Statistical Methods for Machine Learning
Assignment 2: Basic Learning Algorithms

Philip Pickering
pgpick@gmx.at

Marco Eilers
eilers.marco@googlemail.com

Thomas Bracht Laumann Jespersen
ntl316@alumni.ku.dk

1 Regression

1.1 Maximum Likelihood solution

Use linear model

$$y(\mathbf{x}, \mathbf{w}) = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_D x_D$$

and for the D variables we let $\phi_i(\mathbf{x}) = x_i$ for $i = 1, \ldots, D$ and $\phi_0(\mathbf{x}) = 1$.

1.1.1 Selection 1

For our first selection S_1 our design matrix becomes a 200×5 matrix.

$$\boldsymbol{\Phi}_{S_1} = \begin{bmatrix} 1 & \mathbf{x}_{1,1} & \mathbf{x}_{1,2} & \mathbf{x}_{1,3} & \mathbf{x}_{1,4} \\ & & \vdots & & \\ 1 & \mathbf{x}_{i,1} & \mathbf{x}_{i,2} & \mathbf{x}_{i,3} & \mathbf{x}_{i,4} \\ & & \vdots & & \\ 1 & \mathbf{x}_{N,1} & \mathbf{x}_{N,2} & \mathbf{x}_{N,3} & \mathbf{x}_{N,4} \end{bmatrix}$$

where the notation $\mathbf{x}_{i,j}$ indicates the j'th entry in the i'th vector.

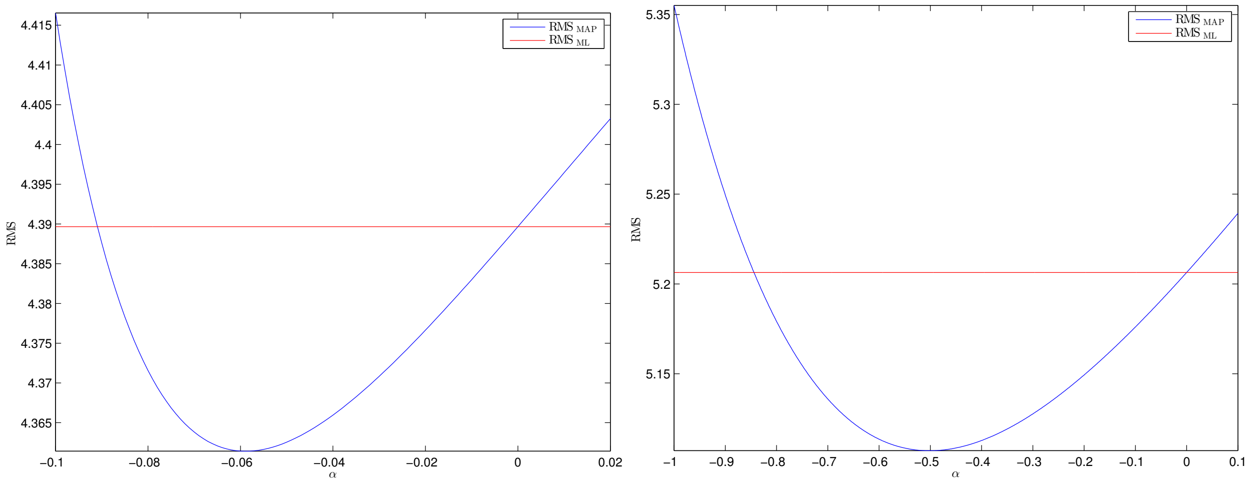
Finding the ML estimate of our parameters for S_1 gives

$$\mathbf{w}_{S_1} = \begin{bmatrix} \text{-43.0947} \\ \text{-0.1299} \\ 0.0352 \\ 0.9335 \\ \text{-0.0433} \end{bmatrix} \quad \text{and} \quad \text{RMS}_{S_1} = 4.3897$$

1.1.2 Selection 2

Our second selection S_2 consists only of the data from the 'Abdomen 2' column, giving a design matrix $\boldsymbol{\Phi}_{S_2}$ of dimensions 200×2. Training the model on the same training data yields:

$$\mathbf{w}_{S_2} = \begin{bmatrix} \text{-37.4085} \\ 0.6133 \end{bmatrix} \quad \text{and} \quad \text{RMS}_{S_2} = 5.2064$$



(a)

(b)

Figure 1: Plot of RMS against varying values of α

1.1.3 Discussion

Just looking at the root mean square values of the two selections, it appears that S_1 performs better than S_2, but not by a lot. This suggests that either the variable 'Abdomen 2' is the most descriptive in terms of body, or that the linear model simply is a poor fit no matter how many variables we include. It could be a combination of the two.

