Machine Learning Worksheet 3

Tomas Ladek, Michael Kratzer
3602673, 3612903
tom.ladek@tum.de, mkratzer@mytum.de

Problem 1

$$\begin{aligned}\frac{\partial}{\partial\theta}\theta^t(1-\theta)^h &= t\theta^{t-1}(1-\theta)^h - \theta^t h(1-\theta)^{h-1} \\ &= \theta^{t-1}(1-\theta)^{h-1}\big(t(1-\theta)-h\theta\big)\end{aligned}$$

$$\begin{aligned}\frac{\partial^2}{\partial\theta^2}\theta^t(1-\theta)^h &= \frac{\partial}{\partial\theta}\theta^{t-1}(1-\theta)^{h-1}\big(t(1-\theta)-h\theta\big) \\ &= -2\,t\,h\,\theta^{t-1}(1-\theta)^{h-1} + t(t-1)\theta^{t-2}(1-\theta)^h + h(h-1)\theta^t(1-\theta)^{h-2}\end{aligned}$$

$$\begin{aligned}\frac{\partial}{\partial\theta}\log\theta^t(1-\theta)^h &= \frac{\partial}{\partial\theta}\big(t\log\theta + h\log(1-\theta)\big) \\ &= \frac{t}{\theta} - \frac{h}{1-\theta}\end{aligned}$$

$$\begin{aligned}\frac{\partial^2}{\partial\theta^2}\log\theta^t(1-\theta)^h &= \frac{\partial}{\partial\theta}(\frac{t}{\theta} - \frac{h}{1-\theta}) \\ &= \frac{h}{(1-\theta)^2} - \frac{t}{\theta^2}\end{aligned}$$

Problem 2

The local extremum of $\log f(\theta)$ can determined by setting the first derivative to zero, i.e.

$$\frac{d}{d\theta}\log\big(f(\theta)\big) \overset{!}{=} 0$$

Evaluating this with the logarithm derivative rule leads to

$$\frac{\frac{d}{d\theta}f(\theta)}{f(\theta)} = 0 \iff \frac{d}{d\theta}f(\theta) = 0$$

Which is also the formula for the extremum of a differentiable positive function $f(\theta)$. That means that taking the logarithm of a function preserves its extremum points (in particular its local maximum points). Considering the results from the first exercise, it can be easier to differentiate the logarithm of a function instead of the function itself, when one is interested only in the location of the maximum and not its value.

Problem 3

θ_{MLE} is a special case of θ_{MAP} with the following prior:

$$Beta(\theta|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1} \quad with\ a=b=1$$

That way the posterior

$$p(\theta = x|\mathcal{D}) = \frac{p(\mathcal{D}|\theta = x)\cdot p(\theta = x)}{p(\mathcal{D})}$$

simplifies to

$$\begin{aligned} p(\theta = x|\mathcal{D}) &= \frac{1}{p(\mathcal{D})}\cdot x^{|T|}(1-x)^{|H|}\cdot \frac{\Gamma(1+1)}{\Gamma(1)\Gamma(1)}x^0(1-x)^0 \\ &= \frac{1}{p(\mathcal{D})}\cdot x^{|T|}(1-x)^{|H|}\cdot \frac{1}{1\cdot 1} \\ &= \frac{1}{p(\mathcal{D})}\cdot x^{|T|}(1-x)^{|H|} \\ &\propto x^{|T|}(1-x)^{|H|} \end{aligned}$$

which equals to the maximum likelihood estimate θ_{MLE}. Another way of seeing this is by taking the formula for θ_{MAP} directly (from the slides):

$$\theta_{MAP} = \frac{|T|+a-1}{|H|+|T|+a+b-2} \overset{a=b=1}{=} \frac{|T|+1-1}{|H|+|T|+1+1-2} = \frac{|T|}{|H|+|T|} = \theta_{MLE}$$

Problem 4

Given that the prior is Beta-distributed and the likelihood is Bin-distributed, the posterior distribution has to be Bin as well. The expected value of the posterior therefore has to be in the form $N\theta$.

According to the hint, the posterior mean can be written as

$$\mathbb{E}[\theta|\mathcal{D}] = \lambda\mathbb{E}[\theta = x] + (1-\lambda)\theta_{MLE}$$

The mean of the Beta-distributed prior $\mathbb{E}[\theta = x]$ is given by $\frac{a}{a+b}$ (lecture notes), the maximum likelihood estimate of a Bin-distribution is the maximum of the function $\binom{N}{m}\theta^m(1-\theta)^{N-m}$, which is $\lfloor (N+1)\theta \rfloor$ (since θ can take non-integer values between 0 and 1). Together:

$$\mathbb{E}[\theta|\mathcal{D}] = \lambda\frac{a}{a+b} + (1-\lambda)\lfloor (N+1)\theta \rfloor$$

TODO One can easily see, that with $\lambda = 0$ and some magic,

$$\lambda\frac{a}{a+b} + (1-\lambda)\lfloor (N+1)\theta \rfloor = N\theta$$

Problem 5

The Poisson PDF is given by $\frac{e^{-\lambda}\lambda^x}{x!}$. The maximum likelihood estimate for λ can be found in this way:

$$\begin{aligned}
\frac{\partial}{\partial\lambda}\frac{e^{-\lambda}\lambda^x}{x!} &\stackrel{!}{=} 0 \\
\Rightarrow \frac{e^{-\lambda}(x-\lambda)\lambda^{x-1}}{x!} &= 0 \\
\iff (x-\lambda) &= 0 \ (i.e.\ \lambda = x \ and\ x > 0) \quad or \quad \lambda = 0
\end{aligned}$$

TODO One can easily see that this estimate is unbiased.

The posterior distribution given the Poisson likelihood and a Gamma prior looks as follows:

$$p(\theta = x|\mathcal{D}) = \frac{1}{p(\mathcal{D})} \cdot \frac{e^{-\lambda}\lambda^x}{x!} \cdot \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx} = \frac{b^a}{p(\mathcal{D})\Gamma(a)} \cdot \frac{e^{-\lambda-bx}\lambda^x x^{a-1}}{x!}$$

The MAP for λ can be computed this way (**TODO** correct?):

$$\begin{aligned}
\frac{\partial}{\partial\lambda}\Big(\frac{b^a}{p(\mathcal{D})\Gamma(a)} \cdot \frac{e^{-\lambda-bx}\lambda^x x^{a-1}}{x!}\Big) &\stackrel{!}{=} 0 \\
\iff \frac{b^a}{p(\mathcal{D})\Gamma(a)} \cdot \Big(\frac{e^{-\lambda-bx}\lambda^{x-1}x^a}{x!} - \frac{e^{-\lambda-bx}\lambda^x x^{a-1}}{x!}\Big) &= 0 \\
\iff \lambda = x \ and\ x > 0 \quad or \quad \lambda = 0
\end{aligned}$$