# Technical University of Denmark

02450 Introduction to Machine Learning and Data Mining

---

Report 1

---

Egill Ingi Jacobsen s172759
Sigurbjorn Jonsson s172581

Date of submission: February 27th, 2018

# Contents

# 1 Introduction

The chosen data set is called Poker Hand and is from the UC Irvine Machine Learning Repository[1]. Each record is an example of a hand consisting of five playing cards drawn from a standard deck of 52 and a description of what type of Poker Hand it represents. The order of cards is important, which is why there are 480 possible Royal Flush hands as compared to 4. From this data a machine can learn to recognize a poker hand and possibly determine the most likely winner in a poker game from the cards that have been dealt. We believe that all 10 attributes are needed to implement machine learning on the data set because only one more card can have significant effect on a poker hand. We hope that we can use a subset of the attributes, such as suit and rank of 2, 3 or 4 cards to make some predictions about who the winner will be or what the winning chances of each hand is.

# 2 The Data Set

The Poker Hand data set consists of 1,025,010 rows of data with no missing values or corrupted data. The data is split into a set for training purposes (25,010 rows) and one for testing purposes (1,000,000 rows). We decided to do our data analysis on the training set, and then use the testing set to validate the efficiency of our machine learning tasks later on. Each row in the data set contains 11 attributes about a poker hand. 10 attributes belong to the cards in the hand and one attribute classifies what kind of hand it is we are dealing with. These attributes are:

## 2.1 Dataset description

1. Attribute 1 is the suit of card number 1. This attribute is as discrete nominal interval ranging from 1 to 4.

2. Attribute 2 is the rank of card number 1, i.e. is it an ace, a deuce, etc. This attribute is a discrete ordinal interval ranging from 1 to 13.

3. Attribute 3 is for the suit of card number 2. This attribute is a discrete nominal interval ranging from 1 to 4.

4. Attribute 4 is the rank of card number 2, i.e. is it an ace, a deuce, etc. This attribute is a discrete ordinal interval ranging from 1 to 13.

5. Attribute 5 is for the suit of card number 3. This attribute is a discrete nominal interval ranging from 1 to 4.

6. Attribute 6 is the rank of card number 3, i.e. is it an ace, a deuce, etc. This attribute is a discrete ordinal interval ranging from 1 to 13.

7. Attribute 7 is for the suit of card number 4. This attribute is a discrete nominal interval ranging from 1 to 4.

8. Attribute 8 is the rank of card number 4, i.e. is it an ace, a deuce, etc. This attribute is a discrete ordinal interval ranging from 1 to 13.

9. Attribute 9 is for the suit of card number 5. This attribute is a discrete nominal interval ranging from 1 to 4.

10. Attribute 10 is the rank of card number 5, i.e. is it an ace, a deuce, etc. This attribute is a discrete ordinal interval ranging from 1 to 13.

11. Attribute 11 is what hand the 5 card make. This is a discrete nominal interval. It is encoded into a number ranging from 0 to 9. The encoding from a numeric value to a poker hand, and back is shown in table 1.

| **Value** | **Hand** | **Description** |
|---|---|---|
| 0 | Nothing | Not recognized as a poker hand. |
| 1 | One pair | Two cards in the hand are of equal ranks. |
| 2 | Two pairs | The hand consists of two one pairs of different ranks. |
| 3 | Three of a kind | Three cards in the hand are of equal ranks. |
| 4 | Straight | All five cards are sequentially ranked with no gaps. |
| 5 | Flush | All five cards are of the same suit. |
| 6 | Full house | The hand consists of one pair and a different ranked three of a kind. |
| 7 | Four of a kind | Four cards in the hand are of the same rank. |
| 8 | Straight flush | The hand is both a straight and a flush. |
| 9 | Royal flush | The hand is a flush and the straight Ten, Jack, Queen, King, Ace. |

Table 1: The encoding used to map a poker hand to a numerical value, and vice versa.

Variables 1, 3, 5, 7 and 9 are all of the same type. They are nominal ratios where each number represents a suit of card, i.e. the card suits have been encoded into numeric values. The variable is a nominal ratio because in poker one suit is not superior to another and there is no clear definition of the zero point. In other words it does not matter if for example the suit Hearts is number 1, 2, 3 or 4. Table 2 shows the encoding used to convert a suit of card into a numeric value.

| **Value** | **Suit** |
|---|---|
| 1 | Hearts |
| 2 | Spade |
| 3 | Diamond |
| 4 | Clubs |

Table 2: The encoding used to map the suit of a card to a numerical value, and vice versa.

