Machine Learning Worksheet 3

Parameter Inference

1 Optimising Likelihoods: Monotonic Transforms

Usually one considers the *log likelihood*, $\log p(x_1, \ldots, x_n \mid \theta)$. The next problems justify 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.

Problem 1: Compute the first and second derivative of this likelihood w.r.t. θ. Then compute first and second derivative of the log likelihood $\log \theta^t (1-\theta)^h$.

To solve this, we need to apply chain and product rule.

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \theta^t (1-\theta)^h = \theta^{t-1} (1-\theta)^{h-1} \left((1-\theta)\, t - \theta h \right)$$

$$\frac{\mathrm{d}^2}{\mathrm{d}\theta^2} \theta^t (1-\theta)^h = \theta^{t-2} (1-\theta)^{h-2} \cdot \left((1-\theta)\,(t-1) - \theta(h-1) \right) \cdot \left((1-\theta)\, t - \theta h \right) - \theta^{t-1} (1-\theta)^{h-1} \left(t + h \right)$$

Observe that the first factor of the first derivative is structurally the same as the original function. This can be used to ease some of the pain of these computations.

The product rule breaks our necks, which quickly renders the expressions long and confusing.

The logarithm decomposes the product into a sum. We only need to apply the chain rule on each of the summands. Do not forget the change of sign from taking the derivative of $1 - \theta$.

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

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

$$\frac{\mathrm{d}^2}{\mathrm{d}\theta^2} g(\theta) = -\left(\frac{t}{\theta^2} + \frac{h}{(1-\theta)^2} \right)$$

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?

Let θ^* be an arbitrary local maximum of $g(\theta) = \log f(\theta)$, i.e., for any θ in a small neighbourhood of θ^*, we have that $g(\theta^*) \geq g(\theta)$. Since exp is a monotonic transform, we also have

$$f(\theta^*) = \exp(g(\theta^*)) \geq \exp(g(\theta)) = f(\theta).$$

Hence, θ^* is also a maximum of f.

With the help of the previous exercise, we can now safely apply the logarithm and any maximum or minimum remains preserved (its position only, of course). Moreover, we have seen that the logarithmic domain can greatly simplify the computational effort to arrive at critical points. This also leads to improved numerical stability. Thus, it is often worth switching to the log domain when analysing likelihoods.

Notice that the exercise left out a part of the argument: We only showed that a maximum of the log likelihood is also a maximum of the likelihood. We would still need to prove that taking the logarithm does not eliminate maxima of the likelihood. This is done by showing that monotonic transforms preserve critical points and observing that the logarithm is monotonic—we will not do this here.

2 Properties of MLE and MAP

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 $\theta_{\text{MLE}} = \theta_{\text{MAP}}$.

We know that

$$\theta_{\text{MAP}} = \arg\max_{\theta} p(\mathcal{D} \mid \theta)p(\theta),$$
$$\theta_{\text{MLE}} = \arg\max_{\theta} p(\mathcal{D} \mid \theta).$$

If $p(\theta)$ is constant for all θ, then $\arg\max_\theta p(\mathcal{D} \mid \theta)p(\theta) = \arg\max_\theta p(\mathcal{D} \mid \theta)$, and consequently $\theta_{\text{MAP}} = \theta_{\text{MLE}}$. That is the case if we choose the uniform distribution as a prior.

Note that the argument is a little more technical when the prior has unbounded support, e.g., the entire real line. The uniform distribution is not well-defined in Kolmogorov's system of axioms and we would have to talk about *improper priors*. This is beyond the scope of our course—and doesn't give much insight w.r.t. this exercise.

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 $E[\theta \mid \mathcal{D}]$ (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 $0 \leq \lambda \leq 1$. 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 $m \in \{0, 1, \ldots, N\}$ is

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

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

For the given observations (and a, b hyper parameters for the prior beta distribution), we see that, just like in the lecture, the posterior is Beta$(m+a, l+b)$-distributed. One gets the *expected* posterior mean for θ:

$$E[\theta \mid \mathcal{D}] = \frac{m+a}{m+l+a+b} = \frac{m}{m+l+a+b} + \frac{a}{m+l+a+b}$$

But:

$$\frac{m}{m+l+a+b} = \underbrace{\frac{m+l}{m+l+a+b}}_{\lambda} \cdot \frac{m}{m+l}$$

and

$$\frac{a}{m+l+a+b} = \underbrace{\frac{a+b}{m+l+a+b}}_{1-\lambda} \cdot \frac{a}{a+b}$$

producing what was asked, because $\frac{m}{m+l}$ is the maximum likelihood estimate and $\frac{a}{a+b}$ is the prior mean value of θ.

3 Poisson Distribution

Problem 5: Let X be Poisson distributed. Again, for n i.i.d. samples from X, determine the maximum likelihood estimate for λ. Show that this estimate is unbiased!

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.

X is Poisson distributed, i.e. $P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}$. Thus

$$l(\lambda) = \ln \prod_{i=1}^{n} e^{-\lambda} \frac{\lambda^{k_i}}{k_i!} = \sum_{i=1}^{n} \ln e^{-\lambda} + \ln \frac{\lambda^{k_i}}{k_i!} = -n\lambda + \sum_{i=1}^{n} (k_i \ln \lambda - \ln k_i!)$$

$$\frac{\partial l(\lambda)}{\partial \lambda} = -n + \frac{\sum_{i=1}^{n} k_i}{\lambda} = 0$$

Here, k_i is the ith realisation of X. Thus, $\tilde{\lambda} = \frac{\sum_{i=1}^{n} k_i}{n}$.

$$E\left[\tilde{\lambda}\right] = \frac{\sum_{i=1}^{n} E[k_i]}{n} = \frac{\sum_{i=1}^{n} \lambda}{n} = \lambda$$

The prior for λ is

$$p(\lambda \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} \exp(-\beta\lambda).$$

So

$$p(\lambda \mid \mathcal{D}) \propto p(\mathcal{D} \mid \lambda) p(\lambda) = \exp(-n\lambda) \prod_{i=1}^{n} \frac{\lambda^{k_i}}{k_i!} \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} \exp(-\beta\lambda) \propto \lambda^{\sum_{i=1}^n k_i + \alpha - 1} \exp(-n\lambda - \beta\lambda).$$

Therefore the posterior is Gamma$(\sum_{i=1}^n k_i + \alpha, n + \beta)$.

For the MAP solution we consider $\arg\max_\lambda p(\lambda \mid \mathcal{D}) = \arg\max_\lambda \ln p(\lambda \mid \mathcal{D})$.

$$\ln p(\lambda \mid \mathcal{D}) = \left(\sum_{i=1}^{n} k_i + \alpha - 1 \right) \ln \lambda - (n + \beta)\lambda + c,$$

with c being constant with respect to λ. We need the first derivative and its zero point:

$$\frac{d}{d\lambda} \ln p(\lambda \mid \mathcal{D}) = \frac{\sum_{i=1}^n k_i + \alpha - 1}{\lambda} - (n + \beta) = 0 \rightarrow \lambda = \frac{\sum_{i=1}^n k_i + \alpha - 1}{n + \beta}$$