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 b)* Answer the following questions:

(a) *(5 b)* Consider a random vector $\mathbf{X}$. Define the covariance and correlation matrices. What are the elementary properties of these matrices? What can they be used for?

(b) *(5 b)* Explain the meaning of the term confounding variable. Provide an example and indicate its effect on the model.

Analysis of variance. *(10 b)* Answer the following questions:

(a) *(2 b)* What is the purpose of parametric one-way ANOVA? Formulate its null and alternative hypothesis.

(b) *(3 b)* What are the assumptions of this method? How will you test them? What happens if they are not met?

(c) *(3 b)* Describe in detail the ANOVA test output table at the end of the short command sequence below.

```
F<-unlist(mapply(rep,times=c(8,9,10),x=c(1,2,3)))
O<-F+rnorm(n=27,mean=0,sd=2)
summary(aov(O ~ as.factor(F)))
            Df Sum Sq Mean Sq F value Pr(>F)
F            2  11.41   5.707   1.953  0.164
Residuals   24  70.14   2.923
```

(d) *(2 b)* What is the follow-up post-hoc test? On what principle is it based?

Discriminant analysis. *(10 b)* Discuss the features of linear and quadratic discriminant analysis (LDA and QDA).

(a) *(2 b)* What is the basic idea behind both the methods? Write down the general equation that defines them.

(b) *(2 b)* What is the basic difference between LDA and QDA? What does it follow from?

(c) *(1 b)* Assume you are addressing a problem with the linear Bayesian decision boundary. Which method will achieve higher accuracy on training data? Which on test data? Why?

(d) *(1 b)* Assume you are addressing a problem with the non-linear Bayesian decision boundary. Which method will achieve higher accuracy on training data? Which on test data? Why?

(e) *(1 b)* Consider a general classification task. With the increasing number of training examples, the ratio between QDA and LDA classification accuracy observed on test data will increase, will it fall or will it remain the same? Why?

(f) *(3 b)* You are supposed to determine whether a company with the last year's revenue of 4% will pay a dividend. From a stock market analysis of a large number of companies, you know that 80% of companies pay dividends and their average annual revenue is 10%. Companies without dividends have an average revenue of 0%. The distribution of revenues in both groups is normal with variance $\hat{\sigma}^2 = 0.36$. Will you apply LDA or QDA? You do not have to calculate the exact probability, just write down all the necessary formulae and substitute in them.

Multivariate regression. *(10 b)* Suppose you learn a multivariate linear model. There is a large number of independent variables, you search for a model that minimizes the criterion (y_i is the value of the dependent variable in the i-th sample, x_ij is the value of the j-th independent variable in the i-th sample):

$$\sum_{i=1}^{m}(y_i - \hat{\beta}_0 - \sum_{j=1}^{p}\hat{\beta}_j x_{ij})^2 + \lambda \sum_{j=1}^{p}\hat{\beta}_j^2$$

You first set the parameter λ to 0, then gradually increment it. As we increase λ from 0:

residual sum of squares (RSS) it initially grows, but from a certain point it starts to fall and creates an inverted U curve, residual sum of squares (RSS) initially it will drop, but from a certain point it will begin to grow and create a U curve, residual sum of squares (RSS) will continue to grow, residual sum of squares (RSS) will still fall, residual sum of squares (RSS) remains constant.

(a) *(2 b)* the training residual sum of squares (RSS) will

i) increase initially, and then eventually start decreasing in an inverted U shape,
ii) decrease initially, and then eventually start increasing in a U shape,
iii) steadily increase,
iv) steadily decrease,
v) remain constant.

(b) *(2 b)* the test residual sum of squares (RSS) will

i) increase initially, and then eventually start decreasing in an inverted U shape,
ii) decrease initially, and then eventually start increasing in a U shape,
iii) steadily increase,
iv) steadily decrease,
v) remain constant.

(c) *(2 b)* variance will

i) increase initially, and then eventually start decreasing in an inverted U shape,
ii) decrease initially, and then eventually start increasing in a U shape,
iii) steadily increase,
iv) steadily decrease,
v) remain constant.

(d) *(2 b)* bias will

i) increase initially, and then eventually start decreasing in an inverted U shape,
ii) decrease initially, and then eventually start increasing in a U shape,
iii) steadily increase,
iv) steadily decrease,
v) remain constant.

(e) *(2 b)* the irreducible error ϵ

i) increase initially, and then eventually start decreasing in an inverted U shape,
ii) decrease initially, and then eventually start increasing in a U shape,
iii) steadily increase,
iv) steadily decrease,
v) remain constant.

Robust statistics. *(10 b)* Robustly estimate the scale from the sample below. Use two different methods. {-1.84,1.18,0.0499,-0.751,-0.00707,-2.05,-1.47,-0.0520,-0.991,-0.945}.

(a) *(2 b)* Method 1 (description and application to the sample):

(b) *(2 b)* Method 2 (description and application to the sample):

(c) *(2 b)* Compare the estimates calculated above with the common standard deviation estimate (both theoretically and technically).

(d) *(2 b)* Describe the criteria that determine the quality of a robust scale estimate.

(e) *(2 b)* Discuss the advantages and disadvantages of chosen methods according to the criteria described in the previous subtask.