Machine Learning (WiSe 2025/2026)

Author: Suvansh Shukla
Matriculation No. 256245

Assignment 3 Task 3.4

(a) The probabiltiy of a word belonging to spam or not-spam can be estimated by the following:

1. Represent each document (here, an email) by a vector of words. So One attribute per word in the email.
2. We use training examples to estimate the following: P(+), P(-), P(doc|+), P(doc|-)

(b) The probability of the size of the mail in regards to spam or not spam can be calculated using gaussian distribution with the size of emails being distributed around a mean of a size of email which may be looked at to determine if an email is spam or not.

(c) We can encounter the same Zero-frequency problems as well as the Missing values problem that already occurrs. In addition to that we may also realize that the words in the email are not so indepent of each other thus reducing the accuracy of Naive Bayes quite a bit.