Assignment - 2
Machine Learning

Derive maximum likelihood estimators for:

1) parameter p, Bernoulli (p) sample of size n.

Let X be the Bernoulli random variable,

Sample size = n,

random sample is $X = \{x_1, x_2, x_3 \ldots\ldots x_n\}$

Probability density function, with parameter p given by,

$$f(x) = p^{x}(1-p)^{1-x} \qquad x = 0, 1$$

Likelihood of sample is given by,

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

$$= p^{\sum_{i=1}^{n} x_i} (1-p)^{n - \sum_{i=1}^{n} x_i}$$

Taking logs of both sides

$$= \ln p^{\sum_{i=1}^{n} x_i} (1-p)^{n - \sum_{i=1}^{n} x_i}$$

$$= \left(\sum_{i=1}^{n} x_i\right) \ln p + \left(n - \sum_{i=1}^{n} x_i\right) \ln(1-p)$$

Maximum likelihood can be found by differentiating w.r.t. P and equating the result to zero.

$$\frac{d}{dP}\left[\left(\sum_{i=1}^{n} x_i\right)\ln P + \left(n - \sum_{i=1}^{n} x_i\right)\ln(1-P)\right] = 0$$

$$\sum_{i=1}^{n} x_i \times \frac{1}{P} - \left(n - \sum_{i=1}^{n} x_i\right) \times \frac{1}{1-P} = 0$$

$$\frac{\sum_{i=1}^{n} x_i}{P} - \frac{\left(n - \sum_{i=1}^{n} x_i\right)}{1-P} = 0$$

Solving for P,

$$\frac{\sum_{i=1}^{n} x_i}{P} = \frac{n - \sum_{i=1}^{n} x_i}{1-P}$$

$$\frac{P}{1-P} = \frac{\sum_{i=1}^{n} x_i}{n - \sum_{i=1}^{n} x_i}$$

$$\frac{1-P}{P} = \frac{n - \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i}$$

$$\frac{-1}{p} - \frac{p}{p} = \frac{n}{\sum_{i=1}^{n} x_i} - \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i}$$

$$p = \frac{\sum_{i=1}^{n} x_i}{n}$$

This is maximum likelihood estimator (MLE) for p.

2) parameter p based on a Binomial (N, p) sample of size n. Compute your estimators if sample is $\{3, 6, 2, 0, 0, 3\}$ and $N = 10$

Assume x as any Binomial random variable with parameters N & p for size n,
let $x = \{x_1, x_2, x_3 \ldots x_n\}$ be a random sample.

Probability density function for the binomial distribution (N, P) is given

$$f(x) = \binom{N}{p} p^{x} (1-p)^{n-x} \quad (x = 0, 1 \ldots n)$$

Hence, likelihood function of the given sample is given by the product of all the random sample values x_i from 1 to n.

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

Taking the log likelihood,

$$\ln \left[\prod_{i=1}^{n} \binom{N}{x} p^{x_i} (1-p)^{N-x_i} \right]$$

Expanding log $= \sum_{i=1}^{n} \ln \binom{N}{x_i} + \left(\sum_{i=1}^{n} x_i \right) \ln(p)$

$$+ \left(nN - \sum_{i=1}^{n} x_i \right) \ln(1-p)$$

We ~~a~~ can now find the maximum likelihood by differentiating the above expression w.r.t. p & ~~eq~~ equating it to 0.

$$\frac{d}{dp} \left[\sum_{i=1}^{n} \ln \binom{N}{x_i} + \left(\sum_{i=1}^{n} x_i \right) \ln(p) + \ln(1-p) \left(nN - \sum_{i=1}^{n} x_i \right) \right] = 0$$

$$0 + \frac{\sum_{i=1}^{n} x_i}{p} - \frac{nN - \sum_{i=1}^{n} x_i}{1-p} = 0$$

$$\frac{\sum_{i=1}^{n} x_i}{p} = \frac{nN - \sum_{i=1}^{n} x_i}{1-p}$$

$$\frac{1-p}{p} = \frac{nN - \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i}$$

$$\frac{1}{p} - \frac{p}{p} = \frac{nN}{\sum_{i=1}^{n} x_i} - \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i}$$

$$p = \frac{\sum_{i=1}^{n} x_i}{nN}$$

This is the max. likelihood estimate (MLE) for binomial (N, p) for p.

Now we are given a sample (3, 6, 2, 0, 0, 3) and $N=10$, then sample size $n=6$.

Substituting the values in the above derived equation for p,

$$p = \frac{3+6+2+0+0+3}{6 \times 10} = \frac{14}{60}$$

$$p = \frac{7}{30}$$

3) Parameters a & b based on a uniform (a,b) sample of size n.

