Abteilung Maschinelles Lernen
Institut für Softwaretechnik und theoretische Informatik
Fakultät IV, Technische Universität Berlin
Prof. Dr. Klaus-Robert Müller
Email: klaus-robert.mueller@tu-berlin.de

Exercise Sheet 2

Exercise 1: Maximum-Likelihood Estimation (5 + 5 + 5 + 5 P)

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 $\mathcal{D} = ((x_1, y_1), \ldots, (x_N, y_N))$ composed of N independent draws from this distribution.

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

(b) *Derive* a maximum likelihood estimator of the parameter λ based on $\mathcal{D}$.

(c) *Derive* a maximum likelihood estimator of the parameter λ based on $\mathcal{D}$ under the constraint $\eta = 1/\lambda$.

(d) *Derive* a maximum likelihood estimator of the parameter λ based on $\mathcal{D}$ under the constraint $\eta = 1 - \lambda$.

Exercise 2: Maximum Likelihood vs. Bayes (5 + 10 + 15 P)

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

$$\mathcal{D} = (x_1, x_2, \ldots, x_7) = (\text{head}, \text{head}, \text{tail}, \text{tail}, \text{head}, \text{head}, \text{head}).$$

We assume that all tosses $x_1, x_2, \ldots$ have been generated independently following the Bernoulli probability distribution

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

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

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

(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 $P(x_8 = \text{head}\,,\, x_9 = \text{head} \mid \hat{\theta})$.

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

$$p(\theta) = \begin{cases} 1 & \text{if} \quad 0 \leq \theta \leq 1 \\ 0 & \text{else}. \end{cases}$$

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

$$\int P(x_8 = \text{head}\,,\, x_9 = \text{head} \mid \theta)\, p(\theta|\mathcal{D})\, d\theta.$$

Exercise 3: Convergence of Bayes Parameter Estimation (5 + 5 P)

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 $\mathcal{D}$ from the data-generating distribution, the posterior probability distribution over the unknown parameter μ becomes $p(\mu|\mathcal{D}) \sim \mathcal{N}(\mu_n, \sigma_n^2)$, where

$$\frac{1}{\sigma_n^2} = \frac{n}{\sigma^2} + \frac{1}{\sigma_0^2} \qquad \frac{\mu_n}{\sigma_n^2} = \frac{n}{\sigma^2}\hat{\mu}_n + \frac{\mu_0}{\sigma_0^2} \qquad \hat{\mu}_n = \frac{1}{n}\sum_{k=1}^{n} x_k.$$

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

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

Exercise 4: Programming (40 P)

Download the programming files on ISIS and follow the instructions.

$\boxed{1}$

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

b) The logarythm of the propabilit $P(D|\Lambda)$ is

$$\log P(D|\Lambda) = \sum_{k=1}^{N} \log p(x_k, y_k | \lambda, \eta)$$

$$= \sum_{k=1}^{N} \log p(x_k) \cdot \log p(y_k) = \sum_{k=1}^{N} \log(\lambda e^{-\lambda x}) \cdot \log(\eta e^{-\eta y})$$

$$= \sum_{k=1}^{N} \log \lambda - \lambda x_k + \log \eta - \eta y_k$$

$\bar{x} = \frac{1}{N} \sum_{k=1}^{N} x_k$ $\quad \overset{1)}{=} (\log \lambda - \lambda \bar{x} + \log \eta - \eta \bar{y}) \cdot N$

To obtain the maximum likelyhood we derive for λ:

$$\frac{\partial \log P(D|\Lambda)}{\partial \lambda} = N\left(\frac{1}{\lambda} - \bar{x}\right) \overset{!}{=} 0 \quad \Rightarrow \lambda = \frac{1}{\bar{x}}$$

c) $\log P(D|\Lambda) = (\log \lambda - \lambda \bar{x} + \log \eta - \eta \bar{y}) N$

$$\overset{\eta = \frac{1}{\lambda}}{=} (\cancel{\log \lambda} - \lambda \bar{x} + \cancel{\log(\tfrac{1}{\lambda})} - \cancel{\eta} \tfrac{1}{\lambda} \bar{y}) N$$

$$\frac{\partial \log P(D|\Lambda)}{\partial \lambda} = \frac{\partial}{\partial \lambda} N\left(-\lambda \bar{x} - \frac{1}{\lambda} \bar{y}\right) = N\left(-\bar{x} + \frac{1}{\lambda^2} \bar{y}\right) \overset{!}{=} 0$$

$$\Leftrightarrow \cancel{N} \frac{1}{\lambda^2} \bar{y} = \cancel{N} \bar{x} \Leftrightarrow \lambda = \pm \sqrt{\frac{\bar{y}}{\bar{x}}}$$

d) $\log P(D|\lambda) = N(\log\lambda - \lambda\bar{x} + \log(1-\lambda) - (1-\lambda)\bar{y})$

