Danmarks Tekniske Universitet

Data: Feature extraction, and visualization

Report 1 - Group 55

Karen Witness - s196140
Aksel Buur Christensen - s203947
Emil Slente Liljegren - s174036

Contribution

Section	Karen Witness	Aksel Buur	Emil Slente
1. Description of dataset	-	-	100%
2. Detailed explanation of the attributes of the data	80%	20%	-
3. Data visualization (PCA)	40%	60%	-
4. Discussion	10%	90%	-
5. Exam problems	50%	-	50%

October 4, 2022

Contents

1 Description of data set

For this report, a data sample from the study by J. E. Rossouw et. al. 1983 was used. The dataset describes a sample of men from the Western Cape in South Africa and what Coronary Heart Disease (CHD) risk factors they each had. The dataset is an excerpt from a larger study used in the article "Coronary risk factor screening in three rural communities, The CORIS baseline study" by J. E. Rossouw et. al. published in 1983 in "The South African Medical journal. The article starts out by establishing thresholds for how the data should be interpreted such as explaining cutoff points of hypercholesterolemia (high cholesterol) and hypertension (high blood pressure) established by other studies and by the World Health Organization. Similar explanations were also done for the other attributes where it made sense. Overall the study used the data to compare how many participants with CHD had one or more CHD risk factors, finding that 73,5% of males had 1 or more risk factors where 32% had just 1 factor 24,4% had 2 factors, 11.6% had 3 factors, 4,7% had 4 factors and 0.8% had 5 factors.

ID	SBP	Tobacco	LDL	Adiposity	Fam. His.	Type A	Obesity	Alcohol	Age	CHD
1	160	12.00	5.73	23.11	Present	49	25.30	97.20	52	1
2	144	0.01	4.41	28.61	Absent	55	28.87	2.06	63	1
3	118	0.08	3.48	32.28	Present	52	29.14	3.81	46	0
4	170	7.50	6.41	38.03	Present	51	31.99	24.26	58	1
5	134	13.60	3.50	27.78	Present	60	25.99	57.34	49	1
6	132	6.20	6.47	36.21	Present	62	30.77	14.14	45	0
7	142	4.05	3.38	16.20	Absent	59	20.81	2.62	38	0
8	114	4.08	4.59	14.60	Present	62	23.11	6.72	58	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
463	132	0.00	4.82	33.41	Present	62	14.70	0.00	46	1

Table 1: *Dataset sample where* $m = 11$ *and* $n = 463$

By applying classification models on the data set we can make predictions in regards to risk factors that contribute to the overall risk category of a person. If classification is applied on the attributes, we can divide each of these attributes in to 2 or more classes that can help simplify future predictions. One such classification could be applied to intervals of tobacco use, obesity and alcohol consumption where they are given a class label of high, low or no increase in CHD risk based on the numbers from people with or without CHD symptoms. We can then make further class labels that uses the other classifications to determine which risk category a person will fall under.

By using regression models we can quantify the statistics of each attribute in relation to CHD and use it to, for example, predict the number of CHD cases in a separate population based on just a limited lifestyle study. Another possibility is to explore the relationship between blood pressure and different lifestyle attributes such as alcohol or tobacco, and use it to predict threshold values for what values can increase CHD risk.

2 Detailed explanation of the attributes of the data

No.	Attribute	Description	Unit
1	SBP	Systolic blood pressure	[mmHg]
2	CT	Cumulative tobacco	[kg]
3	LDL	Low densiity lipoprotein cholesterol	Number
4	BAI	Body adiposity index	Percent%
5	FAM	Family history of heart disease	(Present/Absent)
6	TA	Type-A behavior	Number
7	BMI	Obesity	[kg/m^2]
8	CAC	Current alcohol consumption	??
9	AGE	Age at onset	Year
10	CHD	Response, coronary heart disease	(Yes/No)

Table 2: *Attributes of dataset*

We notice that our attributes are non-equivalent, most of them are a data number, 'famhist' collects data as a string, and 'chd' is either one or zero. Therefore we advantageously can distinguish the attributes between continuous, discrete, and binary. When an attribute is **continuous**, it can take real numbers as values a and b but also take any value in between them. Furthermore, a **discrete** can take a finite or countably infinite set of values. Lastly, if the attribute only can take two values such as 0 or 1, then it is called **binary**. [1]

