Biomedical Data Science & AI

Exercise sheet 2

SoSe2025

Prof. Dr. Holger Fröhlich, Mohamed Aborageh, Jonas Botz, Diego Valderrama

Due date: Apr 22th

Questions

Exercise 1 - Correlation (total: 4 points)

Load the dataset from the given *dataset.csv* file.

1. Which correlation measure would suit the comparison of feature_1 and feature_3? Calculate the relevant correlation coefficient for the 2 features. (2 points)
2. Plot the correlation heatmap of the entire dataset. (2 points)

Exercise 2 - Probability (total: 14 points)

Assume you have an unfair n-sided dice, which is thrown once. The probability to observe the number i, $i = \{1, \dots, n\}$ is $P_n = \frac{i}{M_n}$ with $M_n = \sum_{k=1}^{n} k$.

Hint 1: M_n is called *arithmetic series*.

Hint 2: Example function to compute expectation value for a fair n-sided dice.

1. Show, that P_n is a probability mass function **(2 points)**
2. Expectation value
   a. Work out the formula to calculate $E_{P_n}[X] = \sum_{i=1}^{n} x_i P_n(x_i)$ **(1 point)***
   b. Compute $E_{P_n}[X]$ for $n = 6$. Show the calculation steps and result. **(2 points)***
   c. Same as (b) for $n = 100$. Write a python function, which takes n as argument and outputs the result. **(3 points)**
3. Variance
   a. Work out the formula to calculate $Var_{P_n}[X] = E_{P_n}[X^2] - (E_{P_n}[X])^2$ **(1 point)***
   b. Compute $Var_{P_n}[X]$ for $n = 6$. Show the calculation steps and result. **(2 points)***
   c. Same as (b) for $n = 100$. Write a python function, which takes n as argument and outputs the result. **(3 points)**

* Pen and paper! Share photos.

Exercise 3 - Understanding Your Dataset (total: 7 points)

Load the *processedClevelandData.csv* dataset. The features for the dataset are described in the *featureDescription.csv* file. Perform data cleaning procedures such that your final dataset is usable in the following questions.

1. Which features are discrete? Which are continuous? Is there any feature which can be described as both? If so, which one and in which way? **(1 point)**
2. Use Spearman's and Kendall correlation measure to quantify the correlation between age and the following.
    a. Resting blood pressure
    b. Serum cholesterol level
    c. Maximum heart rate achieved
    d. Also, which variable(s) are most correlated with age? Illustrate with heatmaps. **(2 points)**
3. Plot the relative frequency of the “Sex” variable in the dataset and describe what you observe in the plot. Similarly plot and describe the 'ca' feature for the male participants. **(2 points)**
4. Detect outlier patients for features "trestbps" and "chol". Illustrate with plots. **(2 points)**