Machine Learning 1, WS 2022/23
Ilina Tuneska
Edgar Eugenio Tovar Placeres
Matthias Jean Théo Personnaz
Muhammad Taqiyuddin Ar Rofi
Hafidz Handyawan Arifin
Maximilian Peters

Machine Learning Homework 2

1. Maximum Likelihood Estimator

We consider the problem of estimating using the maximum-likelihood approach the parameters $\lambda, \eta > 0$ of the probability distribution:

$$p(x,y) = \lambda\eta e^{-\lambda x - \eta y}$$

supported on $\mathbb{R}_+^2$. We consider a dataset $D = ((x_1, y_1), \ldots, (x_N, y_N))$ composed of N independent draws from this distribution.

(a) *Show that x and y are independent.*

Proof. Two random variables are independent if the joint probability density function can be factored as a product of two functions which depend solely on the respective random variables, i.e. $p(x,y) = p(x) \cdot p(y)$.

$$p(x,y) = \lambda\eta e^{-\lambda x - \eta y} = \underbrace{\lambda e^{-\lambda x}}_{p(x)} \cdot \underbrace{\eta e^{-\eta y}}_{p(y)} = p(x) \cdot p(y)$$

(b) *Derive a maximum likelihood estimator of the parameter λ based on D.*

Solution: The maximum likelihood (ML) estimator of λ based on D is the parameter which is obtained as $\hat{\lambda} = argmax_\lambda p(D|\lambda)$.

In order to compute this we use the formula $p(D|\lambda,\eta) = \prod_{i=1}^N p(x_i, y_i|\lambda,\eta) = \prod_{i=1}^N \lambda\eta e^{-\lambda x_i - \eta y_i}$ where η has a fixed value.

If the function is concave, we can find its maximum by setting the derivative equal to 0. Since $p(D,\lambda,\eta)$ is not concave, we can apply a function to make it concave and doesn't change the solution to the maximization problem $argmax_\lambda$.

We apply the logarithm to $p(D,\lambda,\eta)$:

$$\log(p(D|\lambda)) = \log\left(\prod_{i=1}^N p(x_i, y_i|\lambda,\eta)\right) = \log\left(\prod_{i=1}^N \lambda\eta e^{-\lambda x_i - \eta y_i}\right) = \sum_{i=1}^N \log(\lambda\eta e^{-\lambda x_i - \eta y_i}) = \sum_{i=1}^N \log(\lambda) + \log(\eta) - \lambda x_i - \eta y_i$$

Since the logarithm function is concave and the term $-\lambda x_i$ doesn't change the concavity (since it is linear), $(p(D|\lambda))$ is a concave function.

$$0 \overset{!}{=} \nabla_\lambda(p(D|\lambda,\eta)) = \nabla_\lambda\left(\sum_{i=1}^N \log(\lambda) + \cancel{\log(\eta)} - \lambda x_i \cancel{- \eta y_i}\right) = \sum_{i=1}^N \frac{1}{\lambda} - x_i$$

$$\iff \quad 0 = \sum_{i=1}^N \frac{1}{\lambda} - x_i$$

$$\iff \quad \sum_{i=1}^N \frac{1}{\lambda} = \sum_{i=1}^N x_i$$

$$\iff N \cdot \frac{1}{\lambda} = \sum_{i=1}^{N} x_i$$

$$\iff \frac{\lambda}{N} = \frac{1}{\sum_{i=1}^{N} x_i}$$

$$\iff \lambda = \frac{N}{\sum_{i=1}^{N} x_i}$$

The optimal parameter is $\hat{\lambda} = \frac{N}{\sum_{i=1}^{N} x_i}$.

(c) *Derive a maximum likelihood estimator of the parameter λ based on D under the constraint $\eta = \frac{1}{\lambda}$.*
Solution:
We repeat the same steps as in (b) and set $\eta = \frac{1}{\lambda}$.

$$\begin{aligned}\log(p(D|\lambda, \eta)) &= \log(p(D|\lambda, \frac{1}{\lambda})) = \log\left(\prod_{i=1}^{N} p\left(x_i, y_i|\lambda, \frac{1}{\lambda}\right)\right) \\ &= \log\left(\prod_{i=1}^{N} \lambda \frac{1}{\lambda} e^{-\lambda x_i - \frac{1}{\lambda} y_i}\right) = \log\left(\prod_{i=1}^{N} e^{-\lambda x_i - \frac{1}{\lambda} y_i}\right) \\ &= \sum_{i=1}^{N} \log(e^{-\lambda x_i - \frac{1}{\lambda} y_i}) = \sum_{i=1}^{N} -\lambda x_i - \frac{1}{\lambda} y_i\end{aligned}$$

