Homework

Kamiar Rahnama Rad

***Upload** your homework as a **single pdf** on blackboard.*

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 logistic regression model to the training data with $n = 16500$.

      i. For each review in the training set calculate $x_i^\top \hat{\beta} + \hat{\beta}_0$ and two histograms on top of each other with different colors. A histogram of $x_i^\top \hat{\beta} + \hat{\beta}_0$ for positive reviews, and another for negative reviews. (2 points)

      ii. For each review in the test set calculate $x_i^\top \hat{\beta} + \hat{\beta}_0$ and two histograms on top of each other with different colors. A histogram of $x_i^\top \hat{\beta} + \hat{\beta}_0$ for positive reviews, and another for negative reviews. (2 points)

      iii. In the training set, for each observation, using logistic regression, calculate $\Pr[y = 1|X = x]$. 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 type I and type II error? (2 points)

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

2. Fit a logistic regression model to the training data with $n = 16500$ with a twist. Use the "weights" argument in the glmnet function in R (there should be a similar argument in other languages) to place different weights for different observations. Specifically, since the number of positive observations is 40/125 of the number of negative observations, there is danger that the majority of negative reviews will make the classifier more sensitive to positive reviews. **In order to account for this imbalance, use the weight of 1 for positive observations, and use the weight of 40/125 for negative observations.** Let n_+ and n_- stand for the number of positive and negative observations, respectively. Then what we are doing here is essentially giving weight 1 to negative reviews and weight n_+/n_- to positive reviews. The objective here is to see if this procedure of fitting a weighted version can circumvent the problem faced by imbalance, as noted in the unweighted version.

(a) For each review in the training set calculate $x_i^\top \hat{\beta} + \hat{\beta}_0$ and two histograms on top of each other with different colors. A histogram of $x_i^\top \hat{\beta} + \hat{\beta}_0$ for positive reviews, and another for negative reviews. (2 points)

(b) For each review in the test set calculate $x_i^\top \hat{\beta} + \hat{\beta}_0$ and two histograms on top of each other with different colors. A histogram of $x_i^\top \hat{\beta} + \hat{\beta}_0$ for positive reviews, and another for negative reviews. (2 points)

(c) In the training set, for each observation, using logistic regression, calculate $\Pr[y = 1|X = x]$. 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)

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

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