Machine Learning Assignment 3

Shivangi Aneja

13-November-2017

Problem 1

log likelihood $= log\ p(x_1, ..., x_n|\theta)$

Maximize $\underset{\theta\in[0,1]}{\text{argmax}}\ \theta^t(1-\theta)^h$

$f(\theta) = \theta^t(1-\theta)^h$

First Derivative :
$\frac{df}{d\theta} = t\theta^{t-1}(1-\theta)^h - h\theta^t(1-\theta)^{h-1}$

Second Derivative :
$\frac{d^2f}{d\theta^2} = t(t-1)\theta^{t-2}(1-\theta)^h - th\theta^{t-1}(1-\theta)^{h-1} - ht\theta^{t-1}(1-\theta)^{h-1} + h(h-1)\theta^t(1-\theta)^{h-2}$

$\frac{d^2f}{d\theta^2} = \theta^{t-2}(1-\theta)^{h-2}[t(t-1)(1-\theta)^2 - 2th\theta(1-\theta) + h(h-1)\theta^2]$

$g(\theta) = log(f(\theta)) = t\ log(\theta) + h\ log(1-\theta)$

First Derivative :
$\frac{dg}{d\theta} = \frac{t}{\theta} - \frac{h}{1-\theta}$

Second Derivative :
$\frac{d^2g}{d\theta^2} = \frac{-t}{\theta^2} - \frac{h}{(1-\theta)^2}$

Problem 2

Consider a positive differentiable function $f(x)$ and $g(x) = log\ [f(x)]$

To find critical point for $f(x)$,differentiate $f(x)$ w.r.t x , we get
$\frac{df}{dx} = 0 \Rightarrow f'(x) = 0$, we get $x = c$

To find critical point for $g(x)$,differentiate $g(x)$ w.r.t x , we get
$\frac{dg}{dx} = 0 \Rightarrow \frac{f'(x)}{f(x)} = 0 \Rightarrow f'(x) = 0$, we get same value $x = c$

Applying log to a function is a **strictly monotonic transformation**. Thus it has same

maxima location i.e.
$arg\ max\ f(x) = arg\ max\ g(x)$,
but different maxima value i.e
$max\ f(x) \neq max\ g(x)$

Thus every local maxima of $log\ f(x)$ is also a local maxima of $f(x)$

For example, Consider $f(\theta)$ and $g(\theta)$ from Problem 1
$f(\theta) = \theta^t(1-\theta)^h$
$f(\theta)$ is a positive differentiable function
$g(\theta) = log(f(\theta)) = t\ log(\theta) + h\ log(1-\theta)$

Finding local maxima for $g(\theta)$ put $\frac{dg}{d\theta} = 0$

$\frac{t}{\theta} - \frac{h}{1-\theta} = 0$

$t - t\theta = h\theta$
$t = (h+t)\theta$
$\theta = \frac{t}{h+t}$ is a critical point. To find whether it is maximum , substitiute it in $\frac{d^2g}{d\theta^2}$

$\frac{-t}{(\frac{t}{t+h})^2} - \frac{h}{(1-\frac{t}{t+h})^2} = -(h+t)^2[\frac{1}{t} + \frac{1}{h}] =$ Negative value
Thus $\theta = \frac{t}{h+t}$ is the Maxima for $g(\theta)$

Now substitute $\theta = \frac{t}{h+t}$ in $\frac{df}{d\theta}$ we get,
$\frac{df}{d\theta} = t\theta^{t-1}(1-\theta)^h - h\theta^t(1-\theta)^{h-1} = t(\frac{t}{t+h})^{t-1}(\frac{h}{t+h})^h - h(\frac{t}{t+h})^t(\frac{h}{t+h})^{h-1} =$
$(\frac{t}{t+h})^{t-1}(\frac{h}{t+h})^{h-1}[\frac{th}{t+h} - \frac{th}{t+h})] = 0$

This means that $\frac{t}{t+h}$ is also a critical point for $f(\theta)$

Now to check if it is maxima or minima , we need to substitute it in $\frac{d^2f}{d\theta^2}$
$\frac{d^2f}{d\theta^2} = \theta^{t-2}(1-\theta)^{h-2}[t(t-1)(1-\theta)^2 - 2th\theta(1-\theta) + h(h-1)\theta^2]$
$= (\frac{t}{t+h})^{t-2}(\frac{h}{t+h})^{h-2}\frac{ht}{(t+h)^2}[-(h+t)] =$ -Negative Value

Thus this point is also a maxima for $f(\theta)$ as well

Thus to conclude it can be said that **arg max f(x) = arg max log[f(x)]**. So to find the maxima / minima of complex positive differentiable functions we should first compute their log and then find maxima for $log[f(x)]$ as this is easy to compute and gives the same result.

