04250 - Introduction to Machine Learning and Data Mining

Project 1, E17

February 28, 2017

This report is written in collaboration by the following students:

Christian Kirstein Thygesen, s123871

Henriette Steenhoff, s134869

DTU Compute
Institut for Matematik og Computer Science

Contents

1 Description of Data

We have chosen the South African Heart Disease dataset[1] as our dataset to analyse throughout the course projects. The data is distributed freely by Stanford University for courses in Statistical Learning and consists of a retrospective sampling of males in a high-risk heart-disease region of Western Cape, South Africa. The data is part of a larger dataset handled by Rousseau et Al.[2] and contains males both with and without coronary heart disease (CHD). According to WHO[1][3], 5.25% af all deaths in South Africa is due to CHD, which makes it one of the top 20 causes of death in South Africa.

> *Coronary heart disease (CHD)* is a disease in which a waxy substance called plaque builds up inside the coronary arteries. Hardened plaque narrows the coronary arteries and reduces the flow of oxygen-rich blood to the heart. If the plaque ruptures, a blood clot can form on its surface. A large blood clot can mostly or completely block blood flow through a coronary artery. Over time, ruptured plaque also hardens and narrows the coronary Over time, CHD can weaken the heart muscle and lead to heart failure and arrhythmias arteries[4].

Looking into the study by Rosseau et al. They used the data to analyse the globin genes in a group of inherited blood disorders, *beta-thalassaemia major*. They wanted to extend their knowledge about the incidence and variety of the β-thalassaemia alleles (variants) and to investigate the clinical potential. The article states that it should now be possible to offer prenatal diagnosis for South African couples in risk of producing thalassaemic offspring provided that their linkage of β-thalassemia alleles to an RFLP test for partial gene deletion can be demonstrated.

The problem of interest, that we will focus on, is classifying whether a person suffers from CHD or not. This means that the class label to predict will be the CHD – does the person have it or not. The dataset includes the attributes Systolic Blood Pressure (SBP), Tobacco, Low Density Lipoprotein Cholesterol (LDL), Adiposity/Obesity, Age, TypeA, Famhist, Alcohol and the Coronary Heart Disease response (CHD). TypaA describes a personality type and Adiposity is a different way to measure fattiness similar to BMI. Even though the attributes obesity and adiposity both are similar attributes, both are included here. One of the attributes may be removed later on. All attributes from the dataset will be included in the analysis as it is better to have more variables which can then be deemed relevant/irrelevant in the PCA analysis than discarding something that might impact the result prematurely. A thorough description of the attributes will be given in the next section.

Before describing possible work on the data, a short definition of coronary heart disease is given:

It is expected that attributes such as tobacco, adiposity/obesity and alcohol intake will influence the LDL- or SBP-level, and/or the CHD diagnose.

Looking into the attribute of alcohol consumption, there are quite a few outliers (for more information consult box plots in section 3). This value could be subject for anomaly detection but this depends on whether it is plausible to classify extreme intake amounts as anomalies or not.

[1] World Health Organization

The first two principal components does not give us much information about how to separate the data. As can be seen from the variance explained in figure 7, we would need up to 4 attributes to cover 90% of the dispersion in data. In order to identify which variables that are important for separating the subjects into classes, one could apply logistic regression, which could make it possible to predict class belonging of new observations, and estimate the probability. Furthermore, we cannot assume linearity which logistic regression is better at explaining than i.e. Linear Discriminant Analysis. The classes are not separated perfectly, which is why logistic regression is a good tool.

From regression several attributes could be explained, i.e. blood pressure/age or obesity/age a plotting of the variables and their correlation can be found in the next section.

2 Explanation of data attributes

As can be seen in figure 1 the dataset contains a total of 11 unique attributes. Each of the attributes will be described, first in terms of their meaning and afterwards in relation to what machine learning techniques to apply.

	row.names	sbp	tobacco	ldl	adiposity	famhist	typea	obesity	alcohol	age	chd
0	1	160	12.00	5.73	23.11	Present	49	25.30	97.20	52	1
1	2	144	0.01	4.41	28.61	Absent	55	28.87	2.06	63	1
2	3	118	0.08	3.48	32.28	Present	52	29.14	3.81	46	0
...	...	...	...	...	...	...	...	...	...	...	...

