Abgabedokument Exercise 2
Einfhrung in die Mustererkennung 186.840 WS 2013

David Pfahler, Matthias Gusenbauer, Matthias Vigele
1126287, 1125577, 1126171

December 15, 2013

1 Wine Classification - k-NN

This section presents the classification of different types of wines using the k-NN classification algorithm. The dataset [1] contains the results of a chemical analysis of wines made in the same region in Italy but derived from three different cultivars. That means we have a three class problem. Section 1.1 presents a method, how to find the best features for the classification. Section 1.2 shows how to separate the data into a training and a test set. Section 1.3 presents the performance of the classification and Section 1.4 evaluates it.

1.1 Feature Extraction

In a chemical analysis of the wines, the following attributes were extracted:

1. Alcohol
2. Malic acid
3. Ash
4. Alcalinity of ash
5. Magnesium
6. Total phenols
7. Flavanoids
8. Nonflavanoid phenols
9. Proanthocyanins
10. Color intensity
11. Hue

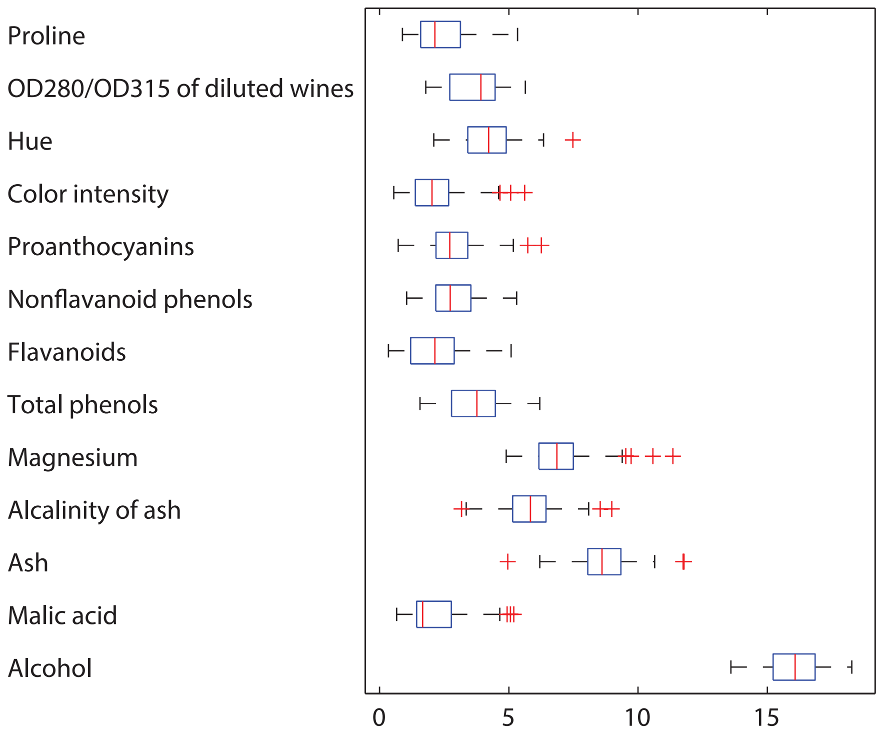


Figure 1: Boxplots of all features of the dataset

12. OD280/OD315 of diluted wines
13. Proline

This features vector could be represented by a 13-dimensional coordinate system. But due to the rule of thumb (1)presented in the lecture this are too many features for the size of the training set.

$$\frac{n}{d} > 10 \tag{1}$$

This rule of thumb says that the size n of the training set divided by the number of features d should be greater than 10.

To reduce the number of features we analyzed the data. Figure 1 shows a boxplot for each feature of the dataset. But first the dataset was standardized by dividing each feature by its standard deviation.

After that we decided to observe the dataset with the Principal Component Analysis (PCA) Figure 2 show a scatterplot of the dataset with its two first principal components and the three classes are represented with different colors. This figure shows that it is possible to separate the three classes into three clusters.

To find this clusters we looked at the main influences of the features to the first three principal components with an interactive `MATLAB` tool shown in Figure 3. After some tests we found out that the best feature vector for the classification contains:

- **Flavanoids**
- **Proline**

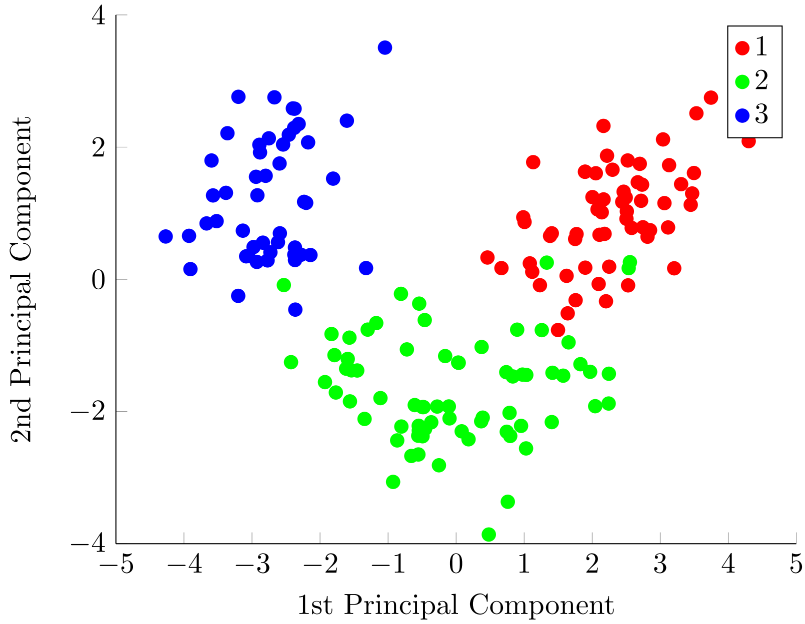


Figure 2: Scatterplot of the dataset with principal components

- **Color Intensity**

1.2 Test and Training Set

To determine a test and a training set we divided the smallest class (class 3) into two parts by a threshold. In our case the threshold is 0.5, which means that the class with 48 samples gets divided into two 24 sample sets (a test set and a training set). All other classes are divided into a set with the same size as the previously calculated training set size and the rest gets into the test set.

To enhance the accuracy and reliability of the classification algorithm we haven't used just one training set, but we randomly shuffled the dataset to use 30 different combinations of training and test sets. With every training and test set we computed a new classification.

1.3 Classification Performance

Figure 4 shows the best result of the k-NN classification with one of the test and training sets. But not all results are that good. Figure 5 presents the performance of all 30 different test and training sets and their used k. Also the Median of the results is shown, which is $91,5094\%$ and the Median of the k is 11. But the median itself gives not a good representation of the derivation of the data. Figure 6 shows this information with boxplots.

