Parameter Inference

Machine Learning - Prof. Dr. Stephan Günnemann

Leonardo Freiherr von Lerchenfeld

November 10, 2017

Contents

1 Optimizing Likelihoods: Monotonic Transforms

1.1 Problem 1

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

1.2 Problem 2

The maximum of an arbitrary positive function $f(\theta)$ is also a maximum of $logf(\theta)$. A maximum occurs when the first derivative is zero and the second the derivative is smaller than zero. The following equations show that if we have a maximum for an arbitrary positive function $f(\theta)$ it is also a maximum for $logf(\theta)$.

$$\begin{aligned}
\frac{\partial}{\partial \theta} f(\theta) &= 0 \\
\frac{\partial}{\partial \theta} logf(\theta) &= 0 \\
&= \frac{1}{f(\theta)} \frac{\partial}{\partial \theta} f(\theta) \\
\frac{\partial^2}{\partial \theta^2} logf(\theta) &= \frac{1}{f(\theta)} \frac{\partial^2}{\partial \theta^2} f(\theta)
\end{aligned}$$

As you can see in Problem 1, it is often more convenient to work with the natural logarithm of the likelihood function, which is called the log-likelihood. An maximum likelihood estimation (MLE) is the same regardless of whether we maximize the likelihood or the log-likelihood function, since log is a monotonically increasing function.

2 Properties of MLE and MAP

2.1 Problem 3

$$\begin{aligned}
\theta_{MAP} &= \underset{\theta}{\operatorname{argmax}}\, p(D|\theta)p(\theta) \\
\theta_{MLE} &= \underset{\theta}{\operatorname{argmax}}\, p(D|\theta)
\end{aligned}$$

If p(θ) = 1, which means every outcome is equally possible (a uniform distribution), which means we have no prior knowledge, then $\theta_{MAP} = \theta_{MLE}$.

2.2 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.

Distribution	**PDF**	**Mode**	**Mean**
Binomial	$\binom{N}{m}\theta^m(1-\theta)^{N-m}$	$\frac{m}{N}$	$N\theta$
Beta	$\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}$	$\frac{a-1}{a+b-2}$	$\frac{a}{a+b}$

$$\begin{aligned}
p(D|\theta) &= \theta^m(1-\theta)^l \\
0 &= \frac{\partial}{\partial\theta}p(D|\theta) = \frac{\partial}{\partial\theta}logp(D|\theta) \\
\theta_{MLE} &= \frac{m}{N} \\
posterior &\propto likelihood * prior \\
p(\theta|D) &\propto \theta^m(1-\theta)^l\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1} \\
&\propto \theta^{m+a-1}(1-\theta)^{l+b-1} \\
\alpha = m+a & \quad \beta = l+b = N \\
\theta_{MAP} &= \frac{m+a-1}{N+a+b-2} \\
p(X=1|D,a,b) &= \int_0^1 p(X=1,\theta|D,a,b)d\theta \\
&= \int_0^1 p(X=1|\theta)p(\theta|D,a,b)d\theta \\
&= \int_0^1 \theta p(\theta|D,a,b)d\theta \\
&= \mathbb{E}[\theta] \; //mean\ w.r.t.\ posterior \\
&= \frac{m+a}{N+a+b}
\end{aligned}$$

The posterior mean value of θ lies between the prior mean of θ and the maximum likelihood estimate for θ.

$$\begin{aligned}
\frac{m+a}{N+a+b} &= \frac{m}{N+a+b} + \frac{a}{N+a+b} \\
\frac{m}{N+a+b} &= \frac{m+l}{N+a+b}\frac{m}{m+l} \\
&= \lambda\ \theta_{MLE} \\
\frac{a}{N+a+b} &= \frac{a+b}{N+a+b}\frac{a}{a+b} \\
&= (1-\lambda)p(\theta) \\
\mathbb{E}[\theta|D] &= \lambda\theta_{MLE} + (1-\lambda)p(\theta)
\end{aligned}$$

3 Poisson Distribution

3.1 Problem 5

X is Poisson distributed. For n i.i.d. samples from X, the maximum likelihood estimation is:

$$\begin{aligned}
p(D|\lambda) &= \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{k_i}}{k_i!} \\
ln\ p(D|\lambda) &= ln \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{k_i}}{k_i!} \\
&= \sum_{i=1}^{n} ln\ e^{-\lambda} + \sum_{i=1}^{n} ln\ \lambda^{k_i} - \sum_{i=1}^{n} ln\ k_i! \\
&= -n\lambda + \sum_{i=1}^{n} [k_i ln\ \lambda - ln\ k_i!] \\
\frac{\partial}{\partial\lambda} = 0 &= -n + \sum_{i=1}^{n} \frac{k_i}{\lambda} \\
\lambda_{MLE} &= \frac{\sum_{i=1}^{n} k_i}{n}
\end{aligned}$$

Distribution	PDF	Mode	Mean
Poisson	$\frac{e^{-\lambda}\lambda^{k}}{k!}$		λ
Gamma	$\frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}$	$\frac{a-1}{b}$	$\frac{a}{b}$

$$\begin{aligned}
posterior &\propto likelihood * prior \\
p(D|\lambda) &\propto \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{k_i}}{k_i!} \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda} \\
&\propto e^{-n\lambda - b\lambda} \lambda^{a-1+\sum k_i} \\
\beta = b + n & \qquad \alpha = a + \sum_{i=1}^{n} k_i \\
\lambda_{MAP} &= \frac{a - 1 + \sum_{i=1}^{n} k_i}{b + n}
\end{aligned}$$