Statistical Methods for Machine Learning
Case 2

Troels Henriksen (athas@sigkill.dk)
Daniel Fairchild (daniel.fairchild@gmail.com)

16th March 2011

We have never used R before, so for the novelty, that is the language we have decided to use for this assignment.

Question 1.1

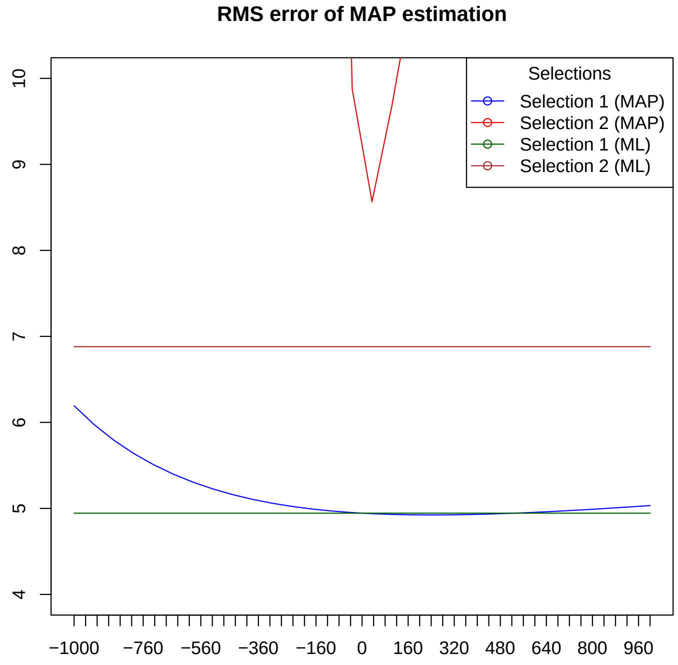
The code for this question is in `src/1112.R`. The pseudoinverse function had to be written from scratch, as we could not find the library supposed to contain it. We obtained the following results.

Selection	RMS error
1	4.944719
2	6.880314

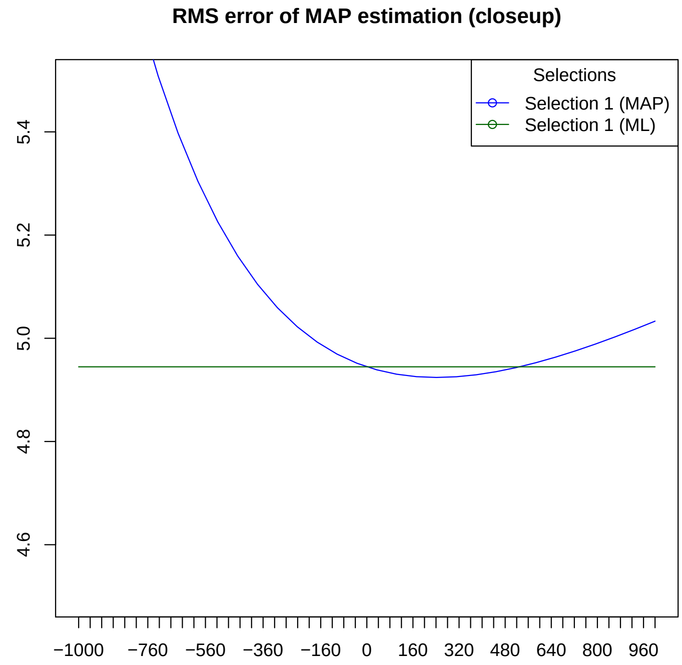
Selection 1 appears to have the smaller error compared to the true results, which is expected as its model incorporates more measurements than selection 2. While body measurements are probably only slightly independent, adding more still constitutes a net information increase.

Question 1.2

The code for this question is in `src/1112.R`.



Below is a close-up of the ML estimate and the MAP estimate.

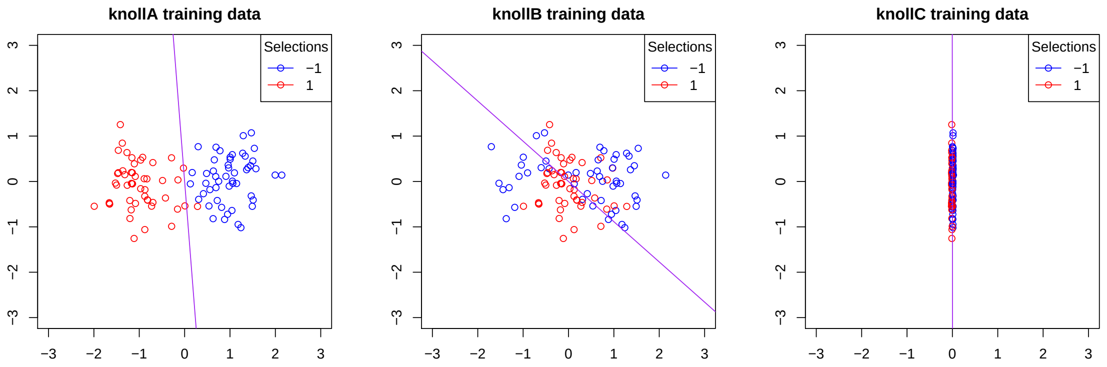


We see that the MAP estimate has a slightly smaller error when the prior precision is in the interval from roughly 0 to 500.

Question 2

The code for this question is in `src/2.R`.

