MACHINE LEARNING HOMEWORK SHEET-2
PARAMETER INFERENCE

NAME: MAHALAKSHMI SABANAYAGAM

TUM ID: ge73yuw

DATE: 05.11.2018

Problem 1:

$$p(x_1, \ldots x_n \mid \theta) = \theta^t (1-\theta)^h$$

$$\theta_{MLE} = \underset{\theta \in [0,1]}{\arg\max}\ \theta^t (1-\theta)^h$$

1st & 2nd derivative of θ_{MLE}:

$$\theta'_{MLE} = t\theta^{t-1}(1-\theta)^h - \theta^t (1-\theta)^{h-1} h = \theta^{t-1}(1-\theta)^{h-1}(t-(t+h)\theta)$$

$$\theta''_{MLE} = (t-1)\theta^{t-2}(1-\theta)^{h-1}(t-(t+h)\theta) - \theta^{t-1}(h-1)(1-\theta)^{h-2}(t-(t+h)\theta) + \theta^{t-1}(1-\theta)^{h-1}(-(t+h))$$

$$\log \theta_{MLE} = \log(\theta^t (1-\theta)^h)$$
$$= \log \theta^t + \log (1-\theta)^h$$
$$= t \log \theta + h \log (1-\theta)$$

1st & 2nd derivative of $\log \theta_{MLE}$

1st derivative $= \dfrac{t}{\theta} - \dfrac{h}{1-\theta}$

2nd derivative $= -\dfrac{t}{\theta^2} - \dfrac{h}{(1-\theta)^2}$

Problem 2: To prove every local max of $\log f(\theta)$ is also local max of differentiable, positive $f(\theta)$.

$f(\theta) \Rightarrow$ differentiable, positive function.

To find local max/min (critical points), we have to differentiate the function ~~f(x)~~ to find the θ where ~~slope~~ slope is 0. These points imply we have reached either maximum or minimum (local)

So, equating $f'(\theta) = 0$ and the solution be $\theta_c = \{\theta_1, \theta_2 \dots \theta_n\}$

The local max./min of $f(\theta)$ occurs at $\theta \in \theta_c$

Now, lets consider $\log f(\theta) = g(\theta)$

The local max/min of $\log f(\theta)$ occurs at,

$$g'(\theta) = \frac{1}{f(\theta)} \cdot \cancel{\log} f'(\theta) = 0.$$

solve $f'(\theta) = 0$ & $f(\theta) \neq 0$ for θ. This satisfies for all $\theta \in \theta_c$. Thus local max/min for $\log f(\theta)$ occurs at every θ in θ_c.

From above, we see that critical points occur at same θ for $f(\theta)$ and $\log f(\theta)$. But we still have to prove local maximum of $f(\theta)$ and $\log f(\theta)$ occurs at θ_{max} & local minimum of $f(\theta)$ and $\log f(\theta)$ occurs at θ_{min}.

As log is ~~a~~ monotonically increasing function, ~~and~~ $\cancel{f(\theta)}$ the critical points ~~that~~ are preserved in the same order as $f(\theta)$.

So, every local maximum of $f(\theta)$ will also be local maximum of $\log f(\theta)$ and every local minimum of $f(\theta)$ will also be local minimum of $\log f(\theta)$.

Considering the above proof, log likelihood enables faster and easier computation in the previous question. The first and second derivatives of likelihood look ugly, whereas the first and second derivatives of the loglikelihood is simple and elagant, and it preserves the critical points!

So, we can make use of log likelihood for finding θ_{MLE}.

Problem 3:

To prove: θ_{MLE} is a special case of θ_{MAP} (ie)

$\theta_{MLE} = \theta_{MAP}$ for some prior $p(\theta)$.

$$\theta_{MLE} = \underset{\theta}{\arg\max}\ P(X|\theta)$$

$$= \underset{\theta}{\arg\max} \prod_i P(x_i|\theta)$$

$$\theta_{MAP} = \underset{\theta}{\arg\max}\ P(X|\theta).P(\theta)$$

$$= \underset{\theta}{\arg\max} \prod_i P(x_i|\theta).P(\theta)$$

Comparing θ_{MAP} and θ_{MLE}, we see that in θ_{MAP} likelihood $P(X|\theta)$ is weighted from the weights given by prior $P(\theta)$.

If the weights are constant, then we can ignore as we are calculating argmax.

So when prior is constant $\theta_{MLE} = \theta_{MAP}$.

$$\theta_{MAP} = \underset{\theta}{\arg\max} \prod_i P(x_i|\theta) . \text{const}$$

$$= \underset{\theta}{\arg\max} \prod_i P(x_i|\theta) = \theta_{MLE}$$

Thus θ_{MLE} is a special case of θ_{MAP} where the prior is constant / uniform.

