Introduction to Machine Learning - Lab 5

Gustav Sternelöv

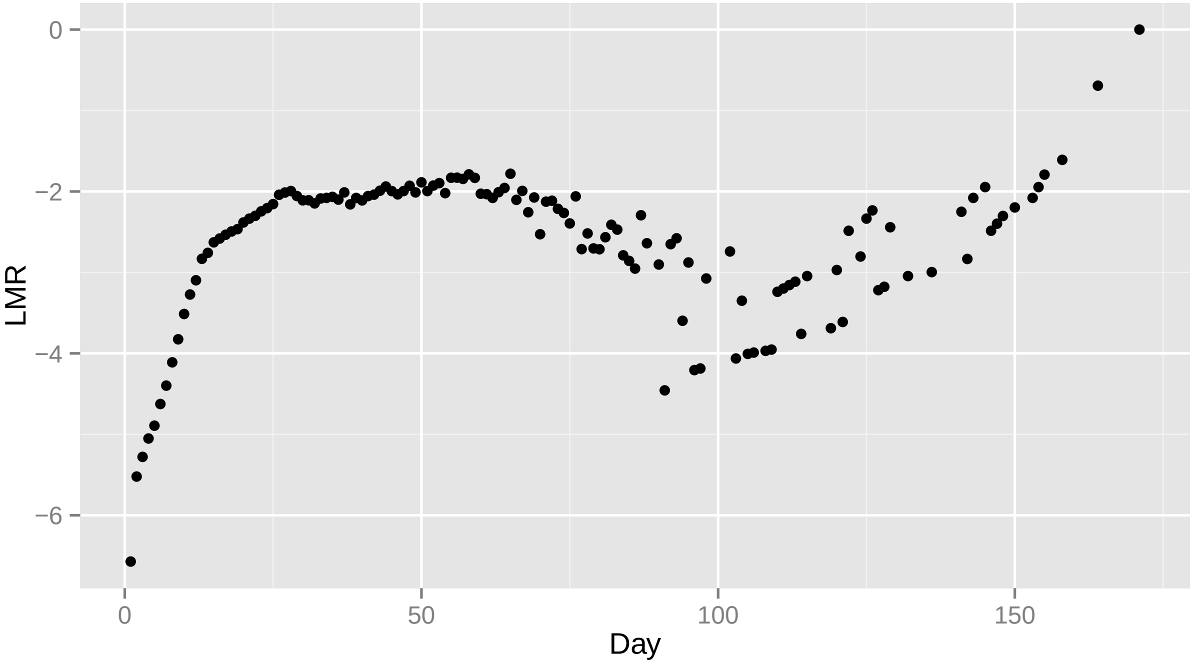
Sunday, November 15, 2015

Assignment 1

The studied data set contains information about the mortality rate for fruit flies for each day. The data comes from a study where the theory that the mortality rates (probability of dying per unit time) of many organisms increase at an exponential rate was tested.

1.1

The variable LMR, that is the logarithm of the variable Rate, is created and plotted against the variable Day.



