N. KAUSIK
CS21M037

Tutorial 3

1. It is a method to estimate the parameters of a probability distribution by which we get maximum probability of observing the given data

$$\hat{\theta}_{MLE} = \arg\max_{\theta} \sum_{x \in X} \log P(x/\theta)$$

2 a) As it is Gaussian dist, we consider parameters μ, σ^2. Let $\theta_1 = \mu$ and $\theta_2 = \sigma^2$
$(\theta_1, \theta_2) = \theta$.

$$L(\theta|x) = P(x|\theta) = \frac{1}{\sqrt{2\pi\theta_2}} e^{\frac{-1}{2\theta_2}(x-\theta_1)^2}$$

$$L(\theta/D) = L(\theta) = \prod_{i=1}^{N} P(x_i/\theta)$$

To maximise $L(\theta)$ wrt θ_1, [Taking log]

$$\frac{\partial \log L(\theta)}{\partial \theta_1} = 0 \Rightarrow \frac{\partial}{\partial \theta_1} \log\left(\prod_{i=1}^{N} P(x_i/\theta)\right) = 0$$

$$\frac{\partial}{\partial \theta_1}\left(\sum_{i=1}^{N} \log\left(\frac{1}{\sqrt{2\pi\theta_2}} e^{-\frac{(x_i-\theta_1)^2}{2\theta_2}}\right)\right) = 0$$

$$\sum_{i=1}^{N} \frac{\partial}{\partial \theta_1}\left(\underbrace{\log\left(\frac{1}{\sqrt{2\pi\theta_2}}\right)}_{\text{Const wrt } \theta_1} - \frac{(x_i-\theta_1)^2}{2\theta_2}\right) = 0$$

$$\Rightarrow \sum_{i=1}^{N} \frac{\partial}{\partial \theta_1}\left(\frac{(x_i-\theta_1)^2}{2\theta_2}\right) = 0 \Rightarrow \sum_{i=1}^{N} \frac{(x_i-\theta_1)}{\theta_2} = 0$$

$$\Rightarrow \left(\sum_{i=1}^{N} x_i\right) - N\theta_1 = 0 \Rightarrow \boxed{\mu = \theta_1 = \frac{1}{N}\sum_{i=1}^{N} x_i}$$

b) To maximise $L(\theta)$ wrt θ_2, [Taking log]

$$\frac{\partial}{\partial\theta_2}\log\left(\prod_{i=1}^{N} p(x_i|\theta)\right) = 0$$

$$\sum_{i=1}^{N}\frac{\partial}{\partial\theta_2}\left(-\log(\sqrt{2\pi\theta_2}) - \frac{(x_i-\theta_1)^2}{2\theta_2}\right) = 0$$ [2π is const. Can be ignored]

$$-\sum_{i=1}^{N}\left(\frac{\partial}{\partial\theta_2}\left(\frac{1}{2}\log\theta_2\right) + \frac{\partial}{\partial\theta_2}\left(\frac{1}{2\theta_2}(x_i-\theta_1)^2\right)\right) = 0$$

$$\left(\sum_{i=1}^{N}\frac{1}{2\theta_2}\right) - \left(\sum_{i=1}^{N}\frac{1}{2\theta_2^2}(x_i-\theta_1)^2\right) = 0$$

$$\frac{N}{\cancel{2}\cancel{\theta_2}} = \frac{\sum_{i=1}^{N}(x_i-\theta_1)^2}{\cancel{2}\theta_2^{\cancel{2}}}$$

$$\Rightarrow \boxed{\sigma^2 = \hat{\theta}_2 = \frac{1}{N}\sum_{i=1}^{N}(x_i-\hat{\theta}_1)^2}$$

c) Since as given prior cannot be ignored, we can do MAP for the mean as,

$$\hat{\mu}_{MAP} = \arg\max_{\mu}\log\left[P(D|\mu)\,P(\mu)\right]$$

$$\Rightarrow \log(P(D|\mu)P(\mu)) = \log\left[\prod_{i=1}^{N} P(x_i|\mu)\cdot N(\mu|\mu_p,\sigma_p^2)\right]$$

