Statistical data analysis

Name: _____

Signature: _____

Labs	
Exam (written + oral)	≥ 25
Total	≥ 50
Grade	

Instructions: the solution time is 150 minutes, clearly answer as many questions as possible, work with the terms used in the course, employ math (notation, expressions, equations) as often as possible, you can use calculators.

Statistical minimum. *(10 b)* Answer the questions below. In each of the questions, just one answer is correct.

(a) *(2 b)* If the p-value of a given test statistic is 0.03 then we can:

i) assume that the extremeness of the test statistic is due to chance,

ii) infer that there is a 3% chance of getting a more extreme test statistic, provided that the null hypothesis is true,

iii) assume that the null hypothesis holds with the probability 0.03,

iv) accept the null hypothesis at the 0.05 significance level.

(b) *(2 b)* If you want to conduct a hypothesis test about a mean from a population with a skewed distribution, you should:

i) use a stratified sample,

ii) use a large significance level α,

iii) have all outcomes classified as success or failure,

iv) use a sample size greater than 30.

(c) *(2 b)* What does the assumption of independence mean? This assumption means that:

i) none of your independent variables are correlated,

ii) the errors in your model are not related to each other,

iii) you must use an independent design rather than a repeated-measures design,

iv) the residuals in your model are not independent.

(d) *(2 b)* If the heights of women are normally distributed with a mean of 64 inches, which of the following is the highest? The probability of randomly choosing:

i) one woman and finding her height is between 63 and 65 inches,

ii) 15 women and finding that their mean height is between 63 and 65 inches,

iii) 100 women and finding that their mean height is between 63 and 65 inches,

iv) all of the above have the same probability.

(e) *(2 b)* You want to estimate the proportion of Czech citizens who support Covid vaccination. How large a sample (the size denoted as n) would be needed to ensure a 95% probability that the actual population proportion p will be no more than 3 percentage points from the sample population? (hints: the number of vaccination supporters in our sample will follow the binomial distribution with the mean np and standard deviation $\sqrt{np(1-p)}$, you should work with the conservative guess that $p = 0.5$ and approximate the binomial distribution with normal one, you know that $z_{0.025} = 1.96$).

i) 512,

ii) 1068,

iii) 2506,

iv) 3152,

v) 6304.

Multivariate regression. *(10 b)* You are building a multivariate linear model. There is a large number of independent variables, greater than the number of training samples you have available. Formally, then: $y \in R$, $\mathbf{x} \in R^p$, $T = \{\langle \mathbf{x_i}, y_i \rangle\}_{i=1}^m$, $p > m$, we assume $y = \mathbf{x}^T\beta + \epsilon$.

(a) *(2 b)* Estimate and comment on the result of directly applying the least squares method to the set T.

(b) *(2 b)* Describe in as much detail as possible the application of the subset selection method in connection with the creation of an optimal linear regression model.

(c) *(2 b)* Explain what problem compared to approach ad a) the method ad b) solves. At the same time explain what pitfalls arise on the contrary.

(d) *(2 b)* Describe the approach to feature selection and regression based on forward and backward stepwise regression. On the basis of which criteria do we compare models? Estimate the complexity of this feature selection method.

(e) *(2 b)* Which of the pair of approaches ad d) would you prefer in this particular task and why? Also compare both approaches with the subset selection method.

Logistic regression. *(10 b)* Discuss the logistic regression.

(a) *(2 b)* Describe the conditions under which the use of logistic regression is appropriate (define the task, including its variants given by different types of variables).

(b) *(2 b)* Write down the definition formula of logistic regression. Explain the meaning of variables.

(c) *(1 b)* Name the method by which the coefficients of the logistics model are determined. Write down its general formula.

(d) *(1 b)* Explain the concept of decision boundary. Make an illustrative picture.

(e) *(2 b)* What is the general shape of the logistic decision boundary? Explain based on the logistic formula that you provided above.

(f) *(2 b)* Compare logistic regression with linear regression. When one method fails and when the other fails?

Robust statistics. *(10 b)* Assume that you learn a multivariate linear model $\hat{y} = \mathbf{x}^T\hat{\beta}$. Your goal is to estimate the vector of model parameters β from training data.

(a) *(2 b)* How do you make this estimate if you assume that the real relationship between the dependent variable Y and the vector of independent variables $\mathbf{X}$ can be described by the generative model $Y = \mathbf{X}^T\beta + \epsilon$, where ϵ is Gaussian noise $N(0, 1)$? Name the method and write down its criterion.

(b) *(2 b)* Justify why the given method is optimal in the given situation.

(c) *(2 b)* What methods would you use if the relationship remains unchanged but the noise ϵ will be a mixture of $\alpha N(0, 1) + (1 - \alpha)N(\gamma, 1)$ where α is close 1 and γ is unknown and finite?

(d) *(2 b)* Will you further change the method if ϵ is a Laplace noise $Laplace(0, 1)$. If so, describe how and why.

(e) *(2 b)* Why are not the methods described in (c) and (d) also appropriate in the case of ad (a)?

Clustering. *(10 b)* Consider the task of clustering.

(a) *(1 b)* Define the clustering task verbally.

(b) *(2 b)* Define the clustering task formally, pose it as an optimization problem. Assign it to the correct complexity class. Justify.

(c) *(3 b)* Is spectral clustering an example of a clustering algorithm that formulates and solves clustering as an optimization problem? Explain.

(d) *(2 b)* Does spectral clustering employ a kernel function? How does the kernel function application differ from kernel k-means? Explain.

(e) *(2 b)* Describe how the optimal number of clusters k can be found in spectral clustering. Assume that k is not known in advance.