

# Homework 7 (due: Apr 01)

## MACHINE LEARNING - COSC 4360

Department of Computer Science and Electrical Engineering

Spring 2025

### Exercises

Create a **New Project** for every exercise. Take a screenshot of the source code along with its output and place the **source code** and the **screenshot** in a **zipped folder** named **LastNameFirstName\_HW7**

#### Exercise 1

You have recently been hired by a major enterprise to filter out spam emails. Your manager has given you access to the following three **Normal** (i.e., not Spam) training emails: *train\_N\_I*, *train\_N\_II*, *train\_N\_III* and the following three **Spam** training emails: *train\_S\_I*, *train\_S\_II*, *train\_S\_III*. Classify the following two emails: **testEmail\_I.txt**, **testEmail\_II.txt** as to whether they are Normal or Spam (use **Naïve Bayes** classifier). In addition, **plot** the frequency of words for Normal and Spam emails. Files are located in: *Files->Homeworks+Quizzes->HW7\_files*.

**Note 1:** The **Prior Probability** for **Normal** emails is: 0.73 while for **Spam** is: 0.27.

**Note 2:** To open a file and place its words into a list:

with `open("train_N_I.txt", "r")` as `f`:

```
train_N_I = f.read().split()
```

**Note 3:** To count the number of words in a file:

```
from collections import Counter
countsN = Counter(train_N_I)
```

**Note 4:** To get the keys/values from a dictionary and convert them to a list:

```
key_listN = list(countsN.keys())
val_listN = list(countsN.values())
```

**Note 5:** You may wish to merge the **Normal** emails into one file and the **Spam** emails into another one.

#### Exercise 2 (Optional)

Given the following five-point dataset:  $x = [1, 2, 3, 4, 5]$  and  $y = [1, 2, 4, 4, 6]$ , develop an **online algorithm** for **simple linear regression**.

**Note:** Submit through **Canvas**