Machine Learning homework 2 solution

Parameter Inference

Wiktor Jurasz - M.Nr. 03709419

November 4, 2018

1 Problem 1

$$\frac{\partial(\theta^t(1-\theta)^n)}{\partial\theta} = t\theta^{t-1}(1-\theta)^n - n\theta^t(1-\theta)^{n-1} \tag{1}$$

$$\begin{aligned}\frac{\partial^2(\theta^t(1-\theta)^n)}{\partial\theta^2} = \frac{\partial(\theta^{t-1}(1-\theta)^n - n\theta^t(1-\theta)^{n-1})}{\partial\theta} = \\ -2tn\theta^{t-1}(1-\theta)^{n-1} + t(t-1)\theta^{t-2}(1-\theta)^n + n(n-1)\theta^t(1-\theta)^{n-2}\end{aligned} \tag{2}$$

$$\frac{\partial(\ln(\theta^t(1-\theta)^n))}{\partial\theta} = \frac{\partial(t\ln(\theta) + nln(1-\theta)))}{\partial\theta} = \frac{t}{\theta} + \frac{n}{\theta-1} \tag{3}$$

$$\frac{\partial^2(\ln(\theta^t(1-\theta)^n))}{\partial\theta^2} = \frac{\partial^2(t\ln(\theta) + nln(1-\theta)))}{\partial\theta^2} = \frac{\partial(\frac{t}{\theta} + \frac{n}{\theta-1})}{\partial\theta} = -\frac{t}{\theta^2} - \frac{n}{(\theta-1)^2} \tag{4}$$

2 Problem 2

Assumptions:

1. $g : \mathbb{R}^+ \mapsto \mathbb{R}$
2. $f : C \mapsto \mathbb{R}^x$
3. $f \ is \ differentiable$
4. $g(x) = ln(x) \Rightarrow \forall x_1, x_2 \in \mathbb{R}^+, g(x_2) > g(x_1) \Leftrightarrow x_2 > x_1$

Proof:

$$\begin{aligned}&g(f(x_o)) \ is \ a \ local \ maximum \Leftrightarrow \\ &\Leftrightarrow \exists\epsilon_1, \epsilon_2 \in \mathbb{R} : \forall x \in (x_0 - \epsilon_1, x_o + \epsilon_2) \ g(f(x_0)) > g(f(x)) \Rightarrow (from \ 4) \\ &\Rightarrow f(x_0) > f(x) \ \forall x \in (x_0 - \epsilon_1, x_o + \epsilon_2) \Rightarrow \\ &\Rightarrow f(x_0) \ is \ a \ local \ maximum\end{aligned} \tag{5}$$

As we can see the log derivatives are much simpler to computer.
We also see that log(f) has maximum in the same place as f.

Because the goal is to find an argument for which function has maximum value (not the value itself), we can use log(f) instead of f to makes computations easier.

3 Problem 3

$$\begin{aligned}\theta_{MLE} = \underset{\theta}{\operatorname{argmax}} P(D|\theta) = \underset{\theta}{\operatorname{argmax}} log(P(D|\theta)) = \underset{\theta}{\operatorname{argmax}}[log(P(D|\theta)) + C] = \\ = \underset{\theta}{\operatorname{argmax}}[log(P(D|\theta)) + log(P(\theta))] = \theta_{MAP} \\ \Leftrightarrow log(P(\theta)) = C \; and \; C \; is \; constant \Rightarrow P(\theta) \; is \; a \; unifrom \; distribution\end{aligned} \tag{6}$$

4 Problem 4

4.1 Determining Posteriori

$$priori = p(\theta|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1} \tag{7}$$

$$likelihood = p(x=m|\theta,N) = \binom{N}{m}\theta^{m}(1-\theta)^{N-m} \tag{8}$$

Now we can plug this in into Bayes' Theorem:

$$p(\theta|a,b,N,x=m) = \frac{p(x=m|\theta,N)p(\theta)}{p(x=m)} = \frac{1}{P(x=m)}\binom{N}{m}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma b}\theta^{a+m-1}(1-\theta)^{b+N-m-1} \tag{9}$$

Now to determine the distribution of posteriori we can do following:
$p(x=m)$ is the marginal likelihood which is defined as $\int p(x=m|\theta,N)p(\theta)\frac{\partial}{\partial\theta}$
Using previous equation we have:

$$p(x=m) = \int \binom{N}{m}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma b}\theta^{a+m-1}(1-\theta)^{b+N-m-1}\frac{\partial}{\partial\theta} \tag{10}$$

Now plugging this back into posteriori equation and cancelling constants we have:

$$p(\theta|a,b,N,x=m) = \frac{\theta^{a+m-1}(1-\theta)^{b+N-m-1}}{\int \theta^{a+m-1}(1-\theta)^{b+N-m-1}\frac{\partial}{\partial\theta}} \tag{11}$$