1.2

1.3

1.4

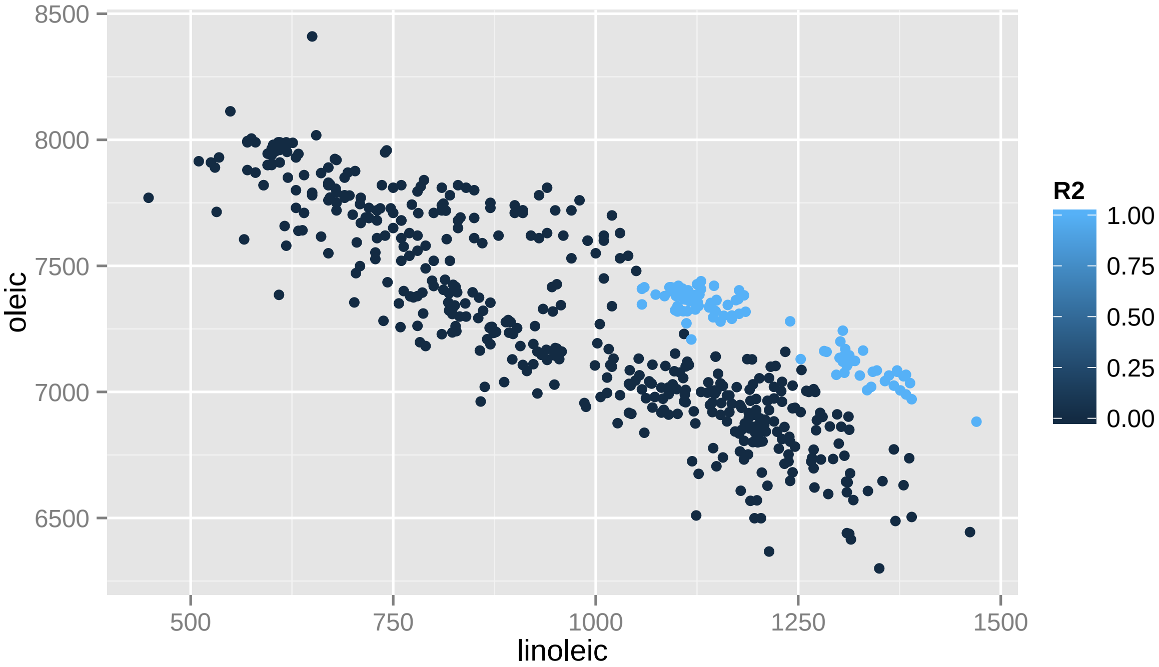
1.5

Assignment 2

The data set analysed in this assignment consists of information about 572 italian olive oils coming from different regions of the country. How much of different acids each olive oil contains and from which region and area the olive oil comes from is the information given.

2.1

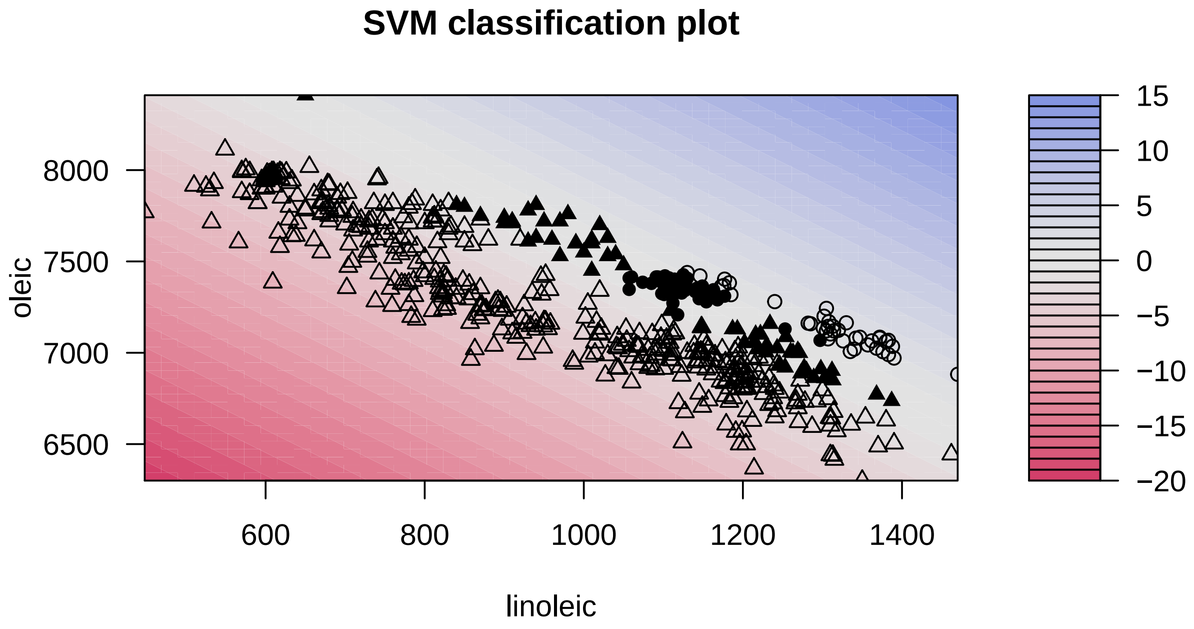
Two of the acids in the data set are *Oleic* and *Linoleic*. In the following graph these acids are plotted against each other and coloured after region where oils from region two are light blue and the others are dark blue.



The oils from region two are quite easy to identify since they lies rather separately from oils from the other regions. At least that is true for the majority of the observations from region two. Some of the dark blue points lies very close to the outer edges of the group of light blue points. For these observations it may be hard for a model to correctly classify an olive oil as coming from region two or from one of the other regions.

