MACHINE LEARNING 1: ASSIGNMENT 2

Tom Nick 340528
Niklas Gebauer 340942

Exercise 1

(a) With the definition $P(x_k \mid \theta)$:

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

We can state the likelihood function $P(\mathcal{D} \mid \theta)$:

$$\begin{aligned} P(\mathcal{D} \mid \theta) &= \prod_{k=i}^{n} P(x_k \mid \theta) \\ &= \theta^5 \cdot (1-\theta)^2 \end{aligned}$$

(b) For this task we define:

$$f(x) = P(\mathcal{D} \mid \theta) = \theta^5 \cdot (1-\theta)^2$$

To obtain the maximum likelihood solution $\hat{\theta}$ we can now use the standard procedure to find local minima and maxima in f. So we will at first set the first derivative to zero and solve for θ:

$$\begin{aligned} f'(x) &= 0 \\ \Leftrightarrow (\theta^5 \cdot (1-\theta)^2)' &= 0 \\ \Leftrightarrow (\theta^7 - 2\theta^6 + \theta^5)' &= 0 \\ \Leftrightarrow 7\theta^6 - 12\theta^5 + 5\theta^4 &= 0 \\ \Leftrightarrow \theta^4(7\theta^2 - 12\theta + 5) &= 0 \\ &\Leftrightarrow \theta = 0 \vee 7\theta^2 - 12\theta + 5 = 0 \end{aligned}$$

So the first possible solution for $\hat{\theta}$ is 0. Now we can use the p-q-formula to calculate the other solutions:

$$\begin{aligned} 7\theta^2 - 12\theta + 5 &= 0 \\ \Leftrightarrow \theta^2 - \frac{12}{7}\theta + \frac{5}{7} &= 0 \\ \Leftrightarrow \frac{12}{14} \pm \sqrt{(\frac{12}{14})^2 - \frac{5}{7}} &= \theta_1, \theta_2 \\ \Leftrightarrow \frac{6}{7} \pm \sqrt{\frac{1}{49}} &= \theta_1, \theta_2 \\ \Leftrightarrow 1 = \theta_1 \wedge \frac{5}{7} &= \theta_2 \end{aligned}$$

So we have three possible solutions for $\hat{\theta} \in [0, 1]$:

$$\begin{aligned} \hat{\theta_0} &= 0 \\ \hat{\theta_1} &= 1 \\ \hat{\theta_2} &= \frac{5}{7} \end{aligned}$$

Now we check more derivatives to see which is the local maximum that we are looking for:

$$f''(0) = f^3(0) = f^4(0) = 0 \neq f^5(0) \Rightarrow \text{saddle point at } \hat{\theta_0}$$
$$f''(1) = 2 \Rightarrow \text{minimum at } \hat{\theta_1}$$
$$f''(\frac{5}{7}) = -\frac{1250}{2401} \Rightarrow \text{maximum at } \hat{\theta_2}$$

So the maximum likelihood solution for $\hat{\theta}$ is $\frac{5}{7}$, which does make sense since it is the number of heads compared to the total number of tosses (which is the sample mean, if we assume a random variable X with $X = 1$ if $x_k = \textit{head}$ and $X = 0$ otherwise):

$$\hat{\theta} = \frac{\#\{x = \text{head} \mid x \in D\}}{\#(D)} = \frac{5}{7}$$

With this and the fact that each sample x_k is generated independently we can compute $P(x_8 = \text{head}, \quad x_9 = \text{head} \mid \hat{\theta})$:

$$P(x_8 = \text{head}, \quad x_9 = \text{head} \mid \hat{\theta}) = P(x_8 = \text{head} \mid \hat{\theta})P(x_9 = \text{head} \mid \hat{\theta}) = \hat{\theta}^2 = \frac{25}{49} \approx 0.51$$

(c) First we want to compute the posterior destribution $p(\theta \mid \mathcal{D})$. With our prior distribution