The denominator is a B function from $Beta$ distribution so in the end we can write:

$$p(\theta|a,b,N,x=m) = \frac{\theta^{a+m-1}(1-\theta)^{b+N-m-1}}{B(a+m,b+N-m)} \tag{12}$$

The other (simpler) approach is to check that for conjugate priori Beta and binominal likelihood, the posteriori is also Beta distribution.
Because we know that:

$$p(\theta|a,b,N,x=m) \propto p(x=m|\theta,N)p(\theta) \propto \theta^{a+m-1}(1-\theta)^{b+N-m-1} \tag{13}$$

(We can omit constants for proportions)
Then the only thing left is to add appropriate normalization constant to change $\propto$ to $=$ which for Beta distribution with given parameters equals to: $\frac{1}{B(a+m,b+N-m)}$.
Thus:

$$p(\theta|a,b,N,x=m) = \frac{\theta^{a+m-1}(1-\theta)^{b+N-m-1}}{B(a+m,b+N-m)} \tag{14}$$

4.2 Mean and MLE

Mean for Beta distrbution is given as:

$$\mu = \mathbb{E}[X] = \frac{\alpha}{\alpha + \beta} \tag{15}$$

For above posteriori:

$$\mu_{posteriori} = \mathbb{E}[\theta] = \frac{a + m}{a + b + N} \tag{16}$$

For above prior:

$$\mu_{prior} = \mathbb{E}[\theta] = \frac{a}{a + b} \tag{17}$$

From equations in Problem 1 we know that:

$$\theta_{MLE} = \frac{t}{t + n} \tag{18}$$

For above likelihood:

$$\theta_{MLE} = \frac{m}{N} \tag{19}$$

4.3 Proof

$$\mathbb{E}[\theta|D] = \frac{a+m}{a+b+N} = \frac{(a+b)\frac{a}{a+b} + m}{a+b+N} = \frac{a+b}{a+b+N}\frac{a}{a+b} + \frac{m}{a+b+N}\frac{N}{N} = \frac{a+b}{a+b+N}\frac{a}{a+b} + \frac{N}{a+b+N}\frac{m}{N} =$$
$$= \lambda\frac{a}{a+b} + (1-\lambda)\frac{m}{N} = \lambda\mu_{prior} + (1-\lambda)\theta_{MLE} \tag{20}$$

5 Problem 5

5.1 MLE

Poisson distribution looks as follows:

$$p(X|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!} \tag{21}$$

For given samples from X likelihood is defined as a product of particular probabilities:

$$L(\lambda|x_1, x_2, ..., x_n) = \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} \tag{22}$$

From this log-likelihood can be obtained:

$$l(\lambda|x_1, x_2, ..., x_n) = ln(\prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}) = \sum_{i=1}^{n} ln(\frac{\lambda^{x_i} e^{-\lambda}}{x_i!}) =$$
$$= nln(e^{-\lambda}) + \sum_{i=1}^{n} ln(\lambda^{x_i}) - \sum_{i=1}^{n} ln(x_i!) = -n\lambda + ln(\lambda)\sum_{i=1}^{n}(x_i) - \sum_{i=1}^{n} ln(x_i!) \tag{23}$$

To find maximum we can take the first derivative $\frac{\partial}{\partial\lambda}$ from equation above:

$$\frac{\partial}{\partial\lambda}(-n\lambda + ln(\lambda)\sum_{i=1}^{n}(x_i) - \sum_{i=1}^{n} ln(x_i!)) = -n + \frac{1}{\lambda}\sum_{i=1}^{n}(x_i) \tag{24}$$

Which compared to 0 gives us estimator:

$$\lambda_{EST} = \frac{1}{n}\sum_{i=1}^{n}(x_i) \tag{25}$$

5.2 Unbiased estimator

$$\mathbb{E}[\lambda_{EST}] = \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n}(x_i)] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[(x_i)] = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu = \lambda \tag{26}$$

5.3 Posteriori

Gamma priori is conjugate to Poisson likelihood. Thus we know that posteriori is also Gamma. From:

$$posteriori \propto likelihood * prior \tag{27}$$

we can derive the posteriori Gamm parameters (or we can check this in some clever book) which gives us:

$$posteriori = \Gamma(\sum x_i + \alpha, n + \beta) \propto \lambda^{\sum x_i + \alpha - 1} e^{-\lambda(n+\beta)} \tag{28}$$

Then to calculate θ_{MAP} we can either derive it in a similar way as we did for MLE or we can look up what is the mode of Gamma distribution and plug in our parameters, thus:

$$\theta_{MAP} = \frac{\sum x_i + \alpha - 1}{n + \beta} \tag{29}$$