Homework #5: Penalized regression and unsupervised learning

Biostat 216: Machine Learning in R for the Biomedical Sciences

Penalized regression

1. Read in the dementia data "dementia2.csv" into a data frame called `dat`
2. Load the `glmnet` package
3. Set the random seed to 4 and then split the data into 2 sets (400 train, and 260 test) – 2 points
4. Convert the training data into components `xTrain` and `yTrain` that are suitable for use in the `glmnet` package. For `xTrain` use all variables except `Dementia` and `MacCohort_kr`. For `yTrain` we will use `Dementia` – 2 points
5. Perform cross-validated lasso in the training data to select the optimal penalty parameter lambda. Set `Dementia` as outcome with all other variables as predictors except `MacCohort_kr`. Use the function `cv.glmnet` to perform the cross-validation. Note also that you will need to tell cv.glmnet that you are fitting a binary outcome – 4 points
6. What is the optimal value of lambda? – 1 point
7. Now refit lasso to the full training data using the optimal value of lambda – 2 points
8. Evaluate the classification accuracy of this final model on the test dataset – 4 points
9. Generate the coefficients for the lasso model using the following `predict` function `predict(lassoMod,type="coefficien ]`
10. Now generate only the non-zero coefficients – 2 points
11. How many coefficients were dropped in the optimal model? – 2 points

Clustering

12. Generate a dataframe `datpred` consisting of all variables in `dat` except the outcomes `Dementia` and `MacCohort_kr` – 2 points
13. Perform K-means clustering for K=1,2,. . . ,20. In each case use 15 different sets of starting configurations. You might find creating a list and looping through elements of the list to save you some typing – 3 points
14. Plot the total within cluster sum of squares against cluster size for the K-means models fitted – 4 points
15. What would your preferred cluster size be? Justify your choice – 2 points