MAD 2024/2025 Exam

Bulat Ibragimov, Sune Darkner

Exam period: 13.01.2025 – 20.1.2025

This is the exam questions for the 8 day take-home exam on the course Modeling and Analysis of Data (MAD). The exam must be solved **individually**, i.e., you are **not allowed to work in teams or to discuss the exam questions with other students**. You have to submit your solution of the exam electronically via the **Digital Exam**[1] system. The submission deadline is **20 January 2025 at 10:00**.
Your solution should consist of

1. a pdf file `answers.pdf` containing your answers, and
2. a `code.zip` file containing the associated code (python files and / or Python Jupyter notebooks).

The grading will be done anonymously; **hence you should not mention your name anywhere in the answers or code**. Instead, you must **add your exam number** at the beginning of your `answers.pdf`.

WARNING: The goal of this exam is to evaluate your individual skills. We have to report any suspicion of cheating, in particular collaboration with other students, to the head of studies. Note that, if proven guilty, you may be expelled from the university. Do not put yourself and your fellow students at risk.

You are allowed to ask questions via the Discussions board in Absalon, but make sure that you do not reveal any significant parts of the solution. In doubt, just contact Bulat or Sune directly via e-mail! Any additional hints given by us will be made available to all of you via Absalon. Some further comments:

1. You **are allowed** to reuse the code that was made available to you via Absalon as well as the code you have developed in the course of the assignments. If you reuse code from the lectures or from the assignments, make sure to put a reference to this in your code, and if your code was developed as part of an assignment, in collaboration with a fellow student, add a corresponding comment to your answers, although keeping anonymity (i.e., just mention which parts stem from team work). In case you reuse code snippets you have found on the internet, please make sure that you provide a reference to this external source as well.
2. In case you notice any inaccuracies in the problem descriptions below, please let us know. If needed, we will provide updates and additional comments via Absalon. Thus, make sure that you check Absalon for announcements and discussions regularly!
3. For the coding tasks, you are given Python files/Jupyter notebook templates. You are supposed to complete these files and notebooks, but you are allowed to convert a Jupyter notebook to a python file or the opposite. Note that you are allowed to import additional Python packages and to make use of the functions provided by, e.g., the Numpy package. However, you should not use built-in functions if we ask you to implement a specific algorithm without using existing implementations (e.g., a single `kmeans` function from some other package that implements the K-means clustering approach). If in doubt, please ask us via the discussion board in Absalon!
4. All code templates and data can be found in the files `code.zip` and `data.zip`.
5. The deadline is hard (late submissions are not allowed), so make sure to submit in good time before the deadline. Up until the deadline it is possible to upload new versions of your solution several times. In the unlikely event that the Digital Exam system fails when you submit just at the deadline, then immediately send an e-mail to `uddannelse@diku.dk` and `bulat@di.ku.dk` with an explanation of what happened and attach your solution (the pdf and zip files) to the exam. Your solution will be assessed if you have a valid excuse and submitted on time.
6. The use of AI is not allowed!
7. Good luck! :-)

[1] `https://eksamen.ku.dk/`

Statistics

In this part, we will test your knowledge and skills in performing statistical analysis.

Question 1 (**Maximum Likelihood Estimation, 2 points).** Background: An exponential distribution can be used to model the time between events in a Poisson process with a probability density function (pdf) given by:

$$f(x;\lambda) = \lambda e^{-\lambda x}$$

for $x \geq 0$, where $\lambda > 0$ is the rate parameter. The mean of the distribution is $\frac{1}{\lambda}$. For the following steps provide a clear, step-by-step report of your analysis, including all mathematical derivations, calculations, and interpretations. Use appropriate software or programming languages for calculations and visualizations if needed. Tasks:

1. Derive the Likelihood Function Given a set of independent and identically distributed (i.i.d.) random variables $X_1, X_2, \ldots, X_n$ drawn from an exponential distribution, derive the likelihood function $L(\lambda)$.

2. Find the MLE: Show how to find the maximum likelihood estimator $\hat{\lambda}$ of the parameter λ by taking the natural logarithm of the likelihood function and finding its derivative.

3. Apply MLE to Sample Data: Use the derived estimator to analyze a given dataset. The dataset contains times (in minutes) between arrivals at a service point:

$$\text{Data: } 2.1, 3.4, 1.8, 4.2, 2.9, 3.1, 5.0, 1.5, 4.3, 2.7$$

4. Estimate the Rate Parameter: - Calculate the estimated rate parameter $\hat{\lambda}$ using the MLE derived in Task 2.

