**Ahmed Shahabaz**
**U 8941-5490**
**shahabaz@usf.edu**

## Ans. to the question 1 (a)

H.W-1

Ans to the Ques 1

(a). $$P(\vec{x} \mid \mu, \sigma^2) = \prod_{n=1}^{N} \frac{e^{\frac{-(x_n-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}$$

$$= \boxed{\frac{e^{\frac{-1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2}}{\left(\sqrt{2\pi\sigma^2}\right)^N}}$$

This a non-tractable function

As it has exponents. To make it tractable we will take the logarithm.

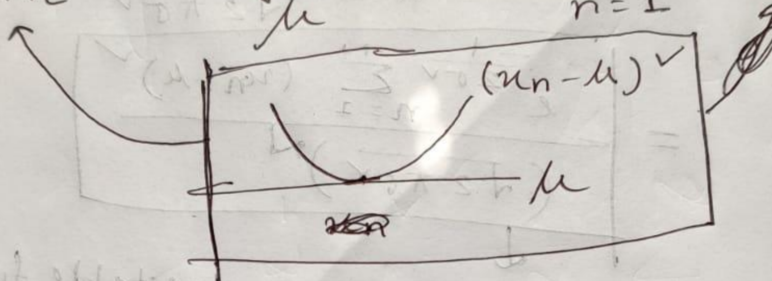
$$\therefore \ln P(\vec{x} \mid \mu, \sigma^2) = \ln\left[\frac{e^{\frac{-1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2}}{(2\pi\sigma^2)^{N/2}}\right]$$

$$= -\frac{1}{2\sigma^2}\sum_{n=1}^{N} e\,(x_n-\mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi)$$

Date :

$$\mu_{ML} = \arg\max_{\mu} \; -\frac{1}{2\sigma^2} \sum_{n=1}^{N} (x_n - \mu)^2$$

$$\mu_{ML} = \arg\min_{\mu} \; \frac{1}{2\sigma^2} \sum_{n=1}^{N} (x_n - \mu)^2$$




* we want to find the minimum. So we will take the derivative with respect to $\mu$.

$$\frac{\delta}{\delta \mu} \frac{1}{2\sigma^2} \sum_{n=1}^{N} (x_n - \mu_{ML})^2 = 0$$

$$\Rightarrow \frac{1}{2\sigma^2} \sum_{n=1}^{N} 2(x_n - \mu_{ML})(-1) = 0$$

$$\Rightarrow \sum_{n=1}^{N} (x_n - \mu_{ML}) = 0$$

Date :

$$\Rightarrow \sum_{n=1}^{N} \mu_{ML} = \sum_{n=1}^{N} x_n$$

$$\Rightarrow N \mu_{ML} = \sum_{n=1}^{N} x_n$$

$$\Rightarrow \boxed{\mu_{ML} = \frac{1}{N} \sum_{n=1}^{N} x_n}$$

Now e ML estimation for $\sigma^2$;

$$\frac{\partial}{\partial \sigma^2} \left[ \frac{1}{2\sigma^2_{ML}} \sum_{n=1}^{N} (x_n - \mu_{ML})^2 + \frac{N}{2} \ln \sigma^2 \right] = 0$$

$$\Rightarrow -\frac{1}{2\sigma^4_{ML}} \sum_{n=1}^{N} (x_n - \mu_{ML})^2 + \frac{N}{2} \frac{1}{\sigma^2_{ML}} = 0$$

$$\Rightarrow -\frac{1}{2\sigma^2_{ML}} \left[ \frac{1}{\sigma^2_{ML}} \sum_{n=1}^{N} (x_n - \mu_{ML})^2 \; N \right] = 0$$

Date :

$\therefore \frac{-1}{2\sigma_{ML}^2} = 0$

it is not possible..

OR,

$\frac{1}{\sigma_{ML}^2} \sum_{n=1}^{N} (x_n - \mu_{ML})^2 - N = 0$

$\Rightarrow \frac{1}{\sigma_{ML}^2} \sum_{n=1}^{N} (x_n - \mu_{ML})^2 = N$

$\Rightarrow \sigma_{ML}^2 = \frac{1}{N} \sum_{n=1}^{N} (x_n - \mu_{ML})^2$

## Ans. to the question 1 (b)

Ans to the Qu No 1

(b) ML estimate $\hat{\mu}_{ML} = \frac{1}{N} \sum_{n=1}^{N} x_n$

$$\Rightarrow \boxed{\mu_{ML} = E[x] = \mu}$$

$$E[\hat{\mu}_{ML}] = \frac{1}{N} \sum_{n=1}^{N} E[x_n]$$

$$\Rightarrow E[\hat{\mu}_{ML}] = \frac{1}{\cancel{N}} \cancel{N} \cdot \mu$$

$$\Rightarrow E[\hat{\mu}_{ML}] = \mu$$

∴ ML estimate ~~of~~ $\hat{\mu}_{ML}$ is unbiased. As it is equal to the expected value of $X$.

