Introduction to Machine Learning - Lab 5

Gustav Sternelöv

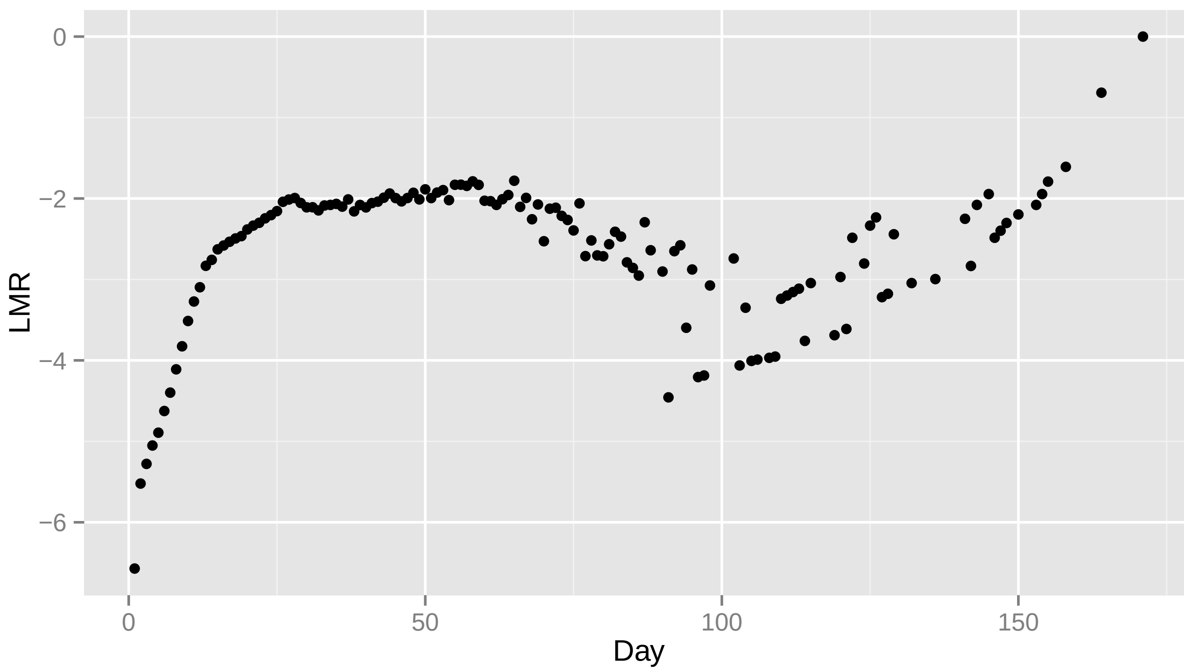
Wednesday, November 18, 2015

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.

