CSCI 347: Project 4

Michael Roduin, Moiyad Alfawwar, Philip Gehde

May 11, 2022

1 Project Plan

In this project we are investigating the correlation between various risk factors such as cholesterol, high-blood pressure, etc. and heart disease. The multivariate heart Disease Data Set we are using contains 270 instances, with 13 attributes that give us insight into the observed patients' biology and lifestyle. The Dataset is said to include Categorical values, Integer, and Real numbers. Some values are missing, so the data will need to be cleaned.

This dataset has been cited in several research papers in the data science field including: Diversity in Neural Network Ensembles (Gavin Brown. The University of Birmingham. 2004.), Overcoming the Myopia of Inductive Learning Algorithms with RELIEFF (Igor Kononenko and Edvard Simec and Marko Robnik-Sikonja), Unanimous Voting using Support Vector Machines (Elena Smirnova and Ida G. Sprinkhuizen-Kuyper and I. Nalbantis and b. ERIM and Universiteit Rotterdam, IKAT, Universiteit Maastricht), Dissertation Towards Understanding Stacking Studies of a General Ensemble Learning Scheme ausgeführt zum Zwecke der Erlangung des akademischen Grades eines Doktors der technischen Naturwissenschaften.

The dataset was uploaded by the University of California, Irvine, and is available at the UCI machine learning archive on their website for the Center for Machine Learning and Intelligent Systems at the following URL: https://archive.ics.uci.edu/ml/datasets/statlog+(heart)

Our data set has already been processed and includes no missing values, so there is no need for a plot to summarize the proportion of missing data as this is non-applicable for all 13 attributes. The 13 attributes (which have been extracted from a larger set of 75) include descriptive variables such as sex, and Chest pain type (4 values), however, these values were already label encoded. As such, we are left with the following attribute types: Real: 1,4,5,8,10,12 Ordered:11,Binary: 2,6,9 Nominal:7,3,13

Given that the categorical values were already label encoded, the choice (label-encoded vs one hot-encoded) was made for us and we assume that the alphabetical ordering of label encoding will not prevent us from making medically relevant inferences from this data. In other words, we assume that the categorical value was ordered alphabetically as to represent the severity of the pain, for example, A-D. This assumption may give us trouble down the road, and should be further investigated.

If we were to work with categorical values for chest pain, one might suggest one-hot encoding to prevent any issues that may arise if there is no obvious ordering, or ranking of our values, and rather solve this potential problem by representing each category as a binary vector. However, in order to avoid the pitfalls of multicollinearity, it would be best to simply determine that categorical data is ranked appropriately and use label-encoding instead. Sex/Gender is binary, and so labelencoding

can be considered appropriate.

We will be using k-means clustering to help identify variables that are most correlated to heart disease, and use dimensionality reduction across all attributes to determine the correlation between respective attributes and k-means.These techniques belong to the class of unsupervised learning and represent highly relevant skills in the toolbelt of every data miner, allowing us to demonstrate our progress throughout the course. At this point we are more concerned with descriptive analysis using the tools that we learned in data mining.

If we were to continue our research into this data set we might concern ourselves with more predictive analysis using tools like classification and regression models.

2 Implimentation

[ ]:
```
# Libraries
import numpy as np
import math
import matplotlib.pyplot as plt
from sklearn.datasets import make_blobs
from sklearn.cluster import DBSCAN
from sklearn.decomposition import PCA
from sklearn.datasets import load_boston
import warnings
from sklearn.cluster import KMeans
import pandas as pd
import random
import networkx as nx
from scipy import *
```

[ ]:
```
# space for implemnetation code from previous projects code
```

3 Report

3.1 What problem were you trying to solve or help solve?

3.2 Describe the data:

3.3 What pre-processing techniques did you apply and when? Make sure to justify the use of each technique you used. For example label vs. one-hot encoding.

3.4 What data mining techniques did you apply and why? Mkae sure to justify the use of each technique you used. For example, why did you use k-means instead of DBSCAN.

3.5 Include relevant visualizations and tables summarizing your data and your findings. This may include:

- a table listing the number attributes, missing values, number of classes, parameter set-tings, etc.
- visualization of a large graph if you are working with graph data.
- one or more visualizations of your data in two dimensions (original dimensions or PCA dimensions).
- for PCA, a plot of r vs. f(r).
- for k-means, a plot of the objective function for various k's.
- for DBSCAN, a plot or table of the precision at various parameters.
- other visualizations or tables that you think will effectively communicate your ideas.

[ ]:
```
# code for the report
```

3.6 What did you learn through your analysis?

3.7 Was anything about your results surprising or unexpected?

3.8 How will your work help with understanding the problem you set out to solve?

3.9 What else would you do if you had more time?

4 Present

Make a 5-10 minute video presentation summarizing your findings. You may use whatever video editing technology you prefer. (The MSU supported tool is TechSmith Relay. See the UIT tutorial for more info.) The video should:

- State your name.
- Summarize your project, including:

– the problem you are interested in.
– what data mining techniques you used to analyze data related to the problem.

- Your key findings and any surprising results.
- What else you would work on if you had more time.

The goal is to summarize the work you have done and what you have learned from the process. Note: any presentation that exceeds 10 minutes or does not reach 5 minutes will be docked 1 point per minute.