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.

- **最大似然估计（MLE）**：仅基于数据本身，不考虑任何先验信息。它假设参数是固定的未知值，通过最大化数据的**似然函数**来找到最有可能的参数值。
- **贝叶斯估计（Bayesian Estimation）**：引入了**先验分布**，假设参数是一个随机变量，而不是一个固定值。通过贝叶斯公式结合**先验分布**和**似然函数**计算**后验分布**，并从中得到参数的估计。

2. 计算公式

- **MLE 的目标**是找到使似然函数 $P(D|\theta)$ 最大化的参数 θ：

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} P(D|\theta)$$

- **贝叶斯估计**利用贝叶斯定理：

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$$

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.

$$P(x) = \int_0^\infty \lambda\eta e^{-\lambda x - \eta y}\, dy = \lim_{b\to\infty} -\lambda e^{-\lambda x - \eta y}\Big|_0^b = \lambda e^{-\lambda x}$$

$$P(y) = \int_0^\infty \lambda\eta e^{-\lambda x - \eta y}\, dx = \lim_{b\to\infty} -\eta e^{-\lambda x - \eta y}\Big|_0^b = \eta e^{-\eta y}$$

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

$\rightarrow$ x, y independent

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

the joint density: $P(D|\lambda) = \prod_{k=1}^{N} \left(\lambda\eta e^{-\lambda x_k - \eta y_k}\right)$

$J(\lambda) = \log(D|\lambda,\eta) = \ln P(D|\lambda) = \sum_{k=1}^{N} \ln(\lambda\eta e^{-\lambda x_k - \eta y_k})$

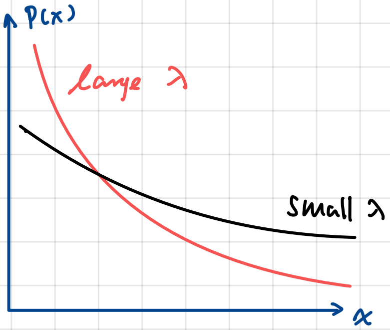
$= \sum \log(P(x_k, y_k|\lambda,\eta))$

$= N \ln(\lambda\eta) - \sum_{k=1}^{N}(\lambda x_k + \eta y_k) = N(\log\lambda + \log\eta - \lambda\bar{x} - \eta\bar{y})$

$\frac{\partial J}{\partial\lambda} = \frac{\partial}{\partial\lambda} P(D|\lambda) = \frac{N}{\lambda\eta}\cdot\eta - \sum_{k=1}^{N}(x_k) \overset{!}{=} 0$

$\frac{N}{\lambda} - \bar{x}N = 0$ $\qquad \lambda = \frac{N}{\sum_{k=1}^{N} x_k} = \frac{1}{\bar{x}}$

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



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

$$\ln P(D|\lambda) = N\ln(\lambda\eta) - \sum_{k=1}^{N}(\lambda x_k + \eta y_k)$$

$$\overset{\eta = \frac{1}{\lambda}}{=} N\ln(1) - \sum_{k=1}^{N}\left(\lambda x_k + \frac{1}{\lambda} y_k\right)$$

$$= -\sum_{k=1}^{N}\left(\lambda x_k + \frac{1}{\lambda} y_k\right)$$

$$\frac{\partial}{\partial\lambda}\ln P(D|\lambda) = -\sum_{k=1}^{N}\left(x_k - \frac{1}{\lambda^2} y_k\right) = 0$$

$$\Rightarrow \lambda^2 = \sum_{k=1}^{N} \frac{y_k}{x_k}$$

$$\hat{\lambda} = \pm\sum_{k=1}^{N}\sqrt{\frac{y_k}{x_k}} = \pm\sqrt{\frac{\bar{y}}{\bar{x}}}$$

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

$$\ln P(D|\lambda) = N \ln(\lambda\eta) - \sum_{k=1}^{N}(\lambda x_k + \eta y_k)$$

$$(\lambda > 0) \wedge (1 - \lambda > 0) \Rightarrow 0 < \lambda < 1$$

$$= N \ln(\lambda - \lambda^2) - \sum_{k=1}^{N}(\lambda x_k + (1-\lambda) y_k)$$

$$\frac{\partial}{\partial \lambda} \ln P(D|\lambda) = \frac{N}{\lambda - \lambda^2} \cdot (1 - 2\lambda) - \sum_{k=1}^{N}(x_k - y_k) \overset{!}{=} 0$$

$$N - 2N\lambda - (\lambda - \lambda^2) \cdot \sum_{k=1}^{N}(x_k - y_k) \overset{!}{=} 0$$

$$\sum_{k=1}^{N}(x_k - y_k) \cdot \lambda^2 + \left(-2N - \sum_{k=1}^{N}(x_k - y_k)\right)\lambda + N = 0$$

$$\sum_{k=1}^{N}(x_k - y_k) \cdot \lambda^2 + \left[\sum_{k=1}^{N}(-2 - (x_k - y_k))\right]\lambda + \sum_{k=1}^{N} 1 = 0$$

$$\lambda_{1,2} = \frac{-b - \sqrt{b^2 - 4ac}}{2a} = \sum_{k=1}^{N} \frac{2 + x_k - y_k - \sqrt{(2 + x_k - y_k)^2 - 4(x_k - y_k)}}{2(x_k - y_k)}$$

$$= \sum_{k=1}^{N} \frac{2 + x_k - y_k - \sqrt{(x_k - y_k)^2 + 4}}{2(x_k - y_k)}$$