Maximum likelihood estimator for parameters a and b based on a uniform (a, b) sample of size n. Assumming X as my uniform random variable with parameters a & b.
For size n, let, $X = \{X_1, \dots X_2, \dots X_n\}$ be a random sample.

Probability density function for the uniform distributed for a & b is given by.

$$f(x) = \begin{cases} 1/b-a & a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

Hence, the likelihood function of the given sample is given by L

$$L(X|a,b) = \left(\frac{1}{b-a}\right)^n$$

Now to find the maximum likelihood function for L, we need to minimize the denominator of the function i.e

$$\min(b-a)$$

Even after minimizing the difference between the parameters 'a' & 'b' we need to keep all the samples, of $X = \{x_1 \dots x_n\}$ in the range a, b.

Maximum likelihood estimator for a & b, would be,

$$\hat{a} = \min(x_i) \qquad \hat{b} = \max(x_i)$$

These values give us minimum length as it is the smallest interval to include all the sample points of X.

Scanned with CamScanner

9) Parameter μ based on a normal (μ, σ^2) sample of size n : known variance σ^2 & unknown mean μ.

Let x be any sample normal random variable such that $(x_1 \dots x_n) \in x$ for n-size.

Probability Density Function for normal distribution with parameters μ & σ^2

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

Hence, likelihood function for normal (μ, σ^2) is given by,

$$L(x_1, \dots x_n | \mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Taking log likelihood on both sides,

$$\log L(x_1, \dots x_n | \mu) = \log \left[\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \right]$$

Solving, $\sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{(x_i - \mu)^2}{2\sigma^2}$

Now, to get maximum likelihood estimator, we need to differentiate L w.r.t. μ & equating the result to zero.

$$\frac{d}{d\mu}\left[\log L(x_1 \ldots x_n \mid \mu)\right] = \frac{d}{d\mu}\left[\sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{(x_i - \mu)^2}{2\sigma^2}\right]$$

$$\sum_{i=1}^{n} \frac{2(x_i - \mu)}{2\sigma^2} = 0 \rightarrow \sum_{i=1}^{n} (x - \mu) = 0$$

$$\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \mu = 0$$

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

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

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

5) Parameter σ based on a normal (μ, σ^2) sample of size n with known mean μ and unknown variance σ^2

Let X be any normal random variable for size n we have $\{X_1, \ldots X_n\} \in X$

Probability density function for normal distribution for parameters μ & σ^2 is:

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

$$= n\sigma^2 + \sum_{i=1}^{n} (x_i - \mu)^2 = 0.$$

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

Substituting $\mu = x$ from eq^n (2),

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

This is maximum likelihood estimator for σ^2

Let X be any normal random variable for size n, we have $\{X_1, \ldots X_n\} \in x$

Probability density function for normal distribution for parameters μ & σ^2 is

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

Likelihood function for normal (μ, σ^2) is given by

$$L(x_i \ldots x_n | \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right.$$

Taking log likelihood on both sides,

$$\log L(x_i, \ldots x_n | \sigma^2) = \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{(x - \mu)^2}{2\sigma^2}$$

For maximum likelihood estimator, we differentiate L w.r.t. σ^2 & equating the result with 0

$$\frac{d}{d\sigma^2}\left[\log L(x_1 \ldots x_n | \sigma^2)\right] = \frac{d}{d\sigma^2}\left[\sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{(x - \mu)^2}{2\sigma^2}\right]$$

$$0 = \frac{-n}{2\sigma^2} + \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2(\sigma^2)^2}$$

Multiplying both sides by $2(\sigma^2)^2$

$$0 = \left[\frac{-n}{2\sigma^2} + \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2(\sigma^2)^2}\right] 2(\sigma^2)^2$$

$$0 = -n\sigma^2 + \sum_{i=1}^{n} (x_i - \mu)^2$$

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

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

now, if u is known we can substitute $\mu = x$ in the above equation,

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

Thus, this is the maximum likelihood estimator for parameter σ^2

6) Parameters (μ, σ^2) based on a Nominal (μ, σ^2) sample of size n with unknown mean μ & variance σ^2

Probability density function for parameter μ & σ^2 is given by,

$$f(x_i \mid \mu, \sigma^2) = \frac{1}{\sqrt{\sigma^2}\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

Here, $-\infty < \mu < \infty$ and $0 < \sigma^2 < \infty$ holds

Now, likelihood function can be written as the

$$L(\mu, \sigma^2) = \prod_{i=1}^{n} \left(\frac{1}{\sqrt{\sigma^2}\sqrt{2\pi}}\right)^n e^{-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}}$$

$$= \prod_{i=1}^{n} (\sigma^2)^{-n/2} (2\pi)^{-n/2} \exp\left[\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right]$$