## Ans. to the question 1 (c)

Ans to the Qu No 1

Date :

(c) $$\hat{\mu}_{MAP} = \arg\max_{\mu} \ln p(\vec{x}|\mu) + \ln p(\mu)$$

$$\hat{\mu}_{ML} = \arg\max_{\mu} \ln p(x|\mu) = \frac{1}{N}\sum_{n=1}^{N} x_n$$

$$p(\mu) = \mathcal{N}(\mu_0, \sigma_0^2)$$

$$= \frac{e^{-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}}}{\sqrt{2\pi\sigma_0}}$$

$$\ln p(\mu) = -\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2 - \frac{1}{2}\ln 2\pi - \frac{1}{2}\ln\sigma_0^2$$

$$\ln p(\vec{x}|\mu) = \frac{-1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi)$$

Ans to the Que 1

(a) $\hat{\mu}_{MAP} = \underset{\mu}{\arg\max}\Big[-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\ln\sigma^2$

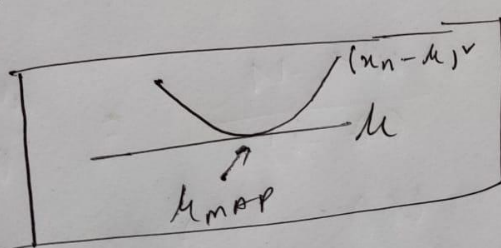
$- \frac{N}{2}\ln(2\pi) - \frac{1}{2\sigma_0^2}(\mu-\mu_0)^2 - \frac{1}{2}\ln(2\pi)$

$- \frac{1}{2}\ln\sigma_0^2\Big]$

* Replacing the values of $\ln p(\vec{x}|\mu)$ and $\ln p(\mu)$

$$\equiv \underset{\mu}{\arg\max}\Big[-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\Big]$$

$$\Rightarrow \hat{\mu}_{MAP} \equiv \underset{\mu}{\arg\min}\Big[\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 + \frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\Big]$$




* So we want to find the value of $\mu$ that minimizes the above eqn. So we will take the derivative w.r.t. $\mu$.

Date :

$$\frac{\delta}{\delta\mu}\left[\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2+\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right]=0$$

$$\Rightarrow \frac{\delta}{\delta\mu}\left[\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2\right]+\frac{\delta}{\delta\mu}\left[\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right]=0$$

$$\Rightarrow \frac{1}{2\sigma^2}\sum_{n=1}^{N}2(x_n-\mu)(-1)+\frac{1}{2\sigma_0^2}\cdot 2\cdot(\mu-\mu_0).(1)=0$$

$$\Rightarrow \frac{-1}{\sigma^2}\sum_{n=1}^{N}(x_n-\mu)+\frac{1}{\sigma_0^2}(\mu-\mu_0)=0$$

$$\Rightarrow \frac{-1}{\sigma^2}\sum_{n=1}^{N}x_n+\frac{1}{\sigma^2}\sum_{n=1}^{N}\mu+\frac{1}{\sigma_0^2}\mu-\frac{1}{\sigma_0^2}\mu_0=0$$

$$\Rightarrow \frac{-1}{\sigma^2}\cdot\frac{1}{N}\sum_{n=1}^{N}x_n+\frac{1}{\sigma^2}\cdot\frac{1}{\cancel{N}}\cdot\cancel{N}\mu+\frac{1}{N\sigma_0^2}\mu-\frac{1}{N.\sigma_0^2}\mu_0=0$$

[Dividing both sides by N]

$$\Rightarrow \frac{-1}{\sigma^2}\cdot\hat{\mu}_{ML}+\frac{\mu}{\sigma^2}+\frac{1}{N\sigma_0^2}\mu-\frac{1}{N\sigma_0^2}\mu_0=0$$

$$\left[\because \hat{\mu}_{ML}=\frac{1}{N}\sum_{n=1}^{N}x_n\right]$$

Date :..............................

