CM 3430 Computational Statistics

Take-Home Assignment

Visualization and Classification Analysis of the Heart Disease Dataset

This assignment is based on Chapter 1: Visualization of Multivariate Data and Chapter 7: Permutation Tests, along with the associated lab sessions. You will explore data visualization techniques, build classification models, and validate their performance using permutation tests..

Assignment guidelines

Dataset

Consider the Heart Disease Dataset (heart.csv). This dataset contains the following attributes:

Attribute Information:

- age
- sex
- chest pain type (4 values)
- resting blood pressure
- serum cholestoral in mg/dl
- fasting blood sugar > 120 mg/dl ((1 = fasting blood sugar > 120 mg/dl, 0 = otherwise)
- resting electrocardiographic results (values 0,1,2)
- maximum heart rate achieved
- exercise induced angina
- oldpeak = ST depression induced by exercise relative to rest
- the slope of the peak exercise ST segment
- number of major vessels (0-3) colored by flourosopy
- thal: 0 = normal; 1 = fixed defect; 2 = reversable defect
- target ; 0 = no heart disease, 1: heart disease

Tasks:

1. Create three visual representations of the dataset using the techniques covered in Chapter 1.

- Use appropriate visualizations (e.g., scatter plots, box plots, or heatmaps, parallel coordinate plot) to explore relationships in the data.
- Provide a detailed interpretation of each visualization.

2. Divide the dataset into training (80%) and testing (20%) subsets.

- Ensure the split is random and stratified based on the target variable to maintain class balance.

3. Build two classification models (e.g., Logistic Regression, Random Forest, or any suitable model).
4. Calculate and report the training accuracy and testing accuracy for both models.
5. Provide a brief discussion on the performance of each model.
6. Calculate the absolute difference in test set accuracy between Model A and Model B using the original test labels. Write a conclusion comparing the models based on this difference.
7. Validate the test set performance difference using a permutation test as covered in Chapter 7.

- Report the distribution of the accuracy differences obtained through permutations.
- Conclude whether the observed difference between Model A and Model B is statistically significant.

8. **Submit a Python notebook containing both the code and the resulting outputs**

Deadline: 3 January 2024

A penalty will be applied for late submission of assignments (0.5 mark for each extra day).