Taking the log likelihood on both sides,

$$\text{Log } L(\mu, \sigma^2) = \frac{-n}{2}\log(\sigma^2) - \frac{n}{2}\log(2\pi) - \sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}$$

Scanned with CamScanner

Now, taking partial derivative w.r.t. μ & equating the result to 0 to get MLE,

$$\frac{\partial}{\partial \mu} \cdot \left[\log L(\mu, \sigma^2) \right] = \frac{-2 \sum_{i=1}^{n} (x_i - \mu)(-1)}{2\sigma^2} = 0$$

$$0 = \sum_{i=1}^{n} \frac{(x_i - \mu)}{\sigma^2}$$

Multiplying both sides by σ^2

$$\sum_{i=1}^{n} (x_i - \mu) = 0$$

$$\sum_{i=1}^{n} (x_i - n\mu = 0$$

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

$$\hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n} \Rightarrow \bar{x} \quad -(2)$$

This is maximum likelihood estimator for unknown μ.

To find MLE, for σ^2; we take the partial derivative of equation ①, w.r.t. σ^2 & equating the result ~~to~~ with 0

$$\frac{\partial}{\partial \sigma^2}\left[\operatorname{Log} L(\mu, \sigma^2)\right] = \frac{-n}{2\sigma^2} + \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{2(\sigma^2)^2} = 0$$

Multiplying by $2(\sigma^2)^2$ we get,

$$0 = \left[\frac{-n}{2\sigma^2} + \frac{\sum (x_i - \mu)^2}{2(\sigma^2)^2}\right] \times 2(\sigma^2)^2$$

$$-n\sigma^2 + \sum_{i=1}^{n} (x_i - \mu)^2 = 0$$

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

Substituting $\mu = x$ from eqn ②,

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

This is the maximum likelihood estimator for σ^2

2) You are given a coin and a thumbtack & you perform the following experiment: toss both the thumbtack & coin 100 times. You get 60 heads & 40 tails for the coin. 70 heads & 30 tails for thumbtack. You put Beta priors Beta (1,1), Beta (40,60), Beta (30,70), Beta (100,100) Beta (1000,1000) & Beta (100,000, 1000,00) on the coin & the thumbtack, respectively.

1) Derive MLE & MAP estimator for both coin & the thumbtack

MLE estimate for coin.

α_H = number of heads = 60

α_T = number of tails = 40

$$\hat{\theta}_{MLE\,(coin)} = \frac{\alpha_H}{\alpha_H + \alpha_T} = \frac{60}{60+40} = \frac{60}{100} = 0.6$$

MLE estimate for thumbtack

$$\hat{\theta}_{MLE\,(thumbtack)} = \frac{\alpha_H}{\alpha_H + \alpha_T} = \frac{70}{70+30} = 0.7$$

Scanned with CamScanner

MAP estimates for coin & thumbtack,

$$\hat{\theta}_{MAP} = \arg\max_{\theta} P(\theta/D) = \frac{\alpha_H + \beta_H - 1}{\alpha_H + \beta_H + \alpha_T + \beta_T - 2}$$

$$= \cancel{\frac{60+40-1}{60+40}}$$

① For $\beta = (1,1)$

$\alpha_H = 60$, $\alpha_T = 40$, $\beta_H = 1$, $\beta_T = 1$

$$\hat{\theta}_{MAP} = \frac{60+1-1}{60+1+40+1-2}$$

$$= \frac{60}{100} = 0.6$$

② For $\beta(40,60)$ $\beta_H = 40$, $\beta_T = 60$

$$\hat{\theta}_{MAP} = \frac{60+40-1}{60+40+40+60-2}$$

$$= \frac{99}{200-2} = \frac{99}{198} = 0.5$$

③ For $\beta = (30, 70)$ $\beta_H = 30$, $\beta_T = 70$
$\alpha_H = 60$, $\alpha_T = 40$.

$$\hat{\theta}_{MAP} = \frac{60+30-1}{60+30+40+70-2} = \frac{89}{198}$$

$$= 0.4494$$

④ For $\beta = (100, 100)$, $\beta_H = 100$, $\beta_T = 100$
$\alpha_H = 60$, $\alpha_T = 40$

$$\hat{\theta}_{MAP} = \frac{100+60-1}{100+60+100+40-2}$$

$$= \frac{160-1}{300-2} = \frac{159}{298} = 0.5335$$

⑤ For $\beta = (1000, 1000)$ $\beta_H = 1000$, $\beta_T = 1000$
$\alpha_H = 60$, $\alpha_T = 40$

$$\hat{\theta}_{MAP} = \frac{1000+60-1}{1000+60+1000+40-2} = \frac{1059}{2098}$$

$$= 0.5$$

⑥ $\beta = (10000, 10000)$, $\beta_H = 10000$, $\beta_T = 10000$

$\alpha_H = 60$, $\alpha_T = 40$

$$\hat{\theta}_{MAP} = \frac{10000 + 60 - 1}{10000 + 60 + 10000 + 40 - 2}$$

$$= \frac{10059}{20,098} = 0.5$$

Now, For thumbtack,

① $\beta = (1, 1)$ $\beta_H = 1$, $\beta_T = 1$

$\alpha_H = 70$, $\alpha_T = 30$

$$\hat{\theta}_{MAP} = \frac{70 + 1 - 1}{70 + 1 + 30 + 1 - 2} = \frac{70}{100} = 0.7$$

