Dataquest Guided Project: Building a Spam Filter with Naive Bayes

Cindy Zhang

2/25/2021

Contents

Introduction

This is my solution to Dataquest's Guided Project from the fourth Probability and Statistics course, which involves building a spam filter using the Naive Bayes theorem.

More details such as the RMD and csv files can be found in the repository in GitHub.

Findings

Exploring the Dataset

```r
spam <- read_csv("spam.csv")

# Calculate percent of messages that are spam and ham
spam_ham_percent <- spam %>%
  group_by(label) %>%
  summarize(Freq=n()) %>%
  mutate(Percentage = Freq / nrow(spam)*100)
```

`spam` has 1000 rows and 2 columns. 15 percent of messages are spam and 85 percent of messages are ham.

Training, Cross-validation and Test Sets

- 84.375 percent of messages in `spam_train` are ham.
- 90 percent of messages in `spam_cv` are ham.
- 85 percent of messages in `spam_test` are ham.

Data Cleaning

Creating the Vocabulary

Calculating Constants First

Calculating Probability Parameters

Classifying a New Message

Calculating Accuracy

Hyperparameter Tuning and Cross-validation

Test Set Performance