use~~ $\cancel{\log(E_{p_0 \ldots \sigma_2^0}(\log L(c) \mid x_1 \ldots x_n))}$

10

~~Set~~ $\cancel{\frac{\partial \log E}{\partial \sigma_1} = 0 \Rightarrow \frac{\partial}{\partial \sigma_1} \log(\log(\sigma_1))}$

11

12 Set $\frac{\partial E}{\partial \sigma_1} = 0 \Rightarrow \frac{n - \sum y_i^0}{\cancel{\sigma_1}} = \left(\frac{1}{\sigma_1}\right)^{\cancel{3}2} \sum (x_i - \mu_1)^2 (1 - y_i^0)$

1

$$(\sigma_1)^2 = \frac{\sum (x_i - \mu_1)^2 (1 - y_i^0)}{n - \sum y_i^0}$$

2

$$\sigma_1^1 = \sqrt{\frac{\sum (x_i - \mu_1^1)^2 (1 - y_i^0)}{n - \sum y_i^0}}$$

3

4

5 Set $\frac{\partial E}{\partial \sigma_2} = 0 \Rightarrow \frac{\sum y_i^0}{\cancel{\sigma_2}} = \left(\frac{1}{\sigma_2}\right)^{\cancel{3}2} \sum (x_i - \mu_2)^2 (y_i^0)$

6

$$(\sigma_2)^2 = \frac{\sum (x_i - \mu_2)^2 (y_i^0)}{\sum y_i^0}$$

$$\sigma_2^1 = \sqrt{\frac{\sum (x_i - \mu_2^1)^2 (y_i^0)}{\sum y_i^0}}$$

8 **Catchup Meeting 6 :- 26 June 2021.**

Discussions
9 → Discussed convergence of EM algorithm
→ Discussed MM Algorithm & its convergence
10 → " convergence of IRWLS for LAD as a special case of MM algorithm

11

Task
→ Task to show that EM algorithm can be viewed
12 as MM algorithm spl case
→ Task to give a ranking metric for cricket teams which have played 1v1 using Bradley Terry Model

2

Bradley - Terry

3
→ Probability model that can predict the outcome of
4 a comparison
→ Given 'i' & 'j' individuals drawn from some population,
5 it estimates the pairwise comparison $i>j$ as

6

$$\boxed{P(i>j) = \frac{p_i}{p_i+p_j}}$$

where p_i is a real-valued score assigned to individual 'i'

→ A Bradley Terry Model can then be used to derive a full ranking.

Estimating the parameters 'p_i' of the Bradley-Terry Model

→ The algorithm that estimates the parameters of the Bradley-Terry model basically computes maximum likelihood estimate.

Ex:- We have 5 teams.

Team	A	B	C	D	E
Score	p_A	p_B	p_C	p_D	p_E

Given set of information :-

Team 1		Team 2		Win Ratio
A	Vs	B	→	$3-1 = W_{AB}-W_{BA}$
B	Vs	C	→	$2-0 = W_{BC}-W_{CB}$
C	Vs	D	→	$4-5 = W_{CD}-W_{DC}$
D	Vs	E	→	$0-4 = W_{DE}-W_{ED}$

Likelihood function → $L = \binom{4}{1}\left(\frac{p_A}{p_A+p_B}\right)^3\left(\frac{p_B}{p_A+p_B}\right)^1 \cdot \binom{2}{2}\left(\frac{p_B}{p_B+p_C}\right)^2 \cdot \binom{9}{5}\left(\frac{p_C}{p_C+p_D}\right)^4\left(\frac{p_D}{p_C+p_D}\right)^5 \cdot \binom{4}{4}\left(\frac{p_E}{p_E+p_D}\right)^4$

log likelihood function → $\log L = C + 3\log p_A + 3\log p_B + 4\log p_C + 5\log p_D + 4\log p_E - 4\log(p_A+p_B) - 9\log(p_C+p_D) - 2\log(p_B+p_C) - 4\log(p_E+p_D)$

(where C is the log of all constants)

E-M Algorithm

loses

wins	A	B	C	D	E
A	X	3	W_{AC}	W_{AD}	W_{AE}
B	1	X	2	W_{BD}	W_{BE}
C	W_{CA}	0	X	4	W_{CE}
D	W_{DA}	W_{DB}	5	X	0
E	W_{EA}	W_{EB}	W_{EC}	4	X

O → known/given/original data

O → unknown data

O + O → Complete data

$$L_C = C \times \prod_{i=1}^{n} \prod_{j=1, j\neq i}^{n} \frac{(p_i)^{w_{ij}}}{(p_i+p_j)^{w_{ij}}}$$

$$\log L_C = c_0 + \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \left(w_{ij} \log p_i - w_{ij} \log(p_i + p_j)\right)$$

Algorithm

→ Start with initial estimates of $p_1^0, p_2^0 \dots p_5^0$ such that keep $p_1 = p_1^0 = 1$. We find $p_2, \dots p_5$ wrt $p_1 = 1$

