Machine Learning Homework 2

Abinav Ravi Venkatakrishnan & Abhijeet Parida

November 5, 2018

1 Optimising Likelihood: Monotonic Transforms

Problem 1

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

$f = log(\theta^t(1-\theta)^h)$
$f = tlog(\theta) + hlog(1-\theta)$
First Derivative:
$\frac{df}{d\theta} = \frac{t}{\theta} + \frac{h}{1-\theta}$
Second Derivative:
$\frac{d}{d\theta}\frac{df}{d\theta} = \frac{-t}{\theta^2} + \frac{h}{(1-\theta)^2}$

Problem 2

Let local maxima for any positive differentiable function $f(\theta)$ be θ^*. Then θ^* satisfies the condition:

1. $\frac{d}{d\theta}f(\theta^*) = 0$
2. $\frac{d^2}{d\theta^2}f(\theta^*) < 0$

For the function $log(f(\theta))$, we find the first derivative and equate to zero,

$$\frac{d}{d\theta}log(f(\theta)) = \frac{1}{f(\theta)}\frac{d}{d\theta}f(\theta)$$

For a positive function $f(\theta)$ is positive and from 1, $\frac{d}{d\theta}f(\theta) = 0$ at θ^*. Also, the second derivative of $log(f(\theta))$ is

$$\frac{d^2}{d\theta^2}log(f(\theta)) = -\frac{1}{f(\theta)^2}\frac{d}{d\theta}f(\theta) + \frac{1}{f(\theta)}\frac{d^2}{d\theta^2}f(\theta)$$

In the above equation we can notice that the first term is negative and from 2 second term is also negative hence proving that θ^* is a local maximum of both $f(\theta)$ and $log(f(\theta))$

From the two problem we understand that Monotonic functions preserve the critical points. Conversion of the problem to a log scale makes the computation easy and is preferred.

2 Properties of MLE and MAP

Problem 3

from Bayes rule we know that

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)} \tag{1}$$

now we know that $p(\theta|D)$ is the prior estimate θ_{MAP} and $p(D|\theta)$ is the Maximum Likelihood Estimate θ_{MLE} so we get the relation

$$posterior \propto likelihood \text{ x } prior$$

for a uniform prior $p(\theta)$ we can say that θ_{MLE} is a special case of θ_{MAP}

Problem 4

$$posterior \propto likelihood \text{ x } prior$$

The prior is beta distribution which is conjugate prior for the binomial distribution of the likelihood. Therefore, from lectures the distribution of the posterior is same as the prior distribution.

$$posterior \propto \ ^{N}C_m \theta^m (1-\theta)^{N-m} * \frac{\Gamma(a+b)}{\Gamma(a)+\Gamma(b)} \theta^{a-1}(1-\theta)^{N-m}$$

by simple pattern matching with Beta distribution we get posterior mean as

$$\mathbb{E}[\theta|\mathcal{D}] = \frac{m+a}{m+l+a+b} \tag{2}$$

From the mean of the beta distribution we get

$$\mathbb{E}[p(\theta)] = \frac{a}{a+b} \tag{3}$$

Also θ_{MLE} From the Eq. 2

$$\frac{m+a}{m+l+a+b} = \tfrac{m}{m+l+a+b} + \tfrac{a}{m+l+a+b} \tag{4}$$

$$= \tfrac{m}{m+l} \tfrac{m+l}{m+l+a+b} + \tfrac{a}{a+b} \tfrac{a+b}{m+l+a+b} \tag{5}$$

$$= (1-\lambda) \tfrac{m}{m+l} + \lambda \tfrac{a}{a+b} \tag{6}$$

Hence the posterior mean value $\mathbb{E}[\theta|\mathcal{D}]$ lies between prior mean of θ and the θ_{MLE}.

3 Poisson Distribution

Problem 5

Part a

The $\mathbb{E}[\lambda]$ is the mean of the Poisson distribution which is λ, which gives the maximum likelihood estimate.

Part b

It is known that the conjugate prior of the Poisson distribution is $Gamma(\alpha, \beta)$. Therefore our posterior is also $Gamma(\alpha, \beta)$.
So,

$$p(\lambda|\mathcal{D}) \propto p(\mathcal{D}|\lambda) * p(\lambda)$$

$$= e^{(-n\lambda)} \textstyle\prod_{i=1}^{n} \tfrac{\lambda^{k_i}}{k_i!} e^{-\beta\lambda}$$

$$\propto \lambda^{\sum_{i=1}^{n} k_i + \alpha - 1} e^{-n\lambda - \beta\lambda}$$

By simple pattern matching the mode gamma distribution, we get

$$\lambda = \frac{\sum_{i=1}^{n} k_i + \alpha - 1}{n+\beta} \tag{7}$$