Problem 3

MLE and MAP both compute point estimates.
Say we have a likelihood function $P(X|\theta)$, then MLE for θ the parameter we want to infer is :

$$\theta_{MLE} = \arg\max_\theta P(X|\theta) = \arg\max_\theta \prod_i P(x_i|\theta)$$

We will instead work in the log space, as logarithm is monotonically increasing, so maximizing a function is equal to maximizing the log of that function.

$$\theta_{MLE} = \arg\max_\theta \, log\ (P(X|\theta)) = \arg\max_\theta \, log(\prod_i P(x_i|\theta)) = \arg\max_\theta \ \sum_i log(P(x_i|\theta))$$

MAP which is the posterior function $P(\theta|X)$ can be expressed in terms of Prior and likelihood as :
$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}$

$P(\theta|X) \propto P(X|\theta)P(\theta)$

If we replace the likelihood in the MLE formula above with the posterior, we get:

$$\theta_{MAP} = \arg\max_\theta P(\theta|X) = \arg\max_\theta P(X|\theta)P(\theta)$$
$$\theta_{MAP} = \arg\max_\theta P(X|\theta)P(\theta) = \arg\max_\theta \prod_i P(x_i|\theta)P(\theta) = \arg\max_\theta \ \sum_i log(P(x_i|\theta)P(\theta))$$

Comparing both MLE and MAP equation, the only thing differs is the inclusion of prior $P(\theta)$ in MAP, otherwise they are identical.

Suppose our prior function is *const.* everywhere in the distribution. In this case , we can ignore the $P(\theta)$ in our estimation as shown:

$$\theta_{MAP} = \arg\max_\theta \ \sum_i log(P(x_i|\theta)P(\theta))$$
$$\theta_{MAP} = \arg\max_\theta \ \sum_i log(P(x_i|\theta)const.)$$
$$\theta_{MAP} = \arg\max_\theta \ \sum_i log\ P(x_i|\theta)$$

$\theta_{MAP} = \theta_{MLE}$
Hence Proved

Problem 4

$p(X|\theta) = Ber(X)$ with m occurences for X=1 and l occurences for X=0

$N = m + l$

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

$p(X|\theta) = \binom{N}{m}\theta^m(1-\theta)^{N-m}$

We have prior distribution as $p(\theta) = Beta(\theta|a, b)$

$p(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}$

We know that for posterior distribution we have,

$p(\theta|X) = \frac{p(X|\theta)p(\theta)}{p(X)}$

$p(\theta|X) \propto p(X|\theta)p(\theta)$

$p(\theta|X) \propto \theta^m(1-\theta)^{N-m}\theta^{a-1}(1-\theta)^{b-1}$

Reverse Engineering we have posterior distribution as
$p(\theta|X) = Beta(\theta|m + a, N - m + b)\frac{\Gamma(N+a+b)}{\Gamma(m+a)\Gamma(N-m+b)}\theta^{m+a-1}(1-\theta)^{N-m+b-1}$

Posterior Mean = $E[\theta|X] = \frac{m+a}{N+a+b} = x(say)$

Prior Mean = $E[\theta] = \frac{a}{a+b} = y(say)$

Max. Likelihood $\theta_{MLE} = \frac{m}{N} = z(say)$

We have,
Posterior Mean $=$ λ Prior Mean $+$ $(1-\lambda)\theta_{MLE}$
$\Rightarrow x = \lambda y + (1-\lambda)z \Rightarrow x = \lambda y + z - \lambda z \Rightarrow \lambda = \frac{x-z}{y-z}$

