CSCE 587 – Big Data Analytics

Homework Assignment 4: Naïve Bayes Classification

Due: Friday, 3/7, by 11:59pm

Deliverables: You will submit an R script (a .R file) with commands to answer the questions below, and comments clearly labelling each answer, to the assignment in Blackboard.

The Data: The csv file attached to the assignment in provides the data to be modeled.

Preparation:

You can download and save the csv file attached to the assignment in Blackboard OR you can use wget at the terminal to download the file from my CSE site to your virtual machine:

wget https://cse.sc.edu/~bhipp/587/penguins.csv --no-check-certificate

Additional Instructions of Note:

Your answer for each question must be clearly labelled in your R script and must include the command(s) necessary to generate the answer. Unlabeled answers and answers that don't contain the R command(s) to justify them will not receive credit.

The R commands must be submitted in an R script (.R) file so that they can easily be run during grading.

All Students

1. (3 points) Following the idea from our Naïve Bayes lab, create a 2D array of indices for each stratified class of penguins (Adelie, Chinstrap, and Gentoo). Each array should have 10 rows (one for each of the folds in our 10-fold stratified cross validation method). The number of columns in each array will depend on the number of records in our dataset for that species of penguin. Note: the number of records for each species is not evenly divisible by 10, so use the closest multiple of 10 that's less than or equal to the number of records. For example, there are 151 Adelie penguins, so create an array of 10 rows of 15 columns for 150 of the Adelie penguin indices. **Set the seed to 587 prior to the creation of each stratified random class of indices.**
2. (3 points) Using the naiveBayes function from package e1071, loop through the 10 folds to train the i^{th} model and build the aggregate confusion matrix (as we did in the Naïve Bayes lab).
3. (1 point) Output the aggregate confusion matrix.
4. (3 points) Calculate and display the sensitivity and specificity for the aggregate confusion matrix created in the loop. Fill in the blank for each of the following in a comment in your R script.
    - Sensitivity for Adelie = _____
    - Specificity for Adelie = _____
    - Sensitivity for Chinstrap = _____
    - Specificity for Chinstrap = _____
    - Sensitivity for Gentoo = _____
    - Specificity for Gentoo = _____

Graduate Students Only

A classmate finds that if you

- set the seed to 500
- create a random sample of 10 of the 342 records for the test set (testpartition = sample(1:342,10))
- create a Naïve Bayes model based on the other 332 records (all records not in the test set)

the model will classify the 10 test records with 100% accuracy, 100% sensitivity and specificity for each classification.

5. (2 points) Verify the classmate's claims (include the R commands to verify the claim in your script).
6. (1 point) Describe the problems with using this method for creating and testing a model, and explain why the 100% accuracy measures overestimate the model's performance. Your explanation must include specifics from this example (e.g. the method, training set used in creation of the model, test set used to test the model).