1.4 Evaluation of the Results

Our results show that the k-NN classifier is a powerful tool to classify the given data. But with a low value of k the classification error enhances. Figure 7 shows the median of the classification error rates of all 30 test sets for different values of k. The smaller the error rate, the better the performance. Like we mentioned above the minimum is at a $k = 11$.

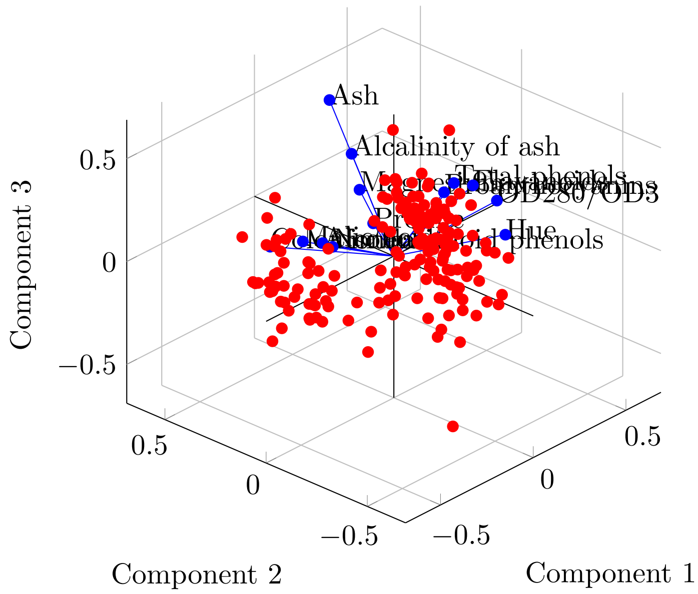


Figure 3: The main influences of the features to the first three principal components represented in an interactive 3D-graph

2 Wine Classification - Mahalanobis Distance

In the year 1936 Prasanta Chandra Mahalanobis introduced the Mahalanobis Distance which is a statistical measurement of distances between data points from a sample point. Mahalanobis Distance measures the similarity of samples with a known data set. Mathematically it is a measurement tool that is scale invariant and pays attention to data correlations.

2.1 Test and Training Set

The Test- and Training Sets for the classification with Mahalanobis Distance are equal to the sets used for the k-NN classification to create the same data basis. With an equal data basis it is easier to directly compare the results to each other. As state before the smallest class is divided by a threshold and then the same amount of data is taken from the other classes to get a training set. The remaining data samples not in the training sets are then accumulated into a single testset. This procedure is repeated 30 times to get a descriptive result of high quality.

2.2 Classification Performance

It was found that with our seed for trainings set randomisation the performance of the Mahalanobis Distance Classifier got worse if a full covariance matrix was used during distance calculation. Interestingly the simpler the covariance matrix got the more accurate the results were. The code was double checked and even a test with the built in Matlab function($mahal()$) was conducted. The results remained the same and it is not possible to fully explain these results. The figures 8, 9, 10 and 11 will show the performance of the Mahalanobis Distance classifier. Another advantage of the Mahalanobis Distance classifier over kNN is the runtime performance. Compared to kNN the calculations done for the Mahalanobis classifier are fast which needs to be considered for the choice between kNN and the Mahalanobis Distance.

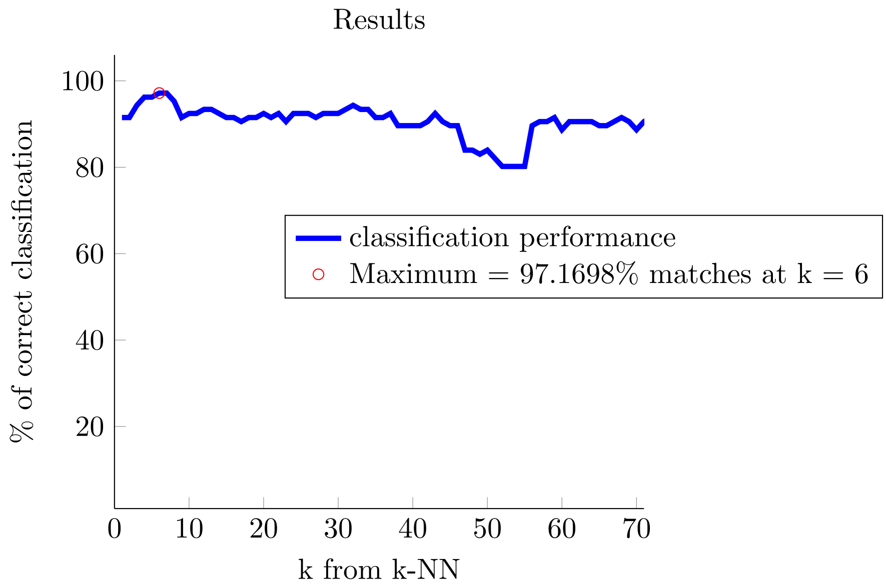


Figure 4: The best result from the k-NN classification

2.3 Comparison with k-NN

Compared to the kNN classifier the results show that the Mahalanobis classifier has better top performance with a success rate of 99.056%. As seen in the figures of the success rate the Mahalanobis classifier is more sensible to the choice of trainings samples than kNN. On the other hand the performance drop of kNN at $2 * trainingsSetSize$ is not given for the Mahalanobis classifier.

3 Discriminant Functions for the Normal Density

This section presents the computation of the discriminate function per hand and compares its results with the ones provided by MATLAB's classify function.

3.1 Computation of the Discriminate Function per Hand

See Figure 12 to Figure 16.

3.2 Computation of the Discriminate Function in `MATLAB`

As shown in Figure 17, the handwritten computation of the discriminate function differs slightly from the one computed by MATLAB's classify function. It is also shown that the connection line between the mean values of the two classes is not perpendicular to the decision border. This is the case because the two classes' covariance matrices are different, and are not multiples of the identity matrix.

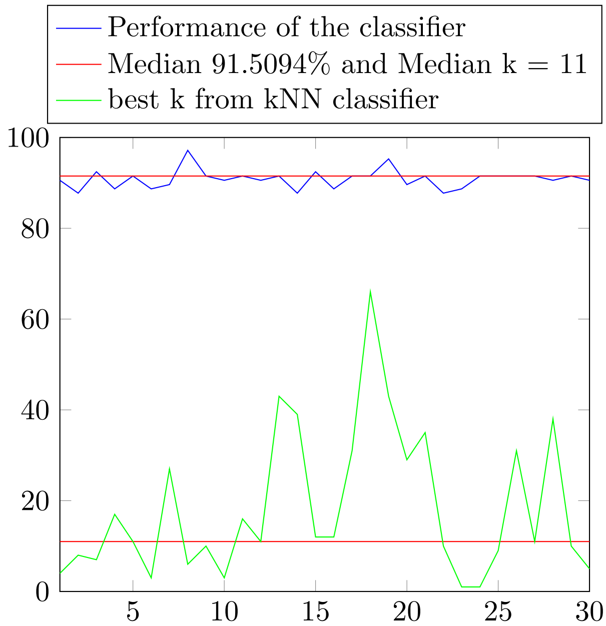


Figure 5: All results from the k-NN classification with 30 different test and training sets

References

[1] PARVUS Forina, M. et al. *An Extendible Package for Data Exploration, Classification and Correlation.* Institute of Pharmaceutical and Food Analysis and Technologies, Via Brigata Salerno, 16147 Genoa, Italy, 1990.

