

Lab #4: Support Vector Machines

Biostat 216: Machine Learning in R for the Biomedical Sciences

The purpose of this lab is to get you working with the Support Vector Machines algorithm, and also to sharpen your R skills.

1. Read in the dementia data "dementia.csv" into a data frame called `dat`
2. Load the package `e1071`
3. Explore the data – use `names`, `head`, `pairs` (you will need to look at just a subset of columns for that function), `summary`
4. Examine a plot of `X3rd_Ventricle` by `Left_MFC_medial_frontal_cortex`. Generate the plot with different colors for each of the values of `dementia`
5. Subset the data to only the variables `X3rd_Ventricle`, `Left_MFC_medial_frontal_cortex`, and `Dementia`
6. Rename the variables to `ventricle`, `frontalCortex`, and `Dementia`
7. Change the values for `Dementia` such that `FALSE` = "-1" and `TRUE` = "1", and convert to a factor
8. Fit an SVM with `Dementia` as outcome, `ventricle` as one predictor and `frontalCortex` as a 2nd predictor (use `cost = 1` and `scale = FALSE`, and `kernel = "linear"`)
9. Why do you think it is not necessary to standardize the predictors, e.g. by using `scale = TRUE`?
10. Look at the resulting fit with the standard `svm` plotting function
11. Look at the resulting fit with the `svmPlot` function that you can download from the CLE as the file "svmPlotFunction.R"
12. Try to get a rough idea of how the `svmPlot` function works by looking at the code
13. Set the random seed to value 4
14. Find the best value of `cost` among possible values of 1, 5, 10, 15 using cross-validation for the same predictors and outcome with the linear kernel (use the `tune` function as shown in the lecture -- don't worry about the warnings)
15. Take a look at the plot for the best fitting `cost` using `svmPlot`
16. Fit an SVM with `Dementia` as outcome, `ventricle` as one predictor and `frontalCortex` as a 2nd predictor (use `cost = 1` and `scale = FALSE`, but this time with `kernel = "radial"`)
17. Look at the resulting fit with the standard `svm` plotting function

18. Set the random seed to value 4
19. Find the best value of cost among possible values of 1, 5, 10, 15 using cross-validation for the same predictors and outcome with the radial kernel
20. Take a look at the plot for the best fitting cost using the standard svm plotting function