ML 4375 – Intro to Machine Learning

Homework 3: Logistic Regression and Naive Bayes in R

In this homework you will create an R script to run logistic regression and naive bayes on the BreastCancer data set, which is part of package mlbench.

BreastCancer Data

If you type "?BreastCancer" at the console after loading package mlbench, you will see that this data has been collected clinically and reported in a paper by Wolberg and Mangasarian, "Multisurface method of pattern separation for medical diagnosis applied to breast cytology", published in 1990 in the National Academy of Sciences.

Instructions:

1. Start with a new Rstudio Rmd file, add headings for Homework 3, your name and a brief description of the purpose of the script. Clearly label each step. Each step should have one or more R code chunks. For step 1, load library mlbench, installing if needed (at the console). You have to load the data frame into memory with data(BreastCancer) Now run str() and head() on BreastCancer and summary() on just the Class column. Use R instructions to calculate the percent in each class, and print them with an appropriate heading using paste(). Answer the questions in step 1:
    a. How many instances are there?
    b. What is your target column?
    c. How many predictors are there? What type of data are the predictors?
    d. What percentage of the observations are malignant?
2. Cell.size and Cell.shape are in one of 10 levels. Build a logistic regression model called glm0, where Class is predicted by Cell.size and Cell.shape. Do you get any error or warning messages? Google the message and try to decide what happened. Run summary on glm0 to confirm that it did build a model. Write a comment about why you think you got this warning message and what you could possibly do about it. List the source of your information in a simple markdown link.
3. Notice in the summary() of glm0 that most of the levels of Cell.size and Cell.shape became predictors and that they had very high p-values. We would need a lot more data to build a good logistic regression model this way. It might be better to just have 2 levels for each variable. In this step, add two new columns to BreastCancer as listed below. Run summary() on Cell.size and Cell.shape as well as the new columns. Comment on the distribution of the new columns. Do you think what we did is a good idea? Why or why not?
    a. Cell.small which is a binary factor that is 1 if Cell.size==1 and 0 otherwise
    b. Cell.regular which is a binary factor that is 1 if Cell.shape==1 and 0 otherwise
4. Create conditional density plots using the original Cell.size and Cell.shape. First attach() the data to reduce typing. Then use par(mfrow=c(1,2)) to set up a 1x2 grid for two cdplot() graphs with Class~Cell.size and Class~Cell.shape. Observing the plots, write a sentence or two comparing size and malignant, and shape and malignant. Do you think our cutoff points for size==1 and shape==1 were justified now that you see this graph? Why or why not?

5. Create plots (not cdplots) with our new columns. Again, use par(mfrow=c(1,2)) to set up a 1x2 grid for two plot() graphs with Class~Cell.small and Class~Cell.regular. Now create two cdplot() graphs for the new columns. Now compute the following and provide a summary in the text portion of this answer. Also indicate based on these results if you think small and regular will be good predictors.
    a. calculate the percentage of small observations that are malignant
    b. calculate the percentage of not-small observations that are malignant
    c. calculate the percentage of regular observations that are malignant
    d. calculate the percentage of non-regular observations that are malignant
6. Randomly divide BreastCancer into two data sets: train (80% of the data) and test (20%). Make sure you first set the seed to 1234 so you get the same results as others.
7. Build a logistic regression classifier to estimate the probability of Class given Cell.small and Cell.regular. Run summary() on your model. Answer the following:
    a. Which predictor(s) seem to be good predictors. Justify your answer.
    b. Comment on the Null deviance versus the Residual deviance.
    c. Comment on the AIC score.
8. Test the model on the test data and compute accuracy. What percent accuracy did you get? Output the confusion matrix and related stats using the confusionMatrix() function in the caret package. Where the mis classifications more false positives or false negatives?
9. Your coefficients from the model are in units of logits. Extract the coefficient of small with glm1$coefficients[]. Answer the following questions:
    a. What is the coefficient?
    b. How do you interpret this value?
    c. Find the estimated probability of malignancy if Cell.small is true using exp().
    d. Find the probability of malignancy if Cell.small is true over the whole BreastCancer data set and compare results. Are they close? Why or why not?
10. Build two more models, each just using Cell.small and Cell.regular and use anova(glm_small, glm_regular, glm1) to compare all 3 models, using whatever names you used for your models. Analyze the results of the anova(). Also, compare the 3 AIC scores of the models. Feel free to use the internet to help you interpret AIC scores.
11. Build a Naive Bayes Model Class ~ Cell.small + Cell.regular on the training data using library e1071. Output the model parameters and answer the following questions:
    a. What percentage of the training data is benign?
    b. What is the likelihood that a malignant sample is not small?
    c. What is the likelihood that a malignant sample is not regular?
12. Predict on the test data with naive bayes model. Compute the accuracy and output the confusion matrix. Are the results the same or different? Why do you think that is the case?

Element	Points
Appropriate comments and use of markdown	16 points
Steps (12)	7 points each