Machine Learning: Homework #3

Due on November 13, 2017 at 09:55am

Professor Dr. Stephan Guennemann

Marc Meissner - 03691673

Problem 1

Usually one considers the log likelihood $\log p(x_1, ..., x_n|\theta)$. The next problems justifies this. In the lecture, we encountered the likelihood maximization problem

$$\arg\max_{\theta \in [0,1]} \theta^t (1-\theta)^h$$

, where t and h denoted the number of tails and heads in a sequence of coin tosses, respectively. Compute the first and second derivative of this likelihood w.r.t. θ. Then compute first and second derivative of the log likelihood $\log\theta^t(1-\theta)^h$.

Solution
First equation:

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

$\frac{\partial f(\theta)}{\partial \theta}$
$= t\theta^{t-1}(1-\theta)^h - \theta^t h(1-\theta)^{h-1}$
$= \theta^{t-1}(1-\theta)^{h-1}(t(1-\theta) - h\theta)$

$\frac{\partial^2 f(\theta)}{\partial \theta^2} = ...$
$= [\theta^{t-2}(1-\theta)^{h-2} \cdot ((t-1)(1-\theta) - (h-1)\theta) \cdot (t(1-\theta) - h\theta)] - \theta^{t-1}(1-\theta)^{h-1}(t+h)$

Second equation:

$g(\theta) = \ln[\theta^t(1-\theta)^h] = t\ln\theta + h\ln(1-\theta)$

$\frac{\partial g(\theta)}{\partial \theta} = \frac{t}{\theta} - \frac{h}{1-\theta}$

$\frac{\partial^2 f(\theta)}{\partial \theta^2} = -\frac{t}{\theta^2} - \frac{h}{(1-\theta)^2}$

Problem 2

Show that every local maximum of $\log f(\theta)$ is also a local maximum of the differentiable, positive function $f(\theta)$. Considering this and the previous exercise, what is your conclusion?

Solution

Problem 2 continued on next page. . .

Given $f(\theta)$ with local maximum $\theta_{\max}$. Then, for a certain ϵ, it locally holds that:

$f(\theta_{\max}) \geq f(\theta)$ for any $\theta \in [\theta_{\max} - \epsilon; \theta_{\max} + \epsilon]$

Since $g(\theta) = \ln f(\theta)$ is a monotonic function, the following properties hold:

$x_2 - x_1 \implies \ln(x_2) > \ln(x_1)$
$g(\theta_{\max}) = \ln f(\theta_{\max}) \geq \ln f(\theta) = g(\theta)$

Thus, one can apply log-likelihood and preserve the position of the maximum while severely reducing the complexity of the solution.

Problem 3

Show that θ_{MLE} can be interpreted as a special case of θ_{MAP} in the sense that there always exists a prior $p(\theta)$ such that $\theta_{\text{MLE}} = \theta_{\text{MAP}}$.

Solution
Any constant prior (uniform distribution) should preserve the position of the maximum, since it just scales the distribution. Given $p(\theta) = c$ and the definitions from the slides:

$$\theta_{\text{MAP}} = \arg\max_{\theta} p(D|\theta)p(\theta) = \arg\max_{\theta} p(D|\theta)c = \arg\max_{\theta} p(D|\theta) = \theta_{\text{MLE}}$$

Problem 4

Consider a Bernoulli random variable X and suppose we have observed m occurrences of $X = 1$ and l occurrences of $X = 0$ in a sequence of $N = m + l$ Bernoulli experiments. We are only interested in the number of occurrences of $X = 1$. We will model this with a Binomial distribution with parameter θ. A prior distribution for θ is given by the Beta distribution with parameters a, b. Show that the posterior mean value $E[\theta|D]$ (not the MAP estimate) of θ lies between the prior mean of θ and the maximum likelihood estimate for θ. To do this, show that the posterior mean can be written as λ times the prior mean plus $(1 - \lambda)$ times the maximum likelihood estimate, with $\theta \leq \lambda \leq 1$. This illustrates the concept of the posterior mean being a compromise between the prior distribution and the maximum likelihood solution. The probability mass function of the Binomial distribution for some $m \in 0, 1, ..., N$ is

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

Hint: Identify the posterior distribution. You may then look up the mean rather than computing it.

Solution
First, calculate the posterior:

$p(\theta) = \text{Beta}(\theta|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}$
$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)} \propto \text{Beta}(m+a, l+b)$

Then, look up the mean of the Beta distribution and compute it for the posterior. For $\lambda = \frac{m+l}{m+l+a+b}$, it holds that:

$E[p(\theta)] = \frac{a}{a+b}$
$E[p(\theta|D)] = \frac{m+a}{m+l+a+b} = \frac{m}{m+l+a+b} + \frac{a}{m+l+a+b} = \frac{m+l}{m+l+a+b}\frac{m}{m+l} + \frac{a+b}{m+l+a+b}\frac{a}{a+b} = \lambda\frac{m}{m+l} + (1-\lambda)\frac{a}{a+b}$

One can see that - for lots of data - $\lambda \to 1$. In this case we almost exclusively trust the experimental data and not the subjective bias that we introduced with our prior. However, for small data samples, the prior has a strong effect on the posterior belief to prevent overfitting.

Problem 5

Let X be Poisson distributed. Again, for n i.i.d. samples from X, determine the maximum likelihood estimate for λ. In class we also talked about avoiding overfitting of parameters via prior information. Compute the posterior distribution over λ, assuming a Gamma(α, β) prior for it. Compute the MAP for λ under this prior. Show your work.

Solution
First, calculate the likelihood:

$p(X=k) = e^{-\lambda}\frac{\lambda^k}{k!}$
$\ln p(D|\lambda) = \ln\prod\limits_{i=1}^{n} e^{-\lambda}\frac{\lambda^{k_i}}{k_i!}$
$= \sum\limits_{i=1}^{n}\ln(e^{-\lambda}) + \sum\limits_{i=1}^{n}\ln(\frac{\lambda^{k_i}}{k_i!}) = -n\lambda + \sum\limits_{i=1}^{n}[k_i\ln(\lambda) - \ln(k_i!)]$

Calculate MLE just for fun:

$\frac{\partial \ln p(D|\lambda)}{\partial\lambda} = -n + \sum\limits_{i=1}^{n}\frac{k_i}{\lambda} \stackrel{!}{=} 0$

$\lambda_{\text{MLE}} = \frac{\sum\limits_{i=1}^{n} k_i}{n}$

Problem 5 continued on next page. . .

Calculate (or rather approximate) the posterior:

$p(\lambda|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda}$
$p(\lambda|D) \propto p(D|\lambda)p(\lambda) = e^{-n\lambda}e^{-\beta\lambda}\frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}\prod_{i=1}^{n}\frac{\lambda^{k_i}}{k_i!} \propto \mathrm{Gamma}(\sum_{i=1}^{n}k_i+\alpha, n+\beta)$

Calculate the maximum w.r.t. λ:

$\ln p(\lambda|D) = -(n+\lambda) + (\sum_{i=1}^{n}k_i + \alpha - 1)\ln\lambda + c$
$\frac{\partial \ln p(\lambda|D)}{\partial\lambda} = \frac{\sum_{i=1}^{n}k_i+\alpha-1}{\lambda} - (n+\beta) \stackrel{!}{=} 0$

From this, the MAP follows:

$\lambda_{\mathrm{MAP}} = \frac{\sum_{i=1}^{n}k_i+\alpha-1}{n+\beta}$