To maximise wrt μ,

$$\frac{\partial}{\partial\mu}\left(\sum_{i=1}^{N}\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right) + \sum_{i=1}^{N}\log\left(\frac{1}{\sqrt{2\pi\sigma_p^2}}e^{-\frac{(\mu-\mu_p)^2}{2\sigma_p^2}}\right)\right) = 0$$

$$\Rightarrow \sum_{i=1}^{N} \left[\frac{\partial}{\partial \mu}\left(-\frac{N}{2}\log\sigma - \frac{(x_i-\mu)^2}{2\sigma^2} \right) + \frac{\partial}{\partial \mu}\left(-\log\sigma_p - \frac{(\mu-\mu_p)^2}{2\sigma_p^2} \right) \right] = 0$$

$$\Rightarrow \sum_{i=1}^{N} \left[\frac{(x_i-\mu)}{\sigma^2} - \frac{(\mu-\mu_p)}{\sigma_p^2} \right] = 0$$

$$\Rightarrow \sum_{i=1}^{N} \frac{x_i}{\sigma^2} - \frac{N\mu}{\sigma^2} - \frac{N\mu}{\sigma_p^2} + \frac{N\mu_p}{\sigma_p^2} = 0$$

$$\Rightarrow \hat{\mu} = \frac{\frac{1}{\sigma^2}\sum_{i=1}^{N} x_i + \frac{\mu_p}{\sigma_p^2}}{\frac{N}{\sigma^2} + \frac{1}{\sigma_p^2}}$$

as $\sum_{i=1}^{N} x_i = N\hat{\mu}_{MLE}$,

$$\boxed{\hat{\mu}_{MAP} = \frac{\frac{N\hat{\mu}_{MLE}}{\sigma^2} + \frac{\mu_p}{\sigma_p^2}}{\frac{N}{\sigma^2} + \frac{1}{\sigma_p^2}}}$$

3. a) Gaussian - μ, σ^2

b) Beta - α, β

c) Exponential - λ

d) Gamma - k, θ

4. $$P(x|\theta) = \begin{cases} 1/\theta & 0 \le x \le \theta \\ 0 & \text{otherwise} \end{cases}$$

MLE for D, $L(\theta/D) = P(D|\theta)$

$$P(D/\theta) = \prod_{i=1}^{n} P(x_i/\theta) = \prod_{i=1}^{n} \frac{1}{\theta} I(0 \le x_i \le \theta)$$

where $I(t)$ is indicator fn $= \begin{cases} 1 & \text{if } t \text{ is true} \\ 0 & \text{otherwise} \end{cases}$

Here, even if one of the x_is fall out of range, it becomes 0.

Saying all x_i are in range $[0, \theta]$ is same as saying
$\min x_i \ge 0$ and $\max x_i \le \theta$.

$$\Rightarrow \frac{1}{\theta^n} I(\theta \ge \max_i x_i) \, I(\min_i x_i \ge 0)$$

Since to maximise $P(D/\theta)$, we must choose minimum possible θ.
as $\theta \ge \max x_i$,
$L(\theta/D)$ is max when $\theta = \max[D]$.

5. True.
In MLE we follow same procedure like in MAP but we assume that all values of θ are equally likely, i.e. $P(\theta) = \text{const}$ (uniform)

6. In MAP estimator, we consider a prior distribution over the parameters, however in MLE estimator we ignore the prior term ($P(\theta)$) as we assume all parameters are equally likely.

7. MLE cannot be used for constrained optimisation problems as MLE does not have facility to optimise while also satisfying constraints.

8. $$P(x|\theta) = \prod_{i=1}^{d} \theta_i^{x_i} (1-\theta_i)^{1-x_i}$$

a) $$P(D|\theta) = \prod_{k=1}^{n} P(x_k|\theta) = \prod_{k=1}^{n} \prod_{i=1}^{d} \theta_i^{x_{ki}} (1-\theta_i)^{1-x_{ki}}$$

$$= \prod_{i=1}^{d} \prod_{k=1}^{n} \theta_i^{x_{ki}} (1-\theta_i)^{1-x_{ki}} = \prod_{i=1}^{d} \theta_i^{\sum_{k=1}^{n} x_{ki}} (1-\theta_i)^{\sum_{k=1}^{n}(1-x_{ki})}$$

$\Rightarrow$ as $s_i = \sum_{k=1}^{n} x_{ki}$,

$$\Rightarrow \prod_{i=1}^{d} \theta_i^{s_i} (1-\theta_i)^{(n-s_i)} \quad // = \text{R+S}$$

b) $P(\theta/D) = \dfrac{P(D|\theta)P(\theta)}{P(D)}$

as uniform aprior dist, we can ignore $P(\theta)$.