$$\Rightarrow \mu \left( \frac{1}{\sigma^2} + \frac{1}{N\sigma_0^2} \right) = \frac{1}{\sigma^2} \hat{\mu}_{ML} + \frac{1}{N\sigma_0^2} \mu_0$$

$$\Rightarrow \mu \left( \frac{N\sigma_0^2 + \sigma^2}{\cancel{N\sigma_0^2 \sigma^2}} \right) = \frac{N\sigma_0^2 \hat{\mu}_{ML} + \sigma^2 \mu_0}{\cancel{N\sigma_0^2 \sigma^2}}$$

$$\Rightarrow \mu = \frac{N\sigma_0^2 \hat{\mu}_{ML} + \sigma^2 \mu_0}{N\sigma_0^2 + \sigma^2}$$

$$\Rightarrow \mu = \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2} \hat{\mu}_{ML} + \frac{\sigma^2}{N\sigma_0^2 + \sigma^2} \cdot \mu_0$$

$$\Rightarrow \mu = a \hat{\mu}_{ML} + b \mu_0$$

$$\therefore \boxed{a = \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}}$$

$$\boxed{b = \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}}$$

## Ans. to the question 2 (a)

Ans to the Que 2(a)

Date :

~~Likelihood~~,

Likelihood $\rightarrow$ $P(D|\theta) = {}^{N}C_{N_H}\ \theta^{N_H} (1-\theta)^{N-N_H}$

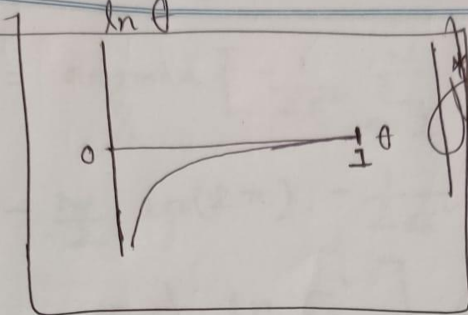
where $P(D|\theta)$ follows Binomial distribution. $N$ - is total # of trials and $N_H$ is the # of hits for a particular player.

$$\ln P(D|\theta) = \ln\left[{}^{N}C_{N_H}\ \theta^{N_H} (1-\theta)^{N-N_H}\right]$$

$$\Rightarrow \ln P(D|\theta) = \ln({}^{N}C_{N_H}) + N_H \ln\theta + (N-N_H)\ln(1-\theta)$$

* we took logarithm of the likelihood to make the problem tractable by ~~eliminati~~ ~~removi~~ simplifying the exponents. Now the derivative of the above eq$^n$ with respect to $\theta$ should be zero. as we want to maximize the value of $\theta$.

Date :




when $\theta = 0$, $\ln\theta$

* when $\theta$ ~~is~~ $0$, $\ln\theta = -\alpha$
* when $\theta = 1$, $\ln\theta = 0$

$$\hat{\theta} \doteq \frac{\partial}{\partial\theta}\Big[\ln\left({}^{N}C_{N_H}\right) + \left(\ln\theta\right)^{N_H} + (N - N_H)\ln(1-\theta)\Big] = 0$$

$$\Rightarrow \frac{1}{\theta}\cdot N_H + \frac{N-N_H}{1-\theta}(-1) = 0$$

$$\Rightarrow \frac{N_H}{\theta} = \frac{N-N_H}{(1-\theta)}$$

$$\Rightarrow N_H - N_H\theta = \theta N - N_H\theta$$

$$\Rightarrow \theta N = N_H$$

$$\Rightarrow \theta = \frac{N_H}{N} = \hat{\theta}$$

## Ans. to the question 2 (b)

Ans to the Ques 2 (b)

(b) $P(\theta \mid a_0, b_0, D) \propto P(D \mid \theta)\, P(\theta \mid a_0, b_0)$

where, $P(\theta \mid a_0, b_0) = Beta(\theta \mid a_0, b_0)$

$P(D \mid \theta)$ follows Binomial distribution.

$$P(\theta \mid a_0, b_0, D) \propto P(D \mid \theta)\, P(\theta \mid a_0, b_0)$$

$$= {}^{N}C_{N_H}\, \theta^{N_H} (1-\theta)^{N-N_H}\, \theta^{a_0-1} (1-\theta)^{b_0-1}$$

$$\propto {}^{N}C_{N_H}\, \theta^{N_H + a_0 - 1} (1-\theta)^{(N-N_H)+b_0-1}$$

Let's, Denote, $a_N = N_H + a_0$

$b_N = (N - N_H) + b_0$

$$\therefore P(\theta \mid a_0, b_0, D) \propto {}^{N}C_{N_H}\, \theta^{a_N - 1} . (1-\theta)^{b_N - 1}$$

$$= {}^{N}C_{N_H}\, Beta(a_N, b_N).$$

Date :..............................

where, ~~$N_{C_{NH}}$~~ $N_{C_{N_H}}$ is a constant term.