Figure 1: Extract of the South African Heart Disease data listing the different attributes of the dataset. The full size is 462×11

The data in figure 1, reveals that most of the attributes have numerical values. The only non-numeric value is the `famhist` attribute with the value set $\{Present, Absent\}$. These values could usefully be amended by binarizing the two values into the representation $\{1, 0\}$, where 1 denotes the cases in which the family history is $Present$, 0 if $Absent$.

In order to standardize the dataset, since the attribute values ranges from 0 to more than 200, it would be preferred to apply feature transformation either by subtracting mean or subtracting mean and dividing by standard deviation, possibly for all attributes that are not already binary. This would suffice since the range from smallest to largest value within a given attribute is relatively small. Had the values differed greatly (> 1000), one could in stead have chosen a logarithmic scaling.

At a first glance, BMI and ABI seems much alike in both calculating a value on a fattiness scale, but, they are in fact linearly independent and keeping both the attributes will not be a problem for the results. The 1st attribute `row.names` is continuosly increasing in steps of 1 from the first observation. The value contributes with no actual meaning to the rest of the data, therefore in the visualization, this attribute will be removed and not be uncluded again. This will reduce the number

of dimensions and the machine learning problem. There are no missing data values, which means that there will be no need to impute values in the dataset.

2.1 Overview of the data attributes

row.names	Technical debt, irrelevant to the analysis.
Systolic Blood Pressure (SBP)	The highest pressure when your heart beats and pushes the blood round your body
Tobacco	Cumulated tobacco consumption (kg)
Low Density Lipoprotein Cholesterol (LDL)	Lipoproteins transfer lipids around the body in the extracellular fluid thereby facilitating fats to be available and taken up by the cells body-wide via receptor-mediated endocytosis.
Adiposity	Body Adiposity Index (BAI) - a method of measuring the amount of body fat in humans. Calculated without using body weight, as BMI. Instead, it uses the size of the hips compared to the person's height.[5]
Famhist	Family history of heart disease (Present, Absent)
TypeA	Type-A behavior score. Type A and Type B personality theory describes two contrasting personality types. The two cardiologists who developed this theory came to believe that Type A personalities had a greater chance of developing coronary heart disease[6].
Obesity	BMI-measure of your health on a scale from approx. 15-40 using your height and weight.
Alcohol	Current alcohol consumption (no unit given)
Age	Age at onset
CHD response	Coronary Heart Disease

Figure 2: Full list of all attributes in the South African Heart Disease dataset and short description and/or meaning of the various abbreviations.

2.2 Attribute types

On the basis of the attributes described in the figure below, the type of each attribute will now be explained and their basic summary statistics will be described in section 3. The figure below gives a brief description of the meaning of the attributes, their units, and the range of the different attributes.

Attribute	Type	Range
SBP	Continuous Interval	$101 - 218$
Tobacco	Continuous Ratio	$0 - 31.11$
Alcohol	Continuous Ratio	$0 - 147.9$
LDL	Discrete Interval	$0.98 - 15.3$
Famhist	Binary Nominal	$Present/Absent$
TypeA	Discrete Ratio	$13 - 78$
Obesity	Continuous Ordinal	$14.7 - 46.58$
Adiposity	Continuous Ordinal	$6.74 - 42.49$
Age	Discrete Interval	$15 - 64$
DHC	Binary	$0/1$

3 Data visualizations

In this section the South African Heart Disease data set will be investigated superficially through data visualizations and principal component analysis (PCA).

3.1 Initial data investigation

One of the first things to investigate in a large dataset like this one, is whether or not there are any significant outliers, that may distort or skew the PCA. This was investigated by generating box plots of the data. Box plots are great for visualizing basic statistics of individual attributes, and figure 3 shows box plots for each feature in the data set both raw and standardized data. Plotting all the features on the same y axis may make the data difficult to interpret, since each feature operates at its own scale. It is obvious that there are some outliers in the data, specifically for sbp and alcohol there are quite a few outliers, but it is difficult to interpret on this scale. By standardizing the data via mean subtraction the data is transformed to be centered around zero, which makes it easier to identify the outliers (Bottom part of figure 3). Considering that there are many outliers for alcohol and sbp, it would be wrong to exclude them from the data set. However, given the few number of outliers in the typea attribute it may be advantageous to exclude these data points at later stages of the analysis.

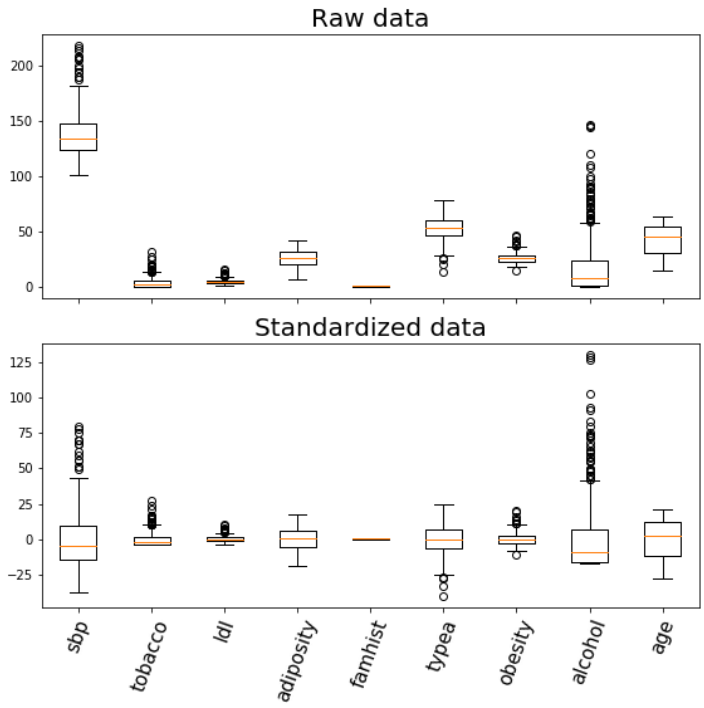

Figure 3: Box plot visualization of each feature in the South African Heart Disease data set. The top plot shows box plots of raw data, while the bottom plot shows box plots of the data after mean-standardization

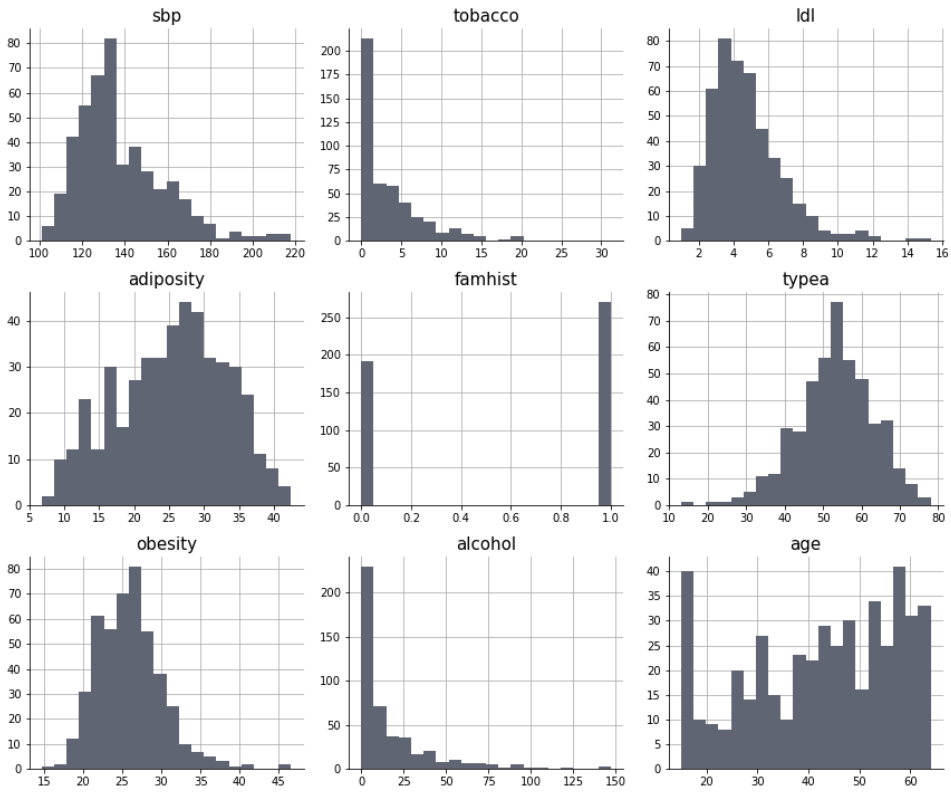

Figure 4: Histograms generated for each feature in the data set. Only obesity, adiposity and typea are close to being normally distributed.