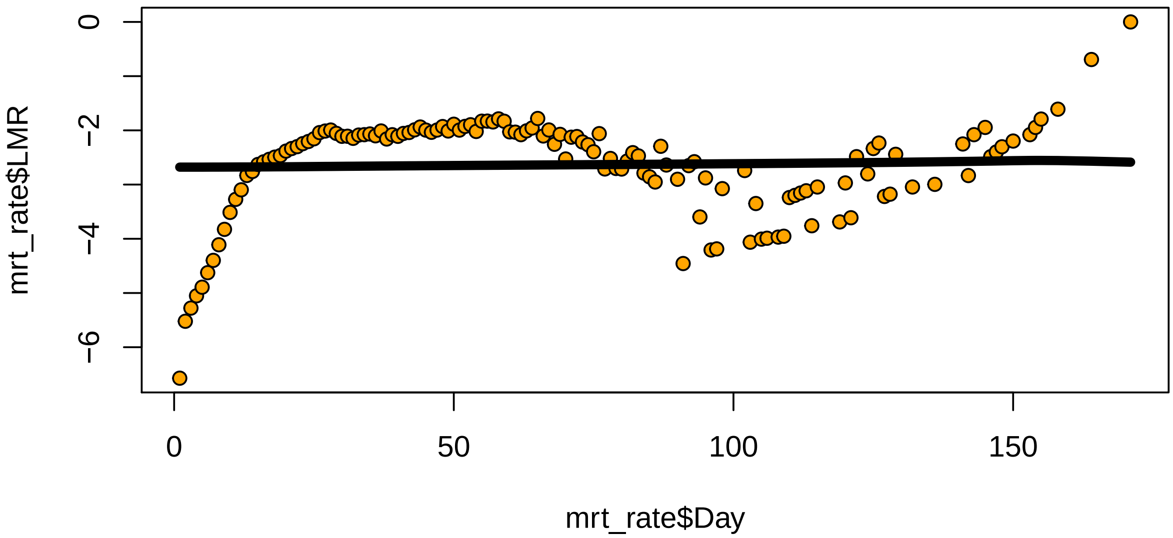


1.2 - 1.3

A function that performs Nadarya-Watson kernel smoothing with Epanecnikov kernel is implemented and tested in this step. The function depends on an x and y vector, a vector Xtest that contains the grid and the value λ that represents the bandwitdh.

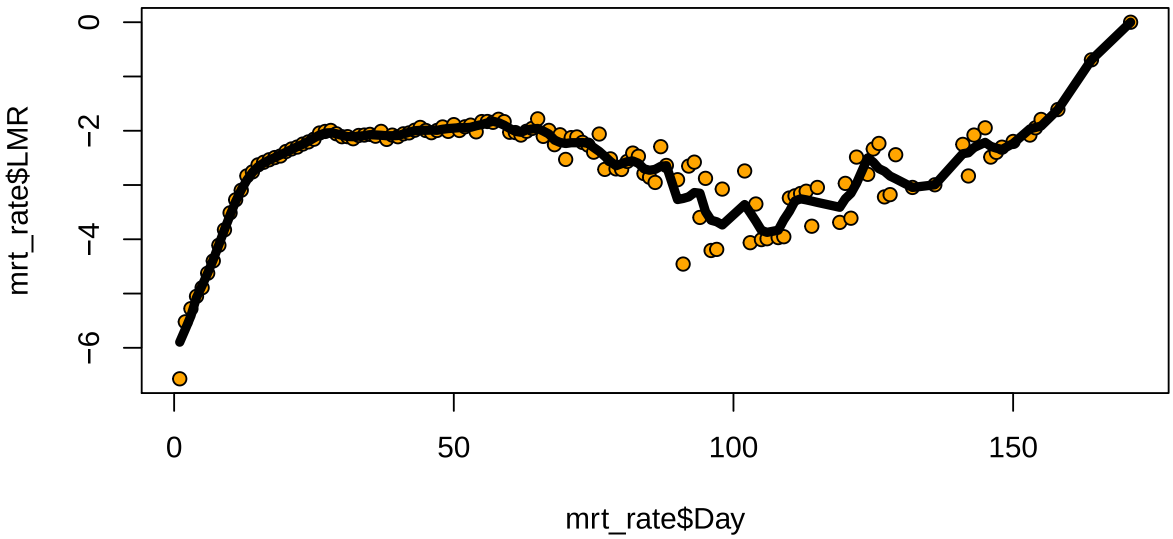
a)

The aim here is to get a very smooth curve and this is achieved by setting λ to 150.