$= N(\log\lambda + \log(1-\lambda) - \lambda\bar{x} + \lambda\bar{y} - \bar{y})$

$\overset{1)}{=} N(\log(\lambda-\lambda^2) - \lambda\bar{d} - \bar{y}$ $\qquad$ 1) $\bar{d} = \bar{x} + \bar{y}$

$$\frac{d\log P(D|\lambda)}{d\lambda} = N\left(\frac{1-2\lambda}{\lambda-\lambda^2} + \bar{d}\right) \overset{!}{=} 0$$

$$\Leftrightarrow \cancel{N}\frac{1-2\lambda}{\lambda-\lambda^2} = -\cancel{N}\bar{d} \quad \Leftrightarrow \lambda = \frac{(\bar{d}-2) \pm \sqrt{\bar{d}^2+4}}{2\bar{d}}$$

$\boxed{2}$

a) $P(D|\theta) = \prod_{k=1}^{7} p(x_k|\theta) = \theta\theta(1-\theta)(1-\theta)\theta\theta\theta$

$= \theta^5(\theta-1)^2$

b) $J = \log p(D|\theta) = 5\log\theta + 2\log(1-\theta)$

$\frac{dJ}{d\theta} = \frac{5}{\theta} - \frac{2}{1-\theta} \stackrel{!}{=} 0 \iff \frac{5}{\theta} = \frac{2}{1-\theta}$

$\iff \theta = \frac{5}{7}$

$p(x_8 = H, x_9 = H | \hat{\theta}) = p(x_8 = H|\hat{\theta})\, p(x_9 = H|\hat{\theta})$

$= \hat{\theta}\hat{\theta} = \frac{25}{49}$

c) $p(\theta|D) = \frac{p(D|\theta)p(\theta)}{\int p(D|\theta)p(\theta)\,d\theta} = \frac{\theta^5(1-\theta)^2 \cdot 1}{\int \theta^5(1-\theta^2)\cdot 1\, d\theta}$

$= 168 \cdot \theta^5(1-\theta)^2$

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

$= \int \theta^7(1-\theta)^2 d\theta \cdot 168\, d\theta = \frac{7}{15}$

$\boxed{3}$

a) ~~[illegible]~~ $n \geq 0$, $\sigma \geq 0$, and $\sigma_0 \geq 0$

Therefore $\frac{1}{\sigma_n^2} = \frac{n}{\sigma^2} + \frac{1}{\sigma_0^2} \geq \max\left(\frac{n}{\sigma^2}, \frac{1}{\sigma_0^2}\right)$

$\Leftrightarrow \left(\frac{1}{\sigma_n^2}\right)^{-1} = \left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\right)^{-1} \leq \left[\max\left(\frac{n}{\sigma^2}, \frac{1}{\sigma_0^2}\right)\right]^{-1} = \min\left(\frac{\sigma^2}{n}, \sigma_0^2\right)$

$\Leftrightarrow \quad \sigma_n^2 \leq \min\left(\frac{\sigma^2}{n}, \sigma_0^2\right)$

b) $\frac{1}{\sigma_n^2}\mu_n = \frac{n}{\sigma_n^2}\hat{\mu}_n + \frac{1}{\sigma_0^2}\mu_0 \leq \frac{n}{\sigma_n^2}\max(\hat{\mu}_n, \mu_0) + \frac{1}{\sigma_0^2}\max(\hat{\mu}_n, \mu$

$= \left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\right)\max(\mu_0, \hat{\mu}_n) = \frac{1}{\sigma^2}\max(\mu_0, \hat{\mu}_n)$

$\Leftrightarrow \quad \mu_n \leq \max(\mu_0, \hat{\mu}_n)$

$\Leftrightarrow \quad \mu_n \geq \min(\mu_0, \hat{\mu}_n)$

~~c)~~