Additionally we can distinguish between the attributes if the variables have different types. If the variable is unique and it belongs to a category, then we call it **nominal**. But if the variable is ordered or ranked we call it **ordinal**. Further, if the variable is ordered and the distance between them can be measured, then it is an **interval**. Lastly, if zero means the absence of what is measured, then it is called **ratio.** Nominal and ordinal variable is qualitative data, and interval and ratio are quantitative data. From these terms, we can now describe all of our attributes. [1]

No.	Attribute	Type
1	SBP	Continuous and Ratio
2	CT	Continuous and Ratio
3	LDL	Continuous and Ratio
4	BAI	Continuous and Ratio
5	FAM	Binary and Nominal
6	TA	Continuous and Nominal
7	BMI	Continuous and Interval
8	CAC	Continuous and Ratio
9	AGE	Continuous and Ratio
10	CHD	Binary and Nominal

Table 3: *Attribute types*

Our methods only operate on numbers and not on text strings, we translate FAM text strings and denote them as 0 (Absent) and 1 (Present).The same applies to 'CHD', which is also a binary attribute and denotes between 0 (No) and 1 (Yes).

When investigating the data for issues as missing values or corrupted data, we don't find any significant issues what we have to take into account. Therefore we can now load the data, and include basic summary statistics of the attributes.

No.	Attribute	Mean	Std	Median	Range
1	SBP	138.33	20.50	134.00	117.00
2	CT	3.64	4.59	2.00	31.20
3	LDL	4.74	2.07	4.34	14.35
4	BAI	25.41	7.78	26.12	35.75
5	FAM	-	-	-	-
6	TA	53.10	9.82	53.00	65.00
7	BMI	26.04	4.21	25.81	31.88
8	CAC	17.04	24.48	7.51	147.19
9	AGE	42.82	14.61	45.00	49.00
10	CHD	-	-	-	-

Table 4: *Statistics of the attributes*

3 Data visualizations (PCA)

As stated earlier we don't identify any significant issues with outliers. To visualize this, we analyze the box-plot diagram. We use ACCENT principles and Tufte's guidelines to visualize the data.

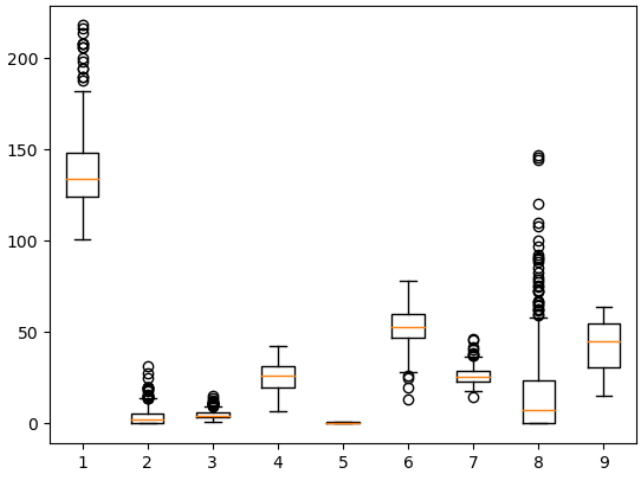

Figure 1: *Attributes Box-plot*

As we see here, all attributes have related data, and none of the data points show issues. Although we wish to do a further investigation of each attribute by visualizing them. When an attribute is continuous, it is suitable to check if they are normally distributed. From table 3 we know that all of our attributes are continuous except for the two binary 'FAM' and 'CHD'. We therefore only analyze the normal distribution for the others, where the mean and variances come from table 4. The histogram is generated by counting how many of the attributes fall within the range covered by each bin of the histogram.[1].