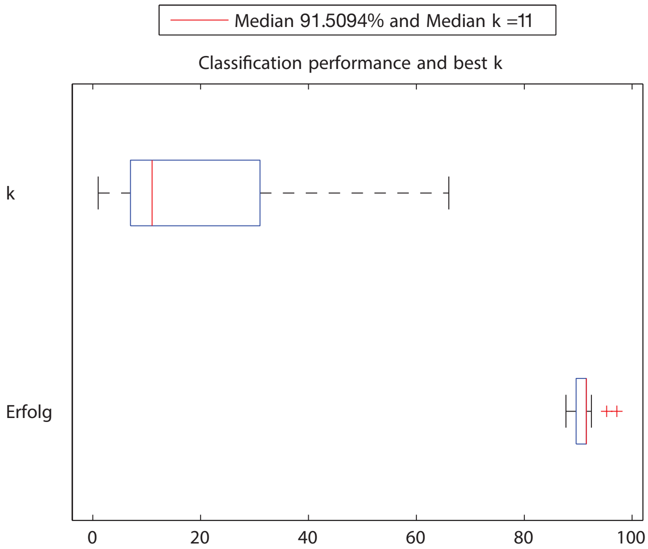


Figure 6: The upper Boxplot shows the derivation of the k from the k-NN classifier. The Boxplot below shows the derivation of the classification performance

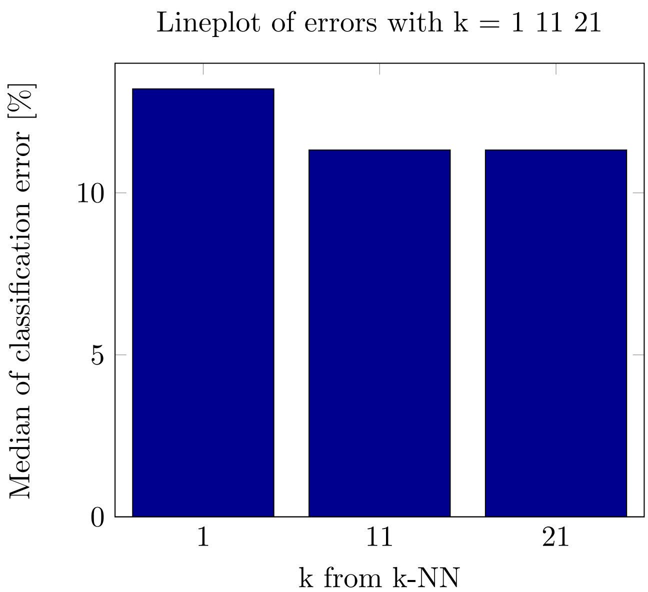


Figure 7: The median of the classification error rates of all 30 test sets for different values of k

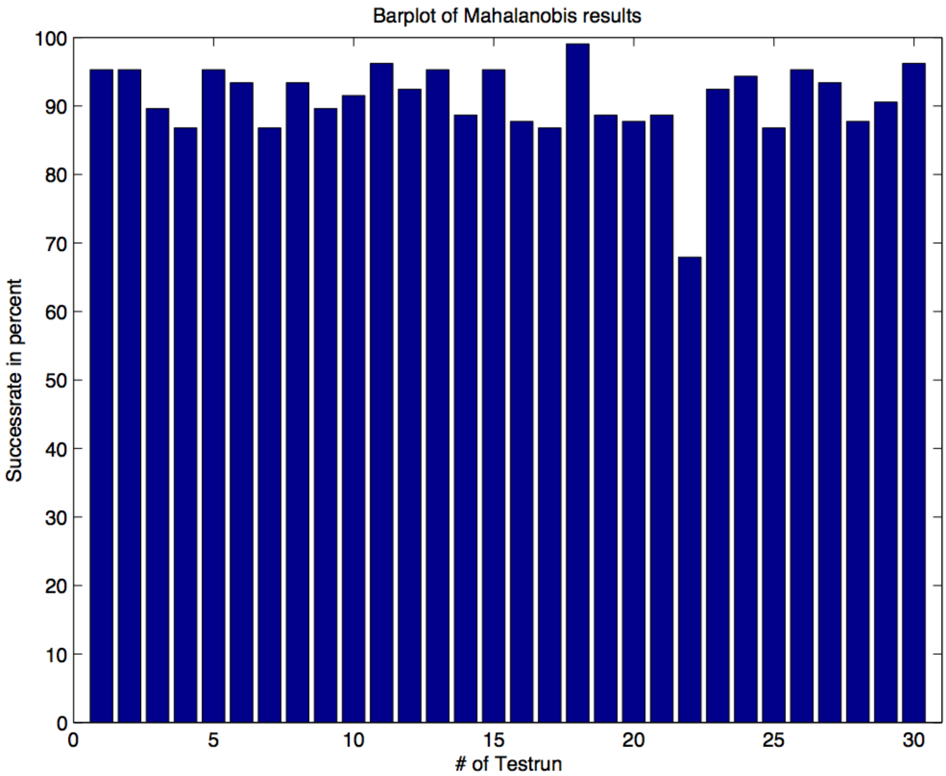


Figure 8: Mahalanobis performance with the diagonal identity matrix.