Since $\frac{1}{\lambda} \sum_{i=1}^{n} y_i$ is a concave function and $-\lambda x_i$ is a linear term, $\log(p(D|\lambda, \eta))$ is a concave function.

$$0 \stackrel{!}{=} \nabla_\lambda \log(p(D|\lambda, \eta)) = \nabla_\lambda \left(\sum_{i=1}^{N} -\lambda x_i - \frac{1}{\lambda} y_i\right) = \sum_{i=1}^{N} -x_i + \frac{y_i}{\lambda^2}$$

$$\iff 0 = \sum_{i=1}^{N} -x_i + y_i \frac{1}{\lambda^2}$$

$$\iff \sum_{i=1}^{N} x_i = \sum_{i=1}^{N} y_i \frac{1}{\lambda^2}$$

$$\iff \sum_{i=1}^{N} x_i = \frac{1}{\lambda^2} \sum_{i=1}^{N} y_i$$

$$\iff \lambda^2 = \frac{\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} x_i}$$

$$\iff \lambda = \sqrt{\frac{\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} x_i}}$$

The optimal parameter is $\hat{\lambda} = \sqrt{\frac{\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} x_i}}$.

(d) *Derive a maximum likelihood estimator of the parameter λ based on D under the constraint $\eta = 1 - \lambda$.*
Solution:
We repeat the same steps as in (a) and (b) and set $\eta = 1 - \lambda$.

$$\begin{aligned}\log(p(D|\lambda, \eta)) &= \log(p(D|\lambda, (1-\lambda))) = \log\left(\prod_{i=1}^{N} p\left(x_i, y_i|\lambda, 1-\lambda\right)\right) \\ &= \log\left(\prod_{i=1}^{N} \lambda(1-\lambda) e^{-\lambda x_i - (1-\lambda) y_i}\right) = \sum_{i=1}^{N} \log(\lambda(1-\lambda) e^{-\lambda x_i - (1-\lambda) y_i}) \\ &= \sum_{i=1}^{N} \log(\lambda(1-\lambda)) - \lambda x_i - (1-\lambda) y_i\end{aligned}$$

The logarithm function is concave and the linear terms λx_i and $(1-\lambda)y_i$ don't not change concavity.
We compute the gradient using Wolfram Alpha:

$$
\begin{aligned}
0 \overset{!}{=} \nabla_\lambda(p(D|\lambda,\eta)) = \nabla_\lambda\left(\sum_{i=1}^N \log(\lambda(1-\lambda)) - \lambda x_i - (1-\lambda)y_i\right) &= \sum_{i=1}^N \frac{1-2\lambda}{\lambda(1-\lambda)} - x_i + y_i \\
\iff \quad 0 &= \sum_{i=1}^N \frac{1-2\lambda}{\lambda(1-\lambda)} - x_i + y_i \\
\iff \quad \sum_{i=1}^N x_i - y_i &= \sum_{i=1}^N \frac{1-2\lambda}{\lambda(1-\lambda)} \\
\iff \quad \sum_{i=1}^N x_i - y_i &= N \cdot \frac{1-2\lambda}{\lambda(1-\lambda)} \\
\iff \quad \frac{1}{N}\sum_{i=1}^N x_i - y_i &= \frac{1-2\lambda}{\lambda(1-\lambda)} \\
\iff \quad \lambda &= \frac{\left(\sum_{i=1}^N y_i - x_i\right) - 2 \pm \sqrt{(\sum_{i=1}^N y_i - x_i)^2 + 4}}{2\sum_{i=1}^N y_i - x_i}
\end{aligned}
$$

The optimal parameter $\hat{\lambda}$ has to satisfy

- $\lambda > 0$ and
- $\eta = 1 - \lambda > 0$ (which implies $\lambda < 1$).

For this to hold, in the last equality for λ we need to take the plus sign, i.e. the optimal parameter is: $\hat{\lambda} = \frac{(\sum_{i=1}^N y_i - x_i) - 2 + \sqrt{(\sum_{i=1}^N y_i - x_i)^2 + 4}}{2\sum_{i=1}^N y_i - x_i}$.

