Shahid Beheshti University
Machine Learning
Spring 2025

Assignment 1

Deadline: 1403.12.24 23:59

1 Theoretical Questions

Exercise 1

Compare t-test and z-test in terms of assumptions, population standard deviation, and use case.

Exercise 2

A tech company is deploying a recommendation system for a music streaming platform. Suppose user ratings for songs (on a scale from 1 to 5) follow a non-normal distribution. If we take a large enough sample, how does CLT justify using a normal approximation for constructing confidence intervals? If the mean predicted rating for a song is 4.2 with a standard deviation of 0.5, how can we use CLT to determine the probability that a randomly selected user will rate the song above 4.5?

Exercise 3

You are given a dataset containing information about customers in an online retail store. The dataset includes the following features:

- **Age** (continuous variable)
- **Annual Income (in USD)** (continuous variable)
- **Customer Satisfaction Score** (ranked from 1 to 10)

- **Preferred Payment Method** (categorical: 'Credit Card', 'PayPal', 'Bank Transfer')

Below is a sample of the dataset with five customers:

Customer	Age	Annual Income (USD)	Customer Satisfaction Score	Preferred Payment Method
1	25	40,000	8	Credit Card
2	32	55,000	6	PayPal
3	40	65,000	7	Bank Transfer
4	50	70,000	5	Credit Card
5	60	85,000	3	PayPal

For each of the following scenarios, determine the most appropriate statistical test from the options: **Pearson's Correlation, Spearman's Rank Correlation, or Chi-Square Test**. Then, apply the test using the dataset provided and interpret the result.

Scenario (a):

The store owner wants to examine whether there is a linear relationship between Age and Annual Income of customers. Which statistical test should be applied? Compute the test statistic and interpret the result.

Scenario (b):

The store owner wants to check whether there is a monotonic relationship between Age and Customer Satisfaction Score. Which statistical test should be applied? Compute the test statistic and interpret the result.

Exercise 4

What is the key difference between the Mann-Whitney U test and the Wilcoxon Signed-Rank test? Suppose you are testing two marketing strategies' effectiveness, but the data is non-normally distributed. Which test would you use?

Exercise 5

We have collected a dataset with three numerical features and one categorical feature called `group`. The `group` feature represents three different experimental conditions (1, 2, or 3). Our goal is to determine whether the groups significantly differ given the numerical feature `X`.

Group	Feature X	Feature Y	Feature Z
1	10.2	7.8	5.4
1	10.8	8.1	5.2
1	11.0	7.5	5.8
2	18.5	6.9	6.2
2	17.9	7.3	6.0
2	18.2	6.8	5.9
3	30.1	6.2	7.0
3	29.8	6.5	7.1
3	30.3	6.0	6.9

a We want to see whether the groups differ given this feature X or not. Now, we should choose between these two tests (ANOVA or Kruskal-Wallis). In your opinion, what should we check (test) to help us choose between these two tests? And once you have made your choice, apply your test. (Consider P-Value less than 0.05.)

b Based on the result of (a), apply the appropriate Statistical Test (ANOVA or Kruskal-Wallis) to determine if there is a significant difference between the groups. Which one would you use and why?

c Assume the normality assumption holds. If you wanted to check whether the groups also significantlly differ given all three features (X, Y and Z), what statistical test could have been used?

d If features X, Y, and Z were the contributing factors in predicting a data sample's group, how could we have used ANOVA or Kruskal-wallis for feature selection?

2 Practical Questions

The purpose of this assignment is for you to practice basic data pre-processing skills and apply different statistical tests. You will be given two datasets: the Sleep Health and Lifestyle Dataset [Link] and the Student Performance Factors dataset [Link].

Sleep Health and Lifestyle Dataset

The Sleep Health and Lifestyle Dataset comprises 400 rows and 13 columns, covering a wide range of variables related to sleep and daily habits. It includes details such as gender, age, occupation, sleep duration, quality of sleep, physical activity level, stress levels, BMI category, blood pressure, heart rate, daily steps, and the presence or absence of sleep disorders.

1. **Exploratory Data Analysis (EDA):**
   You are required to apply basic pre-processing steps on the data and conduct a thorough investigation. This could include:
   - Getting familiar with the dataset. How many features does it have? How many data samples? What are the values for each feature? What is the data type of each feature? Etc.
   - Check whether the dataset contains any invalid or missing values, consider ways to handle them, and explain the reason behind your choices in each step.
   - Further investigate the data using appropriate visualization techniques. You could refer to this source for choosing the best plot for your problem.
2. **Hypothesis Testing:**
   a. Does women's sleep duration follow a normal distribution?
   b. Is having higher daily steps a contributing factor into better sleep? Check the corresponding correlation of Daily Steps and Quality of Sleep.

c. Is stress level different among different occupations? First, check this hypothesis with a test, and then compute the average stress level among different occupations. Use a bar chart or any other desired visualization method to demonstrate the result.

d. Are different BMI categories significantly different given their blood pressure? (Hint: Convert blood pressure into two columns and apply your test given these new two features.)

e. Do people with sleep disorders have higher heart rates than those without any sleep disorder?

3. **Bonus:**
Proposing new hypotheses and applying tests would provide bonus points.

Student Performance Factors Dataset

This dataset provides a comprehensive overview of various factors affecting student performance in exams. It includes information on study habits, attendance, parental involvement, and other aspects influencing academic success.

1. **Exploratory Data Analysis (EDA):** It is mandatory for you to apply EDA and any necessary pre-processing steps as explained in the sleep dataset.
2. **Hypothesis Testing:** Propose at least 5 hypothesis tests about the things you are mostly curious about in this dataset. Try to have a variation of hypotheses requiring different kinds of tests. Then, apply the appropriate test for them.

3 Submission Format

- Your final output should be a Jupyter notebook file containing all your codes and a report (PDF) explaining the overall result of your EDA investigation, your insights into the dataset, and the output of your tests with thorough explanation.
- In general, your submission has to be comprehensive and coherent, showcasing a reasonable approach in your workflow.

Note:

- The purpose of this assignment is mostly to evaluate your insights and analysis skills. Hence, any advanced and creative work would provide bonus points, such as high-quality visual demonstrations of the above hypotheses.
- Since this is your first assignment, it is recommended for you to look through others' submissions for inspiration. However, copying and blind usage of AI is not the goal and will not be accepted.