Machine learning II
Master Data Science
Winter Semester 2023/24 **Prof. Tim Downie**

Workshop 6
Introduction to support vector machines

In this workshop you will need the packages `ISLR2`, `MASS`, `e1071` and `pROC`. Load these packages, installing them if need be.

Exercise 1 Linear Support vector classifier

The package to fit SVM models in R is called `e1071`, which you used in Workshop 5.

Work through Section **9.6.1** in James et al. on page 389, which applies the `svm()` function to a small simulated data set. Read the text carefully! Notice in particular, that in `e1071` the cost is defined differently as in the notes, so a large cost gives a small margin. Also I find that the scatter plot of an object of class `svm` is poor, in particular some points are hard to see, as they are partially cut off. I extend the graphics margins a little using e.g.

```
> plot(svmfit, dat,xlim=c(-2.1,2.4),ylim=c(-1.4,2.7))
```

Stop at the end of Page 392. You will continue with Section 9.6.2 next week.

Exercise 2 SVM Model for the `College` data set

In Moodle there is an R file for you to use as a basis for this exercise. As usual you will need to replace the `???` with appropriate commands.

In Workshop 3 you developed a model for predicting the number of students accepted to different colleges and universities in the USA using GAM modelling. As support vector machines are a binary classifier, we will this time predict the variable `Private`, whether the college is private (yes) or public (no).

Short notes are given in the R file as comments. The overall approach and a few longer comments are given here.

1. You will start by defining a training data set and a test data set. The test data are set aside and used for evaluation at the end.

2. Fit an SVM model using two variables `log(Accept)` and `PhD` to predict `Private`. For now use a cost parameter of $c = 0.1$, no scaling and linear kernel.

3. To obtain the SVM plot, you need to specify which two variables should be plotted in on the y and x axis using `lAccept~PhD`. In James et al, the order of the variables was chosen so that the default plot was appropriate.

4. Obtain the confusion matrix and the accuracy for this and each further model.

5. When fitting a three variable model, we have to choose which two variables to plot. This means we choose a value for the third variable. The default value is 0, which is rarely a good choice. The second plot for this model uses the argument `slice=list(lExpend=9)` which is approximately the mean value of the variable `lExpend`. Try varying this value to see the effect it has on the visual representation of the boundary. Note this argument only changes the graphical representation; the model and the boundary remain unchanged. The scatter plot can be helpful to investigate SVM models for small data sets with just two variables but is not so helpful in practice.

6. Now fit an SVM model with all the variables used in Workshop 3. There is a warning that the maximum number of iterations has been reached, and implying the iterative algorithm has not converged. It would be possible to increase the maximum number of iterations, but we'll try another approach, scaling the input data, which is easily done using the argument `scale=TRUE`.

7. As in Exercise 1, use `tune()` to find the optimal value for the cost parameter. Note: I found two different optimal cost values when running the code several times. The cross validation method uses 10-Fold C.V. so a random number generator is used to define the folds. The results are similar whichever of these "optimal values" are used.

8. Obtain an ROC plot for the final model using the library `pROC` (which you used in Workshop 5 and ML1). The ROC function needs the probabilities of `Private=Yes`, which can be obtained from `svm()` and `predict()` by using the argument `prob=TRUE`. The way of accessing the probabilities is not particularly pretty; we want the first column of the matrix accessed using `attr(ypred, "probabilities")`. Obtaining the probabilities is not actually part of the SVM algorithm, they are calculated afterwards using an adapted type of logistic regression.

9. Finally obtain the predictions on the test data and the corresponding ROC-plot.