02450 Introduction to Machine Learning and Data Mining

Report 1 - Data: Feature Extraction and Visualization

s153355 Leif Førland Schill
s205972 Jens Waaben
s202915 Angelos Daglaroglou

	Leif	Angelos	Jens
Section			
Description of the data	30	30	40
Our investigation: concrete machine learning tasks	30	30	40
Attributes of the data	40	30	30
Data visualization	30	40	30
PCA	40	30	30
Discussion	30	30	40

1 Description of the Data

Data was obtained from Stanford.edu.. This data is a retrospective sample gathered from a original dataset gathered and presented by Roussow*et al.*[1]. The next paragraphs goes through the data in detail: 1) what the data is about (and what purpose of collecting it was) 2) how it was originally analyzed and 3) what machine learning tasks we intend to perform on it.

In their original paper, the authors explain how there is increased incidence of coronary heart disease (CHD) in rural, white, South African (SA) villages. To investigate what is causing this, researchers had 3357 males and 3831 females in the area enrolled into a study mapping different attributes possibly affecting CHD development. Some of the attributes measured include: blood pressure, BMI and smoking habits. The measured attributes were then compared to a threshold value known to be associated with CHD. For example: a systolic blood pressure above 160 mm Hg is known to be a risk factor for CHD, and all the people with a blood pressure above the treshold were marked as having blood pressure as a risk factor. 73 % of men and 67 % of women showed at least one attribute exceeding the threshold cut-off, making the researchers conclude that the population have a high prevalence of CHD risk factors, that could explain the high CHD prevalence.

As mentioned above, the data set used for our machine learning tasks is a subset of the original data - more specifically all the men who had CHD, and approximately 2 ”healthy” controls for each CHD case (total N = 462). The data is without missing values. Not all of the original attributes are in our dataset - those that are can be seen listed in section 2. The only problem with the data seems to be that some CHD patients were getting treatment that could affect the attributes. This is further elucidated when talking about anomaly estimation. Furthermore, some of the attributes seem quite high but are hard to distinguish from faulty measurements so they are left in.

[1] Rossouw, J E et al. “Coronary risk factor screening in three rural communities. The CORIS baseline study.” South African medical journal = Suid-Afrikaanse tydskrif vir geneeskunde vol. 64,12 (1983): 430-6.

1.1 Our investigation: Concrete machine learning tasks

Our problem of interest is not to simply map the prevalence of high risk factors in a binary fashion (as was done in the original study), but to use machine learning and data mining methods to investigate the high risk factors, and use them to predict CHD. In other words: We hope to learn something about the risk factors causing CHD, and their importance in predicting CHD. Following machine learning methods can be considered:

1) Classification: The most exciting application of a classification method would be predicting CHD occurrence based on risk factors. This is our main machine learning aim. In other words, we wish to predict the class label "status of CHD" (binary - {0, 1}) based on the risk factors listed in section 2. Finding at risk individuals could help in treatment.

2) Regression: Regression could be used to elucidate some of the relationships between the risk factors. Understanding risk factors and what causes them can prove important not only for understanding CHD, but also for other diseases. It would be good to see if the lifestyle attributes (e.g. smoking and obesity - look in section 2) could be used to predict some of the attributes correlated with CHD - for example predicting blood pressure with smoking, lifestyle etc.

On the surface it does not look like any special transformations are necessary, other than normalizing variables to not unequally weigh the attributes (subtracting mean and dividing with standard deviation). However it might be a necessary to look for outliers and remove them if deemed necessary. The next section will go into more detail regarding the attributes.

2 Attributes of the Data

A brief explanation of each attribute and some summary statistics are shown below.

- **sdp**: systolic blood pressure (mmHg)
- **tobacco**: cumulative tobacco (kg)
- **ldl**: low density lipoprotein cholesterol
- **adiposity**: body fat measure, comparison of hip size to height
- **famhist**: family history of heart disease (Present, Absent)
- **typea**: type-A behavior measure (competitiveness, aggression, fast-paced lifestyle)
- **obesity**: BMI
- **alcohol**: current alcohol consumption
- **age**: age at onset
- **chd**: coronary heart disease (Present, Absent)

Variable Name	Type I	Type II	$\hat{\mu}$	$\hat{\sigma}$	min	max
sbp	Continuous	Ratio	138.33	20.47	101	218
tobacco	Continuous	Ratio	3.64	4.59	0	31.2
ldl	Continuous	Ratio	4.74	2.07	0.98	15.33
adiposity	Continuous	Ratio	25.41	7.77	6.74	42.49
famhist	Binary	Nominal				
typea	Discrete	Ordinal	53.10	9.81	13	78
obesity	Continuous	Ratio	26.04	4.21	14.7	46.6
alcohol	Continuous	Ratio	17.04	24.45	0	147.2
age	Discrete	Ratio	42.82	14.59	15	64
chd	Binary	Nominal				

As we see, in our dataset there are not any serious issues, since all the attributes are relevant with our main task and there are not any missing data. The only preparation needed, was to simply discard the 'row names' attribute in order to reduce the number of dimensions of the dataset.

