Machine Learning – I

Ramakrishna Mission Vivekananda University

Assignment 2

Posted on 31 Mar 2017 | Seek clarification by 6 Apr 2017 | Submit by 11 Apr 2017

Your solutions for the problems in this assignment should be in a single `R` file `GroupXX_a2.R` or in a single Python file `GroupXX_a2.py`, where `XX` is your group number (i.e., `XX` = 01 to 10). The code must be properly commented. The solution file should be emailed to `sg.sourav@gmail.com`, with a copy to all your group members, *before* the lecture session on the submission deadline.

Properly acknowledge every source of information that you refer to, including discussions with your friends outside the group. Verbatim copy from any source is strongly discouraged, and plagiarism will be heavily penalized. You should try to write the codes completely on your own.

Problem 1 [20 points]

Titanic: Machine Learning from Disaster
`https://www.kaggle.com/c/titanic`

Background: The sinking of the RMS Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew. This sensational tragedy shocked the international community and led to better safety regulations for ships. One of the reasons that the shipwreck led to such loss of life was that there were not enough lifeboats for the passengers and crew. Although there was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others, such as women, children, and the upper-class.

In this challenge, Kaggle asks you to complete the analysis of what sorts of people were likely to survive. In particular, Kaggle asks you to apply the tools of machine learning to predict which passengers survived the tragedy. You may treat it as a classic case-study of Binary Classification.

Problem statement: Take part in the Kaggle Competition – “Titanic: Machine Learning from Disaster” – available at `https://www.kaggle.com/c/titanic`. Submit your predictions online.

Submission: Submit the `R` (or Python) code you wrote for the Kaggle competition as the first part of your submission file `GroupXX_a2.R`. In the commented section, please mention the name of your Kaggle team and your final score/standing on the leaderboard. You may use any public *Kernel* or *Discussion* posted on Kaggle, with proper acknowledgement for each such reference.

Restriction: Please restrict yourselves to the models we have covered so far, that is, linear and tree-based models for binary classification. Do not use SVM or Neural Networks at this stage.

Problem 2 [30 points]

Spam Detection for Text Messages
`http://www.souravsengupta.com/ml2017/evaluation/smsdata.txt`

Background: Text Message Spam are unsolicited text messages (SMS), especially advertising, directed at mobile phones or smartphones. As the popularity of mobile phones surged in the early 2000s, frequent users of text messaging began to see an increase in the number of unsolicited (and generally unwanted) commercial advertisements being sent to their telephones through text messaging (SMS). This can be particularly annoying for the recipient because, unlike in email, some recipients may be charged a fee for every message received, including the spam messages.

In this challenge, we ask you to complete the analysis of what type of text messages are likely to be spam. In particular, we ask you to apply the tools of machine learning to predict which messages in a corpus are spam. You may treat it as a classic case-study of Binary Classification.

Problem statement: Use as training set the labeled (`good`/`spam`) text messages available at `http://www.souravsengupta.com/ml2017/evaluation/smsdata.txt` to build a robust tree-based binary classifier that is capable of distinguishing spam text messages from regular ones.

Submission: Submit the `R` (or Python) code you wrote for this challenge as the second part of your submission file `GroupXX_a2.R`. In the commented section, please mention the main text processing packages you used, and acknowledge any online/offline resources you have consulted.

Dataset: Note that the training set of labeled text messages is structured as follows, where the first element is the label, either `good` or `spam`, and then the text message is posted as raw text.

```
good    Go until jurong point, crazy.. Available only in bugis n great world la e buffet...
good    Ok lar... Joking wif u oni...
spam    Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005.
good    U dun say so early hor... U c already then say...
good    Nah I don’t think he goes to usf, he lives around here though
spam    FreeMsg Hey there darling it’s been 3 week’s now and no word back!
good    Even my brother is not like to speak with me. They treat me like aids patent.
```

Dataset available at — `http://www.souravsengupta.com/ml2017/evaluation/smsdata.txt`