2. Maximum Likelihood vs. Bayes

An unfair coin is tossed seven times and the event (head or tail) is recorded at each iteration. The observed sequence of events is

$$\mathcal{D} = (x_1, \dots, x_7) = (head, head, tail, tail, head, head, head)$$

We assume that all tosses $x_1, x)2, \dots$ have been generated independently following the Bernoulli probability distribution

$$\begin{cases} P(x,\theta) = \begin{cases} \theta, & \text{if } x = head, \\ 1-\theta, & \text{if } x = tail \end{cases} \end{cases}$$

where $\theta \in [0,1]$ is an unknown parameter.

(a) *State the likelihood function $P(D|\theta)$, that depends on the parameter θ.*
Solution:

$$
\begin{aligned}
P(D|\theta) &\overset{\text{def.}}{=} \prod_{i=1}^7 p(x_i|\theta) \\
&= p(x=\text{head}|\theta)\cdot p(x=\text{head}|\theta)\cdot p(x=\text{tail}|\theta)\cdot p(x=\text{tail}|\theta)\cdot p(x=\text{head}|\theta)\cdot p(x=\text{head}|\theta)\cdot p(x=\text{head}|\theta) \\
&= \theta^5 \cdot (1-\theta)^2
\end{aligned}
$$

(b) *Compute the maximum likelihood solution $\hat{\theta}$, and evaluate for this parameter the probability that the next two tosses are "head", that is, evaluate $P(x_8 = head, x_9 = head|\hat{\theta})$.*
Solution:

(a) First, we compute the maximum likelihood $\hat{\theta} = argmax_\theta p(D|\theta) = argmax_\theta \prod_{i=1}^{7} p(x_i|\theta)$.
In order for us to work with a concave function, we apply the logarithm function:

$$\log(p(D|\theta)) = \log(\theta^5 \cdot (1-\theta)^2) = 5\log(\theta) + 2\log(1-\theta)$$

The logarithm function is concave and sum of concave functions is concave.

We set the derivative equal to 0 to obtain the optimal value:

$$\begin{aligned}
& 0 \overset{!}{=} \nabla_\theta log(p(D|\theta)) = \nabla_\theta(5\log(\theta) + 2\log(1-\theta)) = 5\frac{1}{\theta} - 2\frac{1}{1-\theta} \\
\iff \quad & 0 = \frac{5(1-\theta) - 2\theta}{\theta(1-\theta)} = \frac{5-7\theta}{\theta(1-\theta)} \\
\iff \quad & \hat{\theta} = \frac{5}{7}
\end{aligned}$$

The optimal value is $\hat{\theta} = \frac{5}{7}$.

(b) We evaluate $P(x_8 = \text{head}, x_9 = \text{head}|\hat{\theta})$.
Since the tosses are generated independently, it holds

$$P(x_8 = \text{head}, x_9 = \text{head}|\hat{\theta}) = P(x_8 = \text{head}|\hat{\theta}) \cdot P(x_9 = \text{head}|\hat{\theta}) = \hat{\theta} \cdot \hat{\theta} = \hat{\theta}^2$$

(c) *We now adopt a Bayesian view on this problem, where we assume a prior distribution for the parameter θ defined as*

$$p(\theta) = \begin{cases} 1, & \text{if } 0 \leqslant \theta \leqslant 1 \\ 0, & \text{else} \end{cases}$$

Compute the posterior distribution $p(\theta|D)$, and evaluate the probability that the next two tosses are head, that is

$$\int P(x_8 = head, x_9 = head|\theta) p(\theta|D) d\theta$$

Solution:

(a) We compute the posterior distribution $p(\theta|D)$.
We calculate $p(D|\theta) = \prod_{i=1}^{7} p(x = k|\theta) = \theta^5(1-\theta)^2$

$$\begin{aligned}
p(\theta|D) &= \frac{p(D|\theta) \cdot p(\theta)}{\int p(D|\theta) \cdot p(\theta) d\theta} \\
&= \frac{\theta^5(1-\theta)^2 \cdot 1}{\int_0^1 \theta^5(1-\theta)^2 \cdot 1 d\theta} \\
&= \frac{\theta^5(1-\theta)^2}{\frac{1}{168}} \\
&= 168 \cdot \theta^5(1-\theta)^2
\end{aligned}$$

