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)* A call center receives an average of 180 independent calls per hour, 24 hours a day. The number of calls received during any minute is an example of which random variable?

i) normal,
ii) binomial,
iii) uniform,
iv) Poisson,
v) none of the above.

(b) *(2 b)* If two events (both with probability greater than 0) are mutually exclusive, then:

i) they also must be independent,
ii) they also could be independent,
iii) they cannot be independent,
iv) none of the above.

(c) *(2 b)* Suppose that vehicle speeds at an interstate location have a normal distribution with a mean equal to 70 mph and standard deviation equal to 8 mph. What is the z-score for a speed of 64 mph?

i) -0.75,
ii) 0.75,
iii) -6,
iv) 6,
v) none of the above.

(d) *(2 b)* A box of 8 marbles has 4 red, 2 green, and 2 blue marbles. If you select two marbles without replacement, what is the probability that you select a red marble followed by a green marble?

i) 0.0625,
ii) 0.1250,
iii) 0.1429,
iv) 0.6,
v) none of the above.

(e) *(2 b)* 500 random samples of size n=900 are taken from a large population in which 10% are left-handed. The proportion of the sample that is left-handed is found for each sample and a histogram of these 500 proportions is drawn. Which interval covers the range into which about 68% of the values in the histogram will fall?

i) 0.1 ± 0.01,
ii) 0.1 ± 0.0134,
iii) 0.1 ± 0.0167,
iv) 0.1 ± 0.02,
v) none of the above.

Multivariate Regression. *(10 b)* You have a dataset for 1,000 individuals. For each person, a large number of variables such as height, weight, BMI, blood pressure, etc., have been measured. Your goal is to determine which of these variables influence the effectiveness of a drug given to the observed individuals in different doses. The amount of administered drug and its effectiveness are also recorded in the data. Additionally, you aim to predict the drug's effectiveness as accurately as possible for future patients. You will use linear regression with LASSO regularization; the choice of the method is thus part of the assignment.

(a) *(1 b)* Write the optimization criterion of the LASSO method. Explain its variables.

(b) *(2 b)* Explain why the LASSO method is suitable for selecting relevant variables. Why does it outperform, for example, selecting all significant β coefficients in ordinary linear regression? And what does it mean that it will overcome it?

(c) *(1 b)* Explain why the LASSO method (very likely) outperforms ordinary linear regression in terms of accuracy in predicting drug efficacy in future patients for this task.

(d) *(2 b)* The LASSO method requires feature preprocessing in this task. Explain why. Formalize this preprocessing mathematically.

(e) *(2 b)* You use the train/test/validate data distribution to verify the performance of the LASSO method as well as to determine its parameters. Describe in as much detail as possible what you will use each set for and why the procedure cannot be simplified.

(f) *(2 b)* Now focus on the feature preprocessing discussed above. How exactly do you implement it in the train/test/validate split? Explain.

Discriminant analysis. *(10 b)* You solve a binary classification problem in which the number of positive examples is 50 times smaller than the number of negative examples. You used linear discriminant analysis (LDA) and created a classifier. Your task is to evaluate how well your classifier separates the classes.

(a) *(1 b)* Define the confusion matrix used to represent the performance of the classifier. Name its individual items and explain their meaning.

(b) *(1 b)* How do you calculate the classification accuracy from the given matrix? What are the pitfalls of using classification accuracy in our task?

(c) *(2 b)* You have built a specific confusion matrix, your classifier does not classify any of the 1,000 test examples into the positive class. How could this happen? How do you adjust your model so that the number of positive predictions is non-zero? How will this change affect the classification accuracy?

(d) *(1 b)* What will you do if mislabeling a positive example as negative carries a double loss for you compared to mislabeling a negative example as positive?

(e) *(2 b)* Define the ROC curve based on the confusion matrix introduced earlier.

(f) *(1 b)* Is the ROC curve applicable in our unbalanced LDA task? Explain why.

(g) *(2 b)* Explain the term area under the ROC curve (AUC, AUROC). What does it define? What can this area be used for? Draw a picture with several different curves and explain what classifiers they characterize. Do not forget to name the axes.

Dimensionality reduction. *(10 b)* Consider the kernel PCA method.

(a) *(2 b)* What is a kernel function? What are its arguments and what is its output? How do we interpret it in kernel PCA?

(b) *(1 b)* What is a kernel matrix? What are its dimensions? Compared to PCA, which matrix of equal dimensions does it replace?

(c) *(2 b)* Give an example of two well-known kernel functions. Define them mathematically. Derive/estimate the dimension of the transformed space after applying these functions if the dimension of the input space is 2.

(d) *(2 b)* Compare the function of PCA and kernel PCA. When is kernel PCA more appropriate? What are its disadvantages compared to PCA?

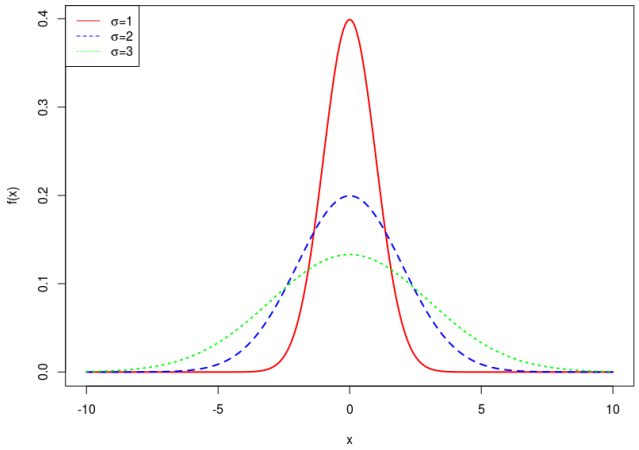
(e) *(2 b)* Compare PCA and kernel PCA method in terms of their computational complexity.

(f) *(1 b)* Compare kernel PCA with the Locally Linear Embedding (LLE) method. Is there any relationship between them?

Clustering. *(10 b)* You have a blood test result for six people. The test result is a real number, here specifically the results are: 10, 12, 13, 17, 18, 21. There are both healthy and sick people, but you do not know who is who. You only know that the test result has a normal distribution in the population of healthy and sick people. You also know that sick people have higher test values and there are half as many of them as healthy people. You want to decide on the statistical properties of the blood test and the division of people into groups using EM clustering.

(a) *(1 b)* Initialize the EM clustering algorithm by setting the initial assignment to the clusters. Work with the numbers from the assignment.

(b) *(3 b)* Perform the first step of the EM algorithm, due to initialization you start with an M step and the output will be specific model parameters. For an approximate estimate of probabilities you can use the normal distributions in the figure below. The calculation does not have to be detailed for all 6 samples ...



(c) *(3 b)* Perform the second step of the EM algorithm, it will be an E step and the output will be the probabilistic distribution of people into groups.

(d) *(3 b)* Describe graphically and verbally the expected result of the EM algorithm after running (what the model will look like, how the objects will be divided into clusters).