Danmarks Tekniske Universitet

Heart attack analysis and prediction data set

02450 Introduction to Machine Learning and Data Mining

Katharina Strauss Søgaard
s214634

Ida Raagaart
s204010

Cecilie Dahl Hvilsted
s214605

Student name	Description	Statistics	PCA	Exam questions
Katharina	30.0 %	30.0 %	40.0 %	33 %
Ida	40.0 %	30.0 %	30.0 %	33 %
Cecilie	30.0 %	40.0 %	30.0 %	33 %

OCTOBER 4TH 2022

Overall problem of interest

The chosen data set holds measures of attributes that attempt to predict whether a patient is in danger of suffering from a heart attack. The dataset is called 'Heart Attack Analysis Prediction Dataset' and has been found on kaggle, link: https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset?resource=downloadselect=heart.csv

A detailed description of the data set

The data set consists of 11 attributes each with 300 records. When working with and making statistics of a data set it's important to know what each attribute represents.
Age is only counted in whole years, therefore the attribute has discrete values. The attribute type can be seen as ratio, because 0 years is interpreted as a person that hasn't lived for a year i.e. 0 years. Sex is a binary attribute as the data only has record from two genders which corresponds to the binary values, 0 and 1. Each gender can be seen as a category therefore the attribute type is nominal. The same goes for the attribute 'exng' which also can take the binary values 1 = 'yes' and 0 = 'no'. 'exng' is also nominal as yes or no is categories and does not refer to a specific number of time or severeness. caa can only be integers in [0,3] as half a vessel isn't possible, so therefore it is discrete. As 0 means physical absence of major vessels the attribute type is ratio. 'cp' is discrete as it only takes integers in [1,4]. In the dataframe it is [0,3] as it is zero indexed. As 'cp' refers to a category of chest pain type and not severeness on a scale, the attribute type is nominal. 'trtbps' can take continuous variables. If 'trtbps' is 0 it is interpreted as an absence of blood pressure therefore the attribute type is ratio. The same goes for 'chol', it' continous, and 0 means absence of cholestoral. 'fbs' is binary as it can only take the binary values 1=True and 0=False for the expression (fasting blood sugar $>$ 120 mg/dl). The two values can be seen as categories, therefore the attribute type is nominal. 'rest_ecg' can take integeres in [0,2] therefore it is discrete. Each value corresponds to a category of a result conclusion. Therefore the attribute type must be nominal. 'thalach' is maximum hear rate achieved. Heart rate is usually measured in beats pr. min. therefore it can be continuous. 'thalach' of 0 would mean a physical absence of heart rate therefore it is ratio. Finally we have the target attribute, which is binary values 0 and 1, which corresponds to not having risk of heart disease or having risk, and therefore its also nominal.

Attribute	**Explanation**	**Attribute type**
age	Age of the person	Discrete, ratio
sex	Gender of the person	Discrete, nominal
exng	exercise induced angina (1 = yes; 0 = no)	Discrete, nominal
caa	number of major vessels (0-3)	Discrete, ratio
cp	Chest Pain type chest pain type	Discrete, nominal
trtbps	resting blood pressure (in mm Hg)	Continuous, ratio
chol	cholestoral in mg/dl fetched via BMI sensor	Continuous, ratio
fbs	(fasting blood sugar $>$ 120 mg/dl) (1 = true; 0 = false)	Discrete, nominal
rest_ecg	resting electrocardiographic results	Discrete, nominal
thalach	maximum heart rate achieved	Continuous, ratio
target	Risk of heart disease	Discrete, nominal

Table 1: Description of attributes

Data quality

The dataset is of high quality due to no missing values and very representative attributes in order to predict the chance of each patient getting attacked by a heart attack. To be sure of no missing values and outliers in the attributes which are the nominal type we search for unique values in each of attributes.

By this it is found that all the data fall into the correct classes meaning no missing values or outliers is found. Nevertheless a bit noise and outliers in some attributes can't be avoided due to the fact that the data set holds data of 303 patients of very different personal details. For instance, 2 patients have measured a high resting blood pressure in the interval 189.40-200.00 mm Hg compared to the mean of all patients that is 132 mm Hg. However there is an unequal distribution of sex, as $sex = 1$ has double as many records compared to $sex = 0$. It's also not stated which numbers corresponds to which specific sex. The data set has features which seem to be evaluated to integers, these features are trtbps, chol, fbs, thalach. But it might not be a huge influence as the total number of records is 300 records.