Hstograms are created for each attribute to investigate the different distributions (figure 4). Of the nine features only three could be approximately identified as being normally distributed, as age has a nice normal distribution, while obesity and adiposity look like slightly skewed normal distributions. This might have been different, had there been more observations.

In order to investigate whether the features are independent, a correlation analysis was carried out for all combinations of attributes by plotting the combinations against each other in a matrix plot (figure 5). For brevity, the attribute `famhist` has been excluded, as it contains binary data. It appears that `adiposity` and `obesity` are slightly correlated, which is natural since these two attributes measure similar things, which indicates that it is perhaps not necessary to use both measures in the later analysis stages. There also seems to be a very slight correlation between `adiposity` and `age`, but not enough to say that any of them would make the other redundant.

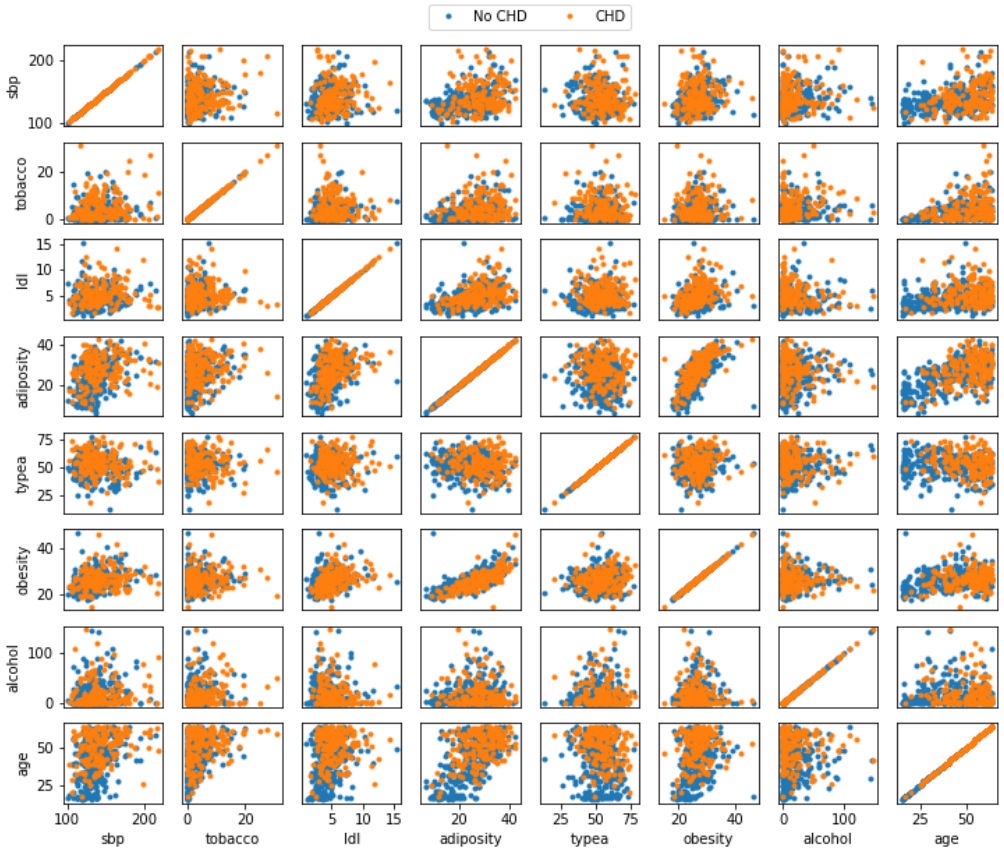

Figure 5: Matrix plot of the correlation between all combinations of the 9 attributes.Orange dots indicate presence of coronary heart disease (CHD), while blue dots indicate the absence of CHD. The binary attribute $famhist$ has been left out.