② $\beta = (40, 60)$, $\beta_H = 40$ $\beta_T = 60$

$\alpha_H = 70$, $\alpha_T = 30$

$$\hat{\theta}_{MAP} = \frac{70 + 40 - 1}{40 + 70 + 60 + 30 - 2} = \frac{109}{198}$$

$$= 0.5505$$

③ $\beta = (30, 70)$, $\alpha_H = 70$, $\alpha_T = 30$,

$$\hat{\theta}_{MAP} = \frac{30 + 70 - 1}{30+70+70+30-2} = \frac{99}{198} = \frac{1}{2}$$

$$= 0.5$$

④ $\beta = (100, 100)$, $\alpha_H = 70$, $\alpha_T = 30$

$$\hat{\theta}_{MAP} = \frac{70 + 100 - 1}{70+100+30+100-2} = \frac{169}{298} = 0.567$$

⑤ $\beta = (1000, 1000)$, $\alpha_H = 70$, $\alpha_T = 30$

$$\hat{\theta}_{MAP} = \frac{1000 + 70 - 1}{1000+70+1000+30-2}$$

$$= \frac{1069}{2098}$$

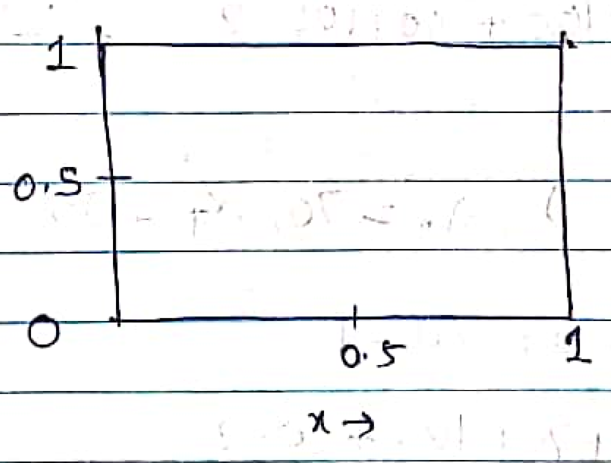
$$= 0.5095$$

(6) $\beta = (10000, 10000)$, $\alpha_H = 70$, $\alpha_T = 30$

$$\hat{\theta}_{MAP} = \frac{10000 + 70 - 1}{10000 + 70 + 10000 + 30 - 2}$$

$$= \frac{10069}{20098} \simeq 0.5$$

2) $\beta = (1, 1)$ Prior

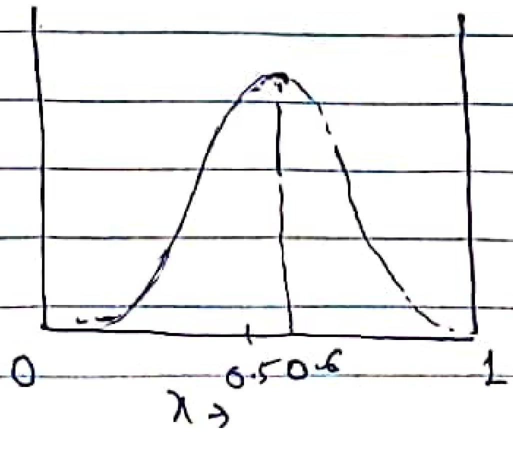


$\beta_H = 1$

$\beta_T = 1$

It is straight line graph ~~a~~ $y = 1$. & ends at $x = 1$.

