Homework 2

Statistical Machine Learning (GR5241)

Instructor: Kamiar Rahnam Rad **TAs:** Navid Ardeshir, Chengliang Tang

1. Imbalanced data refers to a classification problem where the number of observations per class is not equally distributed. In this question we subsample the IMDB data to create imbalanced data. To subsample the data, keep all the negative observations, but only keep the first 4000 (out of the 12500) of the positive observations. Do this separately for both train and test. We end with 12500 +4000 = 16500 observations separately for training and testing. Use the $p = 2500$ most frequent words as predictors.

   (a) Fit a LASSO logistic regression model to the training data and use 10-fold cross-validation using the AUC as a measure of error to tune λ. For the optimal λ answer the following questions.

      i. What are the top 5 words associated with positive reviews? (2 points)

      ii. What are the top 5 words associated with negative reviews? (2 points)

      iii. In the training set, for each observation, using logistic regression, calculate $\Pr[y = 1|X = x]$ for the model fitted using the λ found from 10-fold CV. For a sequence of thresholds $\theta = 0, 0.01, 0.02, 0.03, \cdots, 1$, calculate the the TPR and FPR, and using these plot the ROC curve and calculate the AUC. Note that to calculate the AUC you need the area under the ROC curve. Repeat the same for the test set. Plot the ROC for the train and the test on the same graph. Also in the graph report the train and test AUC. In other words, one figure should show the ROC of the train and test, and values of the AUC. Use color coding and make sure to label the horizontal and vertical axes. (2 points)

      iv. For $\theta = 0.5$, what is the training and testing type I and type II error? (2 points)

      v. For what θ, the training type I error is equal (as much as possible) to the type II error? (2 points)

   (b) Fit a ridge logistic regression model to the training data and use 10-fold cross-validation using the AUC as a measure of error to tune λ. For the optimal λ answer the same questions (i., ii., ...) as asked for LASSO (8 points).

   (c) Fit an Elastic-net logistic regression model to the training data and use 10-fold cross-validation using the AUC as a measure of error to tune λ. For the optimal λ answer the same questions (i., ii., ...) as asked for LASSO (8 points).