3 Data Visualization

Our first goal is to have a visual overview of our data, so we can more easily comprehend attribute associations and data outliers. To look at all continuous attributes simultaneously all attributes were normalized (subtract mean and add standard deviation - the resulting variable is unitless) and plotted as a violin plot - middle bar shows median and side bars show min/max:

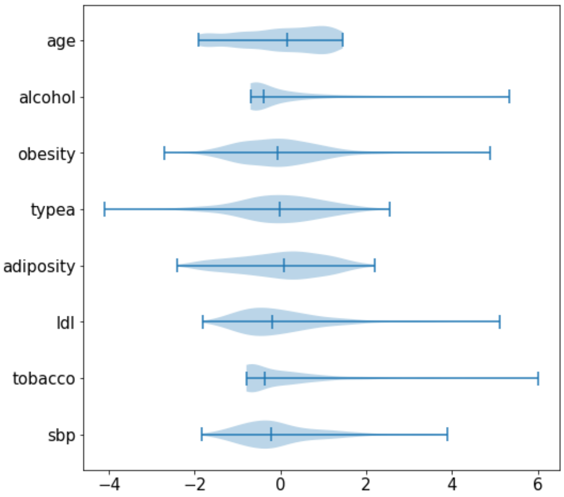


Figure 1: Histogram of disease-wise attributes

Looking at the value distribution, it is noticable that some attributes appear somewhat normally distributed (Obesity, TypeA, ldl and SBP - to some extent also adiposity) while the rest do not. It is noticeable that some values are far from the bulk of the distributions. However, looking at the min/max values in the table in section 2, they seem to all fall within believable values, and we do not think there is a reason to discard them as outliers. We wanted to look into differences in the distributions based whether the patients had CHD or not, and to do this we plotted each attribute distribution for sick and healthy patients:

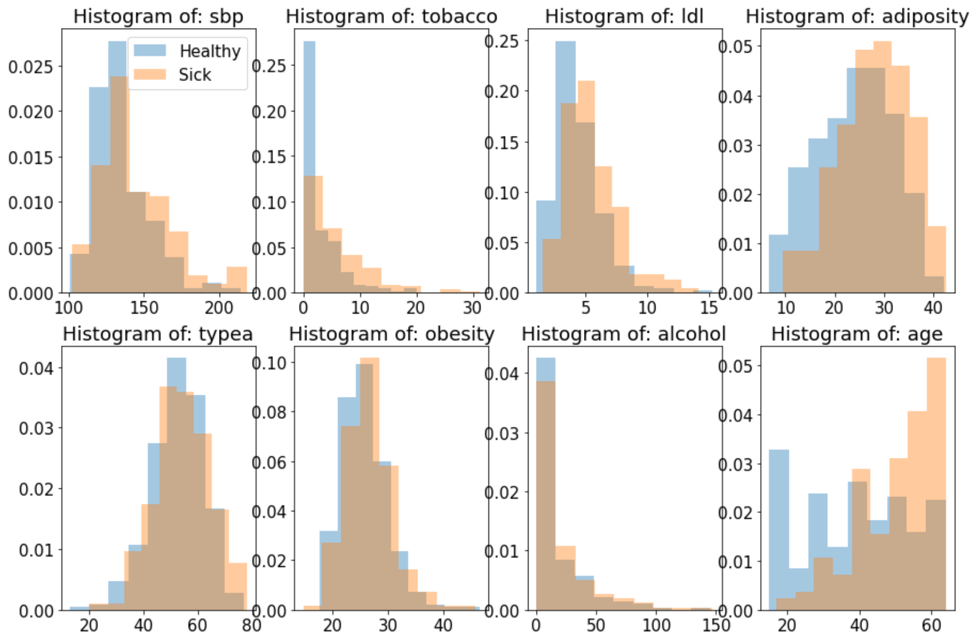


Figure 2: Histogram of disease-wise attributes

Except for age, there did not seem to be real differences in the distributions of attributes. We also looked at the binary family history variable. To see if there was an over-representation of cardiac disease for people with a family history of CHD, the percentage of people with CHD with and without family history of CHD was calculated. 50 % of people who had CHD history in the family developed CHD and only 23 % of people without CHD history developed CHD. The next aspect we wanted to look into was attribute correlation. This was done by investigating the linear correlation between all sets of attributes. Beneath is the correlation matrix:

Table 1: Correlation Matrix

	sbp	tobacco	ldl	adiposity	typea	obesity	alcohol	age
sbp	1	0,21225	0,15830	0,35650	-0,05745	0,23807	0,14010	0,38877
tobacco	0,21225	1	0,15891	0,28664	-0,01461	0,12453	0,20081	0,45033
ldl	0,15830	0,15891	1	0,44043	0,04405	0,33051	-0,03340	0,31180
adiposity	0,35650	0,28664	0,44043	1	-0,04314	0,71656	0,10033	0,62595
typea	-0,05745	-0,01461	0,04405	-0,04314	1	0,07401	0,03950	-0,10261
obesity	0,23807	0,12453	0,33051	0,71656	0,07401	1	0,05162	0,29178
alcohol	0,14010	0,20081	-0,03340	0,10033	0,03950	0,05162	1	0,10112
age	0,38877	0,45033	0,31180	0,62595	-0,10261	0,29178	0,10112	1

As expected, two similar terms - obesity and adiposity - seem to be positively correlated. In this table, we can also see a correlation of adiposity (and obesity in a smaller scale) and ldl cholesterol.

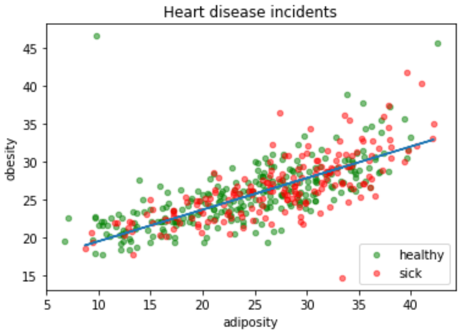


Figure 3: Heart disease incidents in relation to obesity and adiposity.

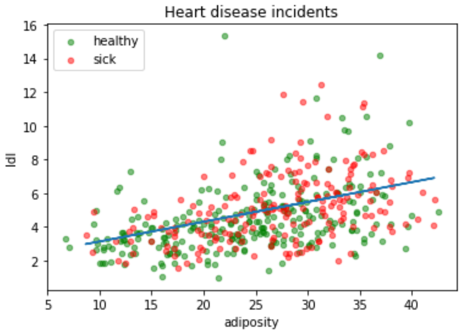
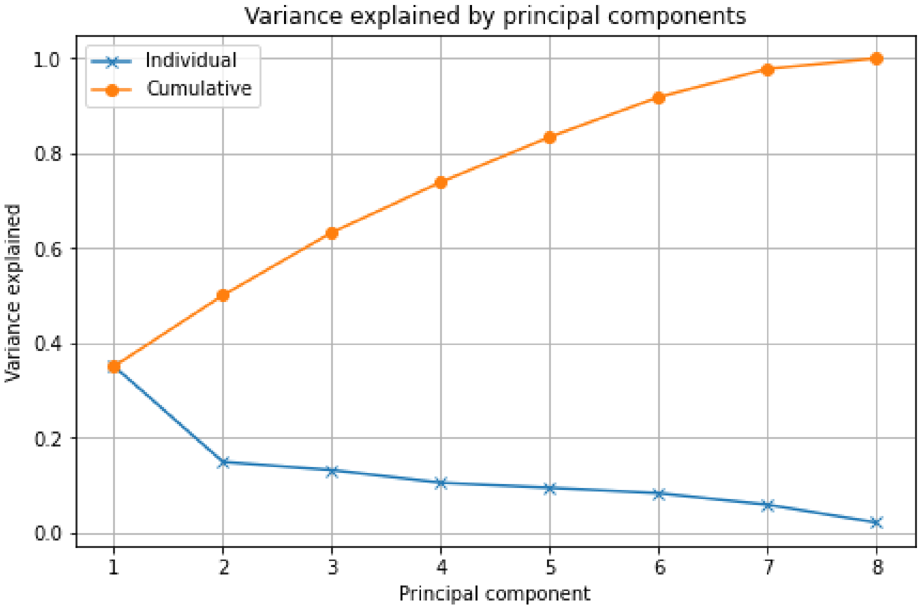


Figure 4: Heart disease incidents in relation to ldl cholesterol and adiposity.