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

we get:

$$\begin{aligned} p(\theta \mid \mathcal{D}) &= \frac{p(\mathcal{D} \mid \theta)p(\theta)}{\int p(\mathcal{D} \mid \theta)p(\theta)d\theta} \\ &= \alpha \prod_{k=1}^{7} P(x_k \mid \theta)p(\theta) \\ &= \begin{cases} \alpha \cdot \theta^5 \cdot (1-\theta)^2 & \text{if } 0 \leq \theta \leq 1 \\ 0 & \text{else} \end{cases} \end{aligned}$$

with:

$$\begin{aligned} \alpha &= \frac{1}{\int\limits_0^1 \prod_{k=1}^{7} P(x_k \mid \theta)p(\theta)d\theta} \\ &= \frac{1}{\int\limits_0^1 \prod_{k=1}^{7} P(x_k \mid \theta)d\theta} \\ &= \frac{1}{\int\limits_0^1 \theta^5 \cdot (1-\theta)^2 d\theta} \\ &= \frac{1}{\int\limits_0^1 \theta^7 - 2\theta^6 + \theta^5 d\theta} \\ &= \frac{1}{[\frac{\theta^8}{8} - \frac{2\cdot\theta^7}{7} + \frac{\theta^6}{6}]_0^1} \\ &= \frac{1}{\frac{28}{168} - \frac{48}{168} + \frac{21}{168}} \\ &= \frac{1}{\frac{1}{168}} = 168 \end{aligned}$$

Now we can evaluate the probability that the next two tosses are head:

$$\begin{aligned}
\int P(x_8 = \text{head}, x_9 = \text{head} \mid \theta) p(\theta \mid D) d\theta &= \int P(x_8 = \text{head} \mid \theta) P(x_9 = \text{head} \mid \theta) p(\theta \mid D) d\theta \\
&= \int_0^1 \theta^2 \cdot \alpha \cdot \theta^5 \cdot (1-\theta)^2 d\theta \\
&= 168 \int_0^1 \theta^7 (1-\theta)^2 d\theta \\
&= 168 \int_0^1 \theta^9 - 2\theta^8 + \theta^7 d\theta \\
&= 168 \cdot [\frac{\theta^{10}}{10} - \frac{2 \cdot \theta^9}{9} + \frac{\theta^8}{8}]_0^1 \\
&= 168 \cdot \frac{1}{360} = \frac{7}{15} \approx 0.4\overline{6}
\end{aligned}$$

Exercise 2

(a) To show

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

we will at first solve the formula from section 3.4.1 of Duda et al. for σ_n^2:

$$\begin{aligned}
\frac{1}{\sigma_n^2} &= \frac{n}{\sigma^2} + \frac{1}{\sigma_0^2} \\
\Leftrightarrow 1 &= \frac{\sigma_0^2 n + \sigma^2}{\sigma^2 \sigma_0^2} \cdot \sigma_n^2 \\
\Leftrightarrow \sigma_n^2 &= \frac{\sigma^2 \sigma_0^2}{\sigma_0^2 n + \sigma^2}
\end{aligned}$$

Now we can show $\sigma_n^2 \leq \min(\frac{\sigma^2}{n}, \sigma_0^2)$. We assume $\frac{\sigma^2}{n} \leq \sigma_0^2$:

$$\begin{aligned}
\sigma_n^2 &\leq \min(\frac{\sigma^2}{n}, \sigma_0^2) \\
\Leftrightarrow \frac{\sigma^2 \sigma_0^2}{\sigma_0^2 n + \sigma^2} &\leq \frac{\sigma^2}{n} \\
\Leftrightarrow \sigma^2 \sigma_0^2 &\leq \sigma^2 \sigma_0^2 + \frac{\sigma^4}{n} \\
\Leftrightarrow \sigma^2 \sigma_0^2 &\leq \sigma^2 \cdot (\sigma_0^2 + \frac{\sigma^2}{n} \\
\Leftrightarrow \sigma_0^2 &\leq \sigma_0^2 + \frac{\sigma^2}{n} \\
\Leftrightarrow 0 &\leq \frac{\sigma^2}{n}
\end{aligned}$$

This holds true since n is the number of features and σ^2 is the variance of a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$, thus $n, \sigma > 0$.

Now we assume the other case ($\sigma_0^2 \leq \frac{\sigma^2}{n}$):

$$\begin{aligned}
\sigma_n^2 &\leq \min\left(\frac{\sigma^2}{n}, \sigma_0^2\right) \\
\Leftrightarrow \frac{\sigma^2\sigma_0^2}{\sigma_0^2 n + \sigma^2} &\leq \sigma_0^2 \\
\Leftrightarrow \sigma^2\sigma_0^2 &\leq \sigma_0^2 \cdot (\sigma_0^2 n + \sigma^2) \\
\Leftrightarrow \sigma^2 &\leq \sigma_0^2 n + \sigma^2 \\
\Leftrightarrow 0 &\leq \sigma_0^2 n
\end{aligned}$$