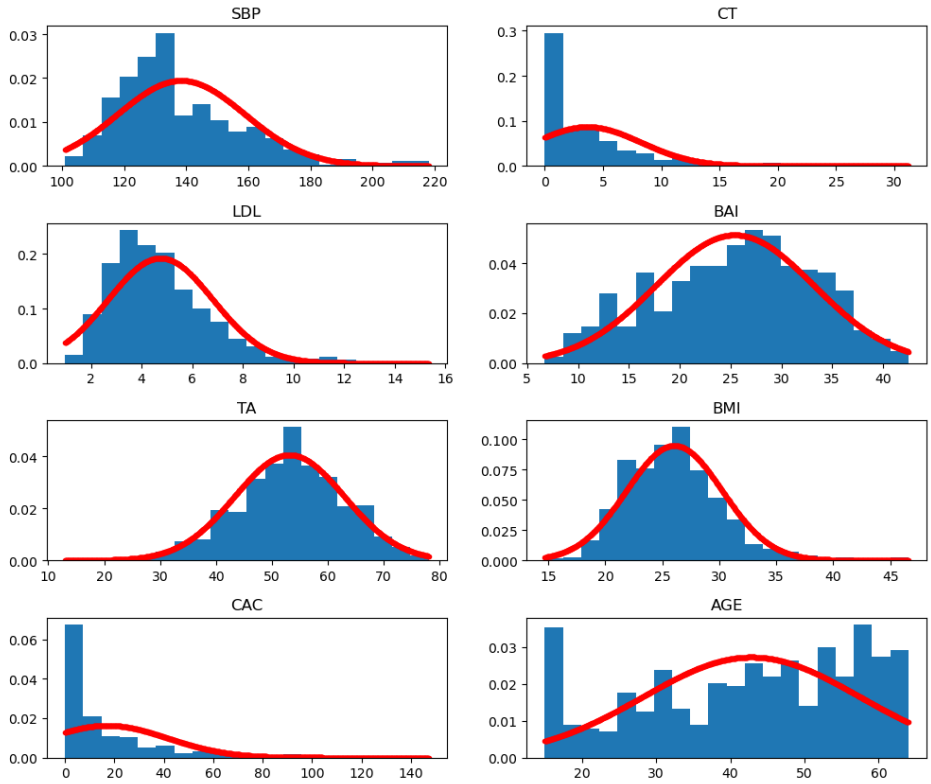

Figure 2: *Normal Distribution*

We observe that all attributes in the figure except 'AGE' are more or less normally distributed. For attributes to be correlated it is important they take on high or low values systematically, even though they may have the same distribution [1]. We see that 'BAI' and 'BMI' are correlated because of their systematic high values. It is troubling to observe the behavior of the attributes related to one another when working in 9 dimensions. The primary machine learning modeling aim appears to be unfeasible. Even though we have enough feasible data which makes it possible to compute and visualize feasible illustrations, we need to use other methods to understand and visualize the 9 dimensions.

3.a PCA

After the general visualization of the data, it is time to perform a Principal Component Analysis (PCA) on the data. This is a useful way to interpret the data and the attributes quickly and easily when there are many attributes in a data set. The PCA analysis performed here is greatly inspired by the theory behind PCA described in [1], chapter 3. The idea behind a PCA is to transform high-dimension data (in our case, 9 attributes equal 9 dimensions) to lower dimensions, making it possible to visualize this data through their principal components.

In the following the PCA, an algorithm will be used to get through all the steps for a PCA, this will first include an evaluation of the data to determine whether it should be standardized or not, after that the variance explained by the PCA will be computed by the single value decomposition (SVD) and then the chosen PCA components will be further investigated.

As already mentioned, the data has 9 attributes related to the test subjects where they all relate to the different body- and health parameters. Naturally, these values have different scales, so when the data is standardized by subtracting the mean there is also divided by the standard deviation:

$$\widetilde{\boldsymbol{x}}_i = \frac{\left(\boldsymbol{x}_i - \frac{1}{N}\sum_{i=1}^{N} x_i\right)}{\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}}} \tag{3.1}$$

Now it is possible to compute the SVD, by collecting all the $\widetilde{\boldsymbol{x_i}}$ into an $N \times M$ matrix, which is denoted $\widetilde{\boldsymbol{X}}$ (here N is number of observations and M is attributes here 9).

$$\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top = \widetilde{\boldsymbol{X}} \tag{3.2}$$

The SVD is used since it is much easier to compute than to select an orthonormal basis and projecting $\boldsymbol{x_i}$ onto the subspace of $\boldsymbol{x_i}$. The equation 3.2 consists of the variance matrix, $\boldsymbol{\Sigma}$ and the matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ consist of the orthornormal basis of the matrix $\widetilde{\boldsymbol{X}}$.
The SVD is carried out and the first three columns of $\boldsymbol{V}$ are the first three principal components:

$$\boldsymbol{v}_1 = \begin{bmatrix} -0.415 \\ -0.057 \\ -0.006 \\ -0.097 \\ -0.003 \\ -0.0005 \\ -0.031 \\ -0.878 \\ -0.207 \end{bmatrix}, \boldsymbol{v}_2 = \begin{bmatrix} 0.774 \\ 0.042 \\ 0.024 \\ 0.164 \\ 0.002 \\ -0.055 \\ 0.056 \\ -0.476 \\ 0.373 \end{bmatrix}, \boldsymbol{v}_3 = \begin{bmatrix} 0.477 \\ -0.115 \\ -0.039 \\ -0.290 \\ -0.007 \\ 0.089 \\ -0.073 \\ 0.008 \\ -0.813 \end{bmatrix} \tag{3.3}$$

If the variance from first three principal components are accumulated, they represent 88.7% of the total variance. Although a threshold of 90 or 95% are most common, the first three principal components will be used here since it is a reasonable amount combined with making the visualizing of the principal components easier. Below is seen the cumulative variance as a function of principal components.

