Machine Learning

Homework 3 solution - Shyam Arumugaswamy

1. $\frac{d}{d\theta}\theta^t(1\text{-}\theta)^h = \theta^t.\,h.\,(1\text{-}\theta)^{h\text{-}1}.\,(\text{-}1) + (1\text{-}\theta)^h.t.\,\theta^{t\text{-}1}$ -- chain rule

$= \text{-}\theta^t.\,h.\,(1\text{-}\theta)^{h\text{-}1} + t.\,\theta^{t\text{-}1}.(1\text{-}\theta)^h$

$= \text{-}h.\theta.\,\theta^{t\text{-}1}.\,(1\text{-}\theta)^{h\text{-}1} + t.\,(1\text{-}\theta)^{h\text{-}1}.\,\theta^{t\text{-}1}\,(1\text{-}\theta)$

$= \theta^{t\text{-}1}.\,(1\text{-}\theta)^{h\text{-}1}\,(t(1\text{-}\theta)\ \text{-}h\theta)$

$\frac{d^2}{d^2\theta}\theta^t(1\text{-}\theta)^h = \theta^{t\text{-}1}.\,(1\text{-}\theta)^{h\text{-}1}\,(\text{-}t\ \text{-}h) + (\,(t(1\text{-}\theta)\text{-}h\theta)\,[\text{-}\theta^{t\text{-}1}.\,(1\text{-}\theta)^{h\text{-}2}.\,(h\text{-}1) + \theta^{t\text{-}2}.\,(1\text{-}\theta)^{h\text{-}1}.\,(t\text{-}1)]\,)$

$= \text{-}\,\theta^{t\text{-}1}.\,(1\text{-}\theta)^{h\text{-}1}\,(t+h) + \theta^{t\text{-}2}.\,(1\text{-}\theta)^{h\text{-}2}.\,(t.(1\text{-}\theta)\text{-}\theta h).((t\text{-}1).(1\text{-}\theta)\text{-}(h\text{-}\ 1)\theta)$

let $f(\theta) = \log\theta^t(1\text{-}\theta)^h = t\log\theta + h.\,\log(1\text{-}\theta)$

$\frac{d}{d\theta}f(\theta) = t/\theta\ \text{-}\ h/(1\text{-}\theta)$

$\frac{d^2}{d^2\theta}f(\theta) = \text{-}t/\theta^2\ \text{-}\ h/\,(1\text{-}\theta)^2$

2. Let $p(\theta) = \log f(\theta)$ and θ^t be any local maximum of $p(\theta)$, so we have $p(\theta^t) >= p(\theta)$ for θ in small neighbourhood of θ^t. for a monotonic transform like exponential function, we can say

   $f(\theta^t) = \exp(p(\theta^t)) >= \exp(p(\theta)) = f(\theta)$ … so θ^t is a max of f

   from the previous problem, we can see that computational becomes simpler using log and maximum or minimum is preserved.

3. $\boldsymbol{\theta}_{MLE} = \arg\max_{\boldsymbol{\theta}} p(D|\boldsymbol{\theta})$ and $\boldsymbol{\theta}_{MAP} = \arg\max_{\boldsymbol{\theta}} p(D|\boldsymbol{\theta})\,p(\boldsymbol{\theta})$

   If we consider uniform distribution as prior, then $p(\boldsymbol{\theta})$ will be constant for all $\boldsymbol{\theta}$.
   So $\arg\max_{\boldsymbol{\theta}} p(D|\boldsymbol{\theta})\,p(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} p(D|\boldsymbol{\theta})$
   therefore, $\boldsymbol{\theta}_{MAP} = \boldsymbol{\theta}_{MLE}$

4. $p(x{=}m\mid N,\boldsymbol{\theta}) = ({}^{N}{}_{m})\,\theta^m(1-\theta)^{N\text{-}m}$
   if a, b are hyper parameters of prior beta distribution, then the posterior is Beta(m+a,l+b) distributed.

so, the posterior mean is

$E(\boldsymbol{\theta}|D) = \frac{m + a}{m + a + l + b}$

$$= \frac{m}{m + a + l + b} + \frac{a}{m + a + l + b}$$

$$= \frac{m}{m + a + l + b} \cdot \frac{m + l}{m + l} + \frac{a}{m + a + l + b} \cdot \frac{a + b}{a + b}$$

$$= \frac{m + l}{m + a + l + b} \cdot \frac{m}{m + l} + \frac{a + b}{m + a + l + b} \cdot \frac{a}{a + b}$$

if $\frac{a + b}{m + a + l + b} = \lambda$

then $1 - \lambda = 1 - \frac{a + b}{m + a + l + b}$

$$= \frac{m + a + l + b - a - b}{m + a + l + b}$$

$$= \frac{m + l}{m + a + l + b}$$

therefore, $E(\boldsymbol{\theta}|D) = (1 - \lambda). \frac{m}{m + l} + \lambda . \frac{a}{a + b}$

the max likelihood estimate is $\frac{m}{m + l}$ and posterior mean is $\frac{a}{a + b}$

5. Let X be Poison distributed

   $Poi\ (x|\lambda) = \frac{e^{-\lambda}}{x!} . \lambda^{x}$

   $\ln(\lambda) = \ln \prod_{x=1}^{n} \frac{e^{-\lambda}}{x!} . \lambda^{x} = \sum_{i=1}^{n} \ln e^{-\lambda} + \sum_{i=1}^{n} \ln \frac{\lambda^{xi}}{x!}$

   $$= -n\lambda + \sum_{i=1}^{n} \left(xiln\lambda - lnx!\right)$$

   $d(\ln \lambda)/d\lambda = -n + n\lambda/\lambda = 0$

   $E[\lambda] = = \sum_{i=1}^{n} \ln \frac{E[x!]}{n} = \lambda$

   prior for λ :

   alpha & beta are denoted as a & b in this problem

   $p(\lambda \mid \alpha, \beta) = \frac{\beta^{\alpha}}{\tau(\alpha)} \lambda^{\alpha\text{-}1} \exp(-\beta\lambda)$

so $P(\lambda \mid D) \propto P(D \mid \lambda)\, P(\lambda)$ and is Gamma $(\sum_{i=1}^{n} xi + \alpha, n + \beta)$

for MAP soln, $\arg\max_{\lambda} p(\lambda|D) = \arg\max_{\lambda} \ln p(\lambda|D)$

$\ln p(\lambda|D) = (\sum_{i=1}^{n} xi + \alpha - 1)\ln \lambda - (n + \beta)\lambda + c$

taking first derivative

$$\frac{d}{d\lambda} \ln p(\lambda|D) = (\sum_{i=1}^{n} xi + \alpha - 1) / \lambda - (n + \beta) = 0$$

$$\lambda = \frac{(\sum_{i=1}^{n} xi + \alpha - 1)}{(n + \beta)}$$