Posterior Beta (61, 40)

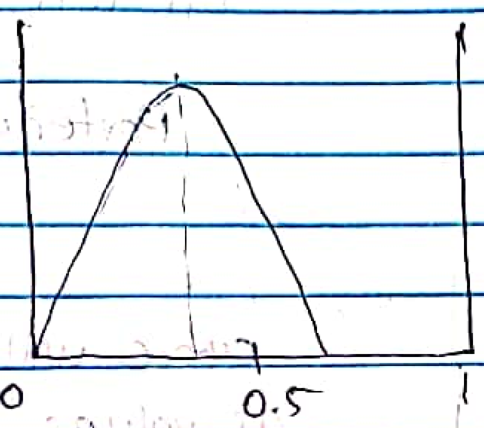


$\alpha_H + \beta_H = 61$

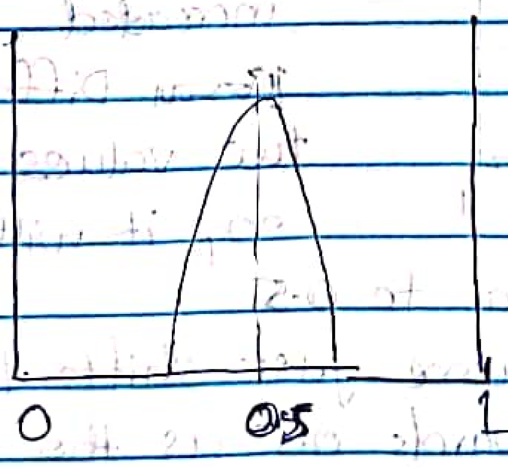
$\alpha_T + \beta_T = 41$

The graph is shifted towards 1. This is because, the priors $\beta(1,1)$ have decreased & the posterior $(61,41)$ has considerable difference.
Hence, as the prior value decreases, the parameter values are affected.

2.

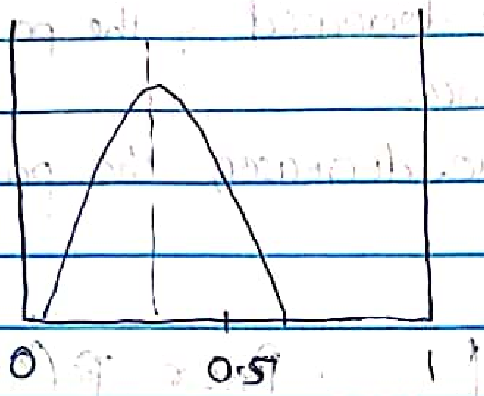


Prior $\beta(40,60)$
$\beta_H < \beta_T$ - graph will be on left of 0.5
graph will be wider compared to values in hundreds



Posterior
$\beta(100,100)$
~~Graph~~ Curve will be on the 0.5 & narrower as the values are higher
Not much dominated by prior.

3 $\beta(30,70)$ Prior

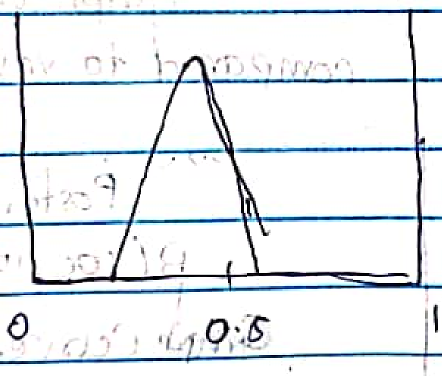


curve will be ~~o~~ wider as values are less than values in hundreds.

$\beta_H < \beta_T$ so, curve will be on left of 0.5

Posterior

$\beta(90,110)$



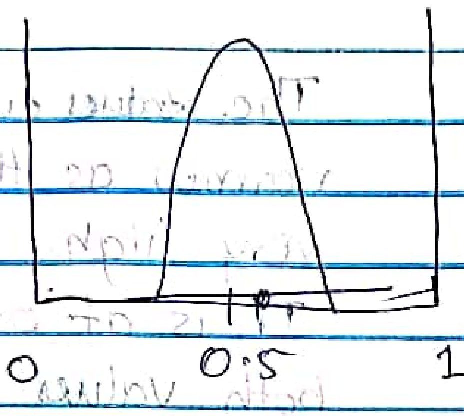
curve will be narrower as values have ~~will increasing~~ increased.

~~The~~ Difference between two values is less so, it will be ~~a~~ little near to 0.5

Here, ~~O~~ the curve just shifts little to right towards 0.5 as the values tend to become same/close.

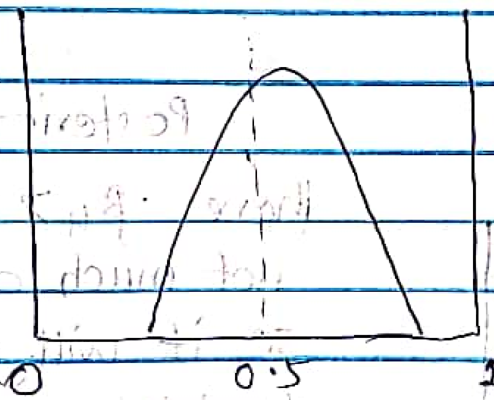
(4) $\beta(100,100)$

prior



curve at 0.5 as the values are same it narrower as values are higher.