It could probably be argued that the linear regression model is a poor predictor, but including more variables should improve the results.

1.2 Maximum a posteriori solution

In fig. 1 two plots are found of the root mean square values for varying values of α. In both plots we set $\beta = 1$. The RMS value of the ML solution is plotted as a straight line.

We can observe for both plots that when $\alpha = 0$, we obtain the same RMS error for the MAP estimate as for the ML solution. This is expected and demonstrates that when our prior precision parameters are set to zero, the MAP estimate becomes the ML estimate.

Fig. 1(a) is the plot for S_1, and it can be seen that the RMS_{MAP} error drops below the RMS_{ML} in the interval $[-0.091, 0]$. In fig. 1(b) the plot for S_2 similarly gives us that the RMS_{MAP} error is lower in the interval $[-0.844, 0]$.

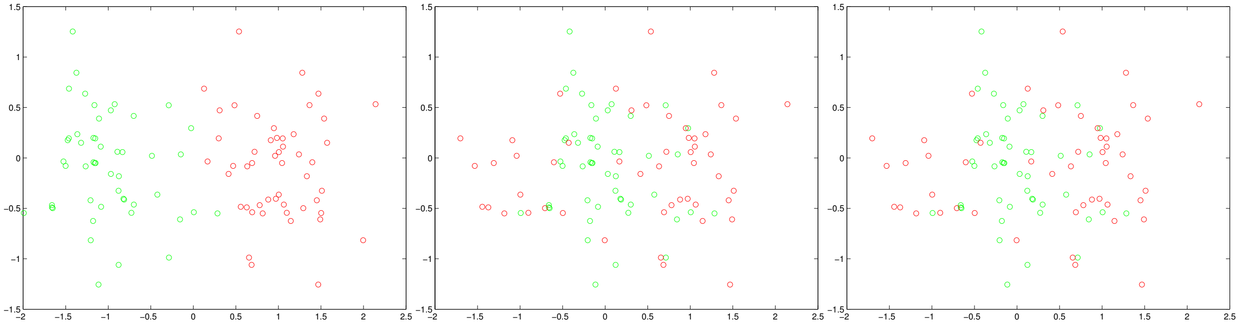
1.3 Theory

Verify result in equation (3.49) for the posterior distribution of the parameters $\mathbf{w}$ in the linear basis function in which $\mathbf{m}_N$ and $\mathbf{S}_N$ are defined

2 Linear Discriminant Analysis

2.1 Observations regarding the data

In fig. 2 you can see plots of all three training data sets. It is obvious that while KnollA and KnollC contain well-separable groups of data points, the distributions in KnollB seem to overlap. We can therefore expect



(a) KNOLLA (b) KNOLLB (c) KNOLLC

Figure 2: Visualisation of the training data for each of the KNOLL problems

	Knoll A		Knoll B		Knoll C	
	Training	Test	Training	Test	Training	Test
Errors	1%	3%	40%	49%	1%	3%
Margin	-0.242763	-0.142026	-1.629922	-2.061411	-0.002435	-0.001425

Table 1: Error percentages and margins for each of the KNOLL problems

that LDA will performa relatively well on KnollA and KnollB, whereas the results for KnollB will likely be more erratic.

2.2 Results of LDA

If we run LDA on all three data sets and plot them with their respective decision boundaries (fig. 3), we find that the algorithm indeed performs well on the A and C sets. In table 1 we can see that although the error count is greater than zero for all training and test sets, which is also implied by the margin always being slightly negative, the number of errors is relatively low for both training and test sets. The training set performs slightly better than the test set in both cases, which is to be expected, but there is not over- or underfitting problem.

For KnollB, however, the results are much worse. The properties of the underlying distribution simply do not allow them to be separated by a straight line, which results in a high error percentage and a highly negative margin. Since the distributions in B overlap to a high degree, it is generally not possible to find a function of any kind to separate both classes accurately. However, a Quadratic Discriminant Analysis might be able to find out that one class concentrates a lot more in a certain range, whereas the other one has a higher variance and is therefore spread out more widely. It could therefore assign all points in this range to one class and the rest to the other. While this will still result in a high error percentage, such an algorithm might actually find the Bayes Optimal Solution.

