Danmarks Tekniske Universitet

DTU

02450 Introduction to Machine Learning and Data Mining

Heart disease

AUTHORS:

Maiken Jakobsen - s193892

Jaime Guzman - s222025

Thomas G. Ewers - s175394

Date: 04.10.2022

Preface

This report was written as a delivery in course 02450 Introduction to Machine Learning and Data Mining, Fall 22. All team members contributed equally to the execution of this report. The table below provides on overview of main responsibilities, though all team members have read and supplemented to all chapters.

Section	Responsible
Abstract	
Data Introduction	
1 - Introducing the heart disease data	Maiken
2 - Detailed explanation of the attributes	Jaime
Analysis	
3 - Summary statistics of the attributes	Maiken
4 - Principal Component Analysis	Jaime
Discussion	
5 - Discussion	Thomas
Exam Questions	
6 - Questions	Thomas

Table 1: Responsibility overview

Contents

1 Introducing the heart disease data

The data contains information about patients registered at a hospital in Cleveland. The data is used to determine which symptoms that can help classify if a patient is suffering from heart disease or not. The overall problem of interest of the dataset is to identify whether a patient suffers from heart disease or not.

The database that will be utilized in this project was provided to the center for machine learning and intelligent systems [1] on the 1^{st} of July 1988. The data is one of four databases where data is extrapolated from a general data file intro a processed file with clean data that is relevant to the prediction of hearth disease. The original data file contained 76 features and the processed data used for the rest of the report contains 14 attributes, which is explained in the following section.

Previous experiments have attempted to distinguish between the presence of heart disease (values greater than 0) and the absence of heart disease (value equal to 0).

In "International application of a new probability algorithm for the diagnosis of coronary artery disease. American Journal of Cardiology" [2] the authors used the data on heart disease to set up a model which predicts the number of patients who suffer from heart disease based on the 14 relevant attributes stated above. The model set up based on the Cleveland data is compared to data collected in Hungary, California and Switzerland.

The data from Cleveland is used as the training data to set up a model, the data from Hungary, California and Switzerland is used as the testing data to estimate the accuracy of the model. The probability derived using the model on the three sets of data was compared to the result of a Bayesian algorithm. Both the proposed model and the Bayesian algorithm overpredicted the probability of having heart disease in Hungary and in California. The same people found that coronary disease probabilities from discriminant functions are reliable and clinically useful when applied to patients with chest pain syndromes and intermediate disease prevalence.

We wish to experiment on transforming attributes using the conversion method one-out-of-k encoding. We are also curious to inspect which kind of attributes that have some correlation between suffering from heart disease and not suffering.

The regression part of the problem will be based on the attribute "num", which determines whether a patient is suffering from heart disease, or not.

The classification part of the problem will be based on the attribute "num", as the attribute

"num" shows different scenarios of heart disease affection.

The attribute "num" is a discrete nominal attribute, as the data does not describe how affected a patient is, if the patient's "num"-value is greater than 0, only that a heart disease is present.

For the regression part of the problem, we will be using the one-out-of-k encoding on attribute "num" to convert it into a binary nominal attribute.

For the classification part of the problem, the attribute "num" will be converted to a discrete nominal.

2 Detailed explanation of the attributes

The data contains 14 columns, which are the following:

- **age**: Integer number that represent the age of the sampled person. This is discrete and ratio.
- **sex**: Boolean value that represents the sex of the sampled person. 1 represents male while 0 represents female. This type of data is discrete and nominal.
- **cp**: Chest pain type. There are four types of pain:
  - 1: typical angina
  - 2: atypical angina
  - 3: non-anginal pain
  - 4: asymptomatic

  This type of data is considered discrete and nominal.
- **trestbps**: Resting blood pressure in mm Hg. The data is continuous and ratio.
- **chol**: Serum cholesterol in mg/dl. This variable is continuous and ratio.
- **fbs**: Boolean value representing if fasting blood sugar above 120 mg/dl happened. 1 = true, 0 = false. This variable is discrete and nominal.
- **restecg**: Integer value representing resting electrocardiographic results. There are 3 different values:

  - Normal
  - Having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
  - showing probable or definite left ventricular hypertrophy by Estes' criteria.

  The values in this variable are discrete and ordinal.