5. Interpret the Results: Compute the mean time between arrivals using the estimated $\hat{\lambda}$.

Deliverables. A detailed report with your derivations, calculations, and interpretations. A script or code file if you've used software for computations.

Question 2 (**Hypothesis testing, 2 points).** 1. Formulate the Hypotheses: State the null hypothesis (H_0) and the alternative hypothesis (H_a). The test should determine whether the mean of the dataset is different from the known population mean. Select an appropriate significance level (α), typically 0.05 or 0.01, for the test.

Test and Distribution: Assume that the data is normally distributed without a known population standard deviation.

2. Calculate the Test Statistic: Compute the sample mean and sample standard deviation for this dataset:

$$\text{Data: } 25.228.731.120.527.932.921.634.824.128.0$$

Calculate the test statistic for the given sample data. Assume $\mu_0 = 30$.

3. Determine the Critical Value: Calculate the critical value(s) for the chosen significance level and justify whether it should be a one-tailed or two-tailed test. Compare the calculated test statistic to the critical value and make a decision regarding the null hypothesis. Interpret the results.

Deliverables. A detailed report with your derivations, calculations, and interpretations. A script or code file if you've used software for computations.

Principal component analysis

Question 3 (**Principal Component Analysis, 4 points).** The aim of this task is to test your understanding of the principal component analysis algorithm. You database is the information about five patients who were treated with drugs A and B. Patients 1 and 3 got drug A, patients 2, 4 got drug B, while patient 5 got both drugs. The survival of the patients was of 0.6 years for patient1, 1.1 years for patient2, 3 years for patient3, 1 year for patient4, and 0.2 years for patient5. Your task is to compute the principal component decomposition for this database. You can use either N or $N-1$ in the denominator during calculation of the covariance matrix.

The second task is to reflect on a slightly modified database. Suppose patient5 would be treated with drug A only. How will this affect PCA analysis?

Deliverables. Step-by-step calculation of the principal components. Your report should include mathematical derivations of the the eigenvalues, eigenvectors and all the intermediate steps that are followed during the principal component analysis.

Classification

In this part, we will test your knowledge and skills in understanding classification and regression topics.

Question 4 (**Classification model selection, 3 points).** Please take a look at these table of patients:

Patient ID	Diagnosis	AFP (ng/mL)	Albumin (g/dL)	Bilirubin (mg/dL)	Age (years)
P1	Healthy	3	4.5	0.8	75
P2	Healthy	5	4.8	0.9	59
P3	Healthy	7	4.2	1.0	83
P4	Healthy	2	4.6	0.7	58
P5	Metastasis	1034	3.2	2.5	60
P6	Metastasis	912	2.8	3.1	65
P7	Metastasis	1579	3.0	2.8	70
P8	HCC	1271	2.5	4.0	88
P9	HCC	2532	2.6	3.5	72
P10	HCC	2198	2.4	4.2	75

Suppose your aim is to predict survival of these patients using the shown above features. Please check MAD lectures and think what kind of regression algorithm can be used here. What kind of issues there are in the data, how can they affect different algorithms, and how to address them.

Deliverables. Explain what kind of algorithm can be used here and why. Explain what kind of issues in the features needs to be addressed and taken into account when selecting the algorithm.

Question 5 (**Random Forests, 7 points).** In this question, you need to implement different components of decision tree in python, and test their performance on `heart_simplified_RandomForest.csv`. You need to implement it yourself without using existing machine learning packages. The following things will be needed in your implementation:

a) A function that evaluates the quality of a threshold for a selected feature. Use either information gain based on entropy or Gini coefficient.
b) A function that identifies possible thresholds for a selected feature and finds the optimal threshold using the the function from a).
c) Find the optimal feature to perform separation using the optimal thresholds for all features from c).

Deliverables. First, don't forget to add appropriate code snippets here with the functions and explanations what they do and why.

To demonstrate that your implementations work, you need to test them on the `heart_simplified_RandomForest.csv` data. The file has 30 data samples with 4 features: `Age, RestingBP, Cholesterol, MaxHR`, and one label `HeartDisease`.

To test function a), you need to print the quality metric you selected using the mean value of each feature. In other words, you need to have four print statements saying something like `For feature "Age", and the threshold <the average value of Age in the database>, <the metric you selected> = ...`. You do the evaluation on the complete database.

To test function b), you can use feature `Age` and the complete database. Print all the age thresholds that need to be tested (don't forget to explain how you decide this). Print the function a) evaluations for these thresholds. Print the statement that indicates which one is the optimal.

To test function c), you need to test function b) on all features `Age, RestingBP, Cholesterol, MaxHR`. Print the optimal thresholds and function a) values for each feature. Print the optimal one and the best feature to choose for the separation.

