Exercises for the course

Machine Learning 1

Winter semester 2024/25

Fachgebiet 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 $P(\mathcal{D} \mid \theta)$, 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 $p(\theta \mid \mathcal{D})$, and *evaluate* the probability that the next two tosses are head, that is,

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

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 $p(x \mid \mu) \sim \mathcal{N}(\mu, \sigma^2)$, 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 $p(\mu \mid \mathcal{D}) \sim \mathcal{N}(\mu_n, \sigma_n^2)$, 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.

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.

Solution:

Since if $p(x,y) = p(x)p(y)$

$\therefore$ we can try to divide the distribution $p(x,y)$ into 2 seperate parts, each of them depends on one variable (x/y).

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

$\therefore$ Here we can derive that $p(x,y) = p(x)p(y)$

$\therefore$ x, y are independent.

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

Solution:

we would like to estimate $p(\lambda|D)$ then we can derive.

$$p(\lambda|D) = \frac{p(D|\lambda)p(\lambda)}{p(D)}$$

$$\downarrow$$

$$\arg\max_{\lambda} p(\lambda|D) \iff \arg\max_{\lambda} p(D|\lambda)p(\lambda)$$

$$\Big\downarrow \text{take log}$$

$$\arg\max \log p(\lambda|D) \iff \arg\max_{\lambda} \log p(D|\lambda) + \log p(\lambda)$$

日期: /

Since $P(\lambda)$ is not related to D

$\downarrow$

The Problem can be rewritten as:

$$\underset{\lambda}{\arg\max} \log P(D|\lambda)$$

$\therefore$ If we define:

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

$$\frac{\partial J}{\partial \lambda} = \frac{N}{\lambda} - \sum_{i=1}^{N} x_i \overset{!}{=} 0$$

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

$$\Downarrow$$

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

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

Solution:

Following the same procedure

$$J(\lambda) = \log p(D|\lambda)$$

$\therefore$ $$J(\lambda) = \log \prod_{i=1}^{N} p(x_i, y_i)$$

日期: /

$$J(\lambda) = N(\log\lambda + \log\eta) - \sum_{i=1}^{N}(\lambda x_i + \eta y_i)$$

$$\Downarrow \eta = \frac{1}{\lambda}$$

$$J(\lambda) = N(\log\lambda - \log\lambda) - \sum_{i=1}^{N}(\lambda x_i + \frac{1}{\lambda} y_i)$$

$$\therefore \frac{\partial J}{\partial \lambda} = -\sum_{i=1}^{N} x_i + \sum_{i=1}^{N} \frac{1}{\lambda^2} y_i \overset{!}{=} 0$$

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

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

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

Solution:

$$J(\lambda) = N(\log\lambda + \log\eta) - \sum_{i=1}^{N}(\lambda x_i + \eta y_i)$$

$$\Downarrow \eta = 1-\lambda$$

$$J(\lambda) = N(\log\lambda + \log(1-\lambda)) - \sum_{i=1}^{N}(\lambda x_i + (1-\lambda) y_i)$$

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

$$\Downarrow$$

$$\frac{1}{\lambda} - \frac{1}{1-\lambda} = \bar{x} - \bar{y}$$

$$\frac{1-2\lambda}{\lambda(1-\lambda)} = (\bar{x} - \bar{y})$$

$$-(\bar{x}-\bar{y})\lambda^2 + (\bar{x}-\bar{y}+2)\lambda - 1 = 0$$

$$(\bar{x}-\bar{y})\lambda^2 - (\bar{x}-\bar{y}+2)\lambda + 1 = 0$$

$$\therefore \lambda = \frac{(\bar{x}-\bar{y}+2) \pm \sqrt{(\bar{x}-\bar{y}+2)^2 - 4(\bar{x}-\bar{y})}}{2(\bar{x}-\bar{y})} \Longrightarrow \lambda = \frac{(\bar{x}-\bar{y}+2) + \sqrt{(\bar{x}-\bar{y}+2)^2 - 4(\bar{x}-\bar{y})}}{2(\bar{x}-\bar{y})}$$

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 $P(\mathcal{D} \mid \theta)$, that depends on the parameter θ.

Solution:

$$P(D|\theta) = \prod_{i=1}^{N} p(x_i|\theta)$$
$$= \theta \cdot \theta \cdot (1-\theta) \cdot (1-\theta) \cdot \theta \cdot \theta \cdot \theta$$
$$= \theta^5 (1-\theta)^2$$

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

Solution:

$$J(\theta) = \log P(D|\theta) = \log \prod_{i=1}^{N} p(x_i|\theta)$$
$$= \log \theta^5 (1-\theta)^2$$
$$= 5 \log \theta + 2 \log (1-\theta)$$

$$\therefore \frac{dJ}{d\theta} = \frac{5}{\theta} - \frac{2}{1-\theta} = 0$$
$$\therefore 5 - 5\theta = 2\theta$$
$$\therefore \hat{\theta} = \frac{5}{7}$$

Since 2 tosses are totally independent:

$$\therefore p(x_8 = \text{head}, x_9 = \text{head} \mid \hat{\theta}) = p(x_8 = \text{head} \mid \hat{\theta}) \cdot p(x_9 = \text{head} \mid \hat{\theta})$$
$$= \hat{\theta} \cdot \hat{\theta} = \frac{5}{7} \cdot \frac{5}{7} = \frac{25}{49}$$

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

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

Solution:

Following the Bayes estimator :

$$p(\theta|D) = \frac{p(D|\theta)\,p(\theta)}{\int p(D|\theta)\,p(\theta)\,d\theta}$$

$$= \frac{\theta^5(1-\theta)^2 \cdot 1}{\int_0^1 \theta^5(1-\theta)^2 \cdot 1\, d\theta}$$

$$= \frac{\theta^5(1-\theta)^2 \cdot 1}{\frac{1}{168}} = 168\,\theta^5(1-\theta)^2$$

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

$$= \int_0^1 \theta^2 \cdot 168\,\theta^5(1-\theta)^2\, d\theta = \int_0^1 168\,\theta^7(1-\theta)^2\, d\theta$$

$$= \frac{7}{15}$$

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 $p(x \mid \mu) \sim \mathcal{N}(\mu, \sigma^2)$, 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 $p(\mu \mid \mathcal{D}) \sim \mathcal{N}(\mu_n, \sigma_n^2)$, 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.

Solution:

we can derive:

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

$$\Downarrow$$

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

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

$\therefore$ proofed.

(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.

Solution:

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

$$\therefore \frac{1}{\sigma_n^2}\mu_n \leq \frac{1}{\sigma_n^2} \cdot \max(\hat{\mu}_n, \mu_0)$$

$$\Downarrow$$

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

日期: /

② $$\frac{1}{\sigma_n^2}\mu_n = \frac{n}{\sigma^2}\hat{\mu}_n + \frac{1}{\sigma_0^2}\mu_0 \geqslant \underbrace{\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)}_{\frac{1}{\sigma_n^2}}\cdot \min(\hat{\mu}_n, \mu_0)$$

$$\therefore \frac{1}{\sigma_n^2}\mu_n \geqslant \frac{1}{\sigma_n^2}\cdot \min(\hat{\mu}_n, \mu_0)$$

$$\therefore \mu_n \geqslant \min(\hat{\mu}_n, \mu_0)$$

Finally:

$$\min(\hat{\mu}_n, \mu_0) \leqslant \mu_n \leqslant \max(\hat{\mu}_n, \mu_0)$$

Proofed.