- **thalach**: Integer value that represents the maximum heart rate archived by patient. This variable is a discrete and ratio.
- **exang**: Boolean value representing if exercise induced angina was done. 1 = yes; 0 = no. This variable is discrete and nominal.
- **oldpeak**: Float value that represents ST depression induced by exercise relative to rest. This variable is continuous and ratio.
- **slope**: Integer value representing the slope of the peak exercise ST segment. There are 3 different values:
  - 1: upsloping
  - 2: flat
  - 3: downsloping

  This variable is discrete and nominal.
- **ca**: Float value representing the number of major vessels (0-3) colored by fluoroscopy. This variable is discrete and ordinal.
- **thal**: Integer value with the following values:
  - 3 = normal
  - 6 = fixed defect
  - 7 = reversible defect

  The values are discrete and ordinal.
- **num**: Represents the presence of heart disease. 0 represents no presence while values 1,2,3,4 represent different scenarios. This variable is discrete and nominal.

The data from the Cleveland data set contained 6 missing values out of the 303 data points. This is approximately 2% of the data, and these values are therefore scrapped, due to the low impact of doing this. We have one value, in "chol", that needs further analysis to determine if it is a corrupted data point.

3 Summary statistics of the attributes

In this section, the basic summary statistics will be described according to the data set used. Firstly the correlation between the 14 attributes are examined, afterwards the distribution of the attributes is analyzed.

To examine how the attributes are correlated, the 14 features are plotted in a correlation matrix plot as viewed on the figure below (Figure (1)).

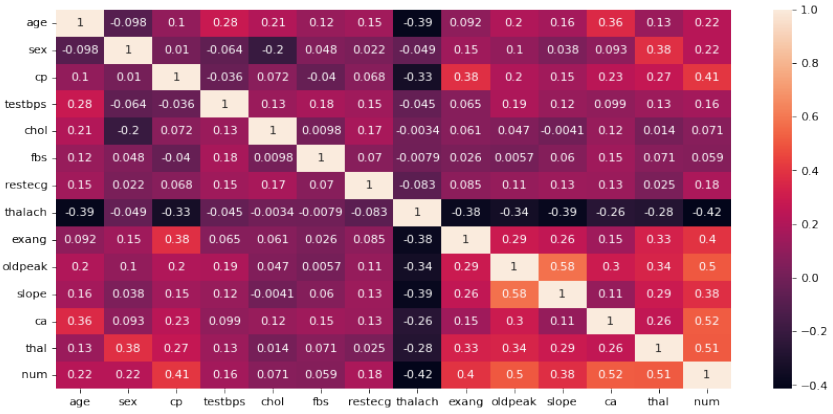

Figure 1: Correlation matrix

The bar on the right views how correlated the attributes are to each other. The closer the color of the cell is to either white or black, the higher a correlation there is between the two attributes. There is a high negative correlation between the attribute "thalach" and the following attributes: "age", "cp", "exang", "oldpeak", "slope", "ca", "thal" and the target attribute "num". There is a high positive correlation between the target variable "num" and "cp", "exang", "oldpeak", "slope", "ca" and "thal".

We expect the variables that a closely correlated to the target attribute "num" to have a higher impact on determining whether a patient suffers from heart disease or not, compared to using attributes that have a low correlation to "num".

5 out of the 14 attributes are binary features, these are "sex", "fbs", "Restecg", "exang" and "thal". The target attribute "num" is left out of the analysis, as we wish to predict the value of this variable based on the other features.

To investigate if the attributes have any outliers, the boxplot of the different attributes is set up (see figure (2)). Another purpose of the boxplot is to illustrate the mean value of each attribute and which intervals the attributes belongs to.

