Intro to ML and Big Data
Assignment 2

CS 4731 — Dr. Mihail
Department of Computer Science
Valdosta State University

September 26, 2019

Do not attempt a Mihail homework the night it's due.

1 Introduction

In this assignment, you will implement a probabilistic classifier, Naïve Bayes, to create an email spam filter. The key idea behind this project is that the simple presence of some words in emails is predictive of spam (or ham) status. This learning is done during the first phase, where you will analyze the words of a dataset, called the training set, and learn conditional probability distributions for a subset of words. In the second phase, you will evaluate your classifier on a separate dataset, called the test set.

2 Background

By Bayes rule, the posterior probability of class given data is

$$p(C_k|x) = \frac{p(C_k)p(x|C_k)}{p(x)} \tag{1}$$

where $x = \{x_1, x_2, \ldots, x_n\}$ is a vector of word occurences, C_k is one of a discrete set of k possible classes, $p(C_k)$ is the unconditional probability of a given class, $p(x|C_k)$ is the likelihood and $p(x)$ is the evidence. Since the posterior is not a function of the class, for the purpose of classification, one can ignore it and only account for the proportionality in the posterior:

$$p(C_k|x) \propto p(C_k)p(x|C_k) \tag{2}$$

where both the unconditional class probability and likelihood can be easily learned from the dataset.

The Naïve part consists of assuming that the joint conditional probability $P(C_k|x_1, x_2, \ldots, x_n)$ can be simplified by assuming independence between features $x_1, \ldots, x_n$. This yields the following result that can be used directly by a classifier:

$$p(C_k|x_1, x_2, \ldots, x_n) \propto p(C_k)\Pi_{i=1}^{n} p(x_i|C_k) \tag{3}$$

3 Classification

For this assignment, you will have two classes: spam and ham (not spam). You will have to decide on some means to classify a new email in light of message content. A simple (but prone to numerical error) way to do this is to compute the quantity in Equation 3 for C_{spam} and C_{ham}, then pick the largest as the predicted class.

It is useful to notice that by taking the log of Equation 3 yields the following:

$$ln\, p(C_k|x) \propto ln\, p(C_k) + \sum_{i=1}^{n} p(x_i|C_k) \tag{4}$$

and furthermore, using the technique of log ratios of the posterior for both classes yields:

$$ln\, \frac{p(C_{spam}|x)}{p(C_{ham}|x)} = ln\, \frac{p(C_{spam})}{p(C_{ham})} + \sum_{i=1}^{n} ln\, \frac{p(x_i|C_{spam})}{p(x_i|C_{ham})} \tag{5}$$

We note that the sign of the ratio indicates whether $p(C_{spam}|x) > p(C_{ham}|x)$.

4 Dataset

You will use a preprocessed dataset from: `https://github.com/tasdikrahman/datasets/tree/master/email/csv`

Preprocessing You will have to perform and document a number of preprocessing steps. For example, if you split an email message by spaces, you will have some words end in commas or periods, etc. You have to document this process in the project write-up.

5 Project requirements

- **Test/train split** You will have to split the dataset into disjoint sets of training and training sets.
- **Evaluation** Once the learning phase is complete, you will evaluate the classifier on the testing set and report the results in confusion matrix. See `https://en.wikipedia.org/wiki/Confusion_matrix`
- **Write-up** You will write-up your project in a LaTeX document with all the decision points along the way documented and discussed. This has to have a section titled “What I’ve learned” where you will discuss what you learned.

6 Due date

The project is due before the end of the day Sunday, October 6th.