Machine Learning and Computational Statistics

Homework 5

Exercise 1

(a) Assuming that $x_1 \dots x_N$ are independent, Likelihood function $p(X;\theta)$ is:

$$p(X;\theta) = \prod_{i=1}^{N} p(x_i;\theta) = \prod_{i=1}^{N} \theta^2 x_i \exp(-\theta x_i)\, u(x_i)$$

- for $x<0$: $u(x)=0$, Thus $p(x;\theta)=0$
- for $x \geq 0$: $u(x)=1$, Thus $p(X;\theta) = \prod_{i=1}^{N} \theta^2 x_i \exp(-\theta x_i)$

Log-Likelihood $L(\theta) = \log(p(X;\theta)$ is:

$$L(\theta) = \log\left(\prod_{i=1}^{N} p(x_i;\theta)\right) = \sum_{i=1}^{N} \log(p(x_i;\theta)) = \sum_{i=1}^{N} \log(\theta^2 x_i \exp(-\theta x_i)) =$$

$$= \sum_{i=1}^{N} \log\theta^2 + \sum_{i=1}^{N} \log x_i + \sum_{i=1}^{N} \log(\exp(-\theta x_i)) \Rightarrow$$

$$\Rightarrow L(\theta) = 2N\log\theta + \sum_{i=1}^{N} \log(x_i) - \theta\sum_{i=1}^{N} x_i$$

To find the Maximum Likelihood estimate of θ (θ_{ML}) we will get the gradient $\frac{\partial L(\theta)}{\partial\theta}$ and equate to 0,

$$\frac{\partial L(\theta)}{\partial\theta} = 0 \Rightarrow \frac{2N}{\theta_{ML}} + 0 + \left(-\sum_{i=1}^{N} x_i\right) = 0 \Rightarrow \theta_{ML} = \frac{2N}{\sum_{i=1}^{N} x_i}$$

(b) For $N=5$, $x_1=2, x_2=2.2, x_3=2.7, x_4=2.4$ and $x_5=2.6$:

- $\theta_{ML} = \frac{2\cdot 5}{2+2.2+2.7+2.4+2.6} = \frac{10}{11.9} = 0.84$

- $\mu_{ML} = 2/\theta = 2/0.84 = 2.38$

- $p(x) = 0.7056 \cdot x \cdot \exp(-0.84x)$ (1)

(c) Using the formula (1):

- $x_1' = 2.1 \Rightarrow p(x_1') = 0.2359$
- $x_2' = 2.3 \Rightarrow p(x_2') = 0.2350$
- $x_3' = 2.9 \Rightarrow p(x_3') = 0.1790$

Exercise 2

(a) $\theta_{MAP} = \arg\max p(X|\theta) \cdot p(\theta)$ or $\theta_{MAP} = \arg\max \log[p(X|\theta)p(\theta)]$ (1)

But $p(X|\theta) = \prod_{i=1}^{N} p(x_i|\theta)$ (as per Exercise 1),

and $p(\theta) = N(\theta_0, \sigma_0^2)$

$$p(\theta) = \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left(-\frac{(\theta-\theta_0)^2}{2\sigma_0^2}\right)$$

$$\log(p(\theta)) = \log\left(\frac{1}{\sqrt{2\pi}\sigma_0}\right) - \frac{(\theta-\theta_0)^2}{2\sigma_0^2} \quad \text{and} \quad \frac{\partial(\log(p(\theta)))}{\partial\theta} = \frac{-2(\theta-\theta_0)}{2\sigma_0^2} = -\frac{\theta-\theta_0}{\sigma_0^2}$$

$$(1) \Rightarrow \theta_{MAP} = \arg\max\left[\log p(X|\theta) + \log p(\theta)\right] = \arg\max\left[\sum_{i=1}^{N} \log(p(x_i|\theta)) + \log(p(\theta))\right]$$

$\Rightarrow$ We will get θ_{MAP} by getting the gradient $\frac{\partial\left[\sum_{i=1}^{N} \log(p(x_i|\theta)) + \log(p(\theta))\right]}{\partial\theta} = 0 \Rightarrow$

$$\Rightarrow \frac{2N}{\theta_{MAP}} - \sum_{i=1}^{N} x_i - \frac{\theta_{MAP}-\theta_0}{\sigma_0^2} = 0 \Rightarrow \quad (2)$$

$$2N\sigma_0^2 - \theta_{MAP}\sigma_0^2 \sum_{i=1}^{N} x_i - \theta_{MAP}^2 + \theta_0\theta_{MAP} = 0 \Rightarrow$$

$$\theta_{MAP}^2 - \left(\theta_0 - \sigma_0^2 \sum_{i=1}^{N} x_i\right)\theta_{MAP} - 2N\sigma_0^2 = 0$$