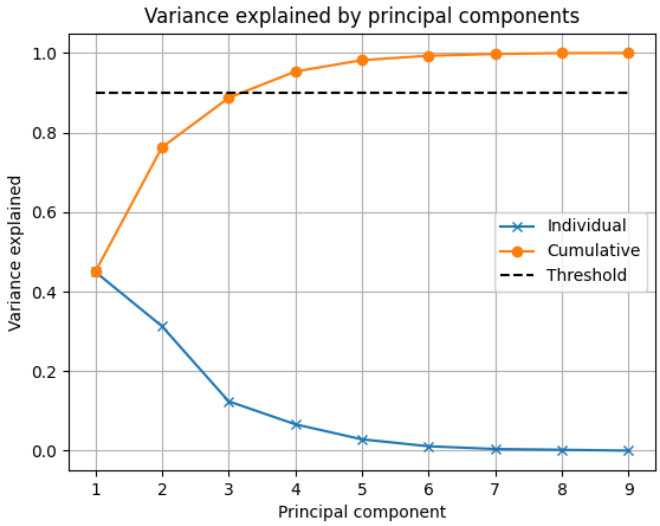

Figure 3: *Cumulative variance as a function of principal components included. Threshold is chosen as 90%.*

With the decision to move on with three principal components it is time to look deeper into the chosen principal components and their respective magnitudes and directions.

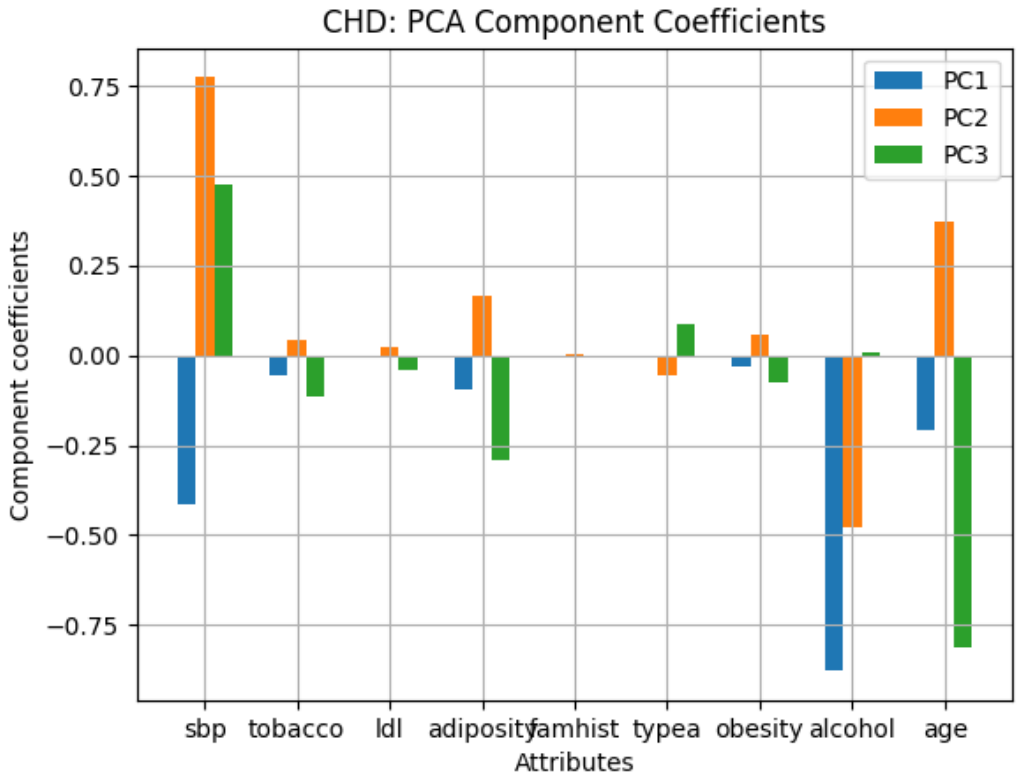

Figure 4: *Component coefficients and directions for the CHD dataset. Notice "famhist" is not visible since it is binary and therefor explained by the very little dot of PC2. It is seen that the direction for PC1 that it is negative, the direction for PC2 is both negative and positive and the direction for PC3 is also negative and positive.*

From figure 4 it is seen that PC1 are the only of the three principal components that purely consists of negative values showing that the direction is negative, with varying magnitude from very close to zero to almost -1. This can also be seen from equation 3.3, but this figure definitely shows how the principal components directs and on which attributes contributes to each principal component.