So, we can write,

$$P(\theta \mid a_0, b_0, D) \propto \text{Beta}(a_N, b_N)$$

So, Beta distribution is a conjugate prior for Binomial distribution.

$a_N$ = updated # of Hits
$b_N$ = updated # of misses.

$a_0$ = prior # of Hits
$b_0$ = prior # of Misses.

in our case, $a_0 = 100$
$b_0 = 300$

Date :

$P(\theta \mid a_0, b_0, D) \propto \theta^{a_N - 1} (1-\theta)^{b_N - 1}$

Again the ~~probl~~ to make the above eq^n easily solveable we will take the logarithm of it.

So,

$\hat{\theta}_{MAP} = \hat{\theta} = \arg\max_{\theta} \ln P(\theta \mid a_0, b_0, D)$

$\propto \arg\max_{\theta} \ln \left[ \theta^{a_N - 1} (1-\theta)^{b_N - 1} \right]$

$= \arg\max_{\theta} (a_N - 1) \ln \theta + (b_N - 1) \ln(1-\theta)$

to maximize the value of $\theta$ we want to ~~get~~ make the derivative of the above eq^n $\hat{\theta}$ equals to 0

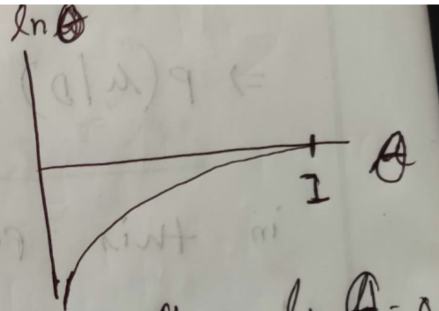
So,

$\hat{\theta}_{MAP} = \frac{\delta}{\delta \theta} \left[ (a_N - 1) \ln \theta + (b_N - 1) \ln(1-\theta) \right] = 0$

$(a_N - 1) \frac{1}{\theta} + (b_N - 1) \frac{1}{1-\theta} (-1) = 0$

$\Rightarrow (a_N - 1)$

* we want to maximize the value of $\theta$. so we want to ~~make~~ get the derivative of the above eq$^n$ to be zero.




* when $\theta = 1$, $\ln\theta = 0$
* when $\theta = 0$, $\ln\theta = -\infty$

$$\Rightarrow (a_N-1)\frac{1}{\theta} - (b_N-1)\frac{1}{1-\theta} = 0$$

$$\Rightarrow \frac{(a_N-1)(1-\theta)-(b_N-1)\theta}{\theta(1-\theta)} = 0$$

$$\Rightarrow \frac{a_N - \theta a_N - 1 + \theta - b_N\theta + \theta}{\theta(1-\theta)} = 0$$

$$\Rightarrow \theta(2-a_N-b_N) = 1-a_N$$

$$\Rightarrow \theta = \frac{a_N-1}{a_N+b_N-2} = \hat{\theta}_{MAP}$$

$$\Rightarrow \hat{\theta}_{MAP} = \frac{N_H+a_0-1}{(N_H+a_0)+(N-N_H+b_0)-2}$$

[Replacing ~~the~~ values of $a_N$ and $b_N$]

$$\Rightarrow \hat{\theta}_{MAP} = \frac{N_H+a_0-1}{a_0+b_0+N-2}$$

Date :

*** point to be Noted:**

when writing the following,

$$P(\theta \mid a_0, b_0, D) \propto Beta(a_N, b_N)$$

we ignored the term $\left({}^{N}C_{N_H}\right)$.