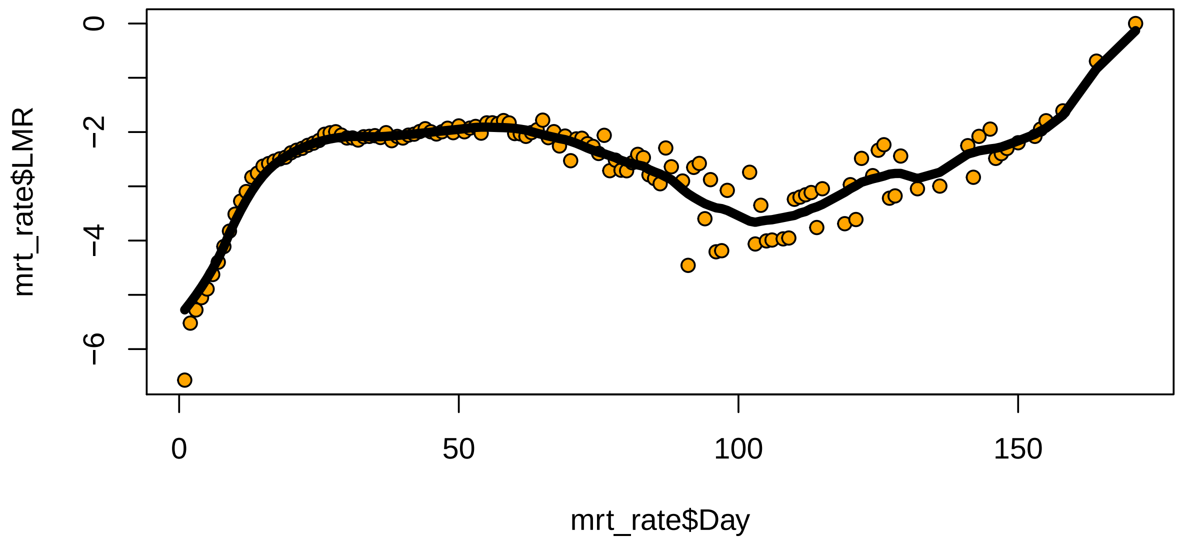
b)

The opposite to a very smooth curve is a very wiggly curve. An example of this type of predicted curve is given for a λ value equal to 3.



c)

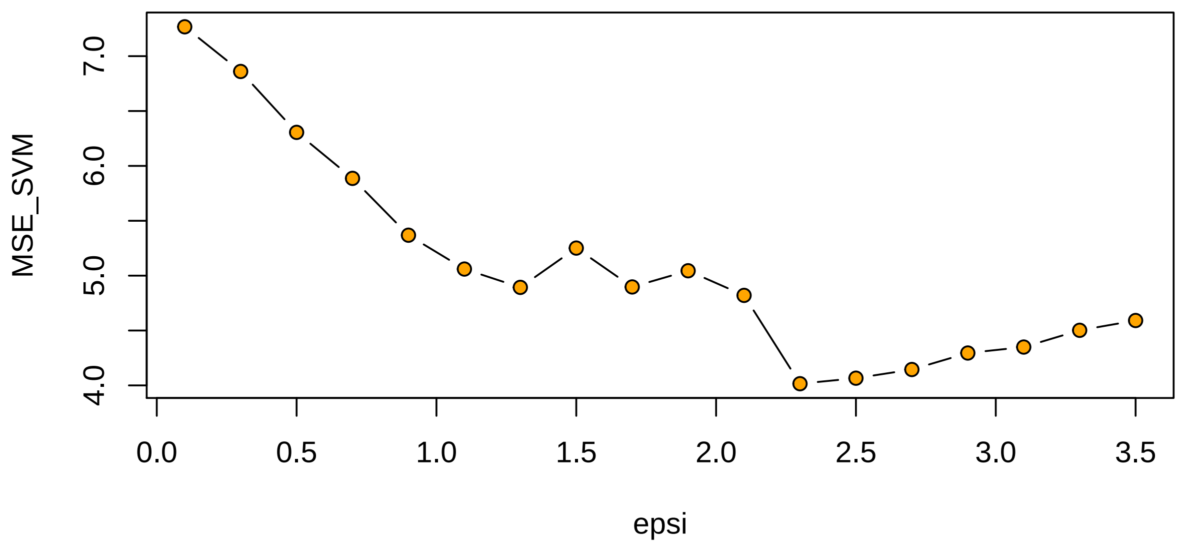
In the last example the aim is to get a predicted curve that seems to be a reasonably good fit. This is achieved by choosing a λ value that neither overfit nor underfit data. A value for λ equal to 8 then seem to be an appropriate choice.