Variables 2, 4, 6, 8 and 10 are also all of the same type. They all represent a rank of a card and are ordinal ratios. **They are ordinal because one rank can be said to be higher than another. A rank of 2 is the lowest and a rank of 1 is the highest.**

They are ratios because there is not a clearly defined zero, i.e. if we were to make a new deck of cards there is nothing that is forcing us to start at 1 and go from there. Table 3 shows how the rank of the cards were mapped to a numeric value.

| **Value** | **Rank** |
|---|---|
| 1 | Ace |
| 2 | 2 |
| 3 | 3 |
| 4 | 4 |
| 5 | 5 |
| 6 | 6 |
| 7 | 7 |
| 8 | 8 |
| 9 | 9 |
| 10 | 10 |
| 11 | Jack |
| 12 | Queen |
| 13 | King |

Table 3: The encoding used to map the rank of a card to a numerical value, and vice versa.

## 2.2 Attribute statistics

The mean and standard deviation of the data is calculated as

$$\mathbf{X_{mean}} = [\ 2.51 \quad 7.00 \quad 2.50 \quad 7.01 \quad 2.51 \quad 7.01 \quad 2.50 \quad 6.94 \quad 2.50 \quad 6.96\ ]$$

and

$$\mathbf{X_{std}} = [\ 1.12 \quad 3.75 \quad 1.12 \quad 3.77 \quad 1.12 \quad 3.74 \quad 1.12 \quad 3.75 \quad 1.12 \quad 3.74\ ].$$

The median and range are

$$\mathbf{X_{median}} = [\ 3.00 \quad 7.00 \quad 2.00 \quad 7.00 \quad 3.00 \quad 7.00 \quad 2.00 \quad 7.00 \quad 3.00 \quad 7.00\ ]$$

and

$$\mathbf{X_{range}} = [\ 3.00 \quad 12.00 \quad 3.00 \quad 12.00 \quad 3.00 \quad 12.00 \quad 3.00 \quad 12.00 \quad 3.00 \quad 12.00\ ]$$

.

From the above statistical attributes we see that there is little to no variation when attributes 1, 3, 5, 7 and 9 are compared. The same is true when attributes 2, 4, 6, 8 and 10 are compared. This was to be expected because those attributes are describing the same property for different cards. The correlation matrix is obtained by calculating the

correlation between every attributes and is listed in equation (1).

$$\rho = \begin{bmatrix} 1.00 & -0.01 & -0.02 & 0.01 & -0.02 & -0.01 & -0.02 & 0.00 & -0.02 & 0.01 \\ -0.01 & 1.00 & 0.00 & -0.01 & -0.00 & -0.03 & 0.00 & -0.01 & 0.00 & -0.02 \\ -0.02 & 0.00 & 1.00 & -0.00 & -0.03 & -0.01 & -0.02 & 0.01 & -0.01 & 0.01 \\ 0.01 & -0.01 & -0.00 & 1.00 & -0.01 & -0.02 & -0.01 & -0.01 & 0.00 & -0.02 \\ -0.02 & -0.00 & -0.03 & -0.01 & 1.00 & 0.02 & -0.01 & 0.00 & -0.03 & -0.00 \\ -0.01 & -0.03 & -0.01 & -0.02 & 0.02 & 1.00 & -0.00 & -0.02 & 0.00 & -0.01 \\ -0.02 & 0.00 & -0.02 & -0.01 & -0.01 & -0.00 & 1.00 & -0.01 & -0.02 & 0.01 \\ 0.00 & -0.01 & 0.01 & -0.01 & 0.00 & -0.02 & -0.01 & 1.00 & 0.00 & -0.01 \\ -0.02 & 0.00 & -0.01 & 0.00 & -0.03 & 0.00 & -0.02 & 0.00 & 1.00 & -0.00 \\ 0.01 & -0.02 & 0.01 & -0.02 & -0.00 & -0.01 & 0.01 & -0.01 & -0.00 & 1.00 \end{bmatrix} \tag{1}$$

By examining the matrix it can be seen that there is very little to none correlation between the attributes and we can in fact say that each of the attributes is independent of each other.