$$\hat{\theta}_{MAP_{1,2}} = \begin{cases} \dfrac{\left(\theta_0 - \sigma_0^2 \sum_{i=1}^{N} x_i\right) + \sqrt{\left(\theta_0 - \sigma_0^2 \sum_{i=1}^{N} x_i\right)^2 + 8N\sigma_0^2}}{2} \\[2ex] \dfrac{\left(\theta_0 - \sigma_0^2 \sum_{i=1}^{N} x_i\right) - \sqrt{\left(\theta_0 - \sigma_0^2 \sum_{i=1}^{N} x_i\right)^2 + 8N\sigma_0^2}}{2} \Rightarrow \theta_{MAP} < 0 \text{, Thus rejected} \end{cases}$$

$$\hat{\theta}_{MAP} = \frac{1}{2}\left[\left(\theta_0 - \sigma_0^2 \sum_{i=1}^{N} x_i\right) + \sqrt{\left(\theta_0 - \sigma_0^2 \sum_{i=1}^{N} x_i\right)^2 + 8N\sigma_0^2}\right]$$

(b) (i) For $N \to \infty$, $\hat{\theta}_{MAP} \approx \hat{\theta}_{ML}$, because $-\frac{\theta_{MAP}-\theta_0}{\sigma_0^2}$ from eq.(2) has small effect on eq.(2)

(ii) $\sigma_0^2 \gg$. That leads $-\frac{\theta_{MAP}-\theta_0}{\sigma_0^2}$ to get value close to 0 (eq.(2)), thus $\hat{\theta}_{MAP} \approx \hat{\theta}_{ML}$

(iii) $\sigma_0^2 \ll$. That leads $-\frac{\theta_{MAP}-\theta_0}{\sigma_0^2}$ to get higher value. The effect of prior knowledge is increased and $\theta_{MAP} \to \theta_0$

(c) $\hat{\theta}_{MAP} = \left[\left(1.1 - \sum_{i=1}^{5} x_i\right) + \sqrt{\left(1.1 - \sum_{i=1}^{5} x_i\right)^2 + 40}\right]/2$ for $\theta_0 = 1.1$, $\sigma_0^2 = 1$, $N=5$

for $x_1 = 2$, $x_2 = 2.2$, $x_3 = 2.7$, $x_4 = 2.4$, $x_5 = 2.6$ $\Rightarrow \hat{\theta}_{MAP} = 0.8577$

and distribution is : $p(x) = (0.8577)^2 \cdot x \cdot \exp(-0.8577 \cdot x)$, $x \geq 0$

(d). continuing*

$x_1' = 2.1 \Rightarrow p(x_1') = 0.2550$

$x_2' = 2.3 \Rightarrow p(x_2') = 0.2353$

$x_3' = 2.9 \Rightarrow p(x_3') = 0.1773$

(e) $\mu_{MAP} = E[x] = \int_0^{+\infty} x\, p(x)\, dx = \int_0^{+\infty} 0.8577^2 x^2 \cdot \exp(-0.8577x)\, dx = \dots = \frac{2}{0.8577} = 2.33$

$\mu_{MAP} = 2.33$

Prior knowledge of θ_0 gives a smaller estimation of μ. It also increases the value of $\theta_{ML} = 0.84$ to $\theta_{MAP} = 0.8577$ (since $\theta_0 = 1.1$)

Exercise 3

Ridge Regression is a way of imposing prior knowledge to the LS method for Regression, by shrinking the norm of NVll. We can apply this method to the Maximum Likelihood computation. Now, the prior knowledge is that μ_0 lies close to μ, and that is in a radius ρ or $(\mu-\mu_0)^2 \leq \rho \Rightarrow (\mu-\mu_0)^2 - \rho \leq 0$. By adding this quantity to the ML estimate (of course multiplying by λ to bias the solution away from the ML case) we regularize the results.

$$L(\mu) = \sum_{n=1}^{N} (x_n-\mu)^2 + \lambda((\mu-\mu_0)^2 - \rho)$$

To minimize $L(\mu)$ we will get the gradient and equate to 0.

$$\frac{\partial L(\mu)}{\partial \mu} = 0 \Rightarrow -2\sum_{n=1}^{N}(x_n - \hat{\mu}) + 2\lambda(\hat{\mu}-\mu_0) = 0 \Rightarrow -\sum_{n=1}^{N} x_n + N\hat{\mu} + \lambda\hat{\mu} - \lambda\mu_0 = 0 \Rightarrow$$

$$\Rightarrow (N+\lambda)\hat{\mu} = \sum_{n=1}^{N} x_n + \lambda\mu_0 \Rightarrow \hat{\mu} = \frac{\sum_{n=1}^{N} x_n + \lambda\mu_0}{N+\lambda}$$