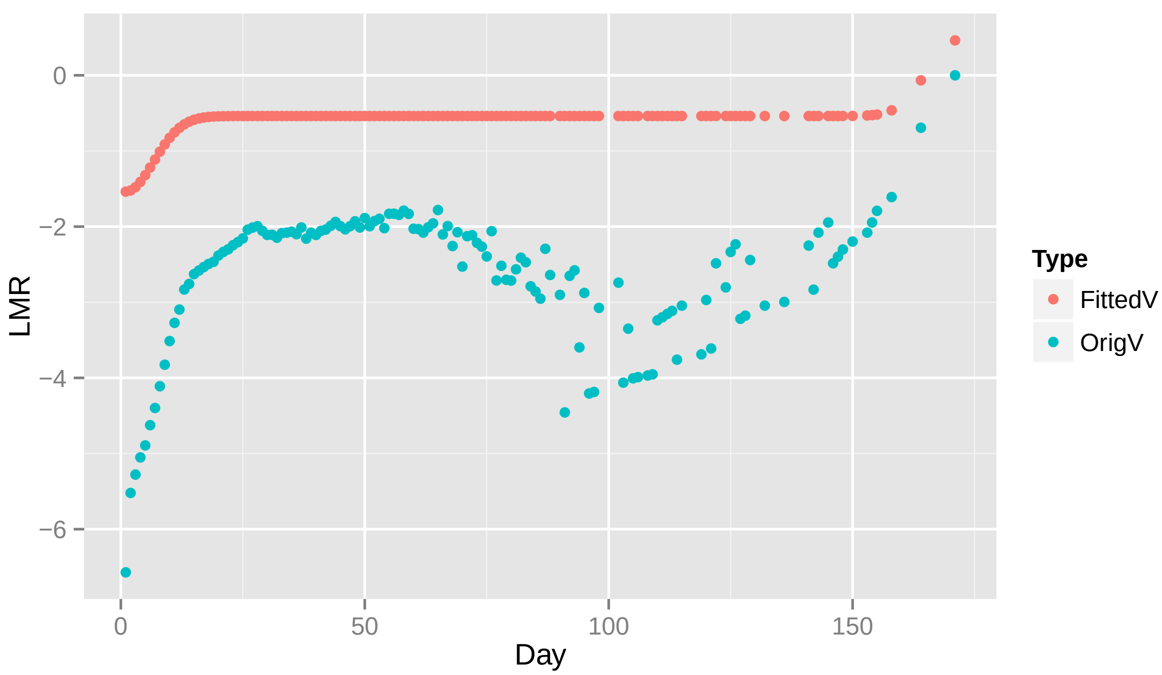


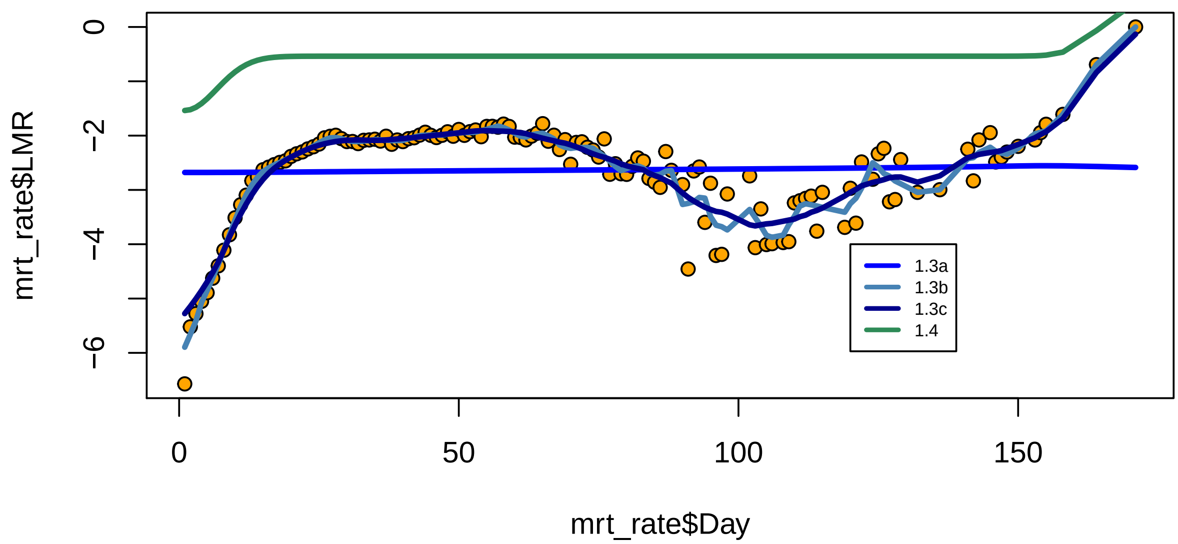
Comparison of the predicted curves

The connection between the bandwidth parameter λ and the form of the predicted curves is straightforward. High values of λ results in more smooth curves and low values results in curves with a more wiggly form. The best model is thought to be the model presented in *3.c* and the estimated MSE value for this model is 0.0918153.

