Technical University of Denmark

02450 Introduction to Machine Learning and Data Mining

Project 1

s191985 Julian Böhm (section 3)
s196119 Emil Chrisander (section 1 & 4)
s192184 Jorge Montalvo Arvizu (section 2)

September 2019

1 Description of our data set (s196119)

Our data set, South African Heart Disease, is a subset of a data set collected by Rousseauw et al (1984)[1]. The original context of the data set is the discovery of white South Africans being excessively prone to ischaemic heart disease (IHD) in the late 1970s. The paper by Rousseauw et al (1984) analyzes a set of biological and socio-economic factors for their potential correlation with IHD. The authors apply simple (at that time advanced) summary statistics to analyze correlation between these factors and IHD. To avoid spoiling the excitement of reading our thrilling reports, we won't summarize their results. The collected data is a combination of self-reported surveys and doctor examinations of citizens from three rural South African communities. The data available to us is a subset of the original data applied by Rousseauw et al (1984). Our ten available variables are:

- **chd:** Our binary outcome variable of whether the respondent is diagnozed with IHD.
- **age:** Age of the respondent.
- **alcohol:** Current alcohol consumption (specific unit not stated).
- **adiposity:** Body adiposity index determines body fat percentage. It is a measure of the percentage of total body mass that is composed of fat.
- **famhist:** Family history of heart disease (Present, Absent).
- **ldl:** Low density lipoprotein cholesterol (often referred to as the 'bad' cholesterol).
- **obesity:** Represented by Body Mass Index (BMI).
- **sbp:** Systolic blood pressure.
- **tobacco:** Cumulative tobacco, kg (time dimension not stated).
- **typea:** Type A personality theory describes personality type that could raise one's chances of developing coronary heart disease[2]. The variable is an index that ranges from 0-100 with a mid-point at 50.

We obtained the data from Stanford's free available data for educational purposes[3]. At this stage we plan to perform classification of the variable *chd* applying decision tree algorithms, neural network, etc. We expect to apply all our 9 attributes for ML analyses as they could all potentially contain valuable variation. Because of our small sample set we plan to examine the data for missing values and outliers thoroughly. Other than standardization of our 9 attributes, we do not have any plans of transforming them. In the next section we describe the attributes of our data set.

[1] paper available here: `https://www.jclinepi.com/article/0021-9681(84)90051-1/pdf`
[2] see: `https://en.wikipedia.org/wiki/Type_A_and_Type_B_personality_theory`
[3] `http://www-stat.stanford.edu/~tibs/ElemStatLearn/datasets/SAheart.data`

2 Attributes of the data (s192184)

The nine attributes available in the data set represent a collection of measurements given a certain characteristic, e.g. obesity is represented as the BMI of certain observation (a person). These attributes are different from one another and can be sorted into attribute types[4]:

Table 1: Attributes sorted by type

type	Nominal	Ordinal	Interval	Ratio
Continuous	–	–	–	**alcohol** **adiposity** **ldl** **obesity** **tobacco**
Discrete	–	**typea**	**age**	**sbp**
Binary	**famhist**	–	–	–

Given the assortment, attributes can be summarized with the following five combinations:

- **Nominal/Binary:** take values "Present" or "Absent"
- **Ordinal/Discrete:** take values from 0 to 100, in integer steps of 1
- **Interval/Discrete:** take values in integer steps of 1
- **Ratio/Continuous:** take values in continuous decimal values
- **Ratio/Discrete:** take values in integer steps of 1

Even though attributes **sbp**, **typea** and **age** could be argued to be continuous, the data show the values in integer steps of 1. Also, for the attribute **age**, the physical meaning of 0 would be that a person is not born yet. This is arguable but for our analysis we'll categorize it as an interval attribute given the records of the data set.

2.1 Missing data and summary statistics

After visual inspection (using boxplot and identifying potential NA values) we conclude that our data set does not contain missing data or corrupted values. However, the units of the variable **alcohol** aren't specified nor the time dimension of the variable **tobacco**. This makes it more difficult for us to interpret

[4]Introduction to Machine Learning and Data Mining. Chapter 2, 2019; Tue Herlau, Mikkel N. Schmidt, Morten Mørup.

the results we obtained based on these variables. Nevertheless, we expect that it won't interfere with our further ML and DM analyses.