Clustering

In this part, we will test your knowledge and skills in performing clustering of data.

Question 6 (**6 points).** This task aims to test your understanding of clustering. The task includes the following:

a) (1 point) Read and normalize data from the comma-separated data file `housing.csv`. The dataset consist of 20640 samples. Each line in the file represent one sample. Each sample is defined with 9 features, however you need to work only with four features **MedInc, Latitude, Longitude, MedHouseVal**. Normalize the data using the mean and standard deviation values computed for each feature. That is, for the i'th sample, the normalized **Latitude** feature value is $Latitude_i^{norm} = (Latitude_i - \bar{Latitude})/(\sigma_{Latitude})$,

where $Latitude_i$ is the original **Latitude** feature value for the i'th sample, $\bar{Latitude}$ is the mean value of feature **Latitude** computed from all samples, and $\sigma_{Latitude}$ is the standard deviation of feature **Latitude** computed from all samples.

b) (4 points) The idea is to compare and analyze linkage strategies for the distance metric between clusters. There are three options to compare: **complete** or **maximum**, **average**, **single** or **minimum**. Run agglomerative clustering using different linkage strategies. Analyze the properties of the resulting clusters when different metrics are used. Compare the intra-, inter-cluster distances and average cluster size and reflect on the resulting observations. Do they agree with the properties we discuss on the lectures? You are allowed to use the existing implementation of agglomerative clustering from **sklearn**.

c) (1 points) Visualize the results for different clustering strategies against two dimensions **Latitude** and **Longitude**. For visualization use 5 clusters from the agglomerative clustering. Color the clusters and show their centers.

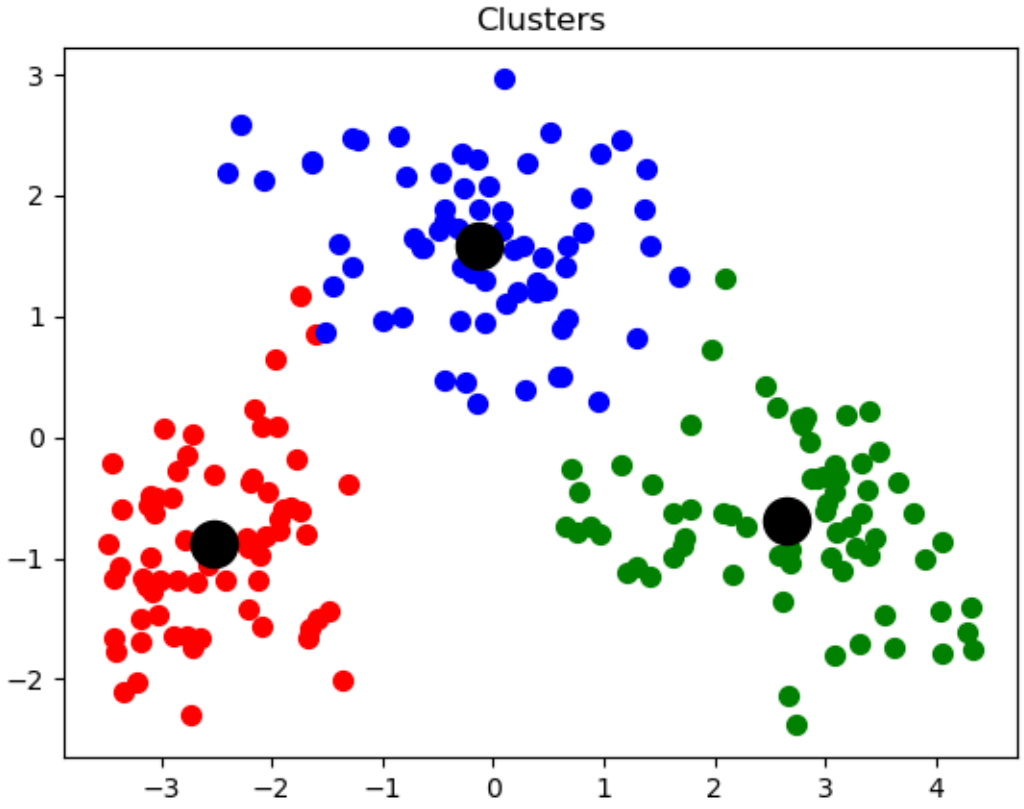

Figure 1: Example of plot we ask for in clustering question. The black dots represents the cluster centers.

Deliverables. a) Provide a code snippet in the report showing how you perform the normalization, b) provide a code snippet in the report showing agglomerative clustering and your analysis of different cluster linkage strategies, e) include a code snippet and the plot. Provide reflection on the results