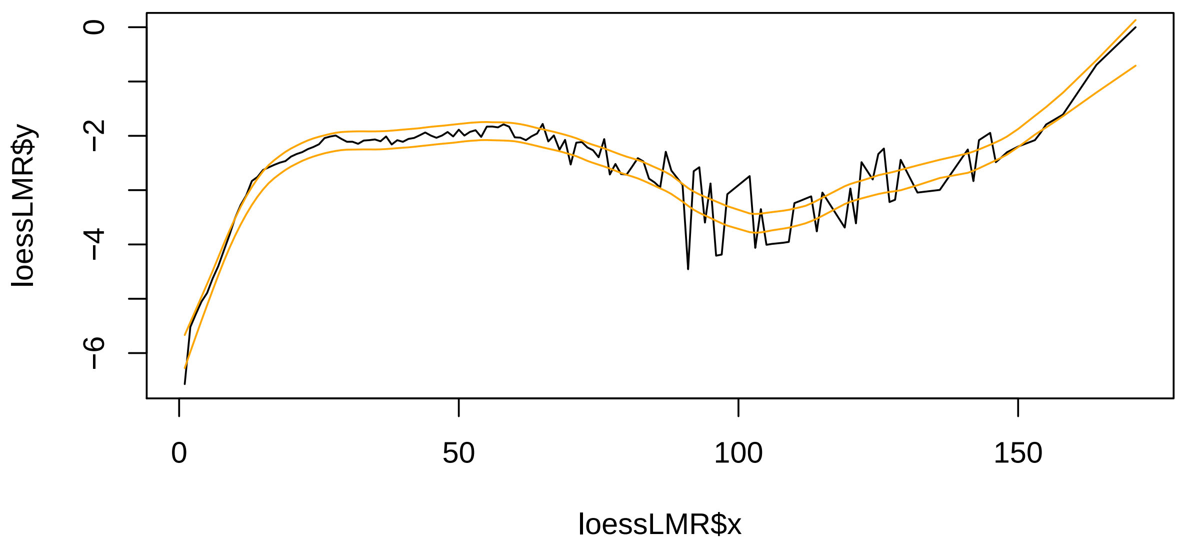
1.4


MSE_SVM
4.0
5.0
6.0
7.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
epsi


LMR
0
−2
−4
−6
0
50
100
150
Day
Type
FittedV
OrigV


0
−2
−4
−6
mrt_rate$LMR
0
50
100
150
mrt_rate$Day
1.3a
1.3b
1.3c
1.4

1.5

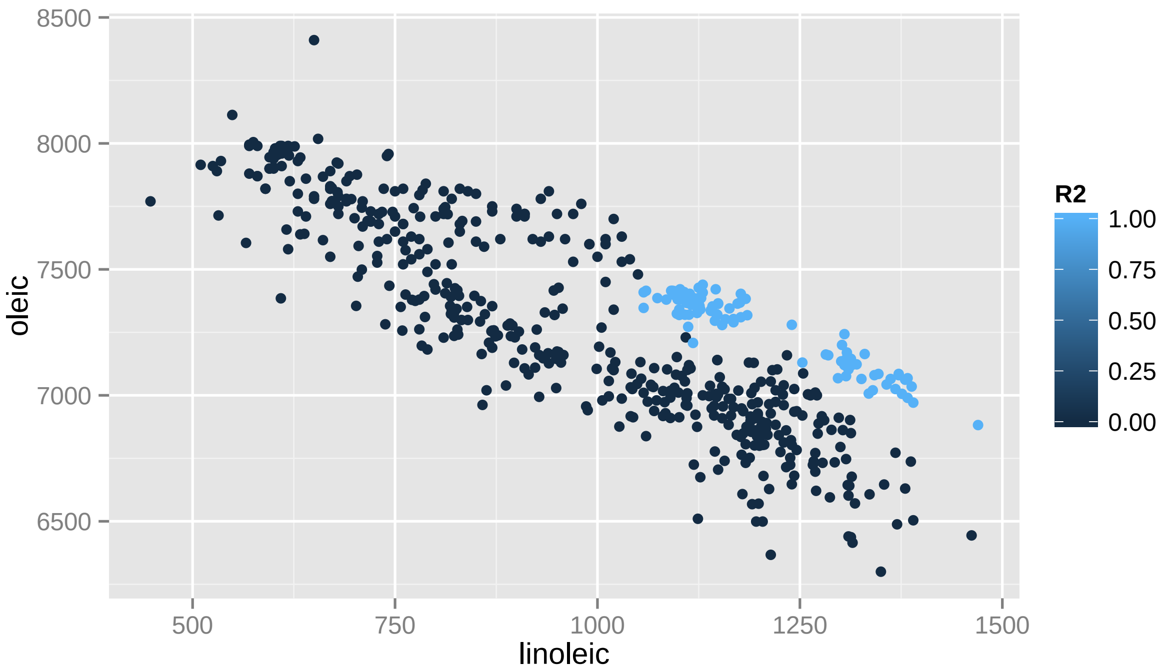

0
−2
−4
−6
loessLMR$y
0
50
100
150