This again holds true since $n > 0$ as stated above and $\sigma_0^2 > 0$ since it is the variance of the Gaussian distribution $\mathcal{N}(\mu_0, \sigma_0^2)$

(b) To show

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

we again first solve the formula from section 3.4.1 of Duda et al. for μ_n, also using our σ_n^2 from above:

$$\begin{aligned}
\frac{\mu_n}{\sigma_n^2} &= \frac{n}{\sigma^2}\hat{\mu}_n + \frac{\mu_0}{\sigma_0^2} \\
\Leftrightarrow \mu_n &= \left(\frac{n}{\sigma^2}\hat{\mu}_n + \frac{\mu_0}{\sigma_0^2}\right) \cdot \sigma_0^2 \\
\Leftrightarrow \mu_n &= \left(\frac{n}{\sigma^2}\hat{\mu}_n + \frac{\mu_0}{\sigma_0^2}\right) \cdot \frac{\sigma^2\sigma_0^2}{\sigma_0^2 n + \sigma^2} \\
\Leftrightarrow \mu_n &= \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\hat{\mu}_n + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\mu_0
\end{aligned}$$

Lets first assume $\mu_0 < \hat{\mu}_n$ (1) and show the first part of the inequality:

$$\begin{aligned}
\min(\mu_0, \hat{\mu}_n) &\leq \mu_n \\
\Leftrightarrow \mu_0 &\leq \mu_n \\
\Leftrightarrow \mu_0 &\leq \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\hat{\mu}_n + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\mu_0 \\
\Leftrightarrow \mu_0 \cdot (n\sigma_0^2 + \sigma^2) &\leq n\sigma_0^2\hat{\mu}_n + \sigma^2\mu_0 \\
\Leftrightarrow \mu_0 n\sigma_0^2 &\leq n\sigma_0^2\hat{\mu}_n \\
\Leftrightarrow \mu_0 &\leq \hat{\mu}_n
\end{aligned}$$

This hold true due to our assumption (1) from above. Let's show the second part of the inequality:

$$\begin{aligned}
\mu_n &\leq \max(\mu_0, \hat{\mu}_n) \\
\Leftrightarrow \mu_n &\leq \hat{\mu}_n \\
\Leftrightarrow \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\hat{\mu}_n + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\mu_0 &\leq \hat{\mu}_n \\
\Leftrightarrow n\sigma_0^2\hat{\mu}_n + \sigma^2\mu_0 &\leq \hat{\mu}_n \cdot (n\sigma_0^2 + \sigma^2) \\
\Leftrightarrow \sigma^2\mu_0 &\leq \hat{\mu}_n\sigma^2 \\
\Leftrightarrow \mu_0 &\leq \hat{\mu}_n
\end{aligned}$$

This again holds true due to our assumption (1) from above.

Now we'll assume that $\hat{\mu}_n < \mu_0$ (2) and show that our inequality still holds. We'll start with the first part again:

$$\begin{aligned}
\min(\mu_0, \hat{\mu}_n) &\leq \mu_n \\
\Leftrightarrow \hat{\mu}_n &\leq \mu_n \\
\Leftrightarrow \hat{\mu}_n &\leq \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\hat{\mu}_n + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\mu_0 \\
\Leftrightarrow \hat{\mu}_n \cdot (n\sigma_0^2 + \sigma^2) &\leq n\sigma_0^2\hat{\mu}_n + \sigma^2\mu_0 \\
\Leftrightarrow \hat{\mu}_n\sigma^2 &\leq \sigma^2\mu_0 \\
\Leftrightarrow \hat{\mu}_n &\leq \mu_0
\end{aligned}$$

This hold true due to our assumption (2) from above. Let's show the second part of the inequality:

$$\begin{aligned}
\mu_n &\leq \max(\mu_0, \hat{\mu}_n) \\
\Leftrightarrow \mu_n &\leq \mu_0 \\
\Leftrightarrow \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\hat{\mu}_n + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\mu_0 &\leq \mu_0 \\
\Leftrightarrow n\sigma_0^2\hat{\mu}_n + \sigma^2\mu_0 &\leq \mu_0 \cdot (n\sigma_0^2 + \sigma^2) \\
\Leftrightarrow n\sigma_0^2\hat{\mu}_n &\leq \mu_0 n\sigma_0^2 \\
\Leftrightarrow \hat{\mu}_n &\leq \mu_0
\end{aligned}$$

This again holds true due to our assumption (2) from above.