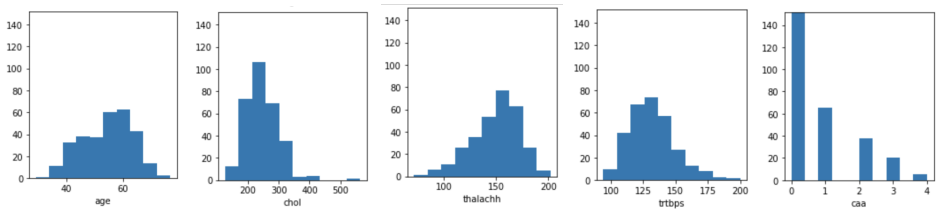
Empirical statistics of the dataset

It only makes sense to take the mean of the attribute of types ratio and interval. Therefore the 5 attributes considered is age, trtbps, chol, thalach and caa. In order to understand the distribution of the data we calculate equally probable intervals called the quantiles.
The quantiles of the attributes is written in the tabel below:

Attribute	q0.025	q0.25	q0.50	q0.75	q0.975
age	37.0	47.5	55.0	61.0	70.0
trtbps	102.0	120.0	130.0	140.0	172.9
chol	165.1	211.0	240.0	274.5	346.95
thalach	101.2	133.5	153.0	166.0	186.0
caa	00.0	00.0	00.0	1.0	3.0

Table 2: Quantiles of attributes

To determine if each attributes contains outliers a histogram of the distribution of values in each attribute are examined:



(a) Histogram for age *(b) Histogram for chol* *(c) Histogram for thalachh* *(d) Histogram for trtbps* *(e) Histogram for caa*

Figure 1: Histograms

For the attribute age (figure 1a) outliers can't be found because the wide histogram describes the large deviation in the age of the patients found in the data. In this case the larger the spread the better the coverage of a whole population in the sense of prediction heart attacks. The histogram of attribute cholesterol (figure 1b) exposes a small number of patients have a measured values of cholesterol 500-550 mg/dl. Eventually these observation can be considered as outliers, but due to the small distance from the largest "usual" values which is less than 450 mg/dl and the outliers in the interval 500-550 mg/dl, the observations can't be ignored. Lastly the histograms for resting blood pressure (trtbps), maximum heart rate achieved (thalachh) and number of major vessels (caa) (figures 1d, 1c and 1e) no outliers is found in the data as no observed values are in a large range from other values meaning that the data in each attribute is coherent if it is considered in slightly bigger intervals of values.

The histograms also revels whether the attributes appear to be normal distributed or not. As the histogram of attribute age approximately fits a bell curve, the attribute can be expected to be normal

distributed. The data of the attributes cholesterol (chol) and resting blood pressure (trtbps) is a skewed right distribution as the tail is longer to the left than to the right. Opposite if this is the histogram for maximum heart rate achieved (thalachh) which is a slightly skewed left distribution. The attribute of number of major vessels (caa) is clearly not normal distributed as the data can't be fitted by a bell curve. The quantiles also reveals that 75 % of the data is distributed in the first two groups ($caa = 0$ and $caa = 1$).

To make the distributions even more clear box plots of each attribute in the predicted class is plotted:

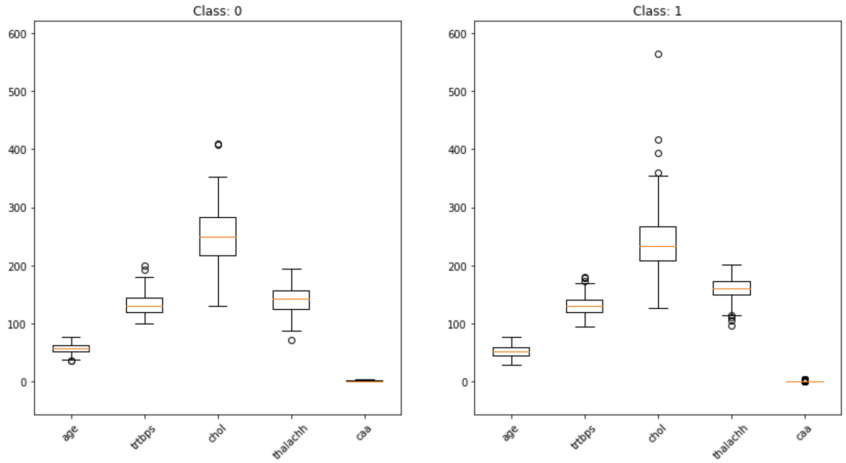


Figure 2: Box plot of attributes in predicted class

Examing the box plots in figure 2 it seems that attribute 'chol' is normal distributed with a small number of outliers detected in both predicted classes. In class 1 (the class that is equal to predicting suffering from a heart attack) the 3 other attributes 'age', 'thalachh' and 'trtbps' are also normal distributed as the data has equal proportions around the median. For class 0 (class of predicting not suffering from a heart attack) the attributes 'age' and 'thalachh' are more negatively skewed as the median is closer to the upper quartile, while 'trtbps' is positively skewed and the median is closer to the lower quartile.

If the box plot should actually be compared to table 2 we must combine the data of the predicted classes. The result is:

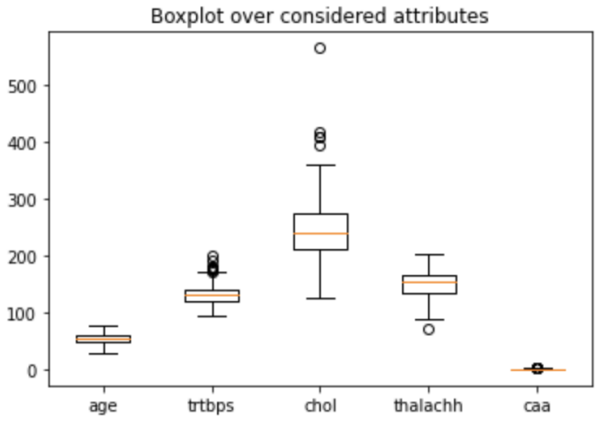


Figure 3: Box plot of attributes in data set

This visualization (figure 3) shows that the attribute 'thalachh' is negatively skewed equal to the values

in table 2 in which the difference between median (153.0) and upper quartile (166.0) is less than between the median and the lower quartile (133.5). The 3 other attributes considered in the statistics are very close to being normal distributed due to equal differences from median to both upper and lower quartiles. From the box plot for attribute caa it is hard to determine the distribution of the values but from the quantiles in table 2 it is clear that as the median is equal to the lower quartile the distribution is positively skewed.

Considering the correlation of the attributes when determining if the patient is in danger of suffering from a heart attack a confusion matrix is shown below:

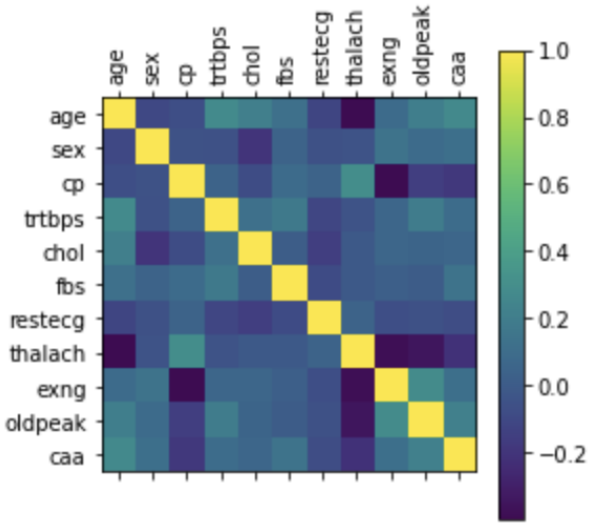


Figure 4: Confusion matrix

Examining the confusion matrix (figure 4) with the variance in the diagonal (meaning the attributes are perfectly correlated) and the correlation between 2 attributes in the rest of the matrix, we are able to see that most of the correlations takes values in range $[-0.2, 0.2]$ meaning that the attributes are either slightly negatively correlated or slightly positively correlated. The most negatively correlated attributes is the pair 'cp' - 'exng' and the pair 'age' - 'thalachh', while the more correlated pairs of attributes are 'trtbps' - 'age' and 'thalachh' - 'cp'. The negative correlation corresponds to inverse correlation meaning when one attribute increases, the other decreases. For the pair 'age' - 'thalachh', when the age of a patient is increasing the smaller the heart rate is achieved. Positive correlation means that the 2 attributes changes in the same direction which in this case is present for the pair 'trtbps' - 'age', so when the age of the patient is increasing, the resting blood pressure is increasing as well.