Problem 4:

Bernoulli random variable X, $\#(X=0) = l$; $\#(X==1) = m$; $N = m+l$

Interested in $X=1$.

prior distribution $P(\theta|a,b) = Beta(\theta|a,b)$

To prove: posterior mean $E[\theta|D]$ lies between the prior mean of θ and θ_{MLE}

prior distribution is beta with a, b parameters

$\Rightarrow$ prior mean $= \dfrac{a}{a+b} \rightarrow$ (1)

Experiments is Binomial distribution and we are interested in $X=1$,

$$P(x=m|N,\theta) = \binom{N}{m} \theta^m (1-\theta)^{N-m}$$

$N = m+l$

$l = N-m$

$$\theta_{MLE} = \frac{m}{N} \rightarrow (2)$$

$$P(\theta|D) \propto P(D|\theta) \cdot P(\theta)$$

$$\propto \text{Binomial}(\underbrace{x=1}_{m}|\theta, N) \cdot \text{Beta}(\theta|a,b)$$

$$\propto \text{Beta}(m+a, N-m+b)$$

$E[\theta|D]$ = expectation of Beta distribution with parameter $m+a$, $N-m+b$

$$= \frac{m+a}{N+a+b} \rightarrow (3)$$

Expectation of Beta(a,b) $= \frac{a}{a+b}$

$$E[\theta|D] = \frac{m+a}{N+a+b}$$

$$= \overbrace{\frac{m}{N+a+b} \cdot \frac{N}{N}}^{(2)} + \overbrace{\frac{a}{N+a+b} \cdot \frac{a+b}{a+b}}^{(1)}$$

$$= \underbrace{\left(\frac{N}{N+a+b}\right)}_{1-\lambda} \theta_{MLE} + \underbrace{\left(\frac{a+b}{N+a+b}\right)}_{\lambda} \text{prior mean}$$

λ & $0 \leq \lambda \leq 1$.

Thus, from above equation, posterior mean lies between prior mean of θ and θ_{MLE}.

Problem 5:

a. $p(X|\lambda)$ - Poisson distribution

λ_{MLE} for n iid samples from X. To prove: Estimate is unbiased.

$$p(X=x_i|\lambda) = \frac{e^{-\lambda} \lambda^{x_i}}{x_i!} \leftarrow \text{Poisson Distribution}$$

$$\lambda_{MLE} = \underset{\lambda}{\text{argmax}} \prod_{i=1}^{n} \frac{e^{-\lambda} \lambda^{x_i}}{(x_i)!}$$

$$= \underset{\lambda}{\text{argmax}} \left[\frac{e^{-\lambda} \cdot \lambda^{x_1}}{x_1!} \cdot \frac{e^{-\lambda} \lambda^{x_2}}{x_2!} \cdots \frac{e^{-\lambda} \lambda^{x_n}}{(x_n)!} \right]$$

$$= \underset{\lambda}{\text{argmax}} \frac{e^{-n\lambda} \lambda^{x_1+x_2+\dots x_n}}{x_1! x_2! \dots x_n!} \leftarrow \text{constant}$$

$$= \arg\max_{\lambda} e^{-n\lambda} \lambda^{x_1+x_2+\dots+x_n}$$

To find the max, differentiate λ_{MLE} w.r.t. λ,

$$\lambda'_{MLE} = \left[e^{-n\lambda + \ln \lambda^{x_1+x_2+\dots+x_n}} \right]'$$

$$0 = \underbrace{\left[e^{-n\lambda + \ln \lambda^{x_1+x_2+\dots+x_n}} \right]}_{\neq 0} \cdot \left[-n + \frac{(x_1+x_2+\dots+x_n)}{\lambda} \right]$$

$$0 = -n + \frac{(x_1+x_2+\dots+x_n)}{\lambda}$$

$$\lambda_{MLE} = \frac{x_1+x_2+\dots+x_n}{n}$$

To prove λ_{MLE} is unbiased,

$$E[\lambda_{MLE}] = \lambda$$

$$E[\lambda_{MLE}] = E\left[\frac{x_1+x_2+\dots+x_n}{n}\right]$$

$$= \frac{1}{n}\left(E[x_1] + E[x_2] + \dots E[x_n]\right)$$

As $x_1, x_2 \dots x_n$ are iid and poisson distributed

$$E[X=x_1] = \lambda$$

$$\therefore E[\lambda_{MLE}] = \frac{1}{n}[\lambda + \lambda + \dots + \lambda] = \frac{n\lambda}{n} = \underline{\underline{\lambda}}$$

$$\Rightarrow E[\lambda_{MLE}] = \lambda$$

Thus the estimate is unbiased.

b. Compute posterior distribution over λ, prior dist - Gamma (α, β)
λ_{MAP} for this prior ?

Prior distribution is Gamma (α, β)

(ie) $P(\lambda \mid \alpha, \beta) = \text{Gamma}(\alpha, \beta)$

Likelihood $P(X \mid \lambda) = \prod_{i=1}^{n} \dfrac{e^{-\lambda} \lambda^{x_i}}{(x_i)!}$

Posterior $P(\lambda \mid X) = \dfrac{P(X \mid \lambda)\, P(\lambda)}{P(X)}$

$$P(\lambda \mid X) \propto P(X \mid \lambda) \cdot P(\lambda)$$

$$\propto \text{Poisson}(x \mid \lambda,n) \cdot \text{Gamma}(\lambda \mid \alpha, \beta)$$ conjugate prior

$$\propto \text{Gamma}\left(\alpha + \sum_{i=1}^{n} x_i, \beta + n\right)$$

$\therefore$ Posterior distribution is $\text{Gamma}\left(\alpha + \sum_{i=1}^{n} x_i, \beta + n\right)$.

λ_{MAP} = mode of the posterior distribution

$$= \frac{\alpha + \sum_{i=1}^{n} x_i - 1}{\beta + n}$$

$$\lambda_{MAP} = \frac{\alpha + \sum_{i=1}^{n} x_i - 1}{\beta + n}$$