Correlation coefficients were found to be very low, almost all across the board, as only one correlation coefficient was found to be higher than 0.7, specifically the correlation coefficient of adiposity and obesity 6.

	sbp	tobacco	ldl	adiposity	typea	obesity	alcohol	age
sbp	1.000000	0.212247	0.158296	0.356500	-0.057454	0.238067	0.140096	0.388771
tobacco	0.212247	1.000000	0.158905	0.286640	-0.014608	0.124529	0.200813	0.450330
ldl	0.158296	0.158905	1.000000	0.440432	0.044048	0.330506	-0.033403	0.311799
adiposity	0.356500	0.286640	0.440432	1.000000	-0.043144	0.716556	0.100330	0.625954
typea	-0.057454	-0.014608	0.044048	-0.043144	1.000000	0.074006	0.039498	-0.102606
obesity	0.238067	0.124529	0.330506	0.716556	0.074006	1.000000	0.051620	0.291777
alcohol	0.140096	0.200813	-0.033403	0.100330	0.039498	0.051620	1.000000	0.101125
age	0.388771	0.450330	0.311799	0.625954	-0.102606	0.291777	0.101125	1.000000

Figure 6: Correlation coefficients of all combinations of attributes. $famhist$ has been excluded.

3.2 Principal Component Analysis

Principal component analysis is a way of boiling down the information in a large dataset to a smaller number of dimensions, more formally known as dimensionality reduction. The South African Heart Disease data consists of 9 different attributes, each serving as a different dimension. By doing a PCA analysis it is possible to find the components that are most important from the data, thus representing the majority of the variation with a fewer number of attributes. From our data we created a 462×9 matrix X, leaving out the class label `CHD`, which we standardized by mean subtraction, creating a new matrix $\tilde{X}$. On this matrix we performed singular value decomposition to compute the eigenvectors of the dataset, i.e. the principal directions of the components, which allowed us to project the data onto the principal components. Singular value decompositions gives three matrices such that

$$U\Sigma V = Y$$

where V contains the principal component directions of $\tilde{X}$.

From the principal components in V and the matrix Σ we calculated the fraction of variance explained by each principal component (Figure 7). The first two principal components only explain about 70% of the variation, and in order to capture approximately 90% of the variance, the top three principal components would need to be used, to be completely sure that all data is captured, one could include the 4th principal component.

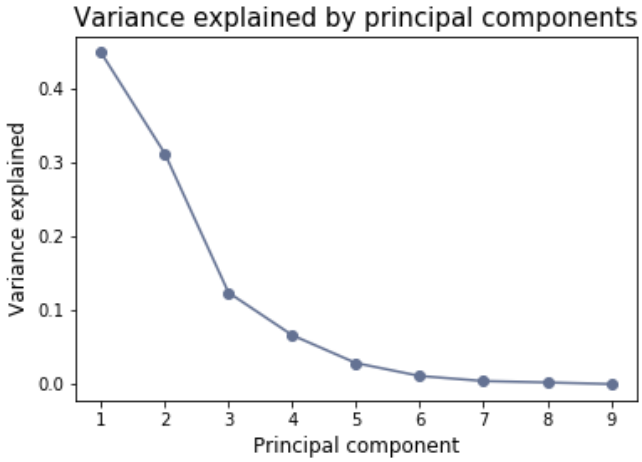

Figure 7: Fraction of the variance explained by the principal components represented by matrix V.

The principal component directions of the top three principal directions capture information about different things (Figure 8). Principal component 1 has a large positive impact on tobaccco intake and LDL levels, while principal component 2 has a large positive impact on the typea-score, while principal component 3 has a large negative impact on alcohol consumption. This goes to show that each principal component captures very different information of the data set.

	sbp	tobacco	ldl	adiposity	famhist	typea	obesity	alcohol	age
1	-0.414583	0.773840	0.477330	-0.003626	0.037536	-0.004177	0.002855	-0.000525	-0.000495
2	-0.056834	0.041824	-0.115293	0.022954	0.089345	0.986222	-0.013337	0.019684	-0.004483
3	-0.006418	0.024449	-0.039320	0.026168	-0.088687	0.022253	0.068668	-0.991708	0.023883

Figure 8: Directions of the top three principal components.