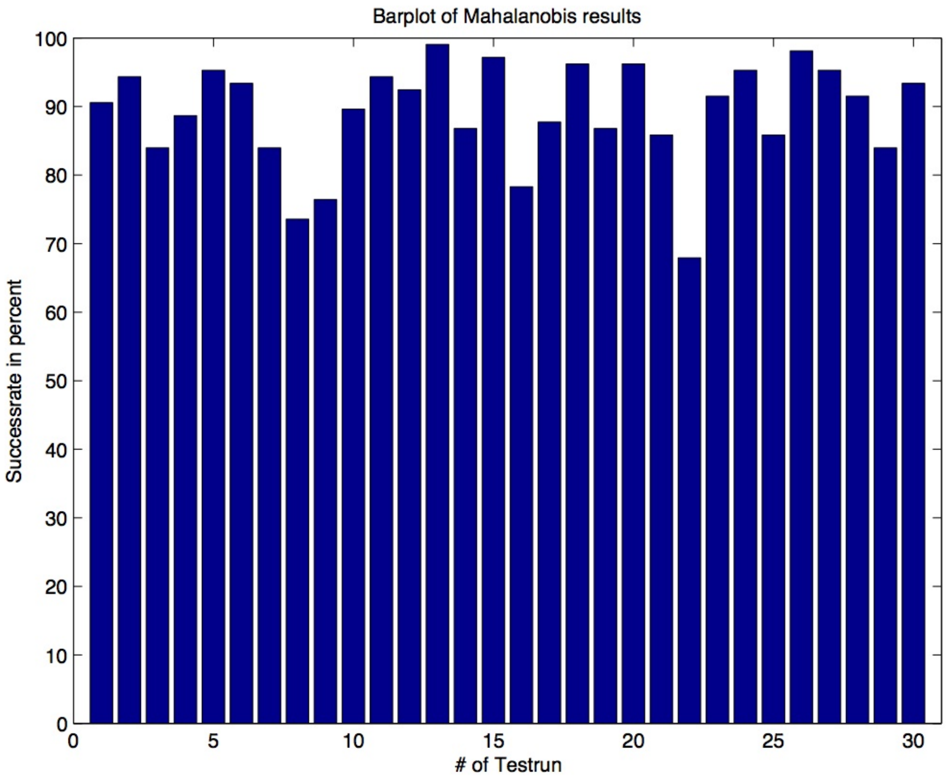


Figure 9: Mahalanobis performance with the diagonal covariance matrix.

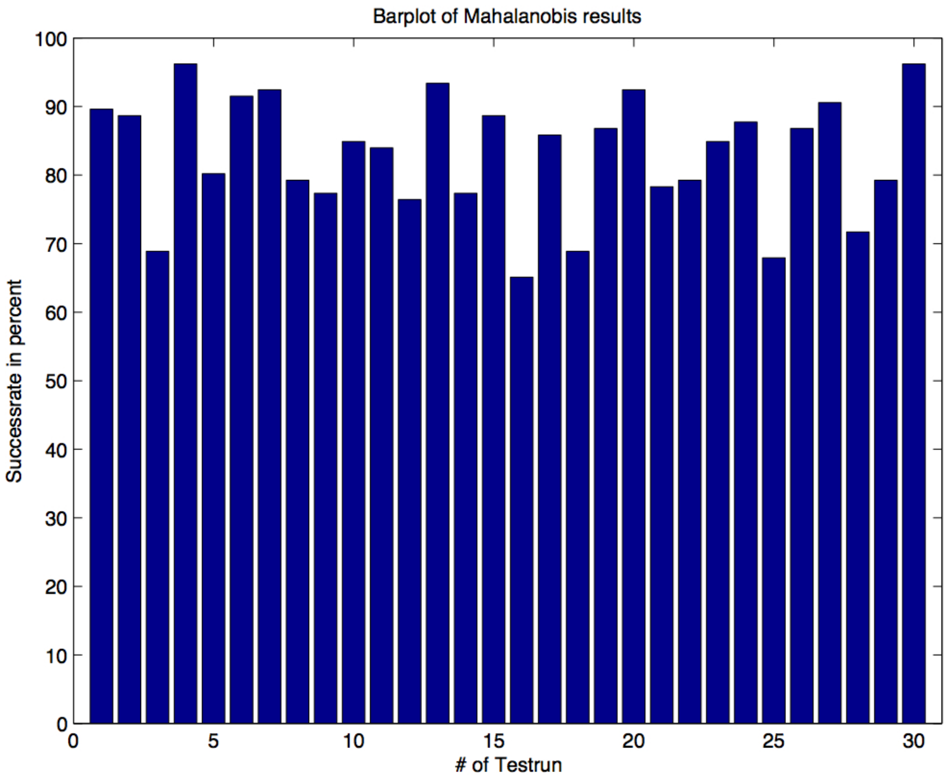


Figure 10: Mahalanobis performance with the full covariance matrix.