Table 2 and Table 3 show the summary statistics of the attributes. The number of observations for each of them is $N = 462$. 35 percent of the 462 persons in our samples are diagnosed with the heart disease. Hence, we do do not have an equal proportion of the two classes we wish to classify later on. The average age is 43 years and the average BMI is 26 (above 25 is defined as over-weight). Hence, we are dealing with a sample of slightly elderly people who are obese. Moreover, as one can see from the tables, the scale (levels) and variance of the variables differ quite a lot. I.e. **tobacco** ranges from 0 to 31.2 with a std of 4.6, whereas **alchohol** ranges from 0 to 147 with a std of 24.5. Since many machine learning algorithms are sensitive to scale and within-variable variance, we performed a standardization of our data set prior to performing machine learning algorithms (including PCA). Also, from the tables, we can see that only the attributes **tobacco**, **alcohol**, and **famhist** contain values of 0, while all the other attributes' observations are in a range of positive values.

Table 2: Summary statistics (1/2)

stats	**tobacco**	**ldl**	**adiposity**	**obesity**	**alcohol**
mean	3.64	4.74	25.41	26.04	17.04
std	4.59	2.07	7.78	4.21	24.48
min	0.0	0.98	6.74	14.7	0.0
25%	0.05	3.28	19.77	22.98	0.51
50%	2.0	4.34	26.12	25.8	7.51
75%	5.5	5.79	31.23	28.5	23.89
max	31.2	15.33	42.49	46.58	147.19

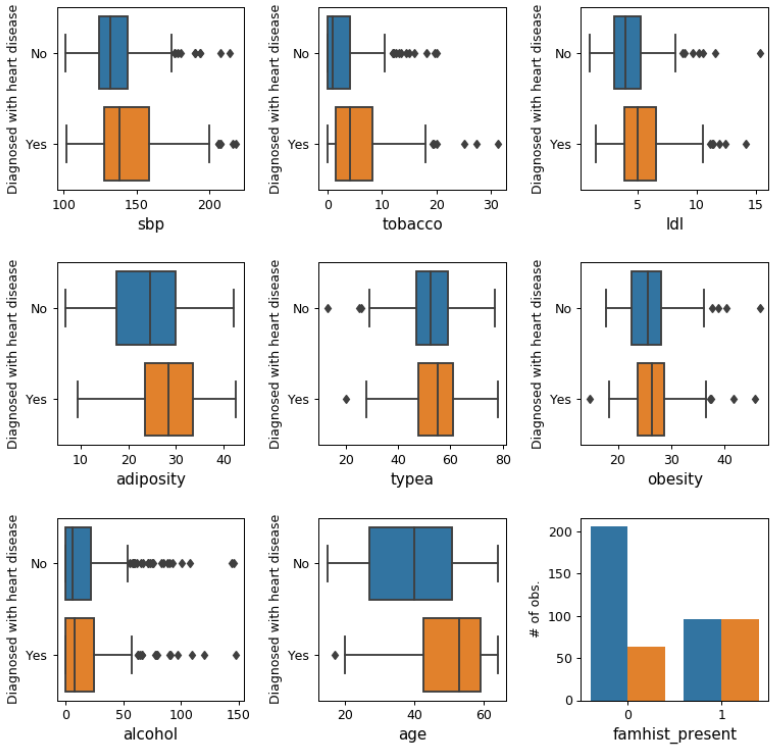
Table 3: Summary statistics (2/2)

stats	**sbp**	**typea**	**age**	**famhist**	**chd**
mean	138.33	53.1	42.82	0.42	0.35
std	20.5	9.82	14.61	0.49	0.48
min	101.0	13.0	15.0	0.0	0.0
25%	124.0	47.0	31.0	0.0	0.0
50%	134.0	53.0	45.0	0.0	0.0
75%	148.0	60.0	55.0	1.0	1.0
max	218.0	78.0	64.0	1.0	1.0

To supplement our summary tables we have also included individual subplots of each of the attributes. We have divided the plots across the outcome variable (**chd**) to make a visual classification. That is, can we reasonably classify a person with a heart disease by simply visually inspecting each attribute? These

plots are shown in Figure 1 as box plots for continuous attributes and a simple bar plot for our single discrete attribute. Ideally, we want our box plots conditional on **chd** to be completely separable. I.e. all the persons with the heart disease should have a BMI of over 30 and all the non-diagnosed should have a BMI under 30. This would make it very easy to classify. However, as we can see from the figure, this is clearly not the case. Generally, there is a large overlap between the diagnosed and non-diagnosed persons. That being said, we do see a tendency towards the diagnosed persons being older, be frequent smokers, have a family history of heart diseases, and higher levels of adiposity and ldl. Nevertheless, we clearly can't classify heart diseases based on visual inspection of a single attribute alone.