After identifying the principal components the data was projected onto the top principal components, to see whether any clear clustering would appear (Figure 9). No clear clustering of the data is observed when projecting onto either of the top three principal components. There is a chance that this is an indication that our machine learning method may be insufficient to do any valid predictions, but it is more likely that we would just need to include more than two principal components in order to observe a clear separation of data points with and without CHD. This will be considered in the analysis at a later stage.

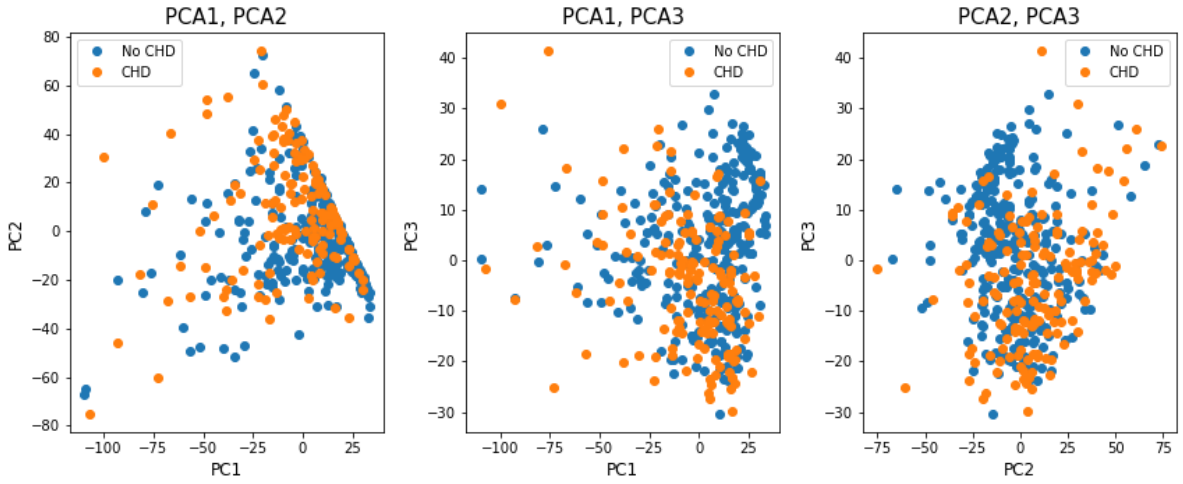

Figure 9: Projection of standardized data, $\tilde{X}$ onto each of the top three principal components.

4 What was learned about the data

The data from the South African Heart Disease data set consists of 10 attributes where two are binary (can be converted into), and the rest is numeric. There were no need to impute/remove any attributes as all data is complete. Before working on the visualization of data, all attributes except the binary values were standardized by mean subtraction. From the attribute histograms, it was concluded that only `age`, `adiposity` and `obesity` can be seen as proper normal distributions, this might have been different had there been more observations.

Most of the attributes in our dataset can not be considered correlated. Only `adiposity` and `obesity`, which are measured on almost the same scale show a slight correlation. Depending on which is more accurate, one of the two will be removed in later analysis.

From the analysis and the related visualizations in this report it has been made clear that in order to successfully account for all the dispersion in the data, more than the first two principal components are needed. In fact we need the first 3 components to account for 90% of the dispersion in the data. After having identified and plotted the first principal components no clear clustering appears, indicating a need for including more principal components.

References

[1] "South African Heart Disease." [Online]. Available: https://statweb.stanford.edu/~tibs/ElemStatLearn/datasets/SAheart.data

[2] J. e. A. Rousseau, "Restriction Endonuclease Mapping of Globin Genes in Beta-Thalassemia," *South African Medical Journal*, vol. 64, no. 11, 1983.

[3] "South africa: Coronary heart disease," 2014. [Online]. Available: http://www.worldlifeexpectancy.com/south-africa-coronary-heart-disease

[4] "What Is Coronary Heart Disease?" [Online]. Available: https://www.nhlbi.nih.gov/health/health-topics/topics/cad

[5] "Body Adiposity Index." [Online]. Available: https://en.wikipedia.org/wiki/Body_adiposity_index

[6] "Type A Behaviour," 1983. [Online]. Available: http://medical-dictionary.thefreedictionary.com/type+A+behavior