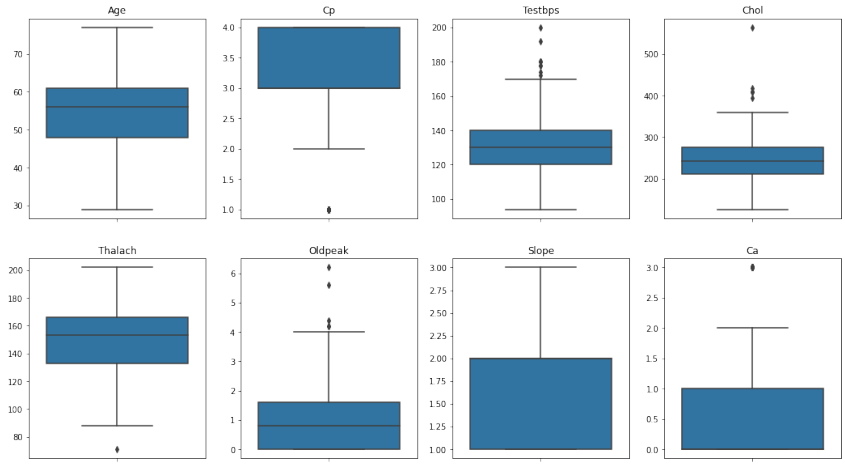

Figure 2: Boxplots of attributes

Looking at the boxplot above (figure (2)), something noticeable is that one of the values of "Chol" is very high, it is above 500. This observation is an extreme value. Some research on the topic of heart disease illness showed that it is possible but not healthy to have a cholesterol level above 500. The observation is therefore not removed from the further analysis.

On the figure below (figure (3)) the 8 non-binary attributes ("Age", "Cp", "Testbps", "Chol", "Thalach", "Oldpeak", "Slope" and "Ca") are plotted.

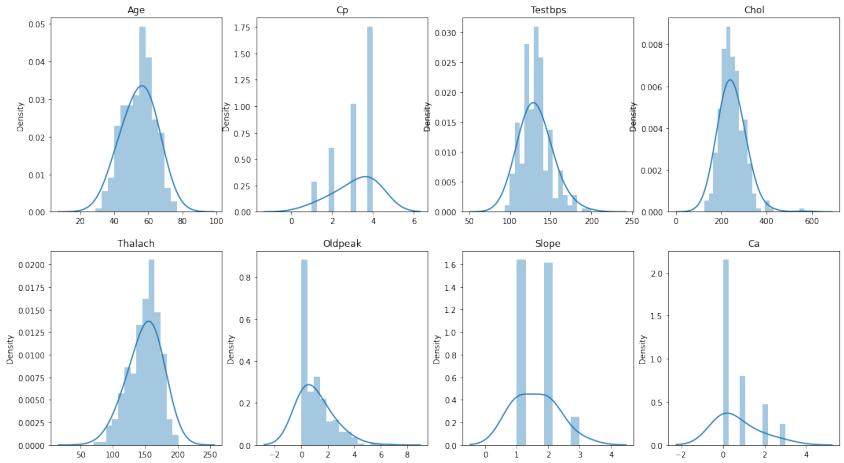

Figure 3: Variable distribution

The attributes which are normally distributed are "Age", "Testbps", "Chol" and "Thalach".

In accordance to our analysis and an initial implementation of a regression model, as well as analyzing the results from previous works, we can say that the data is not suitable for regression, as the highest accuracy achieved in previous works was around 77%, which is not enough to say the model gives satisfactory results. As for our rough implementation, we found an accuracy of no higher than 55% over many iterations and changes, with an MAE of 64.6%, an MSE of 74.6%, and a RMSE of 86.3%.

4 Principal Component Analysis

In this section, the principal component Analysis (PCA) of the heart disease data is explained.

Binary variables are not compatible with continuous variables, and these are therefore removed in the normalization of the data but will be included in the PCA analysis. 3 out of the 14 attributes are therefore removed from the PCA, these are "sex", "fbs", "exang". The target variable "num" is left out of the analysis.

The data is standardized using the mean and standard deviation of the data. The standardized data is afterwards run through the PCA algorithm to identify potential correlations in the data. The amount of variation explained as a function of the number of PCA components is depicted on the graph below (Figure 4).

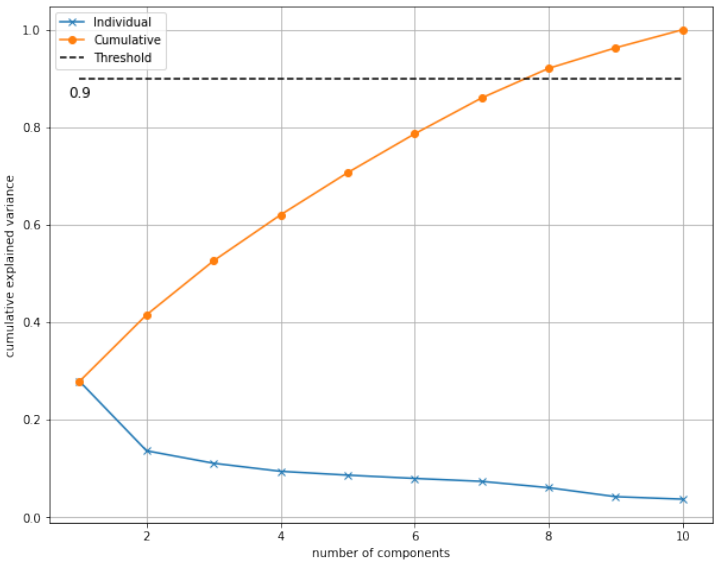

Figure 4: Variance explained by principal components

As seen from the graph above (figure (4)) there are 10 principal components. The figure views that the amount of principal components needed to satisfy a threshold value of at least 90% is 8 principal components.