Figure 1: Distribution of features conditional on outcome variable



As a final description of our attributes we also include a correlation table. This is shown in Figure 2. We added a color layer to make it easier to identify potential strong correlation. The color scale goes from red (correlation equal to -1) to green (correlation equal to 1). Yellow indicates the mid-point of no correlation (correlation equal to 0). Generally, we do not see any strong correlations among our variables. Age is the attribute with the strongest co-movements. If a person is old, the person typically also smokes more, has a higher BMI and higher levels of ldl, adiposity, and sbp. Finally, the correlation table confirms our observations from Figure 1. With the exception of age, the attributes only have a relatively weak correlation with the heart disease outcome variable. In the next section we perform a Principal Component Analysis on our data set.

Figure 2: Correlation Matrix

	sbp	tobacco	ldl	adiposity	typea	obesity	alcohol	age	chd	famhist_present
sbp	1	0.2	0.2	0.4	-0.1	0.2	0.1	0.4	0.2	0.1
tobacco		1	0.2	0.3	0	0.1	0.2	0.5	0.3	0.1
ldl			1	0.4	0	0.3	0	0.3	0.3	0.2
adiposity				1	0	0.7	0.1	0.6	0.3	0.2
typea					1	0.1	0	-0.1	0.1	0
obesity						1	0.1	0.3	0.1	0.1
alcohol							1	0.1	0.1	0.1
age								1	0.4	0.2
chd									1	0.3
famhist_present										1

3 Principal Component Analysis (s191985)

In order to find a lower-dimensional representation of a data set, one can apply the Principal Component Analysis (PCA). The PCA enables the (lossy) compression of data sets, the preprocessing of very high-dimensional data and a powerful visualization (Introduction to Machine Learning and Data Mining, 2019; Tue Herlau, Mikkel N. Schmidt, Morten Mørup).
The intuitive idea to reduce the dimensionality while keeping the maximum variation within observations. We want to keep variation between observations as this may help us to classify person with and without heart disease later on. One can understand this by considering the extreme case in which we have no variation between the observed persons. In this case every person has the same level of attributes which make it impossible to make a qualified classification based on the attributes. Thus, generally we prefer to have more variation in our data set.

Below we explain the steps we have taken to perform the PCA of our South African Heart Disease data set. The data set is stored in a matrix M for this purpose.
Afterwards, the following steps have been executed:

- Substract mean values from X (creates $\tilde{X}$)
- Divide by standard deviation from $\tilde{X}$
- PCA by computing singular value decompositon (SVD) of $\tilde{X}$: Application of *svd*() function of the linear algebra package 'scipy.linalg'
- Calculate the variance ρ explained by principal components
- Visualization of the variance explained by principal components, see figure 3
- Visualization of the PCA projection of the PC1 and PC2, note figure 3
- Direction of attributes coefficients (Eigenvectors): figure 5

3.1 Interpretation of the PCA

To get an overall overview of the explained variance by the principal components we have to take a closer look at figure 3. It can be seen that the cumulative curve of the variance explained reaches the threshold of 0.95 after the seventh principal component (out of nine). In numbers, it can be calculated that PC1 and PC2 respectively represent 32.1% and 13.3% of the variance in the data set (values of ρ). Therefore, it can be said that the reduction of the data set's dimensions comes at a high cost in terms of lost variation. It can hardly be lowered without losing representation of the original data.

Moreover, we can see from figure 4 that a 2-d projection of PC1 and PC2 does not clearly segregate observations based on their class. That being said, we

do see a tendency towards the non-diagnosed persons having high values of the PC1 projection and close to zero value of the PC2 projection. The diagnosed persons seem to be all over the plot with no obvious pattern. This tells us, that we probably need more variation than what we can get from PC1 and PC2 to perform a proper classification.

Finally, figure 5 shows the direction of attributes coefficients or the Eigenvectors of the attributes. This plot can tells us something about the variation each of the attribute 'brings to the table' in PC1 and PC2. Firstly, the length of the vector tells us something about how much variation the attribute contributes to the PC. Secondly, the closer the attributes' vectors are to each other, the more contribution of variation they share. We can see that tobacco and age both have a relatively strong contributions to the PC1 and PC2. However, they are more or less on top of each other indicating that they contribute with the same variation. This fits our finding from section two, in which older people were found to be more likely to smoke. Obesity and typea also contributes fairly strongly - but it's very different variation they contribute with. This is seen from the observation that the angle between their vectors is relatively large. The remaining attributes are closer to origo. Thus, they contribute with less variation to the PC1 and PC2. In the next and final section we briefly sum up and discuss the key findings from our data analysis in Project 1.