Also, $P(D) = \int_0^1 P(D|\theta)\, d\theta = \int_0^1 \prod_{i=1}^{d} \theta_i^{s_i} (1-\theta_i)^{n-s_i} d\theta$

$= \prod_{i=1}^{d} \int_0^1 \theta_i^{s_i} (1-\theta_i)^{n-s_i} d\theta$ using given identity,

$= \prod_{i=1}^{d} \dfrac{(s_i)!\,(n-s_i)!}{(s_i+n-s_i+1)!} = \prod_{i=1}^{d} \dfrac{(s_i)!\,(n-s_i)!}{(n+1)!}$

$\therefore P(\theta|D) = \dfrac{\prod_{i=1}^{d} \theta_i^{s_i} (1-\theta_i)^{n-s_i} (n+1)!}{\prod_{i=1}^{d} (s_i)!\,(n-s_i)!}$

$= \prod_{i=1}^{d} \dfrac{(n+1)!}{s_i!\,(n-s_i)!} \theta_i^{s_i} (1-\theta_i)^{n-s_i} = RHS$

c) $\int_0^1 P(x|\theta)\, P(\theta|D)\, d\theta$

$$= \int_0^1 \left[\prod_{i=1}^{d} \theta_i^{x_i} (1-\theta_i)^{(1-x_i)} \right] \frac{(n+1)!}{(s_i)!\,(n-s_i)!} \theta_i^{s_i} (1-\theta_i)^{n-s_i} d\theta$$

Let $\frac{(n+1)!}{(s_i)!\,(n-s_i)!} = F$,

$$\Rightarrow \prod_{i=1}^{d} \cdot F \cdot \int_0^1 \theta_i^{x_i+s_i} (1-\theta_i)^{n+1-s_i-x_i} d\theta$$

Using given identity in b),

$$\Rightarrow \prod_{i=1}^{d} \cdot F \cdot \frac{(x_i+s_i)!\,(n+1-s_i-x_i)!}{(n+2)!}$$

$$= \prod_{i=1}^{d} \frac{\cancel{(n+1)!}}{(s_i)!\,(n-s_i)!} \cdot \frac{(x_i+s_i)!\,(n+1-s_i-x_i)!}{(n+2)\cancel{!}}$$

Since x_i can either be 0 or 1, $x_i \in \{0,1\}$,

If $x_i = 0$,

$$P(x|D)_{x_i=0} = \prod_{i=1}^{d} \frac{1}{\cancel{(s_i)!}\,\cancel{(n-s_i)!}} \cdot \frac{\cancel{(s_i)!}\,(n-s_i+1)\cancel{!}}{(n+2)}$$

$$= \prod_{i=1}^{d} \frac{n-s_i+1}{n+2} = \prod_{i=1}^{d} \left(1 - \frac{s_i+1}{n+2}\right)$$

If $x_i = 1$,

$$P(x|D)_{x_i=1} = \prod_{i=1}^{d} \frac{1}{\cancel{(s_i)!}\,\cancel{(n-s_i)!}} \cdot \frac{(s_i+1)\cancel{!}\,\cancel{(n-s_i)!}}{n+2}$$

$$= \prod_{i=1}^{d} \frac{s_i+1}{n+2}$$

Combining the both, we get,

$$\therefore P(x|D) = \prod_{i=1}^{d} \left(\frac{S_i+1}{n+2}\right)^{x_i} \left(1-\left(\frac{S_i+1}{n+2}\right)\right)^{1-x_i}$$

= RHS

d) By observation, we can see that

both are of same form.

In $P(x|D)$, we are taking

$\hat{\theta}_i = \frac{S_i+1}{n+2}$ for all i and substituting

in $P(x|\theta)$ formula.

$\therefore$ Effective Bayesian Estimate for $\hat{\theta}_i = \frac{S_i+1}{n+2}$

for each i

in 1 to d.

9. In MLE, we consider the prior distribution ($P(\theta)$) as a constant and ignore it in finding $\hat{\theta}_{MLE}$.

However, in Bayesian Estimator, we consider

θ as a random variable and we dont

ignore prior distribution.