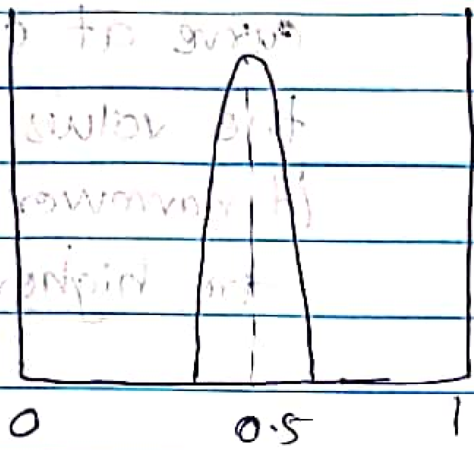
posterior $\beta(160,140)$



curve ~~also~~ near 0.5, ~~more~~ shifts near 1 to right as $\beta_H > \beta_T$

~~Not much~~ dominated by prior as there is little shift in position of curve.

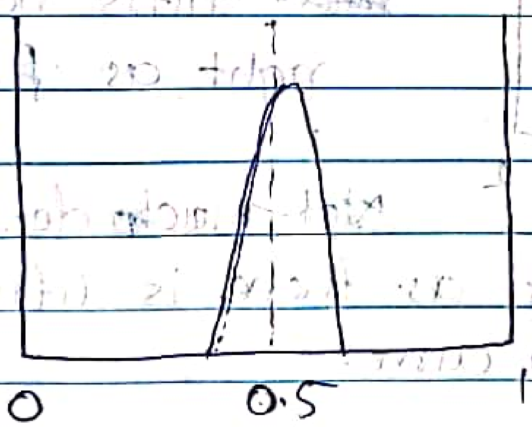
5 $\beta(1000, 1000)$ Prior



The ~~values~~ curve is narrow as the values a[re] very high.
It is at 0.5 as the both values are same

posterior

$\beta(1060, 1040)$ Posterior

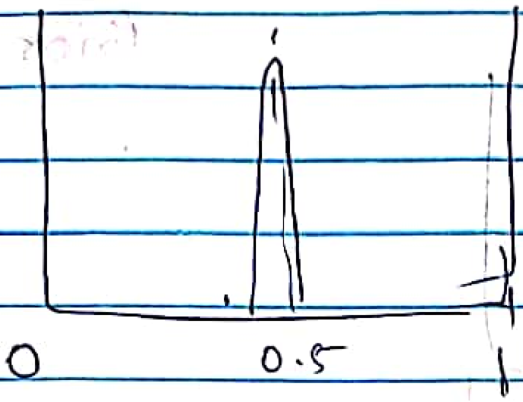


here $\beta_H > \beta_T$ but not much difference so, it will just shift to right.
It is majorly dominate[d] by prior.

11

⑥ $\beta(10000, 10000)$

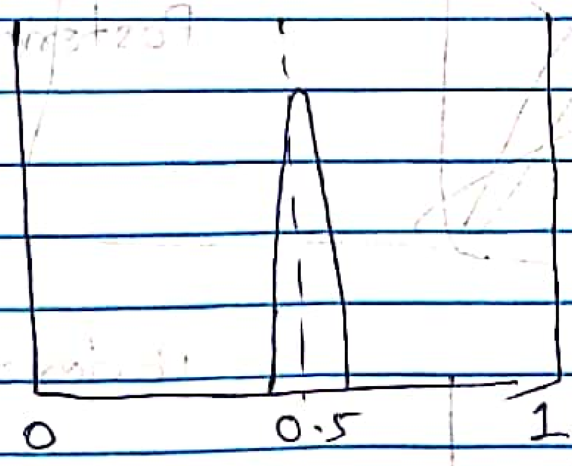
Prior



It is narrower as values are too high
It is at 0.5 as values are same

$\beta(10060, 10040)$

Posterior

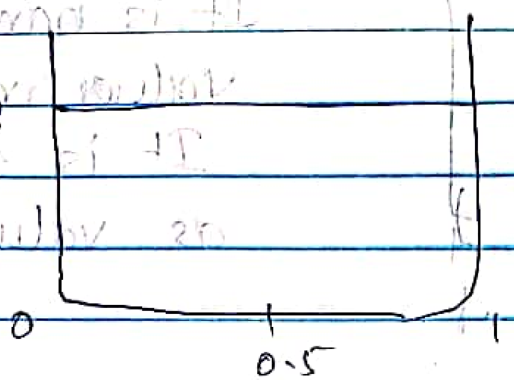


It will same as prior except, it will be little towards right as $P_H > P_T$
Majorly dominated by prior as data values don't make much difference