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

$$= N(\ln(\lambda - \lambda^2) + \lambda\underbrace{(\bar{y} - \bar{x})}_{\bar{d}} + \bar{y})$$

$$\frac{\partial J}{\partial \lambda} = N\left(\frac{1 - 2\lambda}{\lambda - \lambda^2} + \bar{d}\right) \overset{!}{=} 0$$

$$\lambda = \frac{(\bar{d} - 2) \pm \sqrt{\bar{d}^2 + 4}}{2\bar{d}}$$

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

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

$\theta \quad \theta \quad 1-\theta \quad 1-\theta \quad \theta \quad \theta \quad \theta$

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.

$$P(D \mid \theta) = \prod_{k=1}^{N} P(X_k \mid \theta) = \theta^5 (1-\theta)^2$$

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

$$\ln P(D \mid \theta) = \ln\left(\theta^5 (1-\theta)^2\right) = 5 \ln \theta + 2 \ln (1-\theta)$$

$$\frac{\partial}{\partial \theta} P(D \mid \theta) = \frac{5}{\theta} - \frac{2}{1-\theta} = 0$$

$$5 - 5\theta - 2\theta = 0$$

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

$$P(X_8 = \text{head}, X_9 = \text{head} \mid \hat{\theta}) = \hat{\theta}^2 = \frac{25}{49}$$

(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,

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

$$P(\theta \mid D) = \frac{P(D \mid \theta) \cdot P(\theta)}{P(D)}$$

$$= \frac{P(D \mid \theta) \cdot P(\theta)}{\int_0^1 P(D \mid \theta) \cdot P(\theta) \, d\theta}$$

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

$$= \frac{\theta^5 (1-\theta)^2}{\int_0^1 \theta^7 - 2\theta^6 + \theta^5 \, d\theta}$$

$$= \frac{\theta^5 (1-\theta)^2}{\frac{1}{8}\theta^8 - \frac{2}{7}\theta^7 + \frac{1}{6}\theta^6 \Big|_0^1}$$

$$= \frac{\theta^5 (1-\theta)^2}{\frac{42 - 96 + 56}{48 \times 7}}$$

$$= \frac{\overset{24}{\cancel{48}} \times 7}{\cancel{2}} \theta^5 (1-\theta)^2 = 168\, \theta^5 (1-\theta)^2$$

$$\int P(X_8 = head, X_9 = head \mid \theta) P(\theta \mid D) d\theta = \int_0^1 \theta^2 \cdot 168 (\theta^5 (1-\theta)^2) \, d\theta$$

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

$$= 168 \int_0^1 \theta^9 - 2\theta^8 + \theta^7 \, d\theta$$

$$= 168 \cdot \left(\frac{1}{10}\theta^{10} - \frac{2}{9}\theta^9 + \frac{1}{8}\theta^8 \right) \Big|_0^1$$

$$= \cancel{168}^{21} \cdot \frac{72 - 160 + 90}{90 \cdot \cancel{8}}$$

$$= \frac{\cancel{21}^{7} \times \cancel{2}}{\cancel{90}\ \cancel{30}\ 15} = \frac{7}{15}$$

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.

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

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

1. if $\frac{\sigma^2}{n} \leq \sigma_0^2$

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

$$\frac{\cancel{\sigma^2}\sigma_0^2}{n\sigma_0^2 + \sigma^2} \leq \frac{\cancel{\sigma^2}}{n}$$

$$\sigma_0^2 \leq \sigma_0^2 + \frac{1}{n}\sigma^2$$

$$\frac{\sigma^2}{n} \geq 0 \qquad \text{always true because } \sigma^2, n > 0 \text{ with } n \in \mathbb{N}^+$$

2. if $\frac{\sigma^2}{n} \geq \sigma_0^2$

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

$$\frac{\sigma^2 \cancel{\sigma_0^2}}{n\sigma_0^2 + \sigma^2} \leq \cancel{\sigma_0^2}$$

$$\sigma^2 \leq n\sigma_0^2 + \sigma^2$$

$$0 \leq n\sigma_0^2 \qquad \text{always true}$$

(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.

1. if $\hat{\mu}_n < \mu_0$, $\Rightarrow \frac{1}{n}\sum_{k=1}^{n} x_k < \mu_0$

to prove: $\hat{\mu}_n < \mu_n < \mu_0$

$$\hat{\mu}_n < \delta_n^2\left(\frac{n}{\delta^2}\hat{\mu}_n + \frac{\mu_0}{\delta_0^2}\right) < \mu_0$$

$$\frac{\delta^2\delta_0^2}{n\delta_0^2+\delta^2}\left(\frac{n}{\delta^2}\hat{\mu}_n + \frac{\mu_0}{\delta_0^2}\right)$$

$$\hat{\mu}_n = \frac{n\delta_0^2}{n\delta_0^2+\delta^2}\hat{\mu}_n + \frac{\delta^2}{n\delta_0^2+\delta^2}\hat{\mu}_n < \frac{n\delta_0^2}{n\delta_0^2+\delta^2}\hat{\mu}_n + \frac{\delta^2}{n\delta_0^2+\delta^2}\mu_0 < \frac{n\delta_0^2}{n\delta_0^2+\delta^2}\mu_0 + \frac{\delta^2}{n\delta_0^2+\delta^2}\mu_0$$

$$= \mu_0$$

2. if $\hat{\mu}_n \geq \mu_0$, ...

See solution !