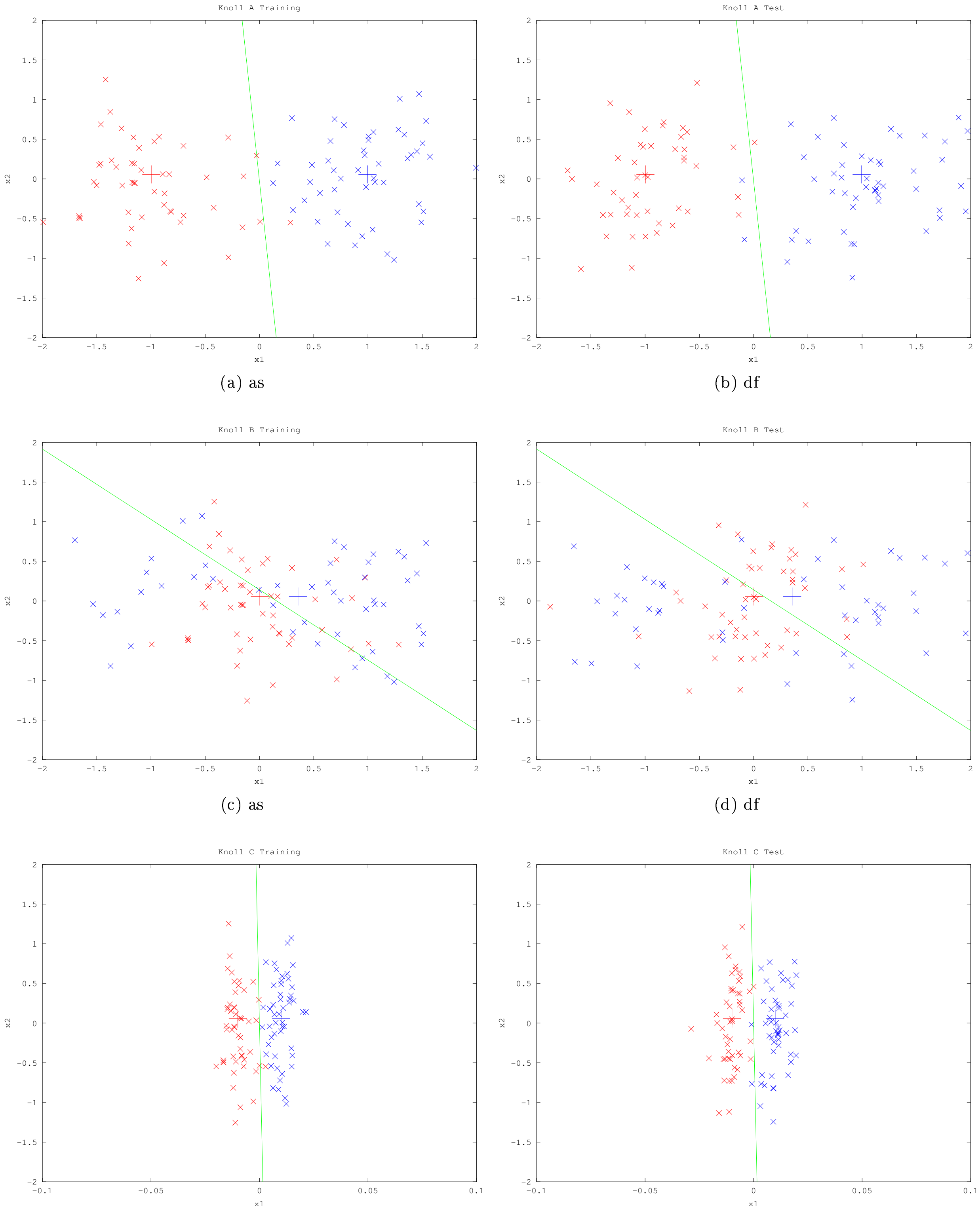
3 Nearest Neighbor Classification

3.1 Nearest Neighbor Classification with Euclidian Metric

Implement k-NN. Train for all three KNOLL problems using training data sets. Report accuracy on corresponding sets for $k = 1, 2, 3, \ldots, 9$. Explain results; discuss similarities and differences in performance on the three data sets. Compare results to LDA results.

Hand in: classifier source code, results, short discussion.

Observations



(a) as

(b) df

(c) as

(d) df

(e) as

(f) df

Figure 3: Visualisation of the training data for each of the KNOLL problems

We can compute an accuracy for the k-NN on the same training data, which for $k = 1$ always gives an error degree of zero, because the nearest neighbour of a given point in the data set is the point itself. But for higher values of k, we can get errors for points surrounded by points from the opposite class.