The 10 attributes are plotted along with a unit circle to view the direction of the 2 principal components on the figure below (figure (5)).

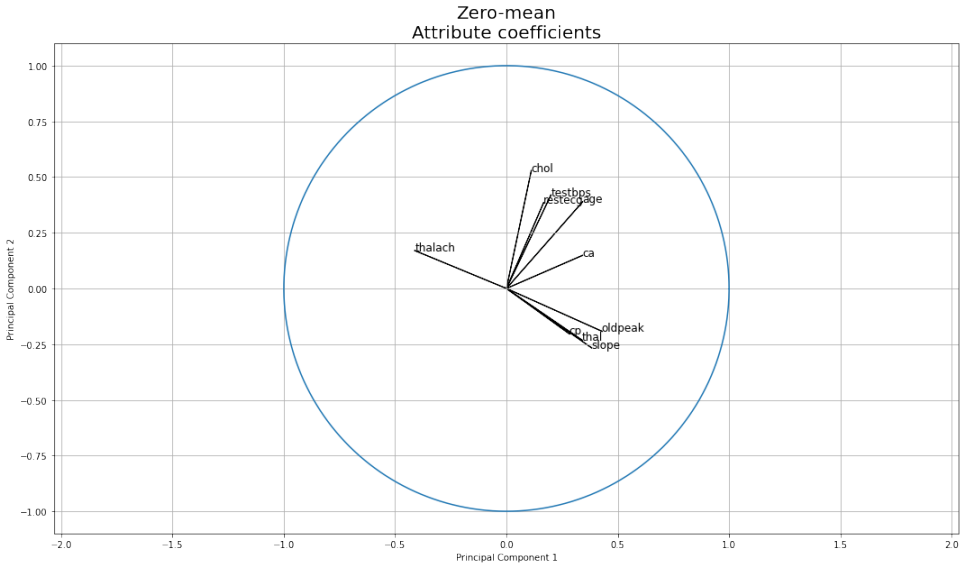

Figure 5: Direction of the two first principal components

On the plot above (figure (5)), the 10 attributes are plotted according to their respective directions along the first two principal components.

Seeing as it is impossible to make a plot in 8 dimensions, the first two principal components will be used in the following analysis. The data projected onto the two considered principal components are displayed on the figure (figure (6)).

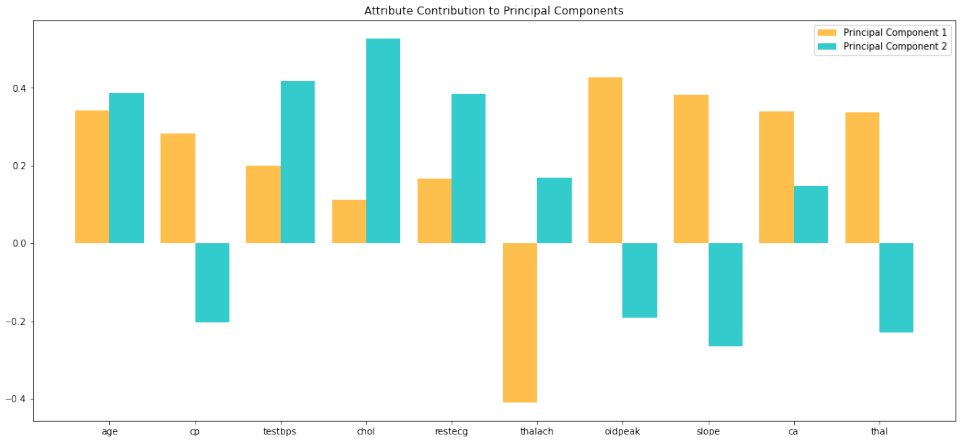

Figure 6: The attributes contributing to the first two PCs

The two principal components are explained by opposite attributes, meaning, the first principal component is affected by almost every other attribute than attribute "testbps", "chol" and "restecg" (corresponding to attribute number 3,4 and 5). The second principal component is mostly affected by attribute "testbps", "chol" and "restecg" (3,4 and 5).

5 Discussion

The 14 attributes from the Cleveland database, had very few missing or corrupted values, which made is easy to manage and transform as the missing values could just be disregarded, as they would have little influence in the overall data processing. By constructing the correlation matrix, we can see that some attributes are more correlated with having heart disease than others, namely "CA", "Thalach" and "Oldpeak". When doing the PCA analysis, we see on figure 4, 8 of the principal components are needed to retain at least 90% of the variance of the data. This indicates, that determining if a patient suffers from heart disease, requires a lot of attributes, and that these attributes aren't heavily correlated. So if too few principal components are chosen, the loss in data is too significant. This is also made clear in figure (5), where it is shown, that the first two principal components,

are influenced by almost every attribute. This is also supported by the bar-plot in figure (6), where the first principal component is mostly influenced by "TestBPS", "Chol" and "RestECG". When looking at the correlation between these attributes, the correlation between them are all below 0.2. This is why so little variance is retained, with only one principal component. Finally, we can also see that a regression model would not be a good fit for this data and what is expected of it, since the accuracy it can produce is found to be around 77% in previous works, and 55% on a rough implementation of our own.

6 Exam Questions

6.1 Question 1

In Question 1 we choose **option C**. This is because x_1 is an ordered variable as a value of $x_1 = 1$ it corresponds to 7:00-7:30 and $x_1 = 27$ corresponds to 20:00-20:30, which makes it an ordinal variable as there is a clear definition of the value of x_1 and this is what distinct it from the other options.

6.2 Question 2

Using the p-norms from chapter 3 we identify the correct answer is **A**. To check if answer A is correct, we use equation (1).

$$p_\infty = \max\{|x_1 - y_1|, |x_2 - y_2|, \ldots, |x_n - y_n|\} \tag{1}$$

To check the answers of B, C and D we use equation (2).

$$||x||_p = (|x_1|^p + |x_2|^p + \cdots + |x_M|^p)^{1/p} \tag{2}$$

When inserting the values of $x_{14,i}$ and $x_{18,i}$ in equation (1) the value obtained is:

$$p_\infty = \max\{|26 - 19|, \ldots, |2 - 0|, \ldots\} = 7$$

6.3 Question 3

The variance explained by the principal components can be estimated using the diagonal of the S matrix. The diagonal of the S matrix contains the variance of the principal com-

ponents. The equation below (equation (3)) demonstrates how the explained variance is computed.

$$\text{Explained variance of PC}_{i:j} = \frac{\sigma_{i:j}^2}{\sum_i^N \sigma_i^2} \tag{3}$$

where the interval of the nominator i:j is the number of principal components examined. Doing so, A is found as the correct answer.

6.4 Question 5

We used the following equation to solve this question.

$$J(x,y) = \frac{f_{11}}{K - f_{00}} \tag{4}$$

Where $K = f_{11} + f_{10} + f_{01} + f_{00}$ and f_{11} are the binaries of values that appear on both texts, f_{00} are values that do not appear on any of the list, which is 0 since there are none.

	the	bag	of	words	represen tation	becomes	less	parsi monious	if	we	do	not	stem
s_1	1	1	1	1	1	1	1	1	0	0	0	0	0
s_2	1	0	0	1	0	0	0	0	1	1	1	1	1

Using (4) with our values we get

$$J(s_1, s_2) = \frac{2}{13} = 0.1538461538$$

Therefore we can conclude that the answer to this question is **A**.

6.5 Question 6

We choose **Option B**. We come to this conclusion, because we are asked to find the probability of $\hat{x}_2 = 0$ when given that $y = 2$. Because the circumstance is just as $y = 2$, we don't need the prior probabilities of the scenarios of y, so we can just look in the table and find

that:

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

References

[1] David W. Aha. *UCI Machine Learning Repository: Heart Disease Data Set*. `https://archive.ics.uci.edu/ml/datasets/heart+disease`. (Accessed on 08/30/2022). 2022 8.

[2] R. Detrano. *International application of a new probability algorithm for the diagnosis of coronary artery disease. American Journal of Cardiology*. 1989. URL: `https://pubmed.ncbi.nlm.nih.gov/2756873/`.