Spam and Ham Email Classifier
Fall 2024

DATA 100

Ajay Sharma

Remark

If you would like to see the project source code (Jupyter Notebook), please contact me at ajay.sharma@berkeley.edu.

Overview

This project applied machine learning techniques, specifically logistic regression, to classify emails as spam or ham. By analyzing features such as email length, punctuation, and token frequencies, and employing advanced preprocessing and feature engineering, the model was designed to robustly detect spam with high accuracy.

Dataset Description

The dataset contained thousands of emails labeled as spam or ham. Key attributes analyzed included:

- **Email Metadata:** Email length, punctuation usage, and capitalized characters.
- **Text Features:** Frequency of specific tokens and presence of stop words.

Methodology

- **Exploratory Data Analysis (EDA):** Analyzed email length, token distributions, and punctuation usage to identify patterns in spam and ham emails.
- **Data Cleaning:** Processed the text data by filtering stop words, handling missing values, and normalizing features.
- **Feature Engineering:** Engineered key features such as word frequencies, capitalized text, and punctuation counts.
- **Model Training:** Built a logistic regression classifier, optimized hyperparameters, and evaluated model performance with cross-validation.
- **Evaluation:** Assessed the model using accuracy, precision, recall, and AUC. Generated an ROC curve with an outstanding AUC score of 0.980.

Results and Insights

The classifier achieved an accuracy (on training & testing) of over 92% and an AUC score of 0.980. Key findings included:

- **Top Predictors:** Email length, punctuation usage, and specific token frequencies were the most influential features.
- **Model Performance:** Placed in the top 2% of a class of 1200 students based on prediction accuracy.