$0 \leq \lambda \leq 1 \Rightarrow 0 \leq \frac{x-z}{y-z} \leq 1$

Using $\frac{x-z}{y-z} \geq 0$
$\Rightarrow x - z \geq 0 \Rightarrow x \geq z...(1)$

Using $\frac{x-z}{y-z} \leq 1$
$\Rightarrow \frac{x-z}{y-z} - 1 \leq 0 \Rightarrow x - y \leq 0 \Rightarrow x \leq y...(2)$

Using (1) and (2), we have

$z \leq x \leq y$
$\Rightarrow \theta_{MLE} \leq E[\theta|X] \leq E[\theta]$
$\Rightarrow \theta_{MLE} \leq Posterior\ Mean \leq Prior\ Mean$

OR

Expected Posterior mean for $\theta =$

$$E[\theta|X] = \frac{m+a}{m+l+a+b}$$

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

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

here

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

and

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

$$E[\theta|X] = \lambda E[\theta] + (1-\lambda)\theta_{MLE}$$

Hence proved, as $\frac{m}{m+l}$ is the maximum likelihood estimate and $\frac{a}{a+b}$ is the prior mean value of θ.

Problem 5

Random Variable X is Poisson distributed
$p(X=x) = \frac{e^{-\lambda}\lambda^x}{x!}$
$\lambda_{MLE} = \arg\max\limits_{\lambda} p(X|\lambda) = \arg\max\limits_{\lambda} \prod\limits_{i=1}^{n} p(x_i|\lambda)$
Taking log we have,
$\lambda_{MLE} = \arg\max\limits_{\lambda} log(\prod\limits_{i=1}^{n} p(x_i|\lambda))$
$f(\lambda) = log(\prod\limits_{i=1}^{n} (\frac{e^{-\lambda}\lambda^{x_i}}{x_i!})) = \sum\limits_{i=1}^{n} log(e^{-\lambda}) + \sum\limits_{i=1}^{n} log(\frac{\lambda^{x_i}}{x_i!})$
$f(\lambda) = -n\lambda + \sum\limits_{i=1}^{n}(log\ \lambda^{x_i} - log\ x_i!)$
$\frac{df}{d\lambda} = 0 \Rightarrow -n + \sum\limits_{i=1}^{n} \frac{x_i}{\lambda} = 0$
$\lambda_{MLE} = \frac{\sum\limits_{i=1}^{n} x_i}{n}$

Given the Prior Distribution, we have

$p(\lambda) = Gamma(\lambda|\alpha,\beta) = \frac{\beta^\alpha \lambda^{\alpha-1} e^{-\beta\lambda}}{\Gamma(\alpha)}$

$p(\lambda|X) \propto p(X|\lambda)p(\lambda) = (\prod\limits_{i=1} n \frac{e^{-\lambda}\lambda^{x_i}}{x_i!})\frac{\beta^\alpha \lambda^{\alpha-1} e^{-\beta\lambda}}{\Gamma(\alpha)}$

Ignoring the constants , we have
$\propto e^{-\lambda(n+\beta)}\lambda^{\sum\limits_{i=1}^{n} x_i+\alpha-1}$

Reverse Engineering , we get Gamma distribution with $\alpha' = \sum\limits_{i=1}^{n} x_i + \alpha$ and $\beta' = n+\beta$

$p(\lambda|X) = Gamma(\lambda|\sum\limits_{i=1}^{n} x_i + \alpha, n+\beta)$

$$\lambda_{MAP} = \arg\max_{\lambda} log(p(\lambda|X))$$

$$g(\lambda) = log(p(\lambda|X)) = log(\lambda^{\sum\limits_{i=1}^{n} x_i + \alpha - 1} e^{-\lambda(n+\beta)}) + C$$
$$g(\lambda) = -\lambda(n+\beta) + (\sum_{i=1}^{n} x_i + \alpha - 1) log\ \lambda$$
$$\frac{dg}{d\lambda} = 0 \Rightarrow -(n+\beta) + \frac{\sum\limits_{i=1}^{n} x_i + \alpha - 1}{\lambda} = 0$$
$$\lambda_{MAP} = \frac{\sum\limits_{i=1}^{n} x_i + \alpha - 1}{n+\beta}$$