Aastha Tiwari
CS22M005

Assignment-3

SPAM or HAM?

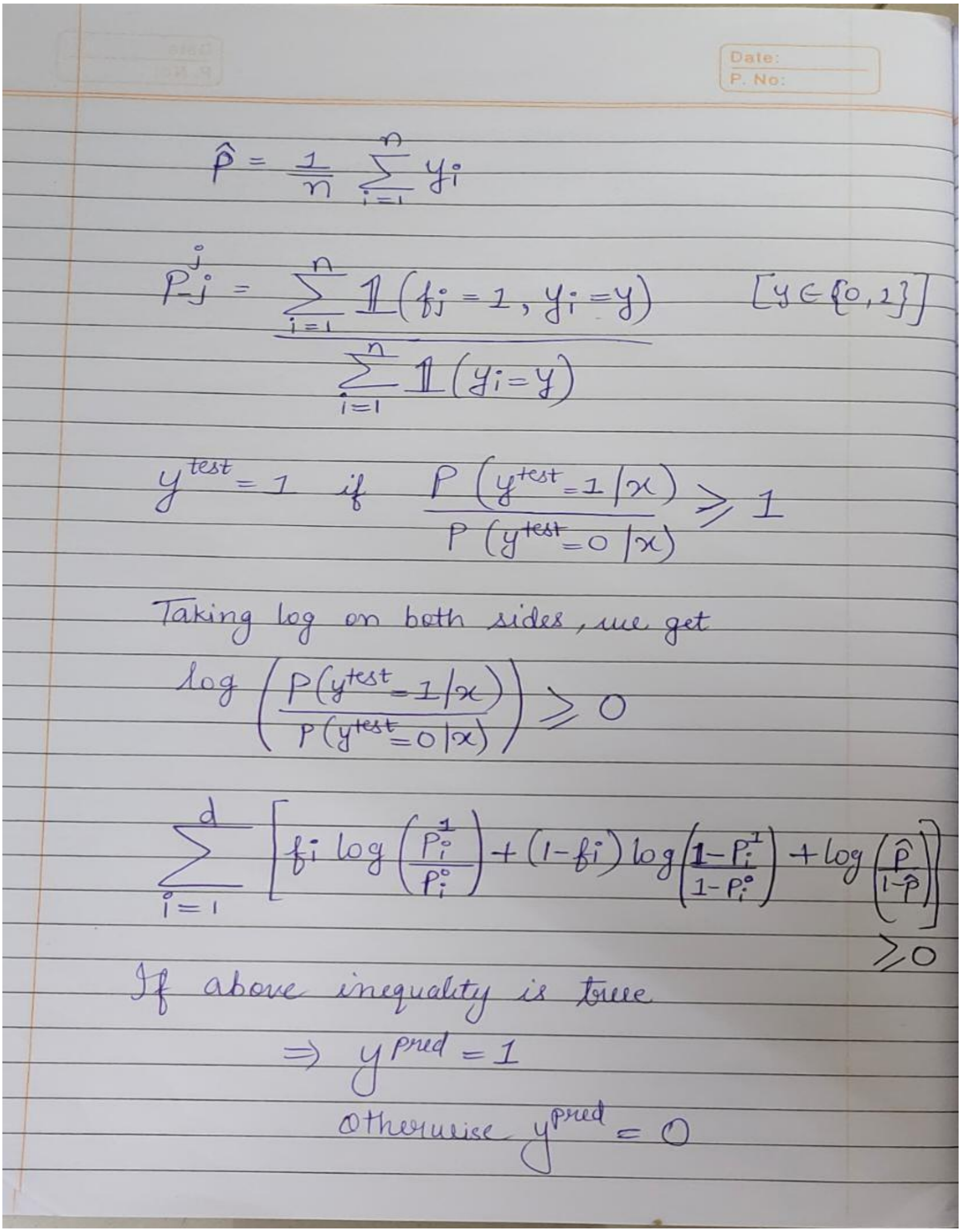
In this assignment, you will build a spam classifier from scratch. No training data will be provided. You are free to use whatever training data that is publicly available/does not have any copyright restrictions (You can build your own training data as well if you think that is useful). You are free to extract features as you think will be appropriate for this problem. The final code you submit should have a function/procedure which when invoked will be able to automatically read a set of emails from a folder titled test in the current directory. Each file in this folder will be a test email and will be named ’email#.txt’ (’email1.txt’, ’email2.txt’, etc). For each of these emails, the classifier should predict +1 (spam) or 0 (non Spam). You are free to use whichever algorithm learnt in the course to build a classifier (or even use more than one). The algorithms (except SVM) need to be coded from scratch. Your report should clearly detail information relating to the data-set chosen, the features extracted and the exact algorithm/procedure used for training including hyperparameter tuning/kernel selection if any. The performance of the algorithm will be based on the accuracy of the test set.

Dataset- I have used dataset from Kaggle.com as it was publicly available and does not have any copyright restrictions. The dataset contains spam and ham emails classified as 70% ham and 30% spam in which label 1 means spam and label 0 means ham. The total mails in the dataset are 5171.

Featured Extracted- Firstly I have used the nltk library in order to create my own dictionary using the words module available in the corpus folder of the nltk library.
The dictionary is created on the basis of words present in the training dataset as well as the words module and the size of the dictionary is around 18007.Now with the help of this dictionary I created the feature vectors for all the emails in the training dataset.The vector size is equal to the size of the dictionary. For every word in the dictionary if the word is present in the email then the corresponding value in the vector is set to 1 otherwise it is set to zero. This will serve as the x(feature vectors) and y is created with the help of labels of the email in the training dataset.

Algorithm- I have used Naive Bayes Algorithm as taught in class to classify the data as spam and non-spam. First I computed p which is the ratio of spam to the total number of mails in my training dataset. Then I created a matrix p_i of size (size of dictionary *2) where each row corresponds to the probability of a word in the training dataset for label value corresponding to 0 and 1 i.e. non-spam and spam respectively. Then I have applied a simple formula to find the predicted values of labels in the test dataset. The test is fetched in a csv file named emails.csv and then test data is generated from this file and the values of labels are predicted with a simple formula as shown below.
The predicted value of the label is stored in a list named y_pred and the value 1 means spam email and 0 means non-spam email.

Aastha Tiwari
CS22M005



$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

$$P_j^{j} = \frac{\sum_{i=1}^{n} \mathbb{1}(f_j = 1, y_i = y)}{\sum_{i=1}^{n} \mathbb{1}(y_i = y)} \qquad [y \in \{0,1\}]$$

$$y^{test} = 1 \quad \text{if} \quad \frac{P(y^{test} = 1 | x)}{P(y^{test} = 0 | x)} \geqslant 1$$

Taking log on both sides, we get

$$\log \left(\frac{P(y^{test} = 1 | x)}{P(y^{test} = 0 | x)} \right) \geqslant 0$$

$$\sum_{i=1}^{d} \left[f_i \log \left(\frac{P_i^1}{P_i^0} \right) + (1 - f_i) \log \left(\frac{1 - P_i^1}{1 - P_i^0} \right) + \log \left(\frac{\hat{p}}{1 - \hat{p}} \right) \right] \geqslant 0$$

If above inequality is true

$$\Rightarrow y^{pred} = 1$$

Otherwise $y^{pred} = 0$