Computