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.

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

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

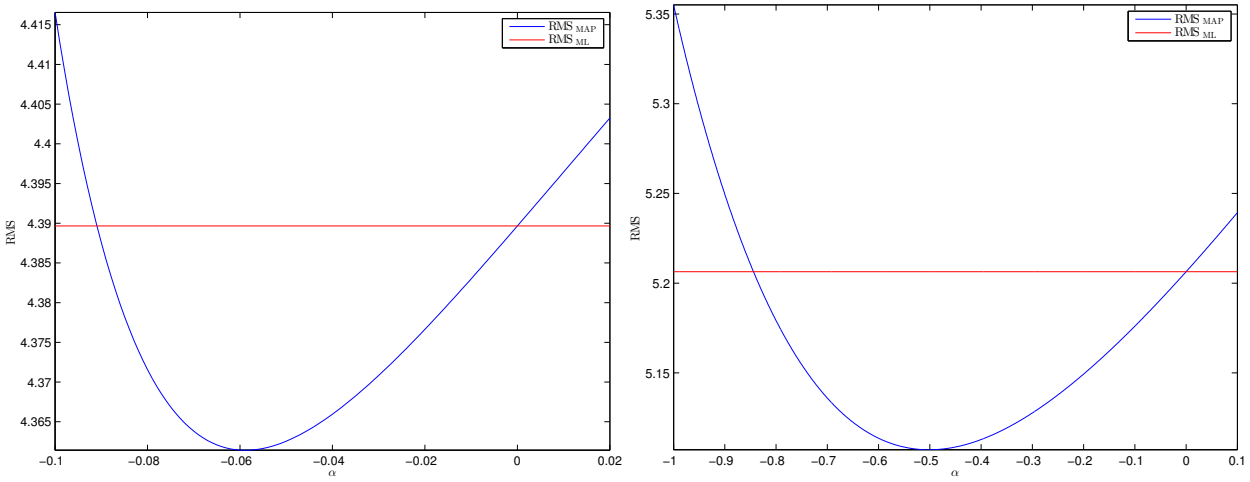
Finding the ML estimate of our parameters for S_1 gives

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

1.1.2 Selection 2

Our second selection S_2 consists only of the data from the 'Abdomen 2' column, giving a design matrix $\mathbf{\Phi}_{S_2}$ 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 could suggest that either the variable 'Abdomen 2' is the most descriptive in terms of body fat, 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

For S_1 the lowest root mean square value is 4.3614, the same as the ML solution differing on the second digit. The difference is larger for S_2 in which we obtain the value 5.1071, differing from the ML solution on the first digit.

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.

1.2.1 Comparison

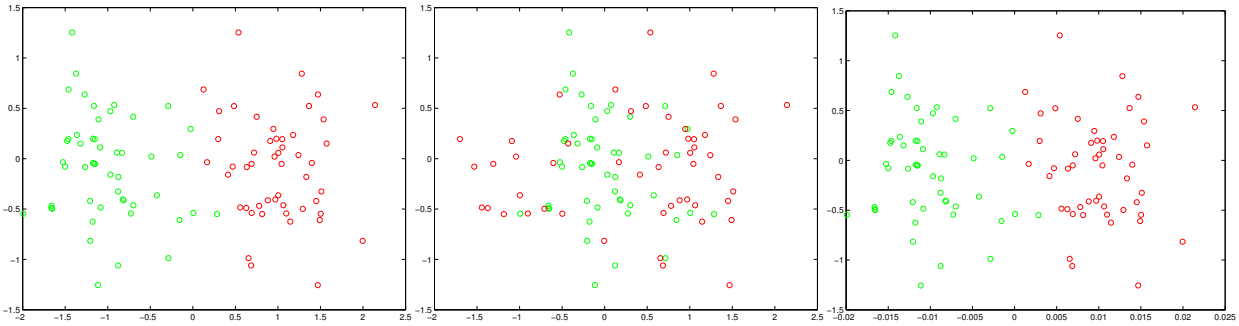
Firstly, we can observe for both plots that when $\alpha = 0$, we obtain the same RMS error for the MAP estimate as for the ML solution. This is expected and demonstrates that when our prior precision parameters are set to zero, the MAP estimate becomes the ML estimate.

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 $[-0.844, 0]$.

The intervals in which the MAP estimate outperforms the ML estimate are very small, and the obtained difference likewise small.

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



(a) knollA (b) knollB (c) knollC

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

2 Linear Discriminant Analysis

Visualize the training data sets in three 2D plots.

Apply LDA to the training data, report accuracies of the classifier on the training as well as on the test sets. Explain the results. Discuss similarities and differences in performance on three data sets. Could a non-linear method do better?

Implement LDA algorithm by hand.

	knollA		knollB		knollC	
k	test	train	test	train	test	train
1	0.04	-	0.23	-	0.20	-
3	0.02	0.01	0.20	0.13	0.32	0.16
5	0.03	0.01	0.19	0.19	0.34	0.21
7	0.04	0.01	0.19	0.19	0.42	0.21
9	0.03	0.02	0.23	0.20	0.46	0.33
	3	3	5	3	1	3

Table 1: Table of results for k-NN classifier on all three knoll problems.

3 Nearest Neighbor Classification

3.1 Nearest Neighbor Classification with Euclidian Metric

The results of the runs of the classifier for different values of k are found in table 1. The bottom field of each column indicates the k for which the classifier gave the lowest error rate.

The error rate is computed by adding up the errors made by the classifier, and then dividing by the test set size. As we have the actual class of a given point (for both test and training sets), this is straightforward to implement.

As a general note, classifying the training set itself will always give us $k = 1$ as the optimal value, which makes sense, because the closest point to a point in both the training and test sets is the point itself, hence the distance will always be zero (and the classification perfect). But it's interesting to see how the classifier behaves when more than one neighbor has to be considered. If a point belonging to one class is surrounded by points in the other class, then it will be misclassified. As can be seen from the results, the optimal value for k in all the training sets is 3, although it should be noted that the error rate for $k = 3, 5$ and 7 in knollA are all 0.1.

In knollA, the error rates are very low (< 0.05) for all the values of k, being lowest when $k = 3$. In knollB and knollC the error rates increase dramatically (≥ 0.19), for knollB the lowest error rates are obtained when $k = 5$ or 7. In knollC the best results are unanimously obtained when $k = 1$.

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 Proof

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:

k	training set	test set
1	0.04	-
3	0.02	0.02
5	0.03	0.03
7	0.04	0.04
9	0.03	0.03
	3	3

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

3.2.2 Results

Using d as metric on KNOLLC we obtain the following results.

The error rates drop dramatically. We could visualize with a plot?