# 3 Data visualization and Principal Component Analysis

Looking at the distribution of each variable independently, as is shown on figure 1, we see that every attribute is uniformly distributed. We expected this since a deck of card contains an equal amount of every suit and an equal amount of every rank. From the same figure outlier detection can also be evaluated. We can see no outlier in the data.

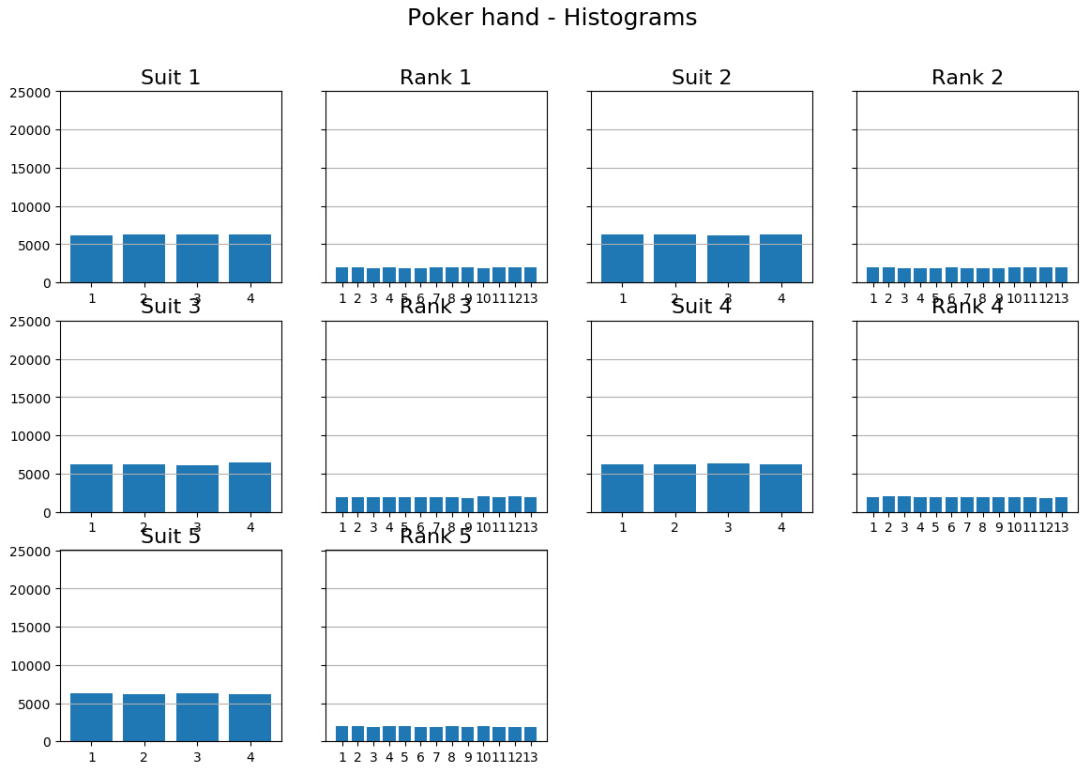


Figure 1: The distribution of each attribute in the form of a histogram.

This is further established whith the boxplot on figure 2, because f any outliers are in the data they would lie outside of the whiskers on the figure.