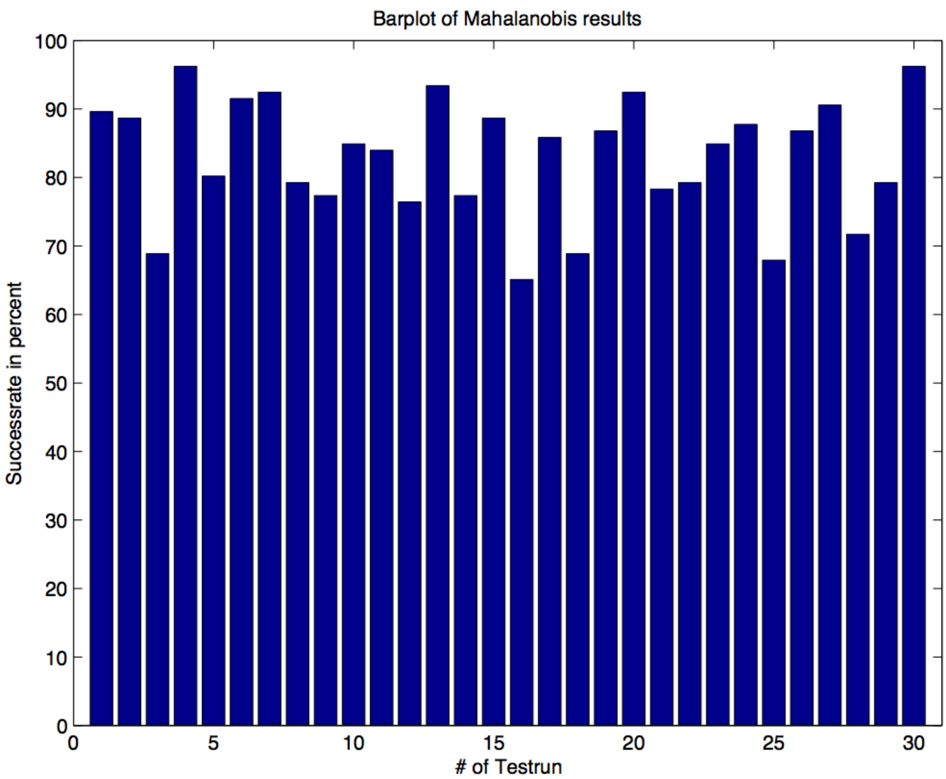
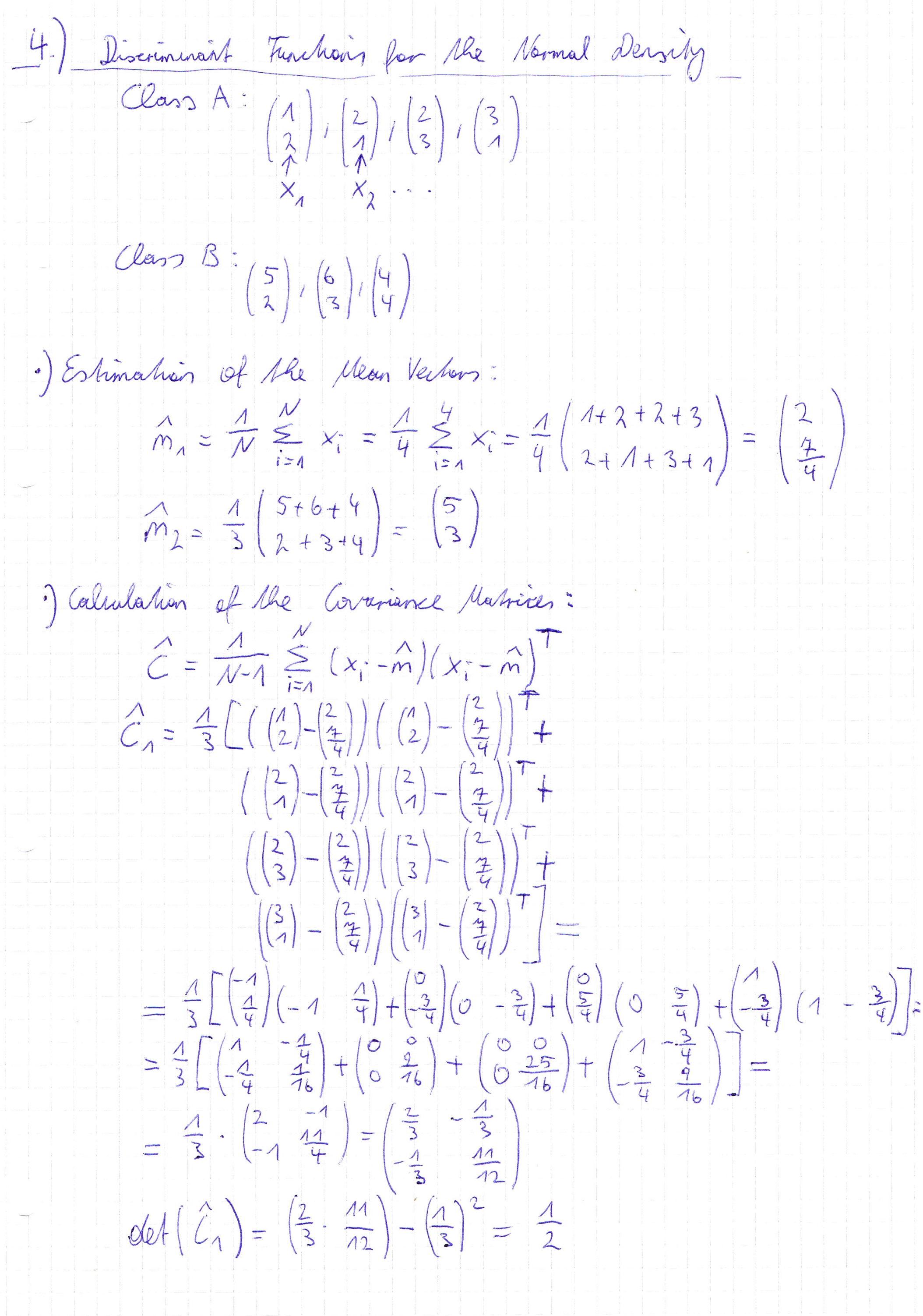


Figure 11: Mahalanobis performance with the built in Matlab function mahal() to validate the custom implementation.



4.) Discriminant Functions for the Normal Density

Class A: $\begin{pmatrix}1\\2\end{pmatrix}, \begin{pmatrix}2\\1\end{pmatrix}, \begin{pmatrix}2\\3\end{pmatrix}, \begin{pmatrix}3\\1\end{pmatrix}$

$\underset{x_1}{\uparrow} \quad \underset{x_2}{\uparrow} \ \ldots$

Class B: $\begin{pmatrix}5\\2\end{pmatrix}, \begin{pmatrix}6\\3\end{pmatrix}, \begin{pmatrix}4\\4\end{pmatrix}$

•) Estimation of the Mean Vectors:

$$\hat{m}_1 = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{1}{4}\sum_{i=1}^{4} x_i = \frac{1}{4}\begin{pmatrix}1+2+2+3\\2+1+3+1\end{pmatrix} = \begin{pmatrix}2\\ \frac{7}{4}\end{pmatrix}$$

$$\hat{m}_2 = \frac{1}{3}\begin{pmatrix}5+6+4\\2+3+4\end{pmatrix} = \begin{pmatrix}5\\3\end{pmatrix}$$

•) Calculation of the Covariance Matrices:

$$\hat{C} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i-\hat{m})(x_i-\hat{m})^T$$

$$\hat{C}_1 = \frac{1}{3}\Bigg[\left(\begin{pmatrix}1\\2\end{pmatrix}-\begin{pmatrix}2\\ \frac{7}{4}\end{pmatrix}\right)\left(\begin{pmatrix}1\\2\end{pmatrix}-\begin{pmatrix}2\\ \frac{7}{4}\end{pmatrix}\right)^T + \left(\begin{pmatrix}2\\1\end{pmatrix}-\begin{pmatrix}2\\ \frac{7}{4}\end{pmatrix}\right)\left(\begin{pmatrix}2\\1\end{pmatrix}-\begin{pmatrix}2\\ \frac{7}{4}\end{pmatrix}\right)^T + \left(\begin{pmatrix}2\\3\end{pmatrix}-\begin{pmatrix}2\\ \frac{7}{4}\end{pmatrix}\right)\left(\begin{pmatrix}2\\3\end{pmatrix}-\begin{pmatrix}2\\ \frac{7}{4}\end{pmatrix}\right)^T + \left(\begin{pmatrix}3\\1\end{pmatrix}-\begin{pmatrix}2\\ \frac{7}{4}\end{pmatrix}\right)\left(\begin{pmatrix}3\\1\end{pmatrix}-\begin{pmatrix}2\\ \frac{7}{4}\end{pmatrix}\right)^T\Bigg] =$$