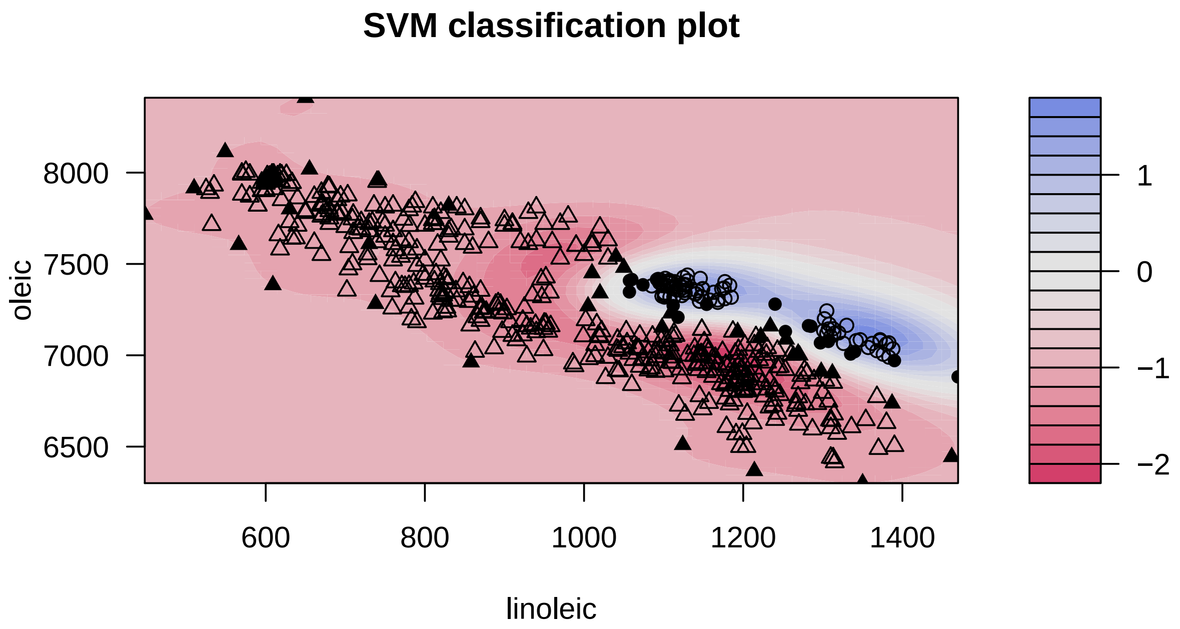
2.2

a)



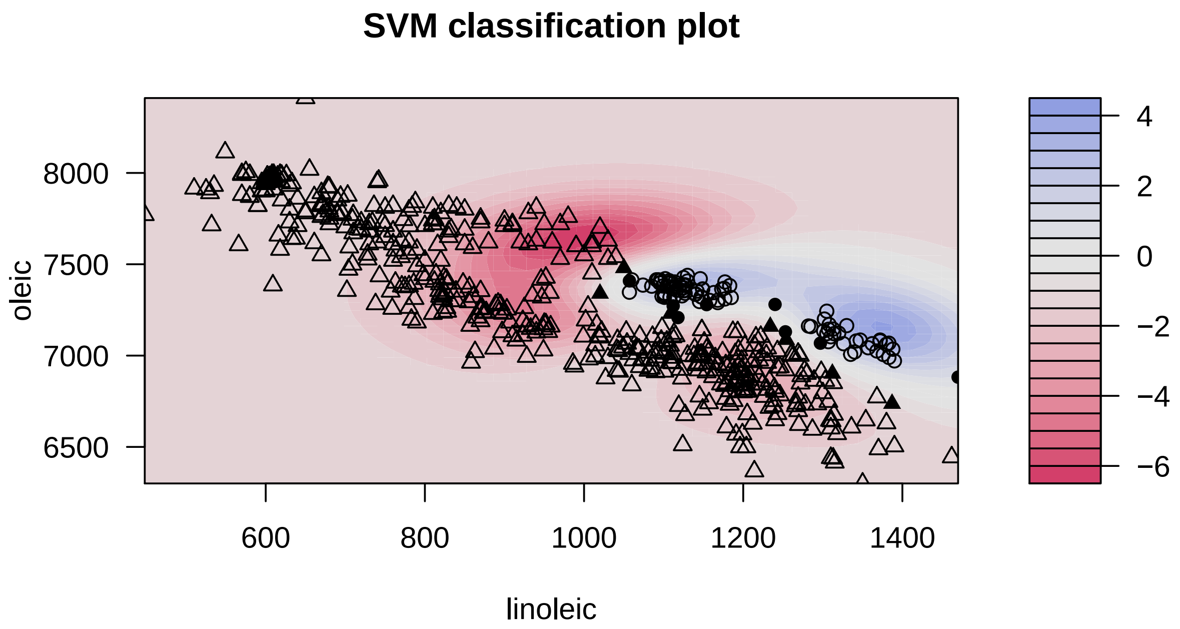
The misclassification rate: 0.0524476
The amount of support vectors: 119

b)



The misclassification rate: 0.0052448
The amount of support vectors: 52

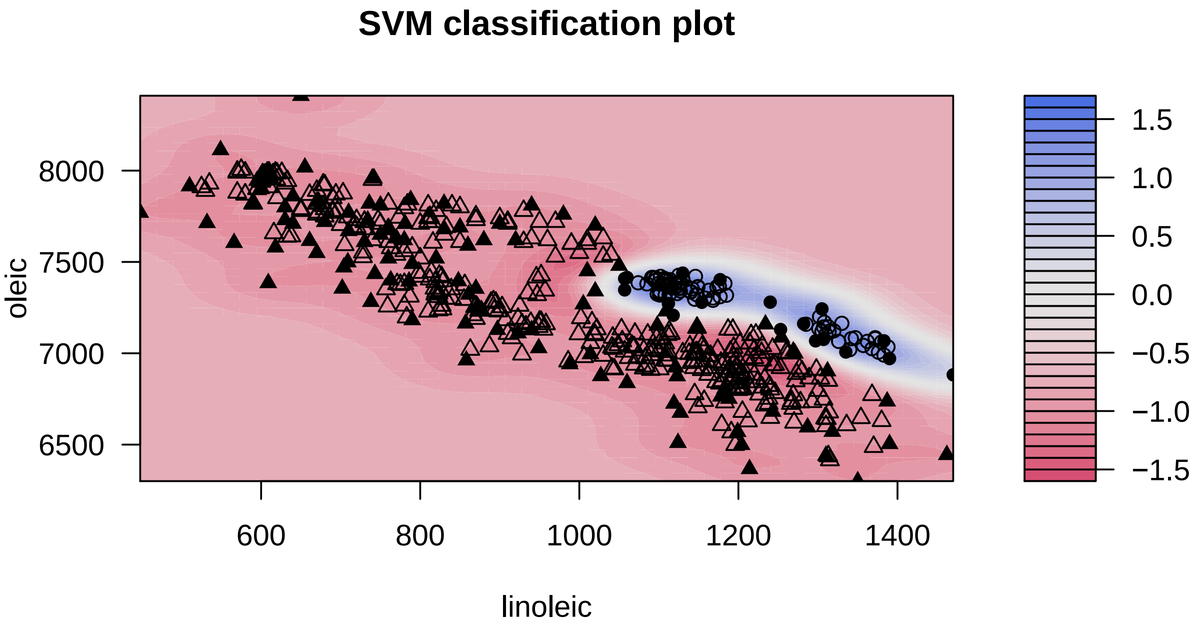
c)



The misclassification rate: 0.0034965

The amount of support vectors: 15

d)



The misclassification rate: 0.0052448
The amount of support vectors: 119

Comparison of models

In terms of misclassification rate model *c*, the model with RBF kernel and penalty for C equal to 100, seem to be the best. It has the lowest misclassification rate, even though it also should be mentioned that the difference between the misclassification rates for the models is very small.

How does the parameters chosen in *c* and *d* influence the classification? The value of C defines the cost of constraints violation.

The amount of support vectors in the models differs significantly. In model *a* and in model *d* 119 of the 572 observated values are used as support vectors. Less than a half of this amount of support vectors are used in model *b*, 52, and in model *c* 15 values are used as support vectors.

2.3

The misclassification rate: 0
The amount of support vectors: 19

How do I find the cross-validation score?