Winter term 2020/21
L. Rendsburg / U. von Luxburg

# Assignment 12

## Mathematics for Machine Learning

Submission due on **15.02.21, 8:00**
Assignment 13 will be the last assignment and consists only of bonus points. This means that you need at least 120 points in total to be admitted to the exam.

Justify all your claims.

**Exercise 1** (**Make your own exam questions**, 6 points (1 per question))**.** In this exercise, you are supposed to go through the lectures and assignments and think of possible exam questions. For each of the three topics Linear algebra, Calculus, and Probability theory/Statistics, state one easy question and one harder question. Remember that students are allowed to use a cheat sheet during the exam, so don't just ask for definitions! Also state the solution for each of your questions.
To submit this exercise, please use the template "questions.tex" provided on the webpage. In addition to your usual submission, upload a .tex and .pdf version of your questions "questions_lastname1_lastname2.tex" and "questions_lastname1_lastname2.pdf" on moodle. (You may solve this exercise by hand if you are unfamiliar with LaTeX.)

**Exercise 2** (**Multiple choice questions**, 5 points (1 per question))**.** For each of the following questions, choose exactly one answer.

1. When testing the null hypothesis $H_0$ against $H_1$, the p-value of an observation is

   A. $P(H_0 \text{ is true})$

   B. $1 - P(H_1 \text{ is true})$

   C. the smallest level of a test, which would reject $H_0$

2. Consider two different tests of level $\alpha$ with power functions $\beta_1(\theta) = \theta^5$ (Test 1) and $\beta_2(\theta) = \theta^3$ (Test 2) for $\theta \in \Theta_1 = [1/2, 1)$. Which test would you prefer?

   A. Test 1  B. Test 2

3. Which of the following (95%)-confidence intervals for a parameter $\theta \in \mathbb{R}$ would you prefer?

   A. $I_1 = [-2, 3]$  B. $I_2 = [-1, 0]$  C. $I_3 = [-1, 3]$  D. $I_4 = [-2, 0]$

4. There always exists a test of level $\alpha = 0$.

   A. True  B. False

5. A manufacturer claims that at least 99% of their produced cellphones are flawless. To test this statement, we observe independent samples $x_1, \dots, x_n$ from a Bernoulli distribution $\mathrm{Ber}(\theta)$ for $\theta \in (0, 1)$, where $x_i = 1$ indicates that cellphone $i$ is faulty and $x_i = 0$ indicates that it is flawless. How do we have to formulate a hypothesis test, if we want to control the error of wrongfully accusing the manufacturer of a false statement?

   A. $H_0 : \theta \geq 0.99$ and $H_1 : \theta < 0.99$

   B. $H_0 : \theta < 0.99$ and $H_1 : \theta \geq 0.99$

   C. $H_0 : \theta > 0.01$ and $H_1 : \theta \leq 0.01$

   D. $H_0 : \theta \leq 0.01$ and $H_1 : \theta > 0.01$

**Exercise 3** (**Testing basics**, 1+1+2 points)**.** A new treatment for a disease is claimed to improve the recovery rate of patients. It is known that 20% of all patients recover without any treatment. We now observe for $n \in \mathbb{N}$ independent patients whether they recover after being treated:

$$X_i = \begin{cases} 1, & \text{if patient } i \text{ recovers} \\ 0, & \text{otherwise} \end{cases}, \quad i \in \{1, \dots, n\}.$$

a) Formalize this test setting by stating the distribution of the $X_i$, the null hypothesis $H_0$, and the alternative hypothesis $H_1$.

Let $x = (x_1, \dots, x_n) \in \{0,1\}^n$. Now consider the following three tests $\varphi_j \colon \{0,1\}^n \to \{0,1\}$ ($\varphi = 1$ means reject and $\varphi = 0$ means retain):

$$\varphi_1(x) = x_1, \quad \varphi_2(x) = \prod_{i=1}^n x_i, \quad \varphi_3(x) = \begin{cases} 1, & \text{if } \sum_{i=1}^n x_i \geq m \\ 0, & \text{otherwise} \end{cases} \quad \text{for some } m \in \mathbb{N}_0$$

b) Reformulate each of the three tests in terms of a test statistic and a rejection region such that $\varphi_j(x_1, \dots, x_n) = \mathbb{1}_{\{T_j(x_1,\dots,x_n) \in R_j\}}$ for $j \in \{1, 2, 3\}$.

c) Compute power function and level for each test. Additionally, state the level of $\varphi_3$ for $n = 10$ and $m = 5$.
**Hint:** You may use that the sum of independent Bernoulli distributions is Binomially distributed ($X_1, \dots, X_n \sim \mathrm{Ber}(\theta) \Rightarrow \sum_{i=1}^n X_i \sim \mathrm{Bin}(n, \theta)$). Additionally, use that $\sum_{k=m}^n \binom{n}{k} \theta^k (1-\theta)^{n-k}$ is increasing in $\theta$.

**Exercise 4** (**Two-tailed testing**, 2+1+2 points)**.** Consider i.i.d. samples from a normal distribution $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ with $\sigma > 0$ known. We want to test the hypothesis $H_0 : \mu = 0$ against $H_1 : \mu \neq 0$. Consider the test with test statistic $T(x_1, \dots, x_n) = |1/n \sum_{i=1}^n x_i|$ and rejection region $R = [c, \infty)$ for some $c > 0$.

a) Show that the power function $\beta(\mu)$ satisfies

$$\beta(\mu) = 2\left(1 - \Phi\left(\frac{\sqrt{n}}{\sigma}(c - \mu)\right)\right),$$

where $\Phi$ is the cumulative distribution function of a standard normal distribution, that is, $\Phi(x) = P(Z \leq x)$ for $Z \sim \mathcal{N}(0,1)$ and $x \in \mathbb{R}$.
**Hint:** Use the following properties of normal distributions:

$$\begin{aligned} X \sim \mathcal{N}(\mu_X, \sigma_X^2),\, Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2) \text{ independent} \quad &\Rightarrow \quad X + Y \sim \mathcal{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2) \\ X \sim \mathcal{N}(\mu_X, \sigma_X^2) \text{ and } a, b \in \mathbb{R} \quad &\Rightarrow \quad aX + b \sim \mathcal{N}(a\mu_X + b, a^2\sigma_X^2) \\ &\Phi(-x) = 1 - \Phi(x) \quad \forall x \in \mathbb{R}. \end{aligned}$$

b) Let $\alpha \in (0,1)$. Show that the value $c = c_\alpha$, which leads to a test level of $\alpha$, is given by

$$c_\alpha = \frac{\sigma}{\sqrt{n}} \Phi^{-1}\left(1 - \frac{\alpha}{2}\right).$$

c) Show that the p-value of a sample $x = (x_1, \dots, x_n) \in \mathbb{R}^n$ is given by

$$p(x) = 2\left(1 - \Phi\left(\frac{\sqrt{n}}{\sigma} T(x)\right)\right).$$

.

**Hint:** $\alpha \mapsto c_\alpha$ is decreasing in $\alpha$.