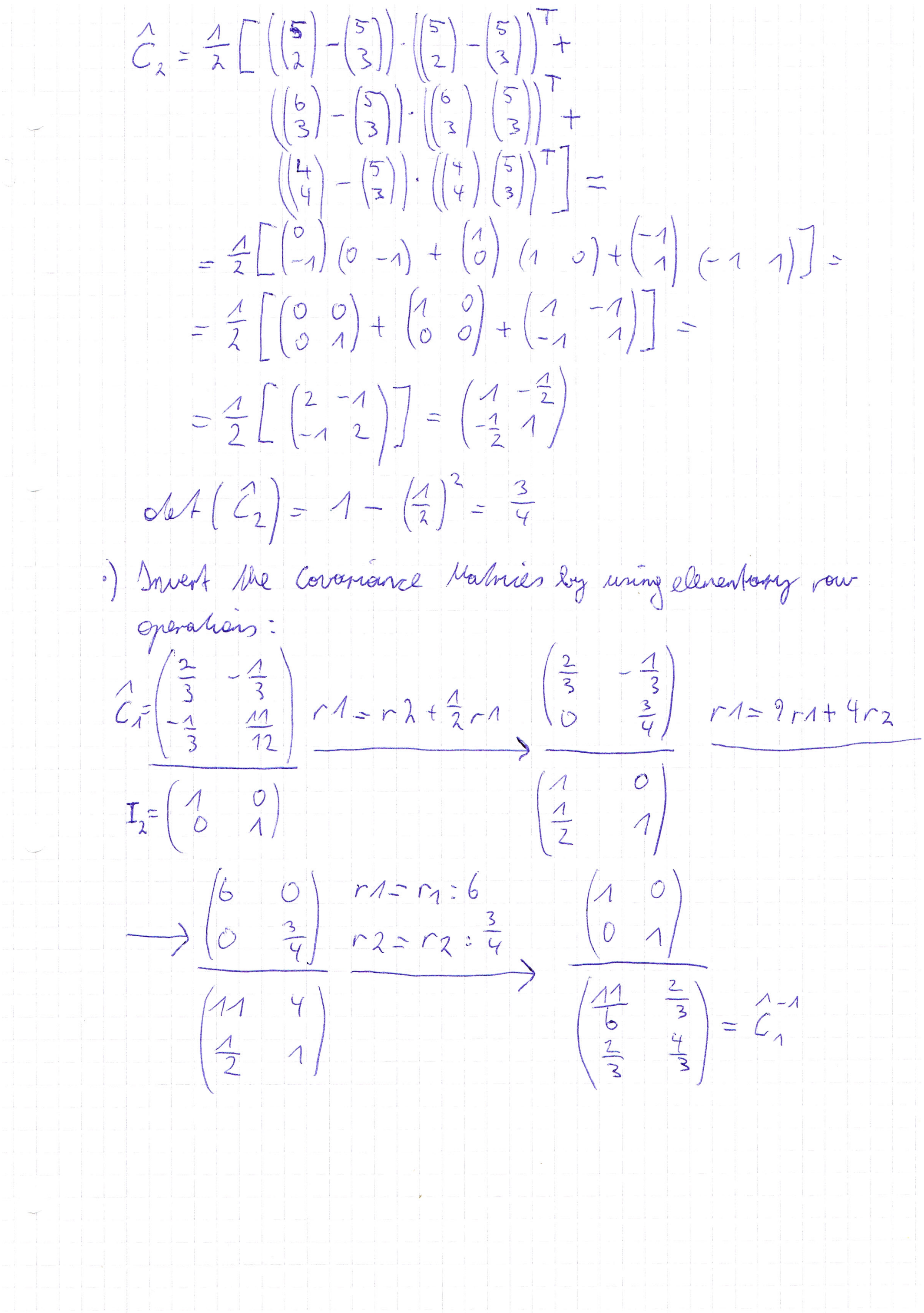
$$= \frac{1}{3}\left[\begin{pmatrix}-1\\ \frac{1}{4}\end{pmatrix}\begin{pmatrix}-1 & \frac{1}{4}\end{pmatrix} + \begin{pmatrix}0\\ -\frac{3}{4}\end{pmatrix}\begin{pmatrix}0 & -\frac{3}{4}\end{pmatrix} + \begin{pmatrix}0\\ \frac{5}{4}\end{pmatrix}\begin{pmatrix}0 & \frac{5}{4}\end{pmatrix} + \begin{pmatrix}1\\ -\frac{3}{4}\end{pmatrix}\begin{pmatrix}1 & -\frac{3}{4}\end{pmatrix}\right] =$$

$$= \frac{1}{3}\left[\begin{pmatrix}1 & -\frac{1}{4}\\ -\frac{1}{4} & \frac{1}{16}\end{pmatrix} + \begin{pmatrix}0 & 0\\ 0 & \frac{9}{16}\end{pmatrix} + \begin{pmatrix}0 & 0\\ 0 & \frac{25}{16}\end{pmatrix} + \begin{pmatrix}1 & -\frac{3}{4}\\ -\frac{3}{4} & \frac{9}{16}\end{pmatrix}\right] =$$

$$= \frac{1}{3}\cdot\begin{pmatrix}2 & -1\\ -1 & \frac{11}{4}\end{pmatrix} = \begin{pmatrix}\frac{2}{3} & -\frac{1}{3}\\ -\frac{1}{3} & \frac{11}{12}\end{pmatrix}$$

$$\det(\hat{C}_1) = \left(\frac{2}{3}\cdot\frac{11}{12}\right) - \left(\frac{1}{3}\right)^2 = \frac{1}{2}$$

Figure 12: Handwritten computation of the discriminate function, page 1



$$\hat{C}_2 = \frac{1}{2}\Bigg[\left(\begin{pmatrix}5\\2\end{pmatrix}-\begin{pmatrix}5\\3\end{pmatrix}\right)\cdot\left(\begin{pmatrix}5\\2\end{pmatrix}-\begin{pmatrix}5\\3\end{pmatrix}\right)^T +$$

$$\left(\begin{pmatrix}6\\3\end{pmatrix}-\begin{pmatrix}5\\3\end{pmatrix}\right)\cdot\left(\begin{pmatrix}6\\3\end{pmatrix}\begin{pmatrix}5\\3\end{pmatrix}\right)^T +$$

$$\left(\begin{pmatrix}4\\4\end{pmatrix}-\begin{pmatrix}5\\3\end{pmatrix}\right)\cdot\left(\begin{pmatrix}4\\4\end{pmatrix}\begin{pmatrix}5\\3\end{pmatrix}\right)^T\Bigg] =$$

$$= \frac{1}{2}\left[\begin{pmatrix}0\\-1\end{pmatrix}\begin{pmatrix}0 & -1\end{pmatrix} + \begin{pmatrix}1\\0\end{pmatrix}\begin{pmatrix}1 & 0\end{pmatrix} + \begin{pmatrix}-1\\1\end{pmatrix}\begin{pmatrix}-1 & 1\end{pmatrix}\right] =$$

$$= \frac{1}{2}\left[\begin{pmatrix}0 & 0\\0 & 1\end{pmatrix} + \begin{pmatrix}1 & 0\\0 & 0\end{pmatrix} + \begin{pmatrix}1 & -1\\-1 & 1\end{pmatrix}\right] =$$

$$= \frac{1}{2}\left[\begin{pmatrix}2 & -1\\-1 & 2\end{pmatrix}\right] = \begin{pmatrix}1 & -\frac{1}{2}\\-\frac{1}{2} & 1\end{pmatrix}$$