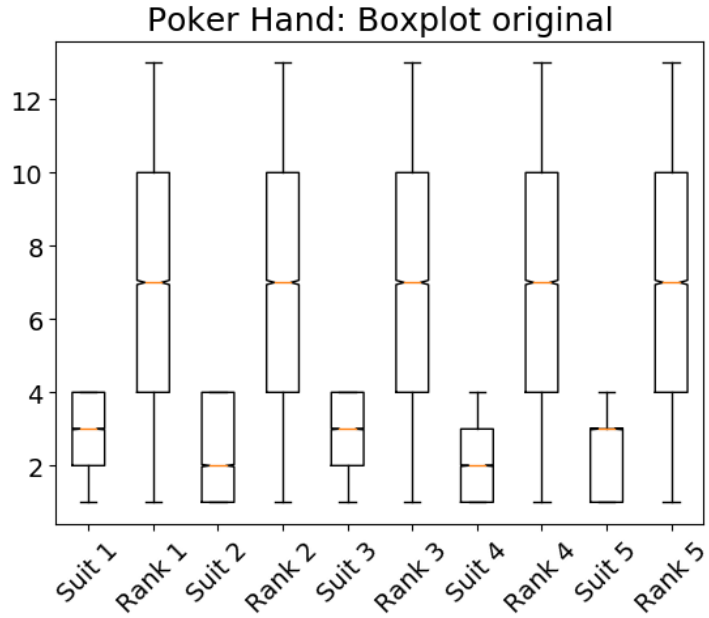


Figure 2: Boxplots for each attribute in the data set. The boxplots show the median, quartiles and max/min of the data attributes.

A Principal Component Analysis (PCA) was performed using Singular Value Decompo-

sition (SVA). From the SVA each Principal Component (PC) is obtained along with how much variance each principal compenent can account for. The variance explained by each principal component is shown on figure 3.

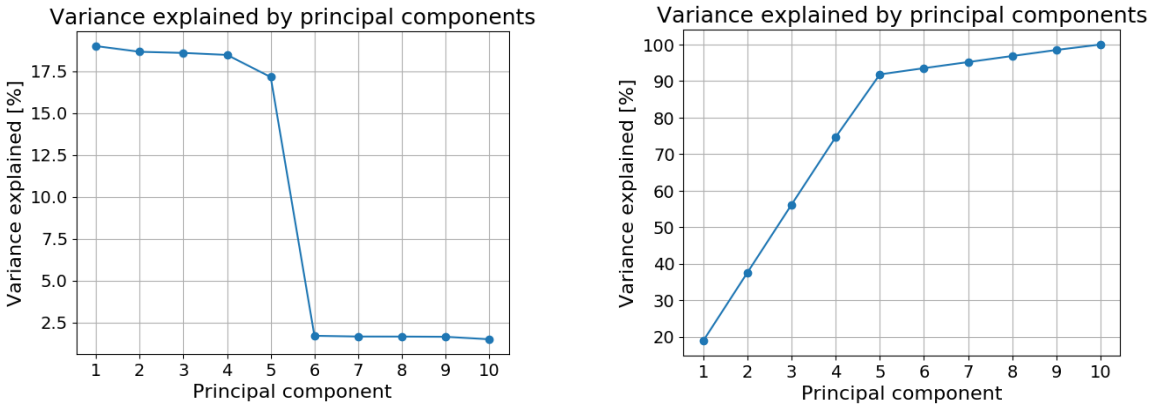


Figure 3: The variance (left) and cumulative variance (right) explained by the ten principal components, starting at principal component 1.

We see from figure 3 that more than 90% of the variance is explained by the first five principal components. We have decided to focus on all principal components because it isn't until we have them all that we have accounted for a variance of more than 99%.

The principal components are listed in equations (2) to (11).

$$PC1 = [\ 0.00 \quad 0.41 \quad 0.00 \quad 0.61 \quad -0.00 \quad -0.63 \quad -0.00 \quad 0.01 \quad 0.00 \quad -0.26\ ]^T \tag{2}$$

$$PC2 = [\ -0.00 \quad 0.70 \quad 0.00 \quad -0.68 \quad 0.00 \quad -0.18 \quad 0.00 \quad 0.11 \quad 0.00 \quad -0.02\ ]^T \tag{3}$$

$$PC3 = [\ 0.00 \quad -0.26 \quad 0.00 \quad -0.09 \quad 0.00 \quad -0.09 \quad -0.00 \quad 0.88 \quad 0.00 \quad -0.37\ ]^T \tag{4}$$

$$PC4 = [\ -0.00 \quad 0.22 \quad -0.00 \quad 0.07 \quad 0.00 \quad 0.54 \quad -0.00 \quad -0.21 \quad 0.00 \quad -0.78\ ]^T \tag{5}$$

$$PC5 = [\ -0.00 \quad 0.47 \quad 0.00 \quad 0.40 \quad 0.00 \quad 0.52 \quad -0.00 \quad 0.41 \quad 0.00 \quad 0.42\ ]^T \tag{6}$$

