ETC3250 - Project 2017

Spam or not spam

Task

The purpose of this project is to make the best model to predict whether an email is spam or not. The competition is hosted on the Kaggle website (https://inclass.kaggle.com/c/spam-or-not-spam).

The data has been splitted into training and test sets. The full training set is available to you. But only the explanatory variables are provided for the test set. You can evaluate your predictions by submitting them to the Kaggle website. Note that only 50% of the test set is used to compute the public leaderboard. Your final score using the full test set will be provided at the end of the competition.

This is a description of the variables:

- numRecipients = the number of people in the To: or Cc: lines
- Domain = domain name of the sender's email address
- replyToSameAddress = is the reply sending an email to the same address
- Weekday = day of the week
- Hour = hour of the day
- percentLowerCase = % of lower case letters in the subject line
- numDigits = number of digits in the sender's email address
- percentNonLetters = % of non-letters in the sender's email address
- Size = size of the email in Kb
- numLinks = number of links
- localSender = was the sender in the local domain as the receiver
- credit, porn, sucker, pharm, prescription, drugs, save, sex, discreet, free, sell, sale, asseenon, discount - binary variables computed by the presence of certain key words newsletter = is this email basically a newsletter
- spam = is this email a spam

The evaluation metric for this competition is the classification accuracy computed using the test set as follows:

$$\frac{1}{n}\sum_{i=1}^{n} I(y_i = \hat{y}_i),$$

where $I(\cdot)$ is the indicator function.

1. Your first task is to form a team with three people.
2. Create a Kaggle account (using your Monash email address), for each team membmer and form a team.
3. Do some basic exploration of the dataset
4. Build your first model. Predict the test set, and upload your predictions to Kaggle.
5. Try, and try again to improve your model. You can submit one prediction per day.
6. Write up how you built your model, and decided on your best model. Also, describe one or more other intersting things your learned about spam relative to other variables.
7. Turn your report into a 10 minute presentation for the class.

Deadlines

Do not wait until the last minute. Late submissions will not be allowed.

- Sep. 15, 11:55pm: Form your team. Submit the name of your team, as well as the names of each memembe of the team on Moodle.
- Sep. 29, 11:55pm: At least one Kaggle submission needs to have been made.
- Oct. 11, 11:55pm: The Kaggle competition closes at 11:55pm.
- Oct. 13, 11:55pm: Upload to Moodle (i) your project report (max 5 pages), one per group, that describes your model fitting, and at least one interesting observation about the spam data, and (ii) the slides for your presentation.
- Oct. 16 and 19: Present your slides in the lecture period. All members of the team must attend and present part of the work. All students must be present to evaluate the presentations, and if not points will be deducted from the absent individual's score.

Grading

- Total points: 20
- Accuracy of classifier: 6
- Report: 7
- Presentation: 7 (Score will be given by other members of the class. All members of the team must participate by speaking in the presentation.)