Figure 3: Variance explained by principal components

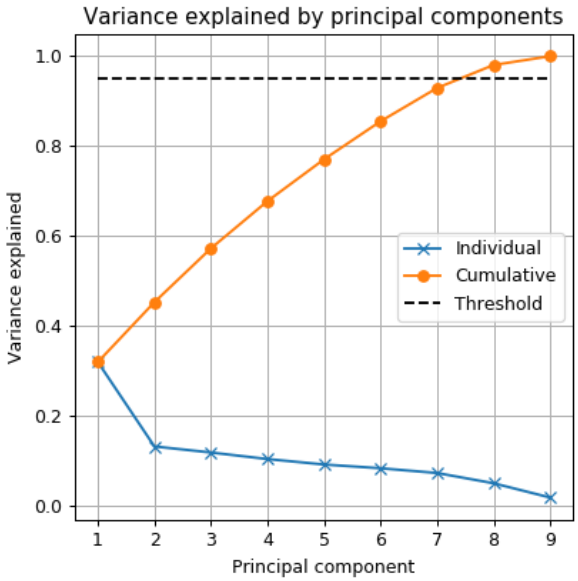


Figure 4: Projection of principal components 1 and 2

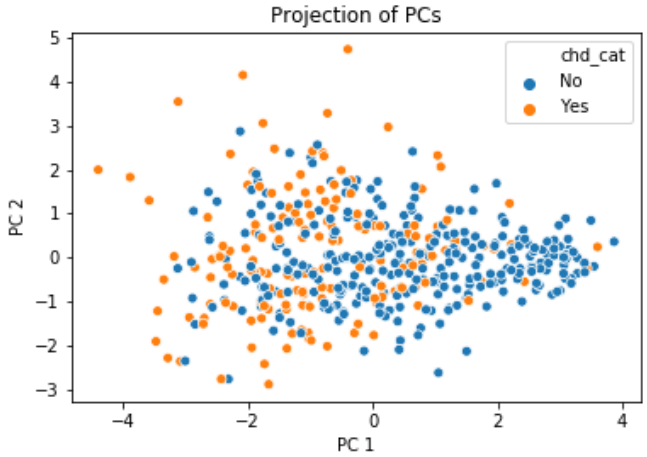
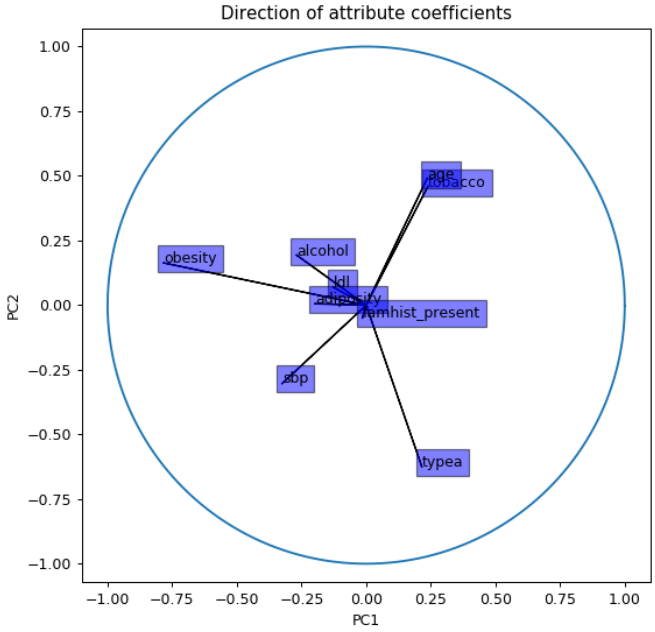


Figure 5: Direction of attributes coefficients



4 Discussion (s196119)

We end by summing up our findings. Firstly, we found that our data set is generally 'good looking'. We do not have any strange observations in our data set and every observation has valid attribute levels. Secondly, we realized that the our attributes vary in scale and variation. This implies that we need to standardize our attributes whenever we perform machine learning analyses which are sensitive to scale and within-variable variation. Thirdly, we learned that the our outcome variable for heart disease diagnose can not easily be classified by visually inspecting individual attributes. However, we did observe a visual tendency towards older persons being more likely to be diagnosed with the heart disease. Fourthly, we learned that Principal Component Analysis is probably not the most feasible tool for classification in our case. We lose quite a lot of variation from reducing the dimension of the attributes - we had to make 7 Principal Components to maintain 95 percent variation. Moreover, we also concluded that a visual classification based on projection of PC1 and PC2 was very difficult. Hence, at this stage we do not see any strong arguments for applying a PCA on our data set prior to a classification task in the next project. Our final conclusion is that we are less confident that we will be able to perform very accurate classification of heart disease based on our 9 attributes. Nevertheless, we are optimistic and look forward to apply machine learning tools to our data set.