$$\det(\hat{C}_2) = 1 - \left(\frac{1}{2}\right)^2 = \frac{3}{4}$$

•) Invert the Covariance Matrices by using elementary row operations:

$$\frac{\hat{C}_1 = \begin{pmatrix}\frac{2}{3} & -\frac{1}{3}\\-\frac{1}{3} & \frac{11}{12}\end{pmatrix}}{I_2 = \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}} \xrightarrow{r1 = r2 + \frac{1}{2}r1} \frac{\begin{pmatrix}\frac{2}{3} & -\frac{1}{3}\\0 & \frac{3}{4}\end{pmatrix}}{\begin{pmatrix}1 & 0\\\frac{1}{2} & 1\end{pmatrix}} \xrightarrow{r1 = 9r1 + 4r2}$$

$$\longrightarrow \frac{\begin{pmatrix}6 & 0\\0 & \frac{3}{4}\end{pmatrix}}{\begin{pmatrix}11 & 4\\\frac{1}{2} & 1\end{pmatrix}} \xrightarrow[r2 = r2 : \frac{3}{4}]{r1 = r1 : 6} \frac{\begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}}{\begin{pmatrix}\frac{11}{6} & \frac{2}{3}\\\frac{2}{3} & \frac{4}{3}\end{pmatrix}} = \hat{C}_1^{-1}$$

Figure 13: Handwritten computation of the discriminate function, page 2

$$\begin{array}{c} \hat{C}_2 = \begin{pmatrix} 1 & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{pmatrix} \\ \hline I_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \end{array} \xrightarrow{r1 = r1 + \frac{1}{2} r2} \begin{array}{c} \begin{pmatrix} \frac{3}{4} & 0 \\ -\frac{1}{2} & 1 \end{pmatrix} \\ \hline \begin{pmatrix} 1 & \frac{1}{2} \\ 0 & 1 \end{pmatrix} \end{array} \xrightarrow{r2 = 3r2 + 2r1}$$

$$\rightarrow \begin{array}{c} \begin{pmatrix} \frac{3}{4} & 0 \\ 0 & 3 \end{pmatrix} \\ \hline \begin{pmatrix} 1 & \frac{1}{2} \\ 2 & 4 \end{pmatrix} \end{array} \xrightarrow[r_2 = r_2 : 3]{r_1 = r_1 : \frac{3}{4}} \begin{array}{c} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \\ \hline \begin{pmatrix} \frac{4}{3} & \frac{2}{3} \\ \frac{2}{3} & \frac{4}{3} \end{pmatrix} \end{array} = \hat{C}_2^{-1}$$

-) Calculate Mahalanobis Distances:

$$d_1^2(\vec{x}) = (\vec{x} - \hat{m}_1)^T \hat{C}_1^{-1} (\vec{x} - \hat{m}_1) =$$

$$= \left[\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} - \begin{pmatrix} 2 \\ \frac{7}{4} \end{pmatrix} \right]^T \begin{pmatrix} \frac{11}{6} & \frac{2}{3} \\ \frac{2}{3} & \frac{4}{3} \end{pmatrix} \left[\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} - \begin{pmatrix} 2 \\ \frac{7}{4} \end{pmatrix} \right] =$$

$$= \begin{pmatrix} x_1 - 2 & x_2 - \frac{7}{4} \end{pmatrix} \begin{pmatrix} \frac{11}{6} & \frac{2}{3} \\ \frac{2}{3} & \frac{4}{3} \end{pmatrix} \begin{pmatrix} x_1 - 2 \\ x_2 - \frac{7}{4} \end{pmatrix} =$$

$$= \begin{pmatrix} \frac{11}{6}(x_1 - 2) + \frac{2}{3}(x_2 - \frac{7}{4}) & \frac{2}{3}(x_1 - 2) + \frac{4}{3}(x_2 - \frac{7}{4}) \end{pmatrix} \begin{pmatrix} x_1 - 2 \\ x_2 - \frac{7}{4} \end{pmatrix}$$

$$= \begin{pmatrix} \frac{11}{6} x_1 - \frac{11}{3} + \frac{2}{3} x_2 - \frac{7}{6} & \frac{2}{3} x_1 - \frac{4}{3} + \frac{4}{3} x_2 - \frac{7}{3} \end{pmatrix} \begin{pmatrix} x_1 - 2 \\ x_2 - \frac{7}{4} \end{pmatrix} =$$

$$= \begin{pmatrix} \frac{11}{6} x_1 + \frac{2}{3} x_2 - \frac{29}{6} & \frac{2}{3} x_1 + \frac{4}{3} x_2 - \frac{11}{3} \end{pmatrix} \begin{pmatrix} x_1 - 2 \\ x_2 - \frac{7}{4} \end{pmatrix} =$$

$$= (x_1 - 2)\left(\frac{11}{6} x_1 + \frac{2}{3} x_2 - \frac{29}{6}\right) + \left(x_2 - \frac{7}{4}\right)\left(\frac{2}{3} x_1 + \frac{4}{3} x_2 - \frac{11}{3}\right) =$$

$$= \frac{11}{6} x_1^2 + \frac{2}{3} x_1 x_2 - \frac{29}{6} x_1 - \frac{11}{3} x_1 - \frac{4}{3} x_2 + \frac{29}{3} +$$

$$+ \frac{2}{3} x_1 x_2 + \frac{4}{3} x_2^2 - \frac{11}{3} x_2 - \frac{7}{6} x_1 - \frac{7}{3} x_2 + \frac{77}{12} =$$

$$= \frac{11}{6} x_1^2 - \frac{29}{3} x_1 + \frac{4}{3} x_1 x_2 + \frac{4}{3} x_2^2 - \frac{22}{3} x_2 + \frac{193}{12}$$