3.2 Changing the Metric

To prove that d is a metric, given

$$d(\mathbf{x}, \mathbf{z}) = \|\mathbf{Mx} - \mathbf{Mz}\|, \text{ where } \mathbf{M} = \begin{pmatrix} 100 & 0 \\ 0 & 1 \end{pmatrix}$$

and $\| \cdot \|$ is the standard L_2-norm (in $\mathbb{R}^2$), we need to verify $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^2$ that 1) $d(\mathbf{x}, \mathbf{y}) \geq 0$; 2) $d(\mathbf{x}, \mathbf{y}) = 0 \Leftrightarrow \mathbf{x} = \mathbf{y}$; 3) $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$ (symmetry) and 4) $\forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^2 : d(\mathbf{x}, z) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z})$.

3.2.1 Revised proof — basic observations

We need only to observe that $\mathbf{M}$ is a projection $m : \mathbb{R}^2 \to \mathbb{R}^2$, i.e. onto $\mathbb{R}^2$ itself, given by $m(\mathbf{x}) = \mathbf{Mx}$. This immediately gives us all the properties we need, because L_2 is itself a (complete) metric on $\mathbb{R}^2$.

For instance, if we let $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^2$ and $\mathbf{x'}, \mathbf{y'}, \mathbf{z'}$ be the result of applying m on $\mathbf{x}, \mathbf{y}, \mathbf{z}$ respectively, we can prove the triangle inequality:

$$\begin{aligned} d(\mathbf{x}, \mathbf{z}) &= \|\mathbf{Mx} - \mathbf{Mz}\| \\ &= \|\mathbf{Mx} - \mathbf{My} + \mathbf{My} - \mathbf{Mz}\| \\ &= \|(\mathbf{x'} - \mathbf{y'}) + (\mathbf{y'} - \mathbf{z'})\| \\ &\leq \|\mathbf{x'} - \mathbf{y'}\| + \|\mathbf{y'} - \mathbf{z'}\| \quad \text{(by property of } L_2 \text{ norm)} \\ &= d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z}) \end{aligned} \tag{1}$$

3.2.2 Original proof — the long way

Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^2$. As it is easier to consider the square of the L_2 norm, we will do so:

$$\begin{aligned} d(\mathbf{x}, \mathbf{y})^2 &= \left\| \begin{pmatrix} 100 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} - \begin{pmatrix} 100 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \right\|^2 \\ &= \left\| \begin{pmatrix} 100(x_1 - y_1) \\ x_2 - y_2 \end{pmatrix} \right\|^2 \\ &= \begin{pmatrix} 100(x_1 - y_1) \\ x_2 - y_2 \end{pmatrix}^T \begin{pmatrix} 100(x_1 - y_1) \\ x_2 - y_2 \end{pmatrix} \\ &= (100(x_1 - y_1))^2 + (x_2 - y_2)^2. \end{aligned} \tag{2}$$

From (2) we observe that $(x_1 - y_1)^2 \geq 0$ and similarly $(x_2 - y_2)^2 \geq 0$ for all values in $\mathbb{R}$, so our first criteria for a metric is fulfilled. We also observe that if $d(\mathbf{x}, \mathbf{y}) = 0$ it implies $x_1 - y_1 = 0$ and $x_2 - y_2 = 0$, which means that $x_1 = y_1$ and $x_2 = y_2$, thus d fulfills our second requirement for a metric.

Our requirement of symmetry requires a little more investigation. Proceeding from (2), we find

$$\begin{aligned} (100(x_1 - y_1))^2 + (x_2 - y_2)^2 &= 100^2(x_1^2 - 2x_1y_1 + y_1^2) + (x_2^2 - 2x_2y_2 + y_2^2) \\ &= 100^2(y_1^2 - 2y_1x_1 + x_1^2) + (y_2^2 - 2y_2x_2 + x_2^2) \\ &= (100(y_1 - x_1))^2 + (y_2 - x_2)^2 \\ &= d(\mathbf{y}, \mathbf{x})^2, \end{aligned} \tag{3}$$

where (3) gives us $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$ and therefore fulfills the symmetry requirement.

Our last requirement is the triangle inequality, i.e. $d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z}), \forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^2$.

Let $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^2$.

$$\begin{aligned} d(\mathbf{x}, \mathbf{z}) &= \|\mathbf{Mx} - \mathbf{Mz}\| \quad (4) \\ &= \|\mathbf{Mz} - \mathbf{My} + \mathbf{My} - \mathbf{Mz}\| \quad (5) \\ &\leq \|\mathbf{Mz} - \mathbf{My}\| + \|\mathbf{My} - \mathbf{Mz}\| \quad (6) \\ &= d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z}) \quad (7) \end{aligned}$$

3.2.3 Results

Use d as metric in k-NN classifier and apply to KNOLLC. Explain results.