STAT 6494 Data Science Project Proposal

Yishu Xue

1 Introduction

The fast development in information technology made communication between people worldwide easier than ever. These advances are always accompanied by challenges. Huge amounts of spam emails and texts are sent everyday. While spam filtering technologies have been widely used by major email service providers such as Gmail and Outlook, its application to mobile SMS is less pervasive. The iPhone, for example, has an "unprotected" inbox. Anybody who knows your mobile phone number or iCloud account can send you messages without being blocked.

There are, however, third-party apps on both iOS and Android platforms that provide spam message filters. What is there filters based on? What algorithms do they use? Will these algorithms be accurate in terms of sensitivity and specificity? In this project, I aim to build different classification models on an SMS Spam Collection[1], compare their performances, and look for the best classification scheme.

2 Data

The dataset is open data from Kaggle. It contains one set of SMS messages in English of 5,572 messages, tagged according being ham (legitimate) or spam. 747 of them are spam, while the rest 4,825 are ham.

Examples:

Spam	Had your mobile 11 months or more? U R entitled to Update to the latest colour mobiles with camera for Free! Call The Mobile Update Co FREE on 08002986030
Ham	Oops, I'll let you know when my roommate's done.

3 Methods

The most frequently used classification method for SMS/email spam detection is Naive Bayes, followed by Support Vector Machine, Ensemble methods, logistic regression and K-Nearest Neighbors(Cormack et al., 2008). I look forward to implementing these methods on the SMS Spam Collection dataset, and see what specific features that they identify, what spam successfully "cheated" the classifiers, etc. I'm also interested in the application of EM for multinomial mixture models in text clustering, i.e., regardless of the class labels, whether the algorithm could successfully assign the SMS messages to two clusters.

References

Cormack, G. V. et al. (2008). Email spam filtering: A systematic review. *Foundations and Trends® in Information Retrieval 1*(4), 335–455.

[1] https://www.kaggle.com/uciml/sms-spam-collection-dataset