By visualizing the correlated values, we will try to see how tight these correlations are. In figure 3, we have most of the values around the regression line, which indicates that there is a strong connection between those attributes. In figure 4, even though we see a more loose connection this time, it is obvious that a less active way of life leads to possibly higher ldl values. Here we also notice a few values that deviate from the regression line. Although those values could be considered as outliers, they are within reason, so we can keep them and they will be normalized in later stages.

3.1 PCA

A principal component analysis was done on the 8 continuous attributes (the binary attributes family history and CHD (response) were left out). Because of the different scales, all variables were standardized by their standard deviations, as well as being normalized around zero. The variance explained as a function of the number of PCA components included is shown below.

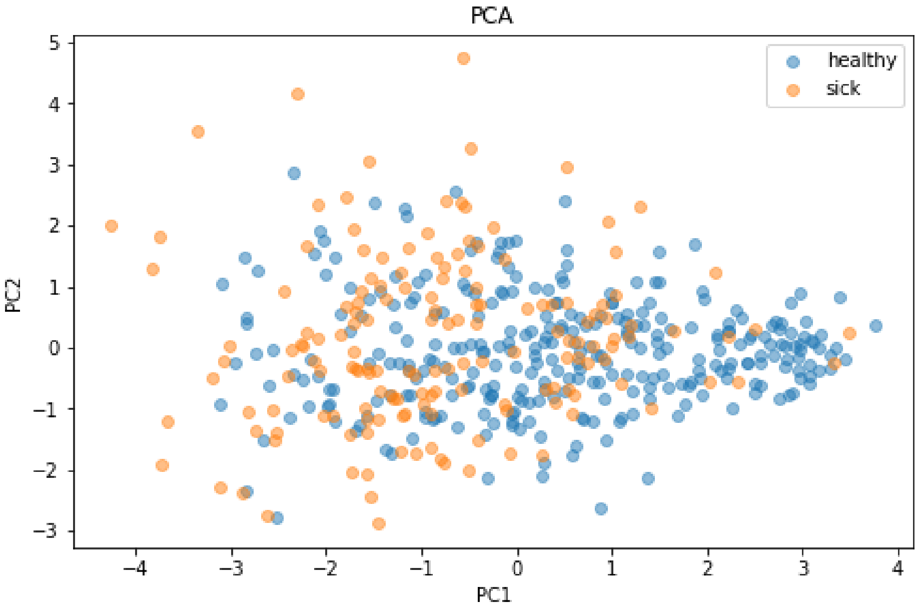


We see that there is one component that explains more than the others, but no component explains a majority of the variance in the data.

	PCA1	PCA2
sbp	-0.33	0.24
tobacco	-0.31	0.46
ldl	-0.34	-0.36
adiposity	-0.53	-0.19
typea	0.02	-0.28
obesity	-0.41	-0.39
alcohol	-0.12	0.54
age	-0.46	0.19
var. explained	35.1 %	15.0 %

Looking at the makeup of each component, it seems that PCA1 carries most information about adiposity, obesity, and age. This component could be interpreted as a measure of healthy body weight. PCA2 carries most information about tobacco and alcohol use, so it could be considered an indicator of how healthy the persons lifestyle is in terms of drug use.

The data is shown projected onto these two principal components.



4 Discussion

The data was originally collected to map risk factors for coronary heart disease. By investigating the individual attributes, we found that the data is of high quality with no missing values and no outliers. Looking further into the distributions of the attributes we found that no single parameter was sufficient to predict CHD. We created a correlation matrix for correlation between the individual attributes. This showed few high correlations between the attributes, for example adiposity/obesity, age/adiposity, age/-tobacco, and age/sbp. This is reflected in the makeup of the first principle component.

In terms of solving the classification problem, the PCA does not group the data into neat groups of sick/healthy people. Furthermore, a lot of principle components are needed to explain a decent amount of variance. From the histograms in Figure 2 we see that some attributes may be more useful in others in classifying, namely adiposity, tobacco, cholesterol, and blood pressure. It could be seen as trivial that older people are more likely to have had heart disease. In reality one would be interested in how much more likely older people are to get heart disease, which the histogram doesn't show.

It seems like knowing these variables for a random person would give a better idea of whether they would get CHD or not, but you would not be able to decide with a high degree of certainty.