(b) We compute $\int P(x_8 = \text{head}, x_9 = \text{head}|\theta) p(\theta|D) d\theta$. Since the draws are done independently, we calculate

$$\begin{aligned}
\int P(x_8 = \text{head}, x_9 = \text{head}|\theta) p(\theta|D) d\theta &= \int P(x_8 = \text{head}|\theta) \cdot P(x_9 = \text{head}|\theta) p(\theta|D) d\theta \\
&= \int_0^1 \theta^2 \cdot 168 \cdot \theta^5(1-\theta)^2 d\theta \\
&= \int_0^1 168 \cdot \theta^7(1-\theta)^2 d\theta = \frac{7}{15}
\end{aligned}$$

3. Convergence of Bayes Parameter Estimation

We consider Section 3.4.1 of Duda et al., where the data is generated according to the univariate probability density $p(x|\mu) \sim \mathcal{N}(\mu, \sigma^2)$, where σ^2 is known and where μ is unknown with prior distribution $p(\mu) \sim \mathcal{N}(\mu_0, \sigma_0^2)$. Having sampled a dataset D from the data-generating distribution, the posterior probability distribution over the unknown parameter μ becomes $p(\mu|D) \sim \mathcal{N}(\mu_n, \sigma_n^2)$, where

$$\frac{1}{\sigma_n^2} = \frac{n}{\sigma^2} + \frac{1}{\sigma_0^2} \tag{1}$$

$$\frac{\mu_n}{\sigma_n^2} = \frac{n}{\sigma^2}\hat{\mu}_n + \frac{\mu_0}{\sigma_0^2} \tag{2}$$

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{3}$$

(a) *Show that the variance of the posterior can be upper-bounded as $\sigma_n^2 \leqslant \min(\frac{\sigma^2}{n}, \sigma_0^2)$, that is, the variance of the posterior is contained both by the uncertainty of the data mean and of the prior.*
Solution:

$$\frac{1}{\sigma_n^2} \overset{(1)}{=} \underbrace{\frac{n}{\sigma^2}}_{\geqslant 0} + \underbrace{\frac{1}{\sigma_0^2}}_{\geqslant 0} \geqslant \max\left(\frac{n}{\sigma^2}, \frac{1}{\sigma_0^2}\right)$$

$$\iff \frac{1}{\frac{1}{\sigma_n^2}} \leqslant \frac{1}{\max\left(\frac{n}{\sigma^2}, \frac{1}{\sigma_0^2}\right)}$$

$$\iff \sigma_n^2 \leqslant \min\left(\frac{\sigma^2}{n}, \sigma_0^2\right)$$

(b) *Show that the mean of the posterior can be lower- and upper-bounded as $\min(\hat{\mu}_n, \mu_0) \leqslant \mu_n \leqslant \max(\hat{\mu}_n, \mu_0)$ that is, the mean of the posterior distribution lies somewhere on the segment between the mean of the prior distribution and the sample mean.*
Solution:

(a) To show: $\mu_n \leqslant \max(\hat{\mu}_n, \mu_0)$:
Using equality (2), we have

$$\frac{\mu_n}{\sigma_n^2} \overset{(2)}{=} \frac{n}{\sigma^2}\hat{\mu}_n + \frac{\mu_0}{\sigma_0^2} \leq \frac{n}{\sigma_n^2}\max(\hat{\mu}_n, \mu_0) + \frac{1}{\sigma_0^2}\max(\hat{\mu}_n, \mu_0) = \left[\frac{n}{\sigma_n^2} + \frac{1}{\sigma_0^2}\right]\max(\hat{\mu}_n, \mu_0) \overset{(1)}{=} \frac{1}{\sigma_n^2}\max(\hat{\mu}_n, \mu_0)$$

The result is obtained by multiplying the inequality by σ_n^2.

(b) To show: $\mu_n \geqslant \min(\hat{\mu}_n, \mu_0)$:

$$\frac{\mu_n}{\sigma_n^2} \overset{(2)}{=} \frac{n}{\sigma^2}\hat{\mu}_n + \frac{\mu_0}{\sigma_0^2} \geq \frac{n}{\sigma_n^2}\min(\hat{\mu}_n, \mu_0) + \frac{1}{\sigma_0^2}\min(\hat{\mu}_n, \mu_0) = \left[\frac{n}{\sigma_n^2} + \frac{1}{\sigma_0^2}\right]\min(\hat{\mu}_n, \mu_0) \overset{(1)}{=} \frac{1}{\sigma_n^2}\min(\hat{\mu}_n, \mu_0)$$

The result is obtained by multiplying the inequality by σ_n^2.