An observation with large projection onto $\boldsymbol{v}_2$ would have a high value of sbp and low value of alcohol. In the same style, an observation with large projection onto $\boldsymbol{v}_3$ would have a high value of sbp and low value of age. Now the principal components with great magnitude in either direction will have the data projected onto the principal components considered:

Attribute	PC1	PC2	PC3
sbp	0.558	-0.215	0.524
alcohol	0.624	-0.309	0.238
age	0.203	-0.399	0.393

Table 5: *Data for sbp, alcohol and age projected onto the three principal components. Notice how the sign for the projection of the PC1 components changes sign from the coefficient for PC1.*

Here it is seen that the value of the projection of PC1 is positive, while the values of $\boldsymbol{v}_1$ were negative. This can be interpreted as the projection in the opposite direction of $\boldsymbol{v}_1$, which would result in a positive (minus times minus) projection of the data onto $\boldsymbol{v}_1$.

This is hard to make any conclusions on, but serves as a well picture of the data going into a further analysis with extended machine learning methods to analyse the data.

4 Discussion

After the basic data visualization of the CHD Heart Disease data set, a few things has become clear:

- There is no data points found in the data set that contains corrupted data, meaning the data is of very good quality.
- The amount of alcohol the subjects has consumed is widely spread and can almost be transformed into two boxplots instead of 1 (see fig 1), this can probably be used going into an even further analysis of the data. These groups together with the CHD response will be very interesting.
- The component coefficients for the PCA locates 'SBP', 'CAC' and 'AGE' as the greatest in magnitude meaning the projection of the data onto these will become large.

This, together with the other visualization, can be used in the upcoming project where more, and more extended machine learning methods can be applied. Here we will be focusing on regression to try and relate the different attributes to the CHD response. This will be very interesting because the visualization here indicates 'SBP', 'CAC' and 'AGE' have great influence on the subjects. This would correlate pretty well with our expectation, but hopefully a regression based analysis, together with some learning targets can be a good way to explore the CHD Heart Diseases even better.

5 Exam problems for the project

Question 1

1. Option D: We immediate see that congestion/slowdown is a level from low to high, and therefore we know its ordinal. Furthermore we know that a the time of the day is a interval, because the distance between time can me measured.

Question 2

Option A: We calculate the distance for every given p, and we see that none of them gives the correct answer. Therefore we can conclude that $p = \infty$, because the distance measures the largest difference in coordinates, which is $26 - 19 = 7$

Question 3

Option A: The variance explained by each component is given by:

$$\frac{\sigma_i^2}{\sum_i \sigma_i^2} \tag{5.1}$$

From this equation we can conclude that the variance explained by the first four principal components is greater than 0.8.

Question 4

Option C: We can see that compared to the "Broken Truck" attribute; "Time of day", "Accident victim" and "Defects" all have lower magnitudes. All attributes of PC-4 (except "Broken Truck") have negative values in the V matrix which means the projection of the observation onto Principal direction 4 will generally be negative

Question 5

Option A: Running a script to count the terms in the two text lines gives the following Document-term matrix:

1	1	0	0	1	0	1	1	1	0	1	0	1
0	0	1	1	0	1	0	0	0	1	1	1	1

Which gives us the f values:

$$f_{11} = 2$$
$$f_{10} = 6$$
$$f_{01} = 5$$

By inserting these values into the equation for the Jaccard similarity and we get the answer:

$$J(x,y) = \frac{f_{11}}{f_{11} + f_{10} + f_{01}} = 0.1538461538 \tag{5.2}$$

Telling us that option A is the correct answer

Question 6

Option B: Based on the sum-rule (5.5a in [1]), we know the following:

$$p(\hat{x}_2 = 0)|y = 2) = p(\hat{x}_2 = 0|y = 2, \hat{x}_7 = 0) + p(\hat{x}_2 = 0|y = 2, \hat{x}_7 = 1) = 0.81 + 0.03 = 0.84$$

References

[1] Tue Herlau, Mikkel N. Schmidt, and Morten Mørup. Introduction to Machine Learning and Data Mining. Technical University of Denmark, 2016.