Statistical data analysis

Name: ________________________________

Signature: ________________________________

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

Instructions: the solution time is 120 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 p)* Answer the following questions:

(a) *(4 p)* What is a cumulative distribution function? What is a quantile and quantile function? What is a probability density function? Define formally and explain their relationship.

(b) *(4 p)* Formulate the central limit theorem. Where can we use it?

(c) *(2 p)* Define the following terms: mean (expected value) and arithmetic mean. Explain the difference between them.

Dimensionality reduction. *(10 p)* Concern the task of dimensionality reduction.

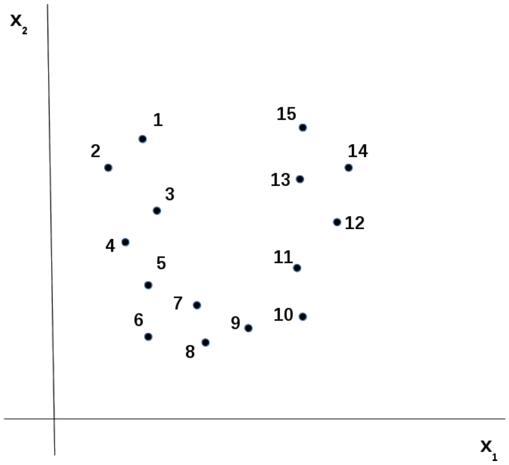
(a) *(2 p)* Define the task of dimensionality reduction formally (input, output, assumptions, criteria).

(b) *(2 p)* Define and explain multidimensional scaling. Is it a linear or non-linear dimensionality reduction method?

(c) *(2 p)* Define the term geodesic distance. How would you calculate it from data? Explain the risks of this calculation.

(d) *(2 p)* Write a pseudo-code of a multidimensional scaling method with a geodesic distance. Name the method (use its common name).

(e) *(2 p)* In the figures below describe the above-defined method (i.e., outline graphically the way in which the objects from the original space map into the reduced space). Can the output look different than you plotted?



Multivariate regression. *(10 p)* Concern the task of multivariate linear model construction. The number of independent variables is large, they differ in relevance, some of them are entirely irrelevant.

(a) *(2 p)* Name the two basic shrinkage methods that could serve to decrease the size of the model. Write down the criterion functions that these shrinkage methods minimize.

(b) *(2 p)* Explain the difference in the outcomes of these two methods. Justify.

(c) *(1 p)* Does any of these two methods ask for data preprocessing? If so, which one and why?

(d) *(2 p)* Does any of these two methods have a parameter to be set? If so, how would you set it? What is the meaning of this parameter?

(e) *(2 p)* Explain why the shrinked/reduced models could reach a smaller test error than the referential full model learned with the least squares method. Proposal: stem from the bias and variance trade-off in models of different sizes.

(f) *(1 p)* Mention the disadvantages of shrinkage methods when compared with the traditional least squares regression?

Robust statistics. *(10 p)* Discuss the term robust regression. Answer the questions below.

(a) *(2 p)* What is robust regression and under which conditions you would recommend its application?

(b) *(1 p)* Demonstrate the idea behind robust regression graphically (it is sufficient to concern a simple case with one independent and one dependent variable, a scatterplot comparing the output of robust and classic regression).

(c) *(2 p)* Change the regression task to become robust. Describe at least two ways of the reformulation. Where are the shortcomings?

(d) *(4 p)* There are two paired samples:

$s_1 = \{293, 311, 331, 295, 337, 328, 291, 306, 323, 316\}$,

$s_2 = \{298, 322, 321, 321, 343, 331, 289, 316, 329, 322\}$.

Statistically compare both the samples through estimation of location (central tendency). Employ one parametric and one non-parametric, i.e. robust method. For both the methods clearly formulate their outcome.

You can stem from the following formulae: $T = \frac{\bar{d}-D_0}{s_d/\sqrt{n}}$, $W = \sum_{i=1}^{n}(sgn(x_i - y_i)R_i)$.

The corresponding tabular values: $t_{0.95,9} = 1.883$, $t_{0.975,9} = 2.262$, $t_{0.99,9} = 2.281$, $t_{0.995,9} = 3.250$;

$w_{0.95,10} = 40$, $w_{0.99,10} = 51$.

(e) *(1 p)* Mention the advantages and disadvantages of both the methods for the given input.

Power analysis. *(10 p)* Answer the questions below.

(a) *(2 p)* Explain the term power of a statistical test.

(b) *(3 p)* Enumerate three basic factors that influence this power (describe the relationship in detail for each factor)?

(c) *(5 p)* How many participants you have to invite for a test if you need to have 90% chance to see problems that affect 30% of all the users? Write down the equation for sample size calculation and obtain the sample size.