PCA

By following the PCA method the explained variance graph (graph 5) shows that over 90% variance is explained by the first 9 principal components. This is clear from the dashed straight line across the plot that holds the threshold equal to 0.90.

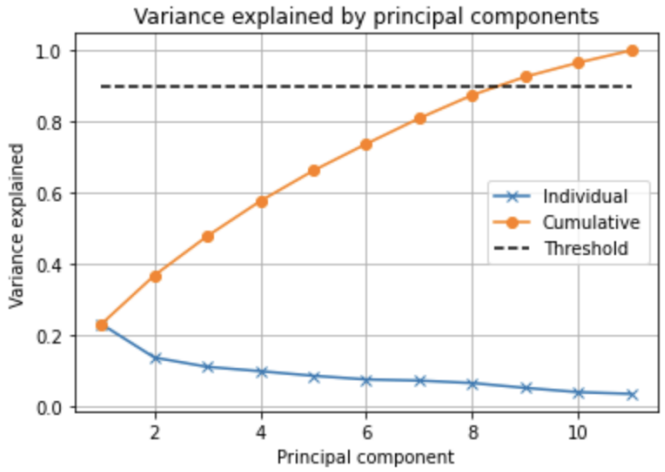


Figure 5: Variance explained by principal components

The calculation of explaned variance is done by using equation (3.18) from the book:

$$\text{Variance Explained } = \frac{\sum_{i=1}^{n} \sigma_i^2}{\sum_{i=1}^{M} \sigma_i^2}$$

As the data has been standardized a larger number of principal components is needed in order to capture most of the variance. This is due to the very different scale of the original attributes, for instance considering the attributes age and cp (chest pain type). Age takes 41 different values while cp can only take 4 different values. The standardization deals with this difference so the attributes are of equal variance.

To see the effect of each principal component on each attribute of the the data set figure 6 is made.

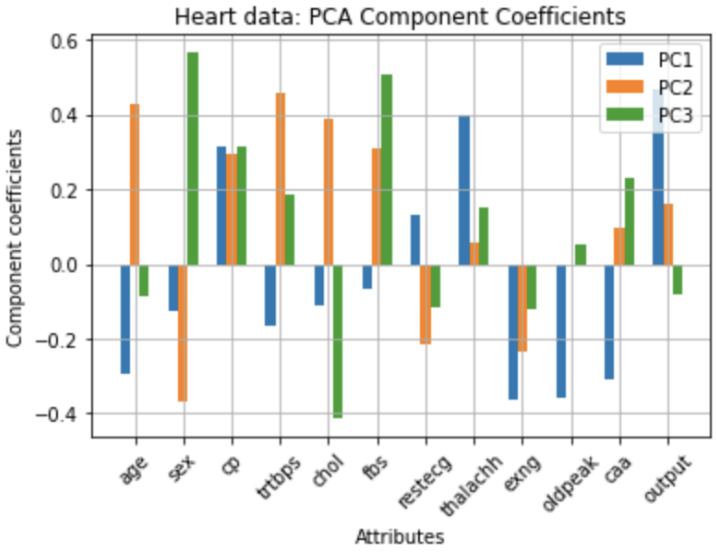


Figure 6: Projection of each attribute and 3 principal components.

This figure shows the direction of the principal components considered for each attribute in the data set. The variance of the attributes is captured differently among the 3 principal components. The first principal component holds most variance for 6 attributes. The second principal component holds most variance for 5 attributes and the third principal component holds most variance for 3 attributes. The figure thereby states that the variance of the attributes is very distributed across the 3 principal components.

In order to display the records in 2 dimensions and get the most information (and thereby highest variance) we consider the 2 first principal components. When the data is displayed in the coordinate system of the principal components, it is expected that the data is classified and grouped somewhat in the two classes (0 and 1). The data is plotted with the first principal component and the second principal component as the axes in the new coordinate system:

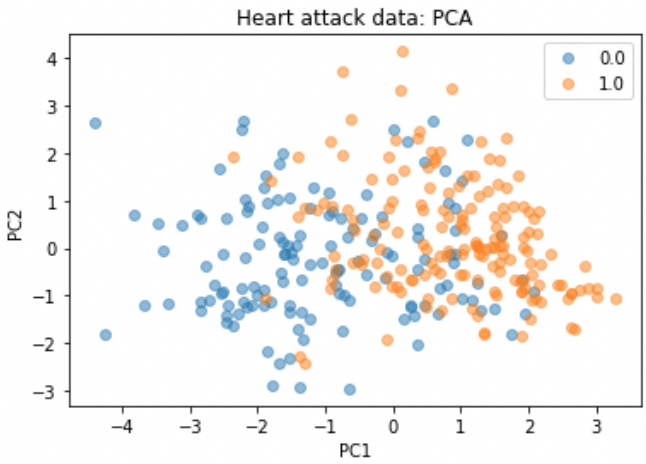


Figure 7

In figure 7 the classification has found that 1 is mostly to the right. Therefore by doing PCA it is possible to put high dimensional data onto only 2 dimensions while still keeping information. 2-dimensional data is also a lot easier to visualize, and fewer dimensions makes computations faster when classifying.

Discussion

We wanted to investigate the heart attack data set to predict if a patient is in danger of suffering from a heart attack. The data set consists of both discrete and continuous attributes. By applying PCA algorithm to the data set, we can conclude that 9 principal components keep 90% of the information in the data set. For visualization in 2 dimensions (figure 7) some division into the two categories we want to predict is visible. There is still a lot of overlapping in the middle which is not ideal for the classification problem. The data could be visualized in a coordinate system of 3 principal components, but it would be more difficult to comprehend, and more than two principal components should therefore only be used for computation purposes.

Exam exercises

Question 1. Spring 2019 question 1

A is the correct answer, as x1(time) of day is split into 30 minute intervals, which cannot be ranked as one time of day is not better than another. x2 and x7 are both a number of instances that are measured, where 0 would indicate no recorded instances. Therefore it is ratio. y is ordinal as it is ranked, but has no distance.

Question 2. Spring 2019 question 2

Considering the max-norm distance as the p-norm distance $d_{p=\infty}$ between the two observation we consider the maximal absolute value between corresponding attributes in the observations. This is described by the equation:

$$d_{p=\infty}(x, y) = max\{|x_1 - y_1|, |x_2 - y_2|, ..., |x_M - y_M|\} \tag{1}$$

The only attributes where an actual difference in values occur is for the first and third attribute, as every other attributes takes the value 0. The absolute value of these differences is:

$$|x_{14,1} - x_{18,1}| = |26 - 19| = 7 \tag{2}$$

$$|x_{14,3} - x_{18,3}| = |2 - 0| = 2 \tag{3}$$

The maximum p-norm distance is therefore $d_{p=\infty}(x_{14}, x_{18}) = 7$ and the answer is A.

Question 3. Spring 2019 question 3

We can find the principal components from the diagonal in the S matrix

$$\sigma_1 = 13.9,\ \sigma_2 = 12.47,\ \sigma_3 = 11.48,\ \sigma_4 = 10.03,\ \sigma_5 = 9.45$$

Calculated explained variance for each option
A) The explained variance of the first four principal components is 0.87
B) The explained variance by the last tree principal components is 0.48
C) The explained variance of the first two principal components is 0.52
D) The explained variance by the first three principal components is 0.72
Conclusion:
It must be option A since

$$\text{Variance Explained } = \frac{\sum_{i=1}^{4} \sigma_i^2}{\sum_{i=1}^{5} \sigma_i^2} = 0.87 > 0.8$$

Question 5. Spring 2019 question 14

We first make a bag of words matrix:

	bag	becomes	do	if	less	not	of	parsimoneous	representation	stem	the	we	words
s_1	1	1	0	0	1	0	1	1	1	0	1	0	1
s_2	0	0	1	1	0	1	0	0	0	1	1	1	1

To calculate jaccard we need to use the equation (4.23)

$$\mathrm{J}(\boldsymbol{x}, \boldsymbol{y}) = \frac{f_{11}}{f_{11} + f_{10} + f_{01}}$$

we found $f_{11} = 2$, $f_{10} = 6$, $f_{01} = 5$ so

$$\mathrm{J}(\boldsymbol{x}, \boldsymbol{y}) = 0.153846$$

Conclusion: Answer A

Question 6: Spring 2019 question 27

If we only considers the probabilities where $y = 2$ for the two possible outcomes giving $x_2 = 0$, the total probability of $p(x_2 = 0|y = 2)$ is calculated as the sum of the probabilities corresponding to both

attributes take the specific values. Therefore the answer is B because:

$$p(x_2 = 0|y = 2) = 0.81 + 0.03 = 0.84 \tag{4}$$