loessLMR$x

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

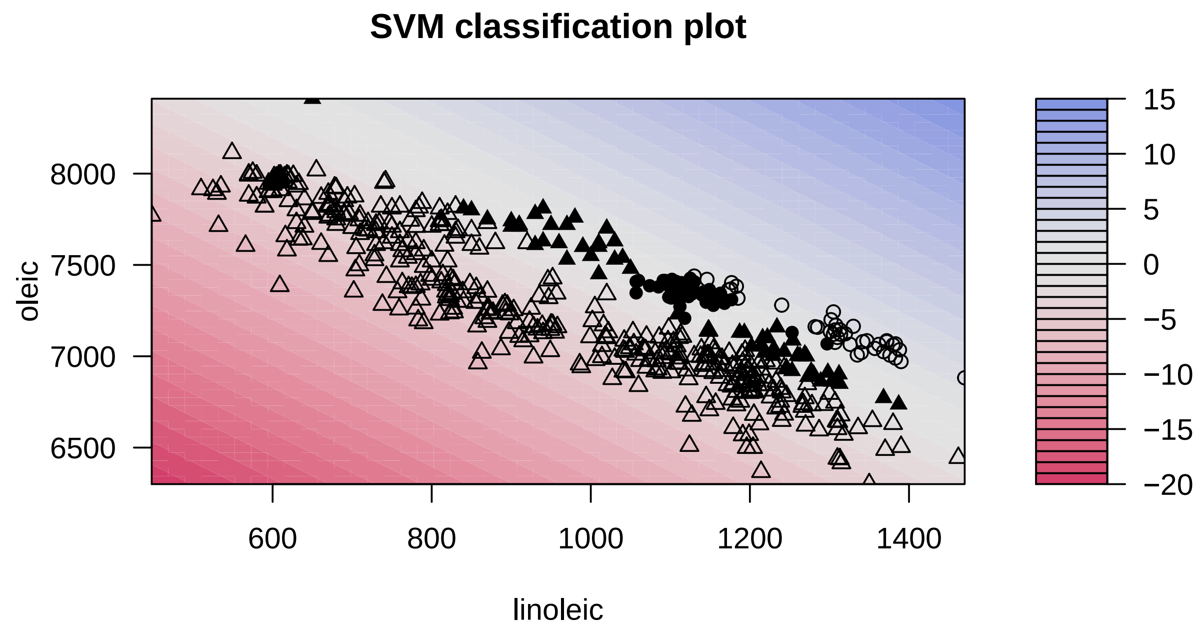
Two of the acids in the data set are *Oleic* and *Linoleic*. In the following graph these acids are plotted against each other and coloured after region where oils from region two are light blue and the others are dark blue.



