Machine learning II
Master Data Science
Winter Semester 2019/20

BEUTH HOCHSCHULE FÜR TECHNIK BERLIN
University of Applied Sciences

Prof. Tim Downie

Workshop 6
Support vector machines

In this workshop you will need the packages `e1071`, `ISLR`, `rpart`, `ROCR` and `MASS`.

Exercise 1 Tutorial

A good website explaining the subject of SVMs with out too much maths can be found here:*Support Vector Machines in R* https://www.datacamp.com/community/tutorials/support-vector-machines-r Link

Included in this tutorial is a nicer plotting method than the deafult in the package `e1071`. If using more than 2 predictor variables then you need to take care defining the grid using the two variables you want to plot.

Read through this tutorial and run the R code yourself.

Exercise 2 Non-linear SVMs: using different kernels

If you have not yet worked through Section 9.6.1 in James page 330 on linear SVMs, do this now.

Section 9.6.2 fits the the SVM algorithm using non-linear kernels. When you get to the `svm` command using the radial kernel function, fit SVM models with the following kernels:

- Linear kernel

```
svmfit=svm(y~., data=dat[train,], kernel="linear",   cost=1)
plot(svmfit, dat[train,])
```

  No hyperplane is found, and all the red points are misclassified. Varying the cost does not help.

- Polynomial kernel with degree 2

```
svmfit=svm(y~., data=dat[train,], kernel="polynomial",  degree=2,
   gamma=1, cost=0.1)
plot(svmfit, dat[train,])
```

  With a quadratic polynomial kernel two distinct borders are possible. Try increasing the cost using a few values between 1 and 10. The boundary now becomes an ellipse.

Remember that the parameter cost is the reverse of the parameter C in the lecture notes. A high value for `cost` penalises heavily each support vector. In the notes the parameter was an allowance and a high value allowed more support vectors.

- Try increasing the degree. With d=3 no sensible boundary is found. With d=4 there is a reasonable fit but with an unrealistic boundary.

Using a radial kernel the boundary can become more irregular.

Continue to work through James until the end of section 9.6.4.

Exercise 3 Using SVMs on a practical data set
In Section 9.6.5 you will analyse the Khan *gene expression* data (in the ISLR package). These data have few observations (63 in the training set) but many variables for 2308 genes, with values corresponding to how much of each was "expressed". Many statistical learning methods have algorithmic problems when $p \gg n$.

When you have finished section 9.6.5 fit an `rpart` classification tree to the Khan data. Obtain a classification matrix for the training and for the test data. Compare the results with the SVM result in the last section.
Hint: the syntax for `predict` for an `rpart` object is a liitle different as for an `svm` object:
`predict(Khan.tree,newdata=dat.te,type="class")`

1 Exercise to do at home

Work through Exercise 3 in James et al. Section 9.7 page 368. This could be typical of an exam question on this subject.
N.B. In Part (g) the task is to find a hyperplane which is separating but not optimal, rather than any non-optimal hyperplane.