$$PC6 = [\ 0.13 \quad 0.00 \quad -0.53 \quad 0.00 \quad 0.68 \quad -0.00 \quad 0.17 \quad 0.00 \quad -0.46 \quad 0.00\ ]^T \tag{7}$$

$$PC7 = [\ 0.82 \quad 0.00 \quad 0.08 \quad -0.00 \quad -0.34 \quad 0.01 \quad -0.16 \quad -0.00 \quad -0.42 \quad -0.00\ ]^T \tag{8}$$

$$PC8 = [\ 0.23 \quad 0.00 \quad -0.70 \quad -0.00 \quad -0.34 \quad 0.00 \quad 0.31 \quad 0.00 \quad 0.49 \quad 0.00\ ]^T \tag{9}$$

$$PC9 = [\ 0.19 \quad 0.00 \quad -0.17 \quad -0.00 \quad 0.31 \quad -0.00 \quad -0.83 \quad -0.00 \quad 0.40 \quad 0.00\ ]^T \tag{10}$$

$$PC10 = [\ -0.47 \quad 0.00 \quad -0.43 \quad -0.00 \quad -0.46 \quad 0.00 \quad -0.41 \quad 0.00 \quad -0.46 \quad 0.00\ ]^T \tag{11}$$

Looking at the PCA we can identify the principal directions of each principal component as the indices in the PC that has the highest absolute values. Table 4 shows the principal direction of every one of the ten principal components.

| **Principal component** | **Principal direction** |
|---|---|
| PC 1 | Rank of cards number 2 and 3 |
| PC 2 | Rank of card number 1 and 2 |
| PC 3 | Rank of card number 4 |
| PC 4 | Rank of card number 5 |
| PC 5 | Rank of every card in the hand |
| PC 6 | Suit of card number 3 |
| PC 7 | Suit of card number 1 |
| PC 8 | Suit of card number 2 |
| PC 9 | Suit of card number 4 |
| PC 10 | Suit of every card in the hand |

Table 4: The principal direction of each principal component.

Figure 4 shows the original data plotted as an image on the top, and the data projected onto the considered principal components on the bottom. Comparing the two images we see that the first 5 principal components are similar to the rank attributes in the original data, and the last 5 components are similar to the suit attributes in the original data. This is consistent with our PCA and the principal directions listed in table 4.

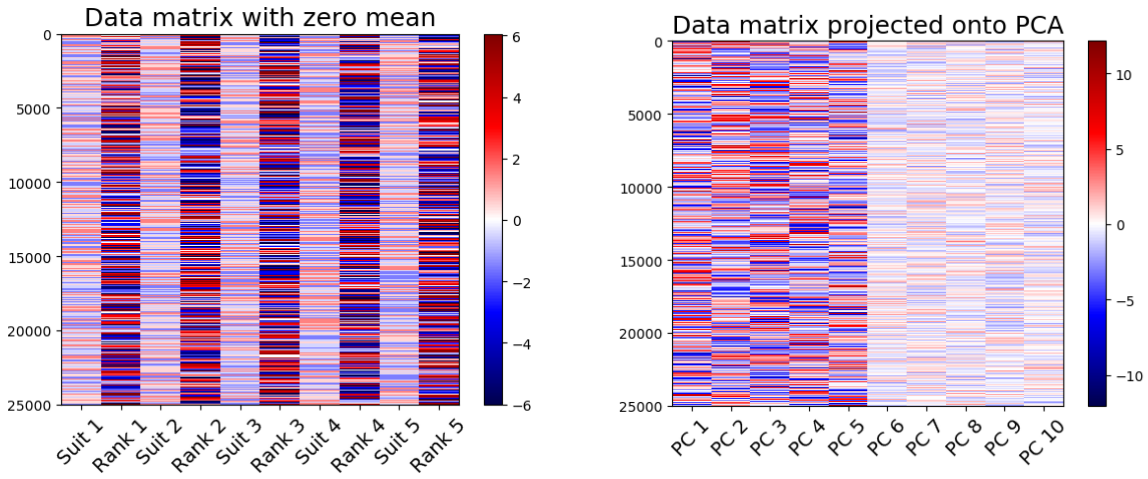


Figure 4: Visualization of (upper) the original data and (lower) the data projected onto the considered principal components.