So even if we kept that term. while taking the derivative we would have gotten zero as there is no $\theta$ in that term. i.e.

$$\frac{\delta}{\delta \theta}\left({}^{N}C_{N_H}\right) = 0$$

**Comparison between MLE and MAP:**

| First name | Last name | N | $N_H$ | MLE | MAP |
|---|---|---|---|---|---|
| Clifford | Bartosh | 1 | 1 | 1 | 0.25062656641604 |
| Adam | Bernero | 1 | 1 | 1 | 0.25062656641604 |
| Bronson | Arroyo | 1 | 0 | 0 | 0.24812030075188 |
| James | Baldwin | 1 | 0 | 0 | 0.24812030075188 |
| Clint | Barmes | 350 | 101 | 0.288571428571429 | 0.267379679144385 |
| Michael | Barrett | 424 | 117 | 0.275943396226415 | 0.262773722627737 |
| Jason | Bartlett | 224 | 54 | 0.241071428571429 | 0.245980707395498 |
| Jayson | Werth | 337 | 79 | 0.234421364985163 | 0.242176870748299 |

MAP estimator works as a regularizer that prevents overfitting of model to the train data. It does so by introducing prior knowledge about the data.

Let's look at the first 4 rows in the table shown above. Those are 4 corner cases. Where the MLE estimator predicted 1 or 0 which is very unlikely in real life. Because the probability of a player hitting or the hit rate of a player can not be zero. Neither can it be 1 which means no chance of missing. From the N column we can see that the number of available data for these cases were only 1. So the decision or prediction was made based on only one observation. Now look at the MAP estimation for those people. It is around 25%. As soon as we introduce prior knowledge our prediction changes to something that is more likely.

But let's look at the last 4 rows of the same table.For those rows we had a higher number of observations available to us. So the MLE estimation made based on these observations should be reliable. Which is also indicated by the MAP estimator for those same observations. There is not much difference between MAP and MLE for the last 4 observations/data. So it means the prediction for MLE and MAP stays more or less the same if we have enough observations. So introducing prior knowledge doesn't have little to no effect on the prediction made based on the current/available observations. Again if we see the MAP estimator formula given above, we can see that if the number of observations (N) goes to infinity (∞) the MAP estimator becomes,

$$\Theta_{MAP} = N_H / N = \Theta_{MLE} \text{ , when } N \asymp \infty$$

## Ans to the question 3

Ans to the Que 3

Steps of PCA:

1. First compute sample mean $\bar{x}$ of the input data. where $x$ is input data.

2. Then compute sample co-variance matrix $S = \frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{x})(x_n - \bar{x})^T$

where both $x_n$ and $\bar{x}$ are D-dimensional vectors. So the covariance matrix would be a D×D.

$$S = \begin{bmatrix} cov(x^1, x^1) & cov(x^1, x^2) & \cdots & cov(x^1, x^D) \\ cov(x^2, x^1) & cov(x^2, x^2) & \cdots & cov(x^2, x^D) \\ \vdots & & & \\ cov(x^D, x_1) & cov(x^D, x^2) & \cdots & cov(x^D, x^D) \end{bmatrix}$$

3. Now we appear at the maximizing variance stage (for projected data). So max $u^T S u$

Date :

But there can be infinitely many sol$^n$ for the above eq$^n$. Inorder to restrict the search space we want to look for vectors with unit magnitude. So,

$$u^T u = 1$$

As we are mainly interested in the direction of projection

But as this constrained problem is not tractable. we use lagrange multiplier to make the problem easily solveable by taking the derivative equals to zero.

$$\max u^T S u + \lambda(1 - u^T u)$$

$$\frac{\delta}{\delta u} = 0 = 2Su - 2\lambda u$$

$$\Rightarrow \boxed{Su = \lambda u} \Rightarrow (S - \lambda I)u = 0$$

so we calculate the eigen vectors here!

Dimension of the matrices : we will have M u's

$S = D \times D$ , $u$ = $\boxed{D \times 1}$ for each u.

I = identity matrix.

As

Ans to the Ques

Date :

for eqn (i) to be true

$S - \lambda I$ must be equal to zero.

As $u$ is a non-zero vector.

$\therefore S - \lambda I = 0$

$\Rightarrow S = \lambda I$

from here we will get the eigen values ($\lambda$). we will get a ~~D~~ D eigen values. From the D eigen values we will choose the top M.
Highest.

By plugging each of those M eigen values to the $(S - \lambda I)u = 0$ eqn we will get M eigen values,

So we will get M, u vector. Each u vector will be $D \times 1$ dimensional.

4. Now we will use the eigen vectors to get the projected data.

$$\tilde{X} = U^T X$$

$$\Rightarrow \tilde{X} = X^T U$$

$\tilde{X}$: $\boxed{N \times M}$ $X^T$: $\boxed{N \times D}$ $U$: $\boxed{D \times M}$

where, $\tilde{X}$ = Transformed data

$X$ = Original Data

$U$ = Transformation Matrix / Eigen Vectors.

PCA Reference