Homework

Kamiar Rahnama Rad

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

1. This question is about the ROC and AUC of logistic regression applied to the IMDB review classification problem. Use the $p = 2500$ most frequent words as predictors. Fit a logistic regression model to the training data with $n_{train} = 25000$ and use this model to answer the following questions.

   (a) What are the top 10 words associated with positive reviews? (1 point)

   (b) What are the top 10 words associated with negative reviews? (1 point)

   (c) How can we identify the review in the training set that is hardest to classify? Find it and present the review. (1 point)

   (d) Find and show the most positive review and the most negative review. (1 point)

   (e) 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. See `https://rb.gy/8xhuor` for an example of overlapping histograms for two different categories. (8 points)

   (f) 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. (8 points)