MLDS 400 Lab Assignment 2

Due Nov 12 at 11:59pm

Instructions: Please submit a report file that includes: a short answer, related code, printouts, etc. for each problem (where necessary). Push your answers to Github or Canvas. All programming must be in R (or R Markdown).

Problem 1

You will analyze the `gradAdmit.csv` dataset (the same dataset from Lab 4). This dataset contains a list of students (rows), along with whether or not they were admitted to graduate school (`admit`), their GRE score (`gre`), their GPA (`gpa`), and the prestige of their undergraduate university (`rank`). Your goal is to predict whether or not a student was admitted to graduate school based on the other data, using Support Vector Machines (SVM).

Problem 1a

Split the data into testing, training, and validation datasets for cross validation (CV). First, hold out 20% for your test dataset. On the remaining 80%, split it into 5 folds. Make sure to seed your random number generator and/or store the indices to ensure your datasets remain consistent through runs.

Problem 1b

Train a number of SVM models (using different hyperparameters) on the training set for each CV fold. For each run, report the accuracy on both the training and validation datasets, averaged over the folds. Use the same split as Problem 1a. Try using various kernel functions, such as `linear`, `polynomial`, `radial basis` (or Gaussian), etc. Also, try to tune their respective hyperparameters (`degree`, `gamma`, and `coef0`), and the value for `cost` (or C), based on the validation accuracy. For a full list of the SVM arguments, visit https://www.rdocumentation.org/packages/e1071/versions/1.7-2/topics/svm. Which kernel(s) perform better than others for this problem? Which hypereparameter values are optimal? Which model performed best on the validation dataset? Do NOT use parameter fitting functions from other R packages.

Problem 1c

For your best model from Problem 1b, retrain it on the full training set (the 80% that was used for training and validation) and compute the accuracy on the test dataset (the 20% that until this point was untouched).