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.

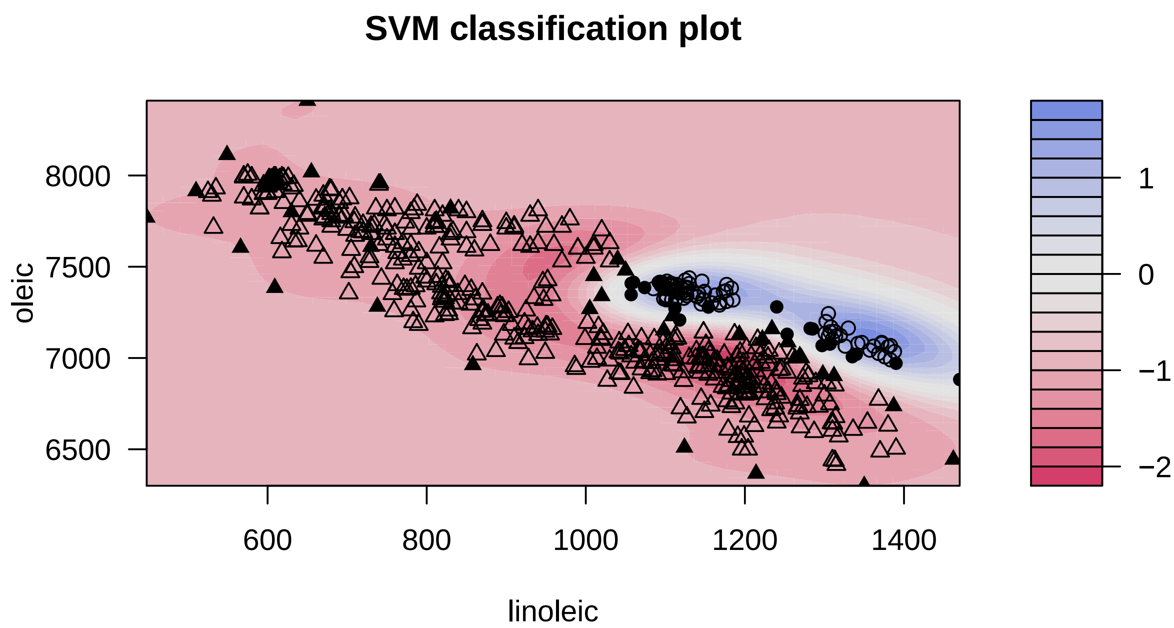
2.2

a)



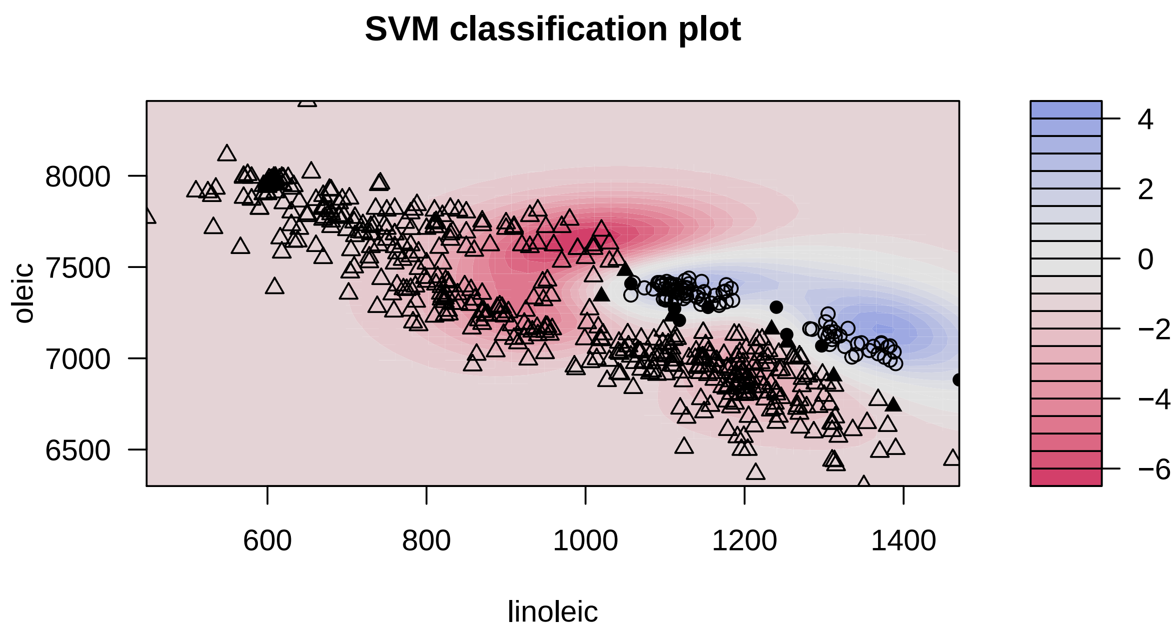
The misclassification rate: 0.0524476
The amount of support vectors: 119

b)



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

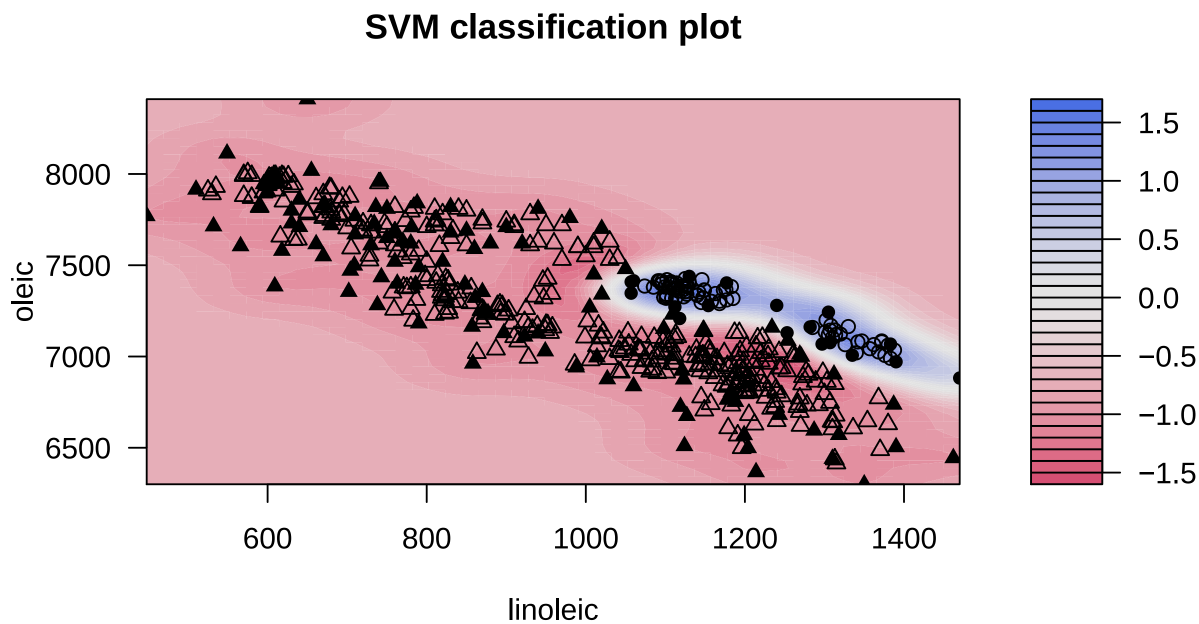
c)



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?