COMP 6721 Project 2

SPAM Detector

Project Members:
Jayaprakash Kumar (40083709)
Darwin Anirudh G (40093368)
Reethu Navale (40124459)

Group Name: G07
TA: Pouria Chalangari

Contents

1 Project Description

A python program to detect whether the mail is a SPAM or not based on Naïve Bayes approach.

2 Analysis

2.1 Steps

1. Train the data using the training set of files
    a. Calculate the unique number of words in training data as your vocabulary count.
    b. Calculate the frequency of each word in both HAM and SPAM set. (Note: add 0.5 smoothing value for each word)
2. Now classify the testing file set using Naïve Bayes classification.
3. Calculate the probability of HAM and SPAM for each file, which ever probability is higher the file is classified to that group.

2.2 Accuracy

Calculate the accuracy for each class i.e., for HAM and SPAM. Below mentioned equation is used to calculate the accuracy for each class.

HAM Accuracy = Total of correctly identified HAM files (TP+TN)/Actual number of HAM files (TP+FP+FN+TN)

SPAM Accuracy = Total of correctly identified SPAM files (TP+TN)/Actual number of SPAM files (TP+FP+FN+TN)

☞ **Note**: Please refer the confusion matrix for HAM and SPAM in the section 2.6

2.2.1 Program Output

HAM Accuracy = (394 +336) / (394 + 64 +6 + 336) = 91.25%

SPAM Accuracy = (336 +394) / (336 + 6 +64 + 394) = 91.25%

2.3 Precision

Calculate precision for each class i.e., for HAM and SPAM. Below mentioned equation is used to calculate the precision.

HAM Precision = TP for HAM / (TP for HAM + FP for HAM)

SPAM Precision = TP for SPAM / (TP for SPAM + FP for SPAM)

☞ **Note**: Please refer the confusion matrix for HAM and SPAM in the section 2.6

2.3.1 Program Output

HAM Precision = 394 / (394 + 64) = 0.86026 → 86.03%

SPAM Precision = 336 / (336 + 6) = 0.982456 → 98.25%

2.4 Recall

Calculate Recall for each class i.e., for HAM and SPAM. Below mentioned equation is used to calculate the recall.

HAM Recall = TP for HAM / (TP for HAM + FN for HAM)

SPAM Recall = TP for SPAM / (TP for SPAM + FN for SPAM)

☞ **Note**: Please refer the confusion matrix for HAM and SPAM in the section 2.6

2.4.1 Program Output

HAM Recall = 394 / (394 + 6) = 0.985 → 98.5%

SPAM Recall = 336 / (336 + 64) = 0.84 → 84%

2.5 F1-Measure

Calculate f1-measure for each class i.e., for HAM and SPAM. Below mentioned equation is used to calculate the f1 measure.

☞ Note: Here we are considering both precision and recall with same priority.

HAM F1-measure = 2 * HAM Precision * HAM Hall / (HAM Precision + HAM Recall)

SPAM F1-measure = 2 * SPAM Precision * SPAM Hall / (SPAM Precision + SPAM Recall)

2.5.1 Program Output

HAM Precision = 2 * 0.86026 * 0.985 / (0.86026 + 0.985) = 0.9184 → 91.84%

SPAM Precision = 2 * 0.982456 * 0.84 / (0.982456 + 0.84) = 0.9057 → 90.57%

2.6 Confusion Matrix

Python program generates two confusion matrix one for each class i.e., HAM and SPAM.

2.6.1 Ham Confusion Matrix

Program generates HAM specific confusion matrix. Program classifies 458 are HAM out of 800 test file set, out of which 394 are actual HAM data and 64 are not HAM. Program classifies 342 as non-HAM, however 336 are correctly classified as NON-HAM and. 6 actual HAM files were classified as NON-HAM by the system.

		Actual Data	
Program Data		HAM	NON-HAM
	HAM	394	64
	NON-HAM	6	336

True Positive (TP) – 394
False Positive (FP) – 64
False Negative (FN) – 6
True Negative (TN) – 336

2.6.2 SPAM Confusion Matrix

Program generates the below SPAM confusion matrix. Program classifies as 342 as SPAM out of 800 test data files, out of which 336 are actual SPAM files and 6 are non-SPAM files but program classified them as SPAM. Program also classifies as 458 as non-SPAM files out of 800 test data, out of which 394 are actual non-SPAM files and 64 are actual SPAM files but program classifies them as non-SPAM.

		Actual Data	
Program Data		**SPAM**	**NON-SPAM**
	SPAM	336	6
	NON-SPAM	64	394

True Positive (TP) – 336
False Positive (FP) – 6
False Negative (FN) – 64
True Negative (TN) – 394

2.6.3 Overall Confusion Matrix

Below is the overall confusion matrix classified by the program. Programs classifies correct 394 HAM files as HAM; it fails to classify 6 HAM files as HAM. Program correctly classifies 336 SPAM files as SPAM; it fails to classify 64 SPAM files as SPAM.

		Actual Data Classification	
Program Classification		**HAM**	**SPAM**
	HAM	394	64
	SPAM	6	336

3 Reference

Referred the below book for Naïve Bayes classifier logic, calculation of precision, recall, f1 measure, accuracy and generation of confusion matrix.

Book: Artificial Intelligence – A Modern Approach, Third Edition, Stuart Russell and Peter Norvig