Visualisation

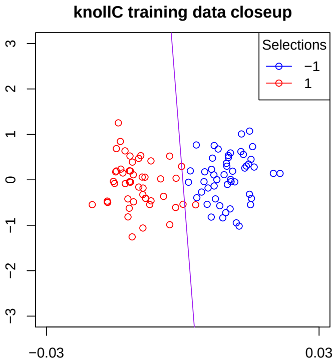


LDA

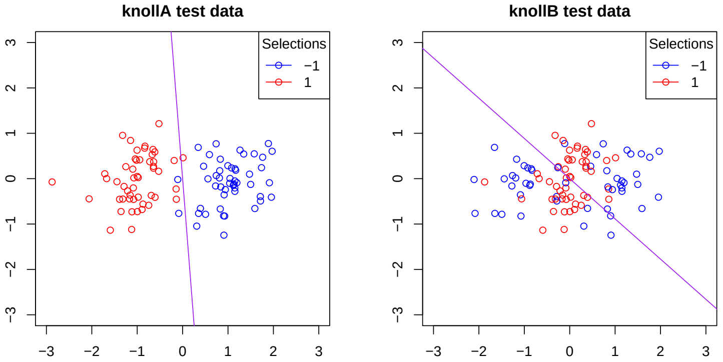
Let class $\mathcal{C}_1$ be the class indexed by 1 and $\mathcal{C}_2$ the class indexed by -1. We have chosen to make use of Fisher's linear discriminant, as sescribed in section 4.1.4 in the textbook, implemented in R. y is classified as class $\mathcal{C}_1$ if $y \geq 0$ and otherwise class $\mathcal{C}_2$ - this threshold is based on the idea that we pick the class that our prediction is closest to. The class dividers are visualised as purple lines in the plots. The accuracy of the linear model is determined by counting the number of correct classifications proportional to the number of data points.

Trained with	Tested on	Accuracy
KnollA	Training set	0.99
	Testing set	0.97
KnollB	Training set	0.63
	Testing set	0.49
KnollC	Training set	0.99
	Testing set	0.97

We note that the first coordinates in the KnollA and KnollC data sets are proportional by a factor of 100, hence their identical behaviour. This is visualised in the plot below. Due to this similarity we will focus only on KnollA and KnollB.



The difference in accuracy between the KnollA and KnollB data sets can be visualised by inspecting the distribution of points in KnollB. It is immediately clear that there is no way to divide the two-dimensional plane with a straight line, such that we will obtain a satisfactory division of the points into their correct classes. On the other hand, it is likely that a nonlinear method could construct a curve that would perform such a division, as the red points do seem to be clustered roughly in the middle.



Question 3.1

Deliverables:

source code of your nearest neighbor classifier; results of k-NN

Question 3.2

Deliverables:

proof that d is a metric; results for the k-NN classifier using the non-standard metric d applied to the knollC data and a short discussion of the results

We check that the given metric fulfills the four requirements for a metric, utilising the properties of norms.

1. $d(\mathbf{x},\mathbf{z}) \geq 0$ (non-negativity): As a norm $||\cdot||$ is always non-negative, this property is immediate.
2. $d(\mathbf{x},\mathbf{z}) = 0 \Leftrightarrow \mathbf{x} = \mathbf{z}$ (identity of indiscernibles): The norm $||\mathbf{Mx} - \mathbf{Mz}||$ is only zero if $\mathbf{Mx} - \mathbf{Mz} = \mathbf{0}$ (by the definition of a norm), which is again only the case if $\mathbf{x} = \mathbf{z}$. Holds in both directions.
3. $d(\mathbf{x},\mathbf{z}) = d(\mathbf{z},\mathbf{x})$ (symmetry):

$$\begin{aligned}
d(\mathbf{x},\mathbf{z}) &= ||\mathbf{Mx} - \mathbf{Mz}|| && \text{(By def.)}\\
&= ||\mathbf{Mx} + (-1)\mathbf{Mz}|| && \text{(Rewriting)}\\
&= ||(-1)\mathbf{Mz} + \mathbf{Mx}|| && \text{(Commutativity of addition)}\\
&= |-1|||(-1)\mathbf{Mz} + \mathbf{Mx}|| && \text{(Multiplying by one)}\\
&= ||\mathbf{Mz} - \mathbf{Mx}|| && \text{(Positive homogeneity of norms)}\\
&= d(\mathbf{z},\mathbf{x}) && \text{(By def.)}
\end{aligned}$$

4. $d(\mathbf{x},\mathbf{z}) \leq d(\mathbf{x},\mathbf{y}) + d(\mathbf{y},\mathbf{z})$ (triangle inequality):

$$\begin{aligned}
||\mathbf{Mx} - \mathbf{My} + \mathbf{My} - \mathbf{Mz}|| &\leq ||\mathbf{Mx} - \mathbf{My}|| + ||\mathbf{My} - \mathbf{Mz}|| && \text{(By triangle ineq. of norms)}\\
||\mathbf{Mx} - \mathbf{Mz}|| &\leq ||\mathbf{Mx} - \mathbf{My}|| + ||\mathbf{My} - \mathbf{Mz}|| && \text{(Simplifying terms)}\\
d(\mathbf{x},\mathbf{z}) &\leq d(\mathbf{x},\mathbf{y}) + d(\mathbf{y},\mathbf{z}) && \text{(By def.)}
\end{aligned}$$

□

Question 3.3

Deliverables:

discussion of average length and distance of random vectors in high dimensional feature spaces