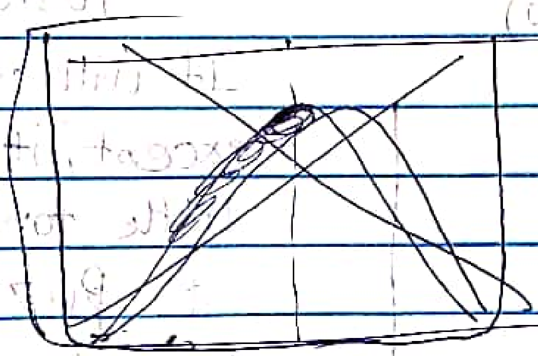
for thumbtack - ~~We will just write~~

① $\beta(1,1)$

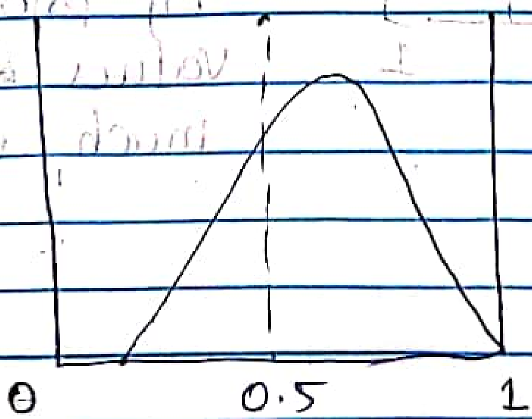
Prior



Posterior



Posterior

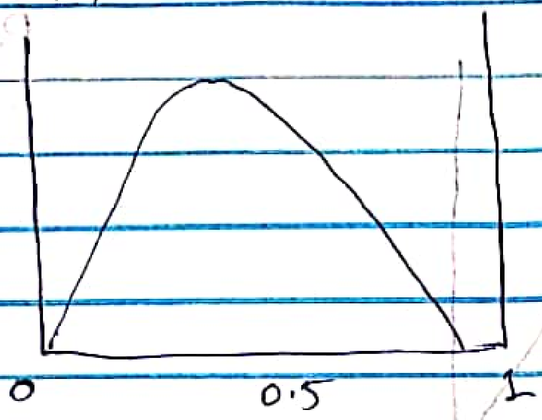


$\beta_H > \beta_T$ so it will be on right of 0.5

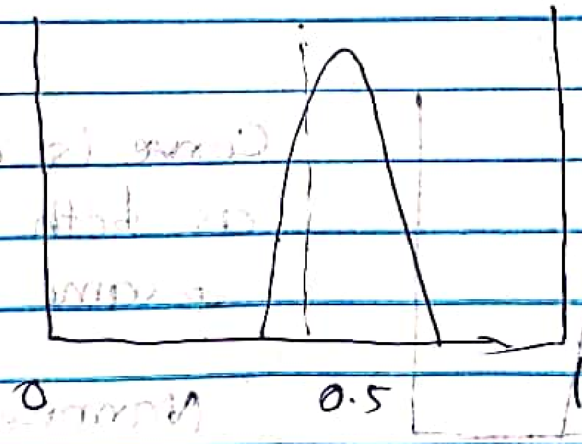
It is dominated by data.

② β(40, 60)

Prior



β(110, 90)



Posterior

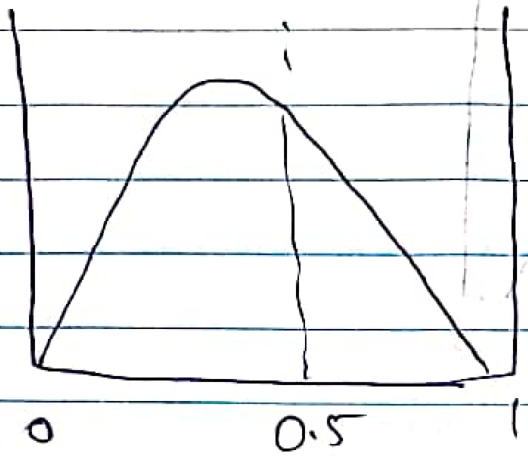
Values are closer so curve is near to ·0.5, it is narrower as the values are much higher

Curve. – Dominated by data

Scanned with CamScanner

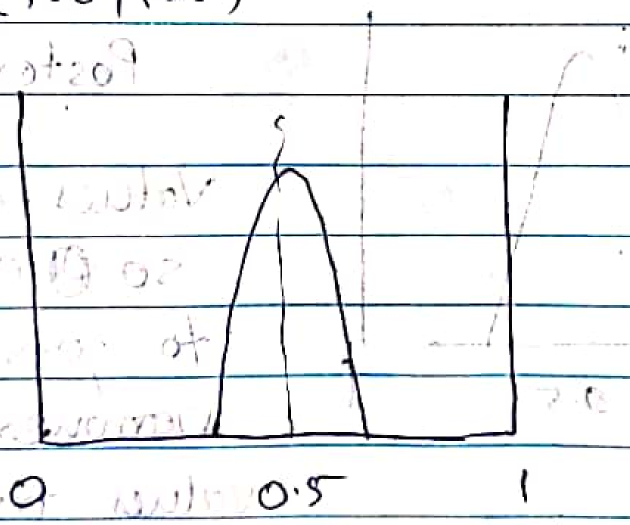
(3) β(30,70)

Prior



β(100,100)

Posterior



Curve is at 0.5 as both values are same.

