CCE – CEBD1260
Big Data Analytics

Data Analytics on UCI Heart Disease Dataset

Hatem M.T. BEN AMOR

June 2018

Outline

- Dataset Overview
- Objectives
- Functional Map
- Wrangling & EDA
- Regression
- Classification
- Clustering
- Data App

Dataset Overview

UCI machine learning archive
(http://archive.ics.uci.edu/ml/datasets/Heart+Disease).
4 different data sets coming from the following sources:

Hungarian Institute of Cardiology. Budapest
University Hospital, Zurich, Switzerland
University Hospital, Basel, Switzerland
V.A. Medical Center, Long Beach and Cleveland Clinic Foundation

More details about data sources are below.

Database	:	0	1	2	3	4	Total
Cleveland	:	164	55	36	35	13	303
Hungarian	:	188	37	26	28	15	294
Switzerland	:	8	48	32	30	5	123
Long Beach VA	:	51	56	41	42	10	200

Overview: Cleveland dataset
Originally have **76** attributes (variables) but only **14** were used in research works.

Attr01. #3 (age)
Attr02. #4 (sex)
Attr03. #9 (cp: chest pain type)
Attr04. #10 (trestbps: restin blood pressure)
Attr05. #12 (chol: cholesterol)
Attr06. #16 (fbs: fasting blood sugar > 120 mg)
Attr07. #19 (restecg: resting electrocardiographic)
Attr08. #32 (thalach: maximum heart rate achieved)
Attr09. #38 (exang: exercised induced angina)
Attr10. #40 (oldpeak: ST depression induced by exercise relative to rest)
Attr11. #41 (slope: slope of the peak exercise segment)
Attr12. #44 (ca: number of major vessels coloured by fluorosopy)
Attr13. #51 (thal: 3=normal, 6=fixed defect, 7=reversible defect)
Attr14. #58 (num)(goal: diagnosis of heart disease

Objectives

- Perform various Big Data Analytics techniques
- Predict the presence of the disease:

  50%+ narrowing of a major vessel

Functional Map

- TCS …

Wrangling and EDA

- Renamed variables: No header
- ‘goal’: values > 1 => Set to 1
- ‘ca’ and ‘thal: values ‘?’ => Set to “no issue” values.
- Transformed categorical values to make more sense
- Histograms for numeric, Bars for categorical, Covariance, etc.
- All numerical variables could not be eliminated.
- Feel the categorical variables have strong dependencies
- ☹ Could not perform multivariable analysis on categorical
- Idea: Use Apriori/FP-Growth to eliminate some => TBC

Regression

- Used all models seen in class (Thanks Ary ☺)
- Substantially different results between models!
- ‘goal’ is binary: Rounded y_pred! => Made more sense
- Linear Regression had best error => Results kind of ‘weird’
- 5 main variables appeared to be dominant in all models
- ‘age’ is not among them!

Classification

- Used all models seen in class (Thanks Ary ☺)
- Les discrepancy between models!
- ‘goal’ is binary: Rounded y_pred! => Made more sense;

  Higher precision
- RF 100 => Best results, then Naïve Bayes
- 6 main variables appeared to be dominant in all models.
- ‘age’ is there ☺

Clustering

- Hierarchical algorithm
- ‘age’ w.r.t. to others
- Several trials: k =10, X_columns reduced
- Changed ‘age’ to have 30-39 in the same category, etc.
- Finally: k=5 + the 6 main variables from previous analysis
- Sklearn model would provide more insight: TBC.
- Results make sense w.r.t to ‘age’ vs health.

Data App

- Predict ‘goal’ based on the various variables
- Should use a classification model because better results
- May use a regression model or more than one.
- Not completed yet ☹

Thank you ☺

Questions/Comments?