Figure 14: Handwritten computation of the discriminate function, page 3

$$d_2^2(\vec{x}) = \left[\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} - \begin{pmatrix} 5 \\ 3 \end{pmatrix}\right]^T \begin{pmatrix} \frac{4}{3} & \frac{2}{3} \\ \frac{2}{3} & \frac{4}{3} \end{pmatrix} \cancel{\begin{pmatrix} x_1 - 5 \\ x_2 - 3 \end{pmatrix}} \left[\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} - \begin{pmatrix} 5 \\ 3 \end{pmatrix}\right] =$$

$$= \begin{pmatrix} x_1 - 5 & x_2 - 3 \end{pmatrix} \begin{pmatrix} \frac{4}{3} & \frac{2}{3} \\ \frac{2}{3} & \frac{4}{3} \end{pmatrix} \begin{pmatrix} x_1 - 5 \\ x_2 - 3 \end{pmatrix} =$$

$$= \begin{pmatrix} \frac{4}{3}(x_1 - 5) + \frac{2}{3}(x_2 - 3) & \frac{2}{3}(x_1 - 5) + \frac{4}{3}(x_2 - 3) \end{pmatrix} \begin{pmatrix} x_1 - 5 \\ x_2 - 3 \end{pmatrix} =$$

$$= \begin{pmatrix} \frac{4}{3}x_1 - \frac{20}{3} + \frac{2}{3}x_2 - 2 & \frac{2}{3}x_1 - \frac{10}{3} + \frac{4}{3}x_2 - 4 \end{pmatrix} \begin{pmatrix} x_1 - 5 \\ x_2 - 3 \end{pmatrix} =$$

$$= \begin{pmatrix} \frac{4}{3}x_1 - \frac{26}{3} + \frac{2}{3}x_2 & \frac{2}{3}x_1 + \frac{4}{3}x_2 - \frac{22}{3} \end{pmatrix} \begin{pmatrix} x_1 - 5 \\ x_2 - 3 \end{pmatrix} =$$

$$= (x_1 - 5)\left(\frac{4}{3}x_1 + \frac{2}{3}x_2 - \frac{26}{3}\right) + (x_2 - 3)\left(\frac{2}{3}x_1 + \frac{4}{3}x_2 - \frac{22}{3}\right) =$$

$$= \frac{4}{3}x_1^2 + \frac{2}{3}x_1x_2 - \frac{26}{3}x_1 - \frac{20}{3}x_1 - \frac{10}{3}x_2 + \frac{130}{3} +$$

$$+ \frac{2}{3}x_1x_2 + \frac{4}{3}x_2^2 - \frac{22}{3}x_2 - 2x_1 - 4x_2 + 22 =$$

$$= \frac{4}{3}x_1^2 - \frac{52}{3}x_1 + \frac{4}{3}x_1x_2 + \frac{4}{3}x_2^2 - \cancel{\frac{4}{3}}\frac{44}{3}x_2 + \frac{196}{3} =$$

$$= \frac{4}{3}\left(x_1^2 - 13x_1 + x_1x_2 + x_2^2 - 11x_2 + 49\right)$$

.) Compute the discriminate function, assuming $P(A) = P(B) = 0{,}5$

$$g_1(\vec{x}) = -\frac{1}{2}d_1^2(\vec{x}) + \left(-\frac{1}{2}\ln\det(\hat{C}_1) + \ln P(A)\right) =$$

$$= -\frac{1}{2}d_1^2(\vec{x}) + \frac{1}{2}\ln(0{,}5)$$

$$g_1(\vec{x}) = -\frac{11}{12}x_1^2 + \frac{29}{6}x_1 - \frac{2}{3}x_1x_2 - \frac{2}{3}x_2^2 + \frac{11}{3}x_2 - \frac{193}{24} + \frac{1}{2}\ln(0{,}5)$$

$$g_2(\vec{x}) = -\frac{1}{2}d_2^2(\vec{x}) + \left(-\frac{1}{2}\ln(\det(\hat{C}_2)) + \ln(0{,}5)\right) =$$

$$= -\frac{2}{3}\left(x_1^2 - 13x_1 + x_1x_2 + x_2 - 11x_2 + 49\right) \cancel{+} - \frac{1}{2}\ln\left(\frac{3}{4}\right) + \ln(0{,}5)$$

Figure 15: Handwritten computation of the discriminate function, page 4

$$f(\vec{x}) = g_1(\vec{x}) - g_2(\vec{x})$$

$$-\frac{11}{12}x_1^2 + \frac{29}{6}x_1 - \frac{2}{3}x_1x_2 - \frac{2}{3}x_2^2 + \frac{11}{3}x_2 - \frac{193}{24} + \frac{1}{2}\ln(0{,}5)$$

$$- \quad -\frac{2}{3}x_1^2 + \frac{26}{3}x_1 - \frac{2}{3}x_1x_2 - \frac{2}{3}x_2^2 + \frac{22}{3}x_2 - \frac{98}{3} - \frac{1}{2}\ln\left(\frac{3}{4}\right) + \ln(0{,}5)$$

$$f(\vec{x}) = -\frac{1}{4}x_1^2 - \frac{25}{6}x_1 - \frac{11}{3}x_2 + \frac{591}{24} - \frac{1}{2}\ln(0{,}5) + \frac{1}{2}\ln\left(\frac{3}{4}\right)$$

Figure 16: Handwritten computation of the discriminate function, page 5

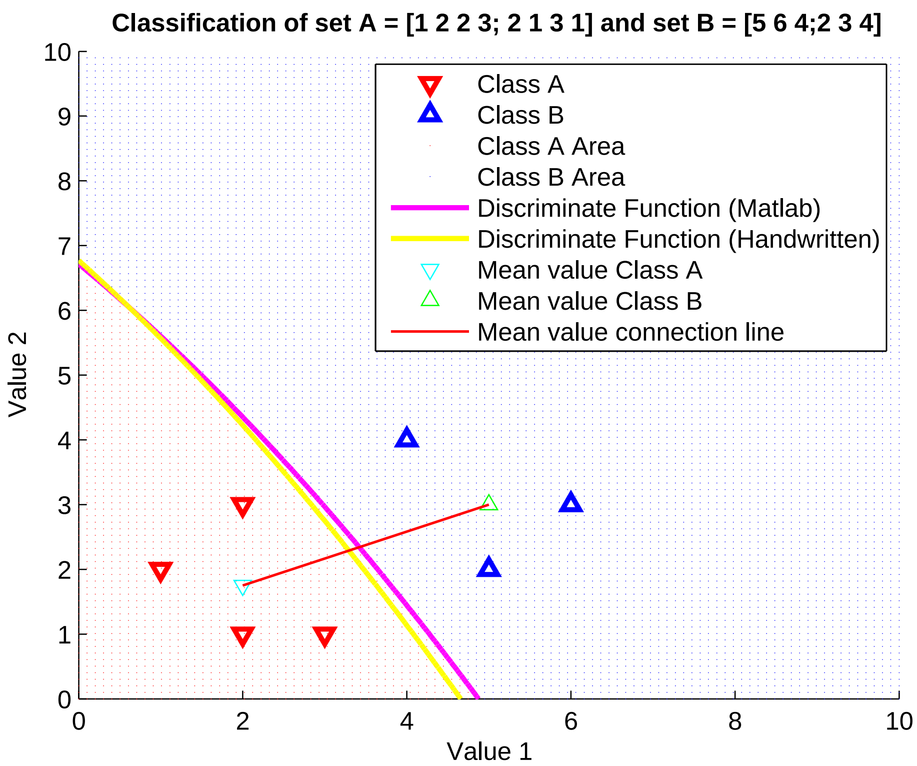


Figure 17: The discriminate function, as computed by MATLAB's classify function (pink line).