Figure 5 shows the projected data in parallel coordinates. We see there that many of the poker hand classes look similar after the projection, especially the top row of the figure. We also see there some poker hand classes distinguish them self, like the flush, straight flush and royal flush that have zero value in principal components 6, 7, 8 and 9.

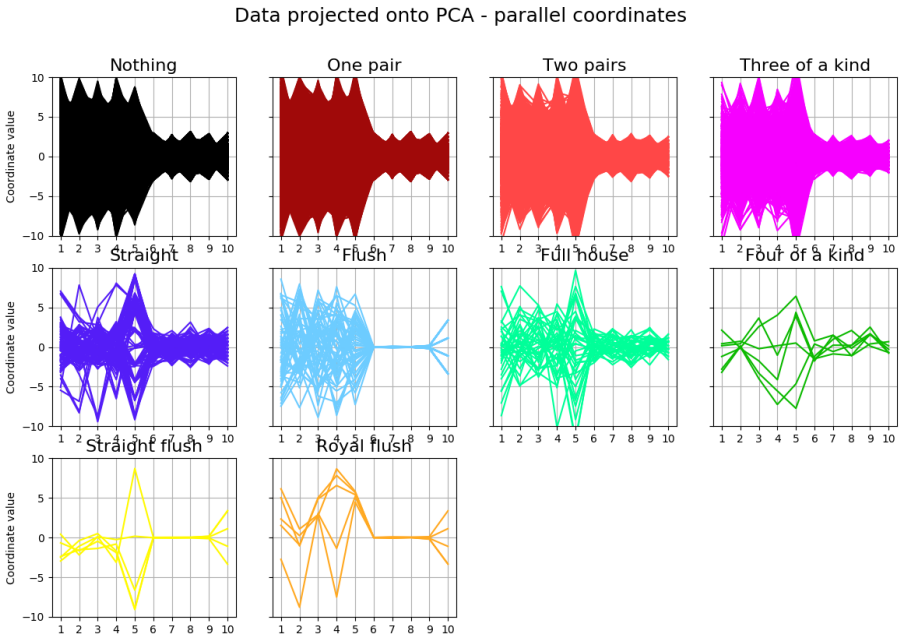


Figure 5: Data projected onto the considered principal components shown in parallel coordinates. The x-axis on each figure is for the principal components in order from 1-10, and the y-axis is the value of the coordinate the projected data takes on each principal component.

# 4 Conclusion

From our analysis and visualizations above we have learned that all of our attributes are independent of each other, i.e. the rank and suit of every card is independent of what other cards in the hand are. The rank is also independent of the suit of the same card. We also learned that our data is error and outlier free. This was expected since each of our attributes only has a predefined number of values it can take. From the principal component analysis we learned that the rank of the five cards in the hand account for more than 90% of the variance in the data, i.e. the ranks of the cards is far more important than the suit of the cards. This can possibly be related to the fact that in poker there are only two hands (flush and royal flush) where the suit of the cards is important. This was also reaffirmed by comparing a matrix plot of the data set with a matrix plot of the data projected onto the PCA.

Looking at figure 5 we see that machine learning tasks should be able to distinguish between at least some of the poker hands. We see that Nothing, One pair, Two pairs and Three of a kind look very similar, but possibly there are some subtle differences that we can not identify with the naked eye.

# A Collaboration

We approached this project by making all of the code required for the report together in a pair programming way, i.e. we worked on one computer and took turnes writing code while the other one watched and pitched in on how to solve the problems before us.

After the code was done and we were satisfied with how our visualizations looked and that we had done a thorough analysis we split into two independent groups, each responsible for writing 2 chapters in the report. Sigurbjörn Jónsson was responsible for writing chapters 1 and 2, while Egill Ingi Jacobsen was responsible for chapters 3 and 4. After each of us was done with our initial writing of our chapters we read over each other work and made comments and suggestions on what we thought should be changed or added. After discussing those points and agreeing on what should be done in each case, we went back and made the relevant modifications to the report. This process was repeted until we were satisfied with our report as a whole.

# References

[1] "Robert Cattral" and "Franz Oppacher". Poker hand data set. `http://archive.ics.uci.edu/ml/datasets/Poker+Hand`, 01 2007.