Data Science Techniques and Applications
Coursework I

Baran Buluttekin
13153116

March 9, 2019

For this coursework, I choose to examine Heart Disease UCI dataset [1] in kaggle [3]. It comprises of data that collected from patients from 4 medical institutions. There is no timestamp associated with data collections but according to UCI website where the kaggle dataset originally acquired, data donated at 1988. Its primarily used for classification machine learning tasks.

Dataset have total of 303 observations (rows) and 14 columns (features). There are 8 categorical columns present at the data but these variables represented in integers assigned to them. For example sex variable denotes male patients with 1 female patients with 0. Below is list of complete feature attributes[1]:

1. age: Age in years (max:77, min:29)
2. sex: Gender of the patient (0:Female, 1:Male)
3. cp: Type of chest pain
   (a) 1: indicate typical angina
   (b) 2: indicate a typical angina
   (c) 3: indicate non-anginal pain
   (d) 4: indicate asymptomatic pain
4. trestbps: Blood pressure recording for resting patient (measured in mm Hg)
5. chol: cholesterol level (mg/dl)
6. fbs: Blood sugar level (fasting) > 120 mg/dl (1:true, 0:false)
7. restecg: Category for electrocardiographic, measured while patient resting

   (a) 0: indicate normal
   (b) 1: indicate abnormal ST-T wave mea sure (T wave inversions and/or ST elevation or depression of > 0.05 mV)
   (c) 2: Estes criteria could be definite or probable ventricular hypertrophy
8. thalach: Highest level of heart rate measurement
9. exang: (0:no, 1:yes) Exercise induced angina
10. oldpeak: Measurement of ST depression
11. slope: St segment slope value recorded at peak excersice
   (a) 1: indicate up-sloping
   (b) 2: indicate constant
   (c) 3: indicate down-sloping
12. ca: Fluoroscopy coloring have 4 category
13. thal: Defect levels
14. target: Outcome (1:Heart disease, 0:Normal)

For the PCA consideration most important dimensions are, age, chol (cholestoral) and thalach (maximum heart rate) because these dimensions have the highest variance which will be important for PCA. There is no indication of data integrity issues. For example age variable have values ranging from 29 to 77 which is expected and contains no abnormal values such as 0 or 250. Similar observation can be obtained from max heart rate measurement that is also have expected values according to general guidence [2].

References

[1] Janosi et al. *UCI Machine Learning Repository*. 1988. URL: `https://archive.ics.uci.edu/ml/datasets/Heart+Disease`.

[2] *Exercise intensity: How to measure it*. URL: `https://www.mayoclinic.org/healthy-lifestyle/fitness/in-depth/exercise-intensity/art-20046887`.

[3] *Heart Disease UCI*. URL: `https://www.kaggle.com/ronitf/heart-disease-uci`.