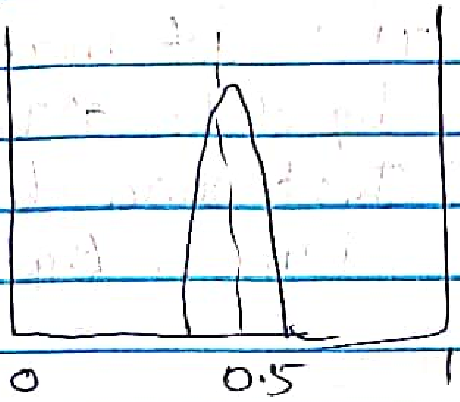
Narrower as the value are higher

Majorly dominated by data as values of data are higher.

Scanned with CamScanner

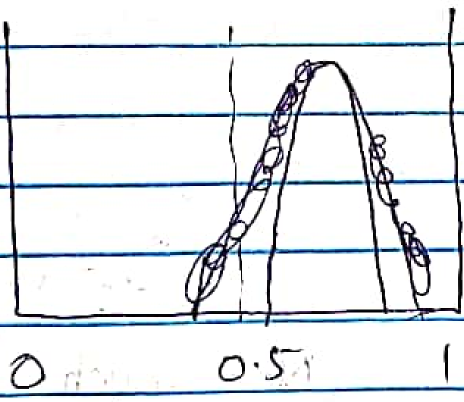
(4) $\beta(100, 100)$

Prior



$\beta(170, 130)$

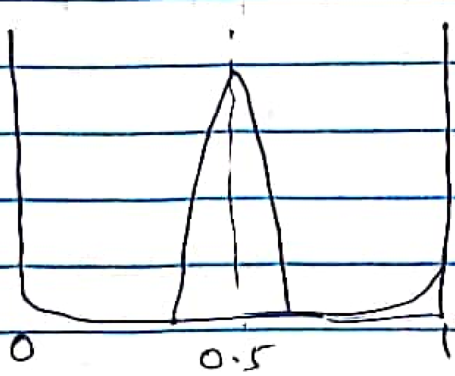
Posterior



It is majorly dominated by data as shifts to right because of change in difference betn both values. It becomes narrower as values increase.

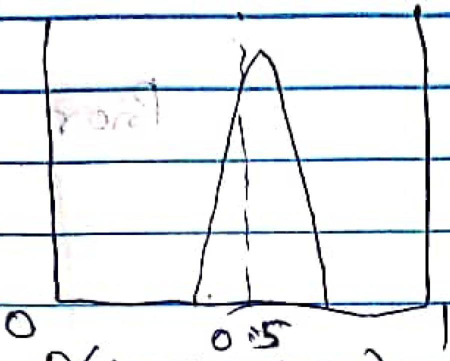
(5) Prior

$\beta(1000, 1000)$



Scanned with CamScanner

(5)

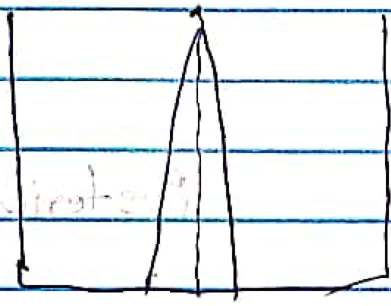


$\beta(10070, 1030)$

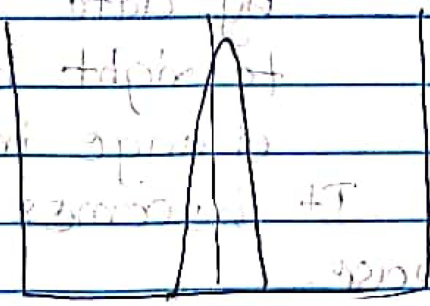
~~Prior~~ Posterior:

It is not much dominated by data. ~~as it is~~ Just curve shifts to right as $\beta_H > \beta_T$

(6)



Prior



Posterior

Not much dominated by data. Just curve shifts to right because $\beta_H > \beta_T$.

3) True ~~False~~. As MLE depends only on data & not on the priors but, as we increase taking the number of trails more & more, the part of the prior will become lesser & lesser dominant. so, MLE will becomes equal to MAP.

5) True, as the priors in the MAP are very high they dominate the values of data which are in MLE, so, their MLE are different but MAPs are same because of dominating priors.