University of the Philippines Cebu
College of Science
Department of Computer Science
CMSC 178DA (Special Topics)
Introduction to Data Analytics

ACTIVITY GUIDE (WEEK 13)

Activity 5
Applying Logistic Regression to Binary Classification Problem

Objective

The activity for this module aims to demonstrate practical knowledge in logistic regression analysis technique by applying it to a binary classification problem as well how to interpret the logistic regression output as to assess its performance, e.g., pseudo-R^2, model parameters, among others.

Tasks

1. Partner with the same classmate as in the previous activity (unless if you want to change partner, with permission from me) since this activity is to be accomplished in pairs to promote collaboration and interaction. (*In case the class size is odd-numbered, one student is to volunteer to work individually.*)

2. Read-analyze together the given case study below (email data adapted from RegressIt: https://regressit.com/):
   The email dataset accompanying this activity guide, 'ActivityGuide-Week13_data.xlsx', contains 9 variables, namely: spam (coded as '1' if email instance is classified as spam, or '0' if otherwise - not a spam), to_multiple, image, attach, password, line_breaks, format, re_subj, and urgent_subj, with total instances (or, rows excluding the row header) of 200 email samples classified equally as a 'spam' or not. A data analyst intends to develop a spam filter using this dataset. He poses this question to himself: ***Can a logistic regression model effectively classifies email messages as spam or not based on its characteristics coded as predictor variables (i.e., from 'to_multiple' to 'urgent_subj'?***

3. Discuss with your partner the given case above. Agree on the responses required by this activity. Active participation by each member of the pair during the discussion *cum* interaction is expected.

4. Create a logistic regression model where 'spam' is the dependent variable and the following are the predictor variables: to_multiple, image, attach, password, line_breaks, format, re_subj, and urgent_subj.

5. Use either of the following application software tools: MS Excel (if it can be done), R, or Python, (or a mix of it) to accomplish task #4.

6. List the steps your pair has undertaken and/or provide the codes (for either R or Python, whichever is applicable) in accomplishing Task #4. Present the narrative in an outline format. Provide appropriate headings/comments above the codes for clarity.

7. Based on your pair's interpretation on the logistic regression output, answer the following questions:

7.1. Is the fitted logistic regression model effective in classifying email messages as spam or not? Justify.
7.2. Which of the predictor variables significantly influence the classification of email messages as spam or not? Identify. Justify.

8. Name the PDF file of your pair's submission with the following file-naming convention: **Class Activity [number]-[Your Family Name]_[Your Partner's Family Name]**. For example, "Class Activity **5**-**Lao_Uy**". The highlighted portion of the example are the ones you supply information for your pair.

9. Structure the document for your responses following required sections below:
   9.1. Cover Page (format accordingly)
      9.1.1. Course Number
      9.1.2. Class Activity Number and Title
      9.1.3. 'Submitted by' information (pair members, course section, and schedule)
   9.2. The Case (in the next page)
   9.3. The Logistic Regression Output
      9.3.1. Fitted Model Summary Output
      9.3.2. Narrative Text (i.e., Task#4, *in outline format*)
   9.4. Pair's interpretation (*refer back to Task 7 for context).*

10. Submit a copy of your pair's Activity 5 responses (in PDF format) to your respective 'Class Activity' submission bin of the Learning Management System (LMS). In certain instances when your Internet connectivity is unstable, you may submit via email, but you must inform me ahead (i.e., not after the deadline).

Assessment Criteria

- **Fitted Model Summary Output - 40%**
  (*correctness in generating appropriate logistic regression analysis summaries*)
- **Narrative Text – 30%**
  - ✓ Fitting logistic regression model (20%)
  - ✓ Data preparation (10%)
- **Interpretation – 20%**
  (*analytical accuracy on observations, descriptions, and implications*)
- **Packaging Submission – 10%**
  (*shows creativity in organizing and presenting the submitted document which contains the pair's responses*)