Machine Learning Exercise Sheet 1

Math Refresher

Group_369

Fan Xue – fan98.xue@tum.de

Xing Zhou – xing.zhou@tum.de

Jianzhe Liu – jianzhe.liu@tum.de

November 10, 2021

Problem 6

First let's compute the first and second derivative of $\theta^t(1-\theta)^h$, According to the derivative calculation rule:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\theta}\theta^t(1-\theta)^h &= t\theta^{t-1}(1-\theta)^h + \theta^t(-1)h(1-\theta)^{h-1} \\ &= \theta^{t-1}(1-\theta)^{h-1} * ((1-\theta)t - \theta h)\end{aligned}$$

$$\begin{aligned}\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\theta^t(1-\theta)^h &= \frac{\mathrm{d}}{\mathrm{d}\theta}\left(t\theta^{t-1}(1-\theta)^h + \theta^t(-1)h(1-\theta)^{h-1}\right) \\ &= \theta^{t-2}(1-\theta)^{h-2} * ((1-\theta)(t-1) - \theta(h-1)) * ((1-\theta)t - \theta h) - \theta^{t-1}(1-\theta)^{h-1}(t+h)\end{aligned}$$

It's quite obvious that as the derivation goes on, the result becomes much more complex, which is really hard for us to read and analyse.

Then let's try computing the first and second derivative of $\log\theta^t(1-\theta)^h$:

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

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\theta}f(\theta) &= \frac{t}{\theta} - \frac{h}{1-\theta} \\ \frac{\mathrm{d}^2}{\mathrm{d}\theta^2}f(\theta) &= \frac{\mathrm{d}}{\mathrm{d}\theta}\left(\frac{t}{\theta} - \frac{h}{1-\theta}\right) \\ &= -\frac{t}{\theta^2} - \frac{h}{(1-\theta)^2}\end{aligned}$$

This time the result is much more clear, and it is of course easier for us to do the further calculation or analysis.

Problem 7

To prove this, let's say we have an arbitrary local maximum θ_a in $\log f(\theta)$, obviously in any small range around θ_a, we have $\log f(\theta_a) \geq \log f(\theta)$.

Then we use function *exp* to transform $\log f(\theta)$ into $f(\theta)$.

Since function *exp* is monotone, we have:

$$f(\theta_a) = e^{\log f(\theta_a)} \geq f(\theta) = e^{\log f(\theta)}$$

Therefore θ_a is also a maximum of $f(\theta)$.

Conclusion: with the result of last problem, we know the using function *log* can help us simplify the calculation and thus the further analysis. And with the result of this problem we know that by using *log* transform, the feature of the original equations remain unchanged, which means we can safely apply this transform in further studies.

Problem 8

To proof this, we need to proof a Lemma first.
Lemma:
For $a, b, c > 0$, $0 < \lambda < 1$, if $c = \lambda a + (1-\lambda)b$, then c lies between a and b.
Proof:
Assuming that $a < b$, we have

$$c - a = (\lambda - 1)a + (1-\lambda)b = (1-\lambda)(b-a) > 0c - b = \lambda a - \lambda b = \lambda(a-b) < 0$$

that indicates that c lies bewteen a and b.
Now we change bach to the problem.
According to the given information, we know that the **prior distribution** $p(\theta) = Beta(a,b)$, and the **likelihood** $p(D|\theta) = Binary(m, N, \theta)$. That is

$$p(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}$$
$$p(D|\theta) = C_N^m \theta^m (1\ \theta)^{N-m}$$

We can know that the **posterior distribution**

$$p(\theta|D) \sim p(D|\theta)p(\theta) \sim \theta^{a+m-1}(1-\theta)^{b+N-m-1}$$

Obviously the **posterior distribution** is a Beta distribution, that is

$$p(\theta|D) \sim Beta(a+m, b+N-m)$$

Let $\log p(D|\theta) = 0$, we can find the maximum likelihood estimate

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

From the konwledge of Beta distribution, we know the posterior mean value of θ

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

and the prior mean of θ

$$\mathbb{E}(\theta) = \frac{a}{a+b}$$

We can rewrite the posterior mean of θ as following

$$\mathbb{E}(\theta|D) = \frac{a+m}{a+b+N} = \frac{N}{a+b+N} \cdot \frac{m}{N} + \frac{a+b}{a+b+N} \cdot \frac{a}{a+b} = \lambda\mathbb{E}(\theta) + (1-\lambda)\theta_{MLE}$$

From the lemma we can proof that $\mathbb{E}\theta$ lies between $\mathbb{E}(\theta)$ and θ_{MLE}.

Problem 9

The MAP estimation of the parameter λ is:

$$\begin{aligned}
\lambda_{MAP} &= \arg\max_{\lambda} \mathrm{p}(\lambda \mid x, a, b) \\
&= \arg\max_{\lambda} \log \mathrm{p}(\lambda \mid x, a, b) \\
&= \arg\max_{\lambda} \log(\mathrm{p}(x \mid \lambda)\, \mathrm{p}(\lambda \mid a, b)) \\
&= \arg\max_{\lambda} \log(\frac{b^a \lambda^{a-1} \exp(-b\lambda)}{\Gamma(a)} \frac{\lambda^x \exp(-\lambda)}{x!}) \\
&= \arg\max_{\lambda} (a-1+x)\log\lambda - (b+1)\lambda + const
\end{aligned}$$

In order to maximize the function, compute the derivative:

$$\frac{\partial}{\partial\lambda}((a-1+x)\log\lambda - (b+1)\lambda + const) = \frac{a-1+x}{\lambda} - b - 1 \overset{!}{=} 0$$

Then we get

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

Hence $\lambda_{MAP} = \frac{x+a-1}{b+1}$.