→ E-step → $E(\log L_C \mid \underbrace{\text{known data}}_{=X}) = E(c_0) + \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} [\log p_i - \log(p_i+p_j)] \, E(w_{ij} \mid X)$

$$E_{p_1^0, \dots p_5^0}(\log L_C \mid X) = E(c_0) + \sum\sum [\log p_i - \log(p_i^0 + p_j^0)] \cdot w_{ij}^0$$

where $w_{ij}^0 = E_{p_1^0, \dots p_5^0}(w_{ij} \mid X)$ for unknow w_{ij}

$w_{ij}^0 = w_{ij}$ for known w_{ij}

NOVEMBER						2021
M	T	W	T	F	S	S
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30					

using $n' =$ constant & $w_{ij} \sim Bin(n, \_)$ is a wrong approach as the unknown data would possibly overpower the known data. Also, if matches are not played then $E(w_{ij})$ should be zero when w_{ij} is unknown

now

$$w_{ij} \sim Bin\left(n, \frac{p_i}{p_i+p_j}\right)$$ where n is the number of matches played b/w i & j

now as we dont know n, let us set it to some +ve integer n' (let n' be avg matches played)

wrong

now $$E_{p_1^0 \dots p_5^0}(w_{ij} \mid X) = E\left(Bin\left(n', \frac{p_i^0}{p_i^0+p_j^0}\right)\right) = w_{ij}^0$$

$$w_{ij}^0 = \frac{n' p_i^0}{p_i^0+p_j^0}, \quad i<j$$

now, $$w_{ji}^0 = n' - \frac{n' p_i^0}{p_i^0+p_j^0} \quad (\text{as } w_{ij}^0 + w_{ji}^0 = n')$$

$$\therefore E_{p_1^0 \dots p_5^0}(\log L_c \mid X) = E_{p_1^0 \dots p_5^0}(C_0 \mid X) + \sum\sum_{i \neq j} [\log p_i - \log(p_i + p_j)] w_{ij}^0$$

wrong

where, w_{ij}^0 (when unknown) $= \begin{cases} n'\left(p_i^0 / p_i^0+p_j^0\right), & i<j \\ n' - n'\left(p_i^0 / p_i^0+p_j^0\right), & i>j \end{cases}$

(here n' = avg matches played by teams acc. to given data)

$w_{ij}^0 = w_{ij}$ (when known)

Another approach is setting the w_{ij}^0 unknown as 'zero' as, if there were matches played, it would have been known.

correct

Facts do not cease to exist because they are ignored. – Aldous Huxley

As there are no matches played b/w i & j, w_{ij}^0 is unknown, hence the expectation is zero.

correct

M step $\longrightarrow$ We need to maximize $E_{p_1^0, \dots p_5^0}(\log L_c \mid X)$ wrt $p_1 \dots p_5$

$$\frac{\partial}{\partial p_i} E_{p_1^0 \dots p_5^0}(\log L_c \mid X) = \sum_{j=1, j \neq i}^{5} \frac{w_{ij}^0}{1}\left(\frac{1}{p_i} - \frac{1}{p_i + p_j}\right) + w_{ji}^0\left(\frac{-1}{p_i + p_j}\right)$$

let us set it to zero

$$\sum \frac{w_{ij}^0}{p_i} = \sum \frac{w_{ij}^0 + w_{ji}^0}{p_i^0 + p_j^0}$$

$$p_i^1 = \frac{\sum w_{ij}^0}{\sum \frac{w_{ij}^0 + w_{ji}^0}{p_j^0 + p_i^0}}, \quad j \neq i, \ i \neq 1$$

$$p_i^1 = p_i^0 = 1 \quad \text{for } i = 1 \text{ (fixed)}$$

$E_{p_i^0}(w_{ij} \mid X) = 0$ when w_{ij} is unknown.

This is the correct approach

Convergence

$$p_i = \phi(p_i) = \frac{\sum w_{ij}}{\sum\left(\frac{w_{ij} + w_{ji}}{p_i + p_j}\right)}$$

$$\phi'(p_i) = \sum w_{ij} \cdot -\left(\sum \frac{w_{ij} + w_{ji}}{p_j + p_i}\right)^{-2} \cdot \left(\sum - \frac{w_{ij} + w_{ji}}{(p_i + p_j)^2}\right)$$

$$= \sum w_{ij} \cdot \frac{\sum\left(\frac{w_{ij} + w_{ji}}{(p_i + p_j)^2}\right)}{\left(\sum \frac{w_{ij} + w_{ji}}{p_i + p_j}\right)^2}$$

$\Rightarrow |\phi'(p_i)| < 1$

Hence convergence is guaranteed

NOVEMBER 2021

M	T	W	T	F	S	S
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30					

Week 47

Labour Thanksgiving Day (Japan)

8 ## Proof that EM is a special case of MM algorithm

9 Parameters to estimate :- θ | complete data :- $[x, z]$

known values/data :- x

10 Unknown " /data :- Z

11 likelihood function (complete) :- $\mathcal{L}_c(\theta)$

log " " " :- $\log \mathcal{L}_c(\theta) = \log p(x|\theta)$

12 $= \log E_{z|\theta}\, p(x|z,\theta)$

1 E-step:- we compute $g(\theta|\theta_t) = E_{\theta_t}(\log \mathcal{L}_c(\theta)|x)$

$= E_{z|x,\theta_t}(\log p(x,z|\theta))$

2 M-step:- we maximize $g(\theta|\theta_t)$ wrt θ to get θ_{t+1}

3 Applying Jensen's inequality

4 $$\log \mathcal{L}_c(\theta) = \log E_{z|\theta}\, p(x|z,\theta)$$

5 $$= \log E_{z|x,\theta_t} \frac{p(x|z,\theta)\, p(z|\theta)}{p(z|x,\theta_t)} \quad ||?$$

6

July 3 2021

Catchup Meeting 7

Discussion
- Discussed Bradley Terry Model for 1v1, Race Case, & Home Advantage case
- Discussed Proof of EM as a spl case of MM algo.
- Discussed the Bulb life problem

Always remember you are braver than you believe, stronger than you seem, and smarter than you think. – Christopher Robin

Task
- To formulize the Bulb life problem based on the given hint.