Machine Learning Project Proposal

Project Title

Email Spam Detection using Hidden Markov Model, Naïve Bayes, and Support Vector Machine

Dataset

We will use the SpamAssassin Public Corpus, a well-known dataset containing both spam and non-spam (ham) emails. The dataset provides a solid benchmark for training and evaluating spam detection models.

Project Idea

The goal of this project is to build an efficient email spam detection system using a hybrid approach that combines Hidden Markov Models (HMMs), Naïve Bayes, and Support Vector Machines (SVMs). Email spam detection remains a critical task in cybersecurity and information filtering systems, and combining different machine learning techniques can significantly boost classification performance.

The approach begins with using Hidden Markov Models to model the sequence and patterns of words in an email, capturing linguistic structure and contextual flow, which is not typically addressed by traditional models. Next, Naïve Bayes will be used as a baseline probabilistic classifier leveraging the bag-of-words model. Finally, SVM will be employed to refine and improve classification accuracy using features extracted from the HMM and Naïve Bayes stages. We will experiment with different feature engineering techniques and assess model performance using accuracy, precision, recall, and F1 score.

Software You Will Need to Write

- Preprocessing scripts to clean and tokenize the email dataset.
- Implementation of Hidden Markov Model for sequential modeling (possibly using hmmlearn or customimplementation).

- Naïve Bayes and SVM classifiers using Scikit-learn.

- Training pipeline and evaluation functions for all three algorithms.

- Visualization tools for comparing classifier performance.

- Optional: Jupyter notebooks for demonstration and interactive experimentation.

Papers to Read

1. A Comparative Study on Email Spam Filtering Techniques - for understanding benchmark models.
2. Machine Learning Approaches for Spam Detection in Email Communication - explores multiple MLtechniques including SVM and Naïve Bayes.
3. Hidden Markov Models and their Applications in NLP - introduces the application of HMMs in sequentialdata modeling.

Teammate

Dann Anthony Astillero and Lucas Barros

What Will You Complete by the Tuesday of 4th Week?

By the fourth week, we aim to:

- Complete preprocessing and feature extraction from the SpamAssassin dataset.

- Implement and evaluate the Naïve Bayes classifier as a baseline.

- Train and evaluate an initial version of the SVM classifier.

- Conduct exploratory data analysis and determine HMM feature integration strategy.

- Produce initial experimental results (accuracy, confusion matrix, etc.) for baseline classifiers (Naïve Bayesand SVM).

Machine Learning Project Proposal

References

- SpamAssassin Dataset: https://spamassassin.apache.org/old/publiccorpus/

- Scikit-learn Documentation: https://scikit-learn.org/

- hmmlearn Documentation: https://hmmlearn.readthedocs.io/

- Related academic papers listed above.