Biomedical Data Science & AI

Exercise sheet 4 - Introduction

SoSe2021

Prof. Dr. Holger Fröhlich, Mohamed Aborageh, Vinay Bharadhwaj, Yasamin Salimi

Due date: May 18th

Questions

Exercise 1 - ANOVA F-test and Hierarchical Clustering (8 points)

Load the leukemia dataset. It contains gene expression data of 1397 genes from 38 tumor mRNA samples. The expression data is organized in a matrix where rows correspond to genes and columns to samples. The tumor class of the columns is given in the file “golub.cl”.

1. ANOVA F-test
    a. What are the assumptions of the ANOVA F-test? **(1 point)**
    b. For each gene in the dataset, perform the ANOVA F-test (assumptions are already met) to see whether the gene is significantly differentially expressed between the two types of Leukemia. **(1 point)**
    c. Due to our analysis, we now know which genes are significantly differentially expressed between groups. These will be the best features to use in order to get good cluster separation. Subset only the rows which represent the top 100 most significant genes. **(1 point)**
2. Plot 2 dendrograms using the 100 selected genes:
    a. One for a single linkage approach and another one for ward approach. **(1 point)**
    b. Which method would you recommend based on the dendrograms for a clustering? Why? **(1 point)**
    c. Familiarize yourself with Cophenetic correlation coefficient and calculate the cophenetic correlation distance for both single linkage as well as ward. **(1 point)**
    d. Based on the cophenetic correlation distance, which clustering method performed better? **(1 point)**
3. Apply two Agglomerative Clustering.
    a. One using single linkage and one using ward method. **(1 point)**

Exercise 2 - PCA (8 points)

Using the same leukemia dataset generate the feature matrix (transposed leukemia dataset) and the class labels (*golub.cl.csv*).

1. Perform a PCA on the feature matrix and answer the following,
    a. How many PC's do you need to explain at least 95% of the variance? **(1 point)**
    b. Make a scatterplot of the projections on the first two PC's with the colouring corresponding to the class labels. **(2 points)**
    c. Based on the scatterplot answer the following questions **(2 point)**
        i. Given the plot, do you think PCA might be a good choice? Why?
        ii. Do you think n=2 components are a good choice? Why?
2. Inform yourself regarding decorrelation of features in a dataset
    a. Identify the correlated features in the dataset **(1 point)**
    b. Decorrelate the correlated datasets **(1 point)**
    c. What is the purpose of carrying out decorrelation of features in a dataset? **(1 point)**

Exercise 3 - Logistic Regression (9 points)

1. Using the reduced dataset from exercise 2.1, carry out the following tasks:
    a. Generate a logistic regression model on the first 5 PCs of the reduced dataset using 80% of the total samples **(2 points)**
    b. Predict the labels for the remaining 20% of the samples and calculate your model's accuracy **(2 points)**
2. Inform yourself about Brier's Score. How can it be used to evaluate the performance of your model? Show by implementation. **(2 points)**
3. Assess the significance of your variables using the likelihood ratio test. **(3 points)**