

Homework #3: Support Vector Machines, Cross-Validation, and Regression Trees

Biostat 216: Machine Learning in R for the Biomedical Sciences

Support Vector Machines

1. Consider the following simple dataset.

```
##   X1 X2   Y
## 1  3  4 Red
## 2  1  2 Red
## 3  3  3 Red
## 4  1  4 Red
## 5  2  1 Blue
## 6  4  3 Blue
## 7  4  1 Blue
```

- (a) Sketch the observations (1 point)
 - (b) Sketch the optimal separating hyperplane (with respect to maximizing the margin in a maximal margin linear classifier) and provide its equation (2 points)
 - (c) Describe the classification rule for the maximal margin classifier. The answer should be along the lines of “classify as Red if $\beta_0 + \beta_1 X_1 + \beta_2 X_2 > 0$ and classify to Blue otherwise” and provide the values of β_0 , β_1 , and β_2 (2 points)
 - (d) On your sketch indicate the margin for the maximal margin hyperplane (1 point)
 - (e) Indicate the support vectors for the maximal margin classifier (2 points)
 - (f) Argue whether a slight movement of the 4th observation would or would not affect the maximal margin hyperplane (1 point)
 - (g) Sketch a separating hyperplane that is not the optimal separating hyperplane and provide the equation for this hyperplane (2 points)
 - (h) Draw an additional observation on the plot so that the two classes are no longer separable by a hyperplane (1 point)
2. Load the `e1071` package
 3. Read in the dementia data “dementia2.csv” into a data frame called `dat`.
 4. Generate a table of the elements in the `MacCohort_kr` variable (1 point)
 5. Set the random seed to 7 and then split the data into 2 sets (440 train, and 220 test)
 6. Fit a support vector machine to the multiclass outcome of `MacCohort_kr` and predictors `Left_MTG_middle_temporal_gyrus`, `Left_Amygdala` and `Left_AIns_anterior_insula`, with a radial basis kernel and no rescaling on the training data. Set `cost = 1` and `gamma = 0.0001`. (3 points)
 7. What kind of approach is `svm` taking to generate a multiple class prediction? Briefly describe the approach and any limitations thereof. (3 points)

8. Determine the predictions of the SVM fit on the test dataset. (1 points)
9. Calculate the predictive accuracy of the SVM fit in the test set (1 points)
10. Is this doing worse than guessing at random? Why/why not? (2 points)
11. Now use cross-validation in the training data to find the best combination of cost and gamma among the possible values of cost = 0.01, 0.1, 1, 10, 100, and of gamma = 0.00001, 0.0001, 0.01, 0.1, 1. First set the seed to 5. (4 points)
12. Now fit the SVM to the full training data with the optimal parameters and determine the classification accuracy in the test set. (2 points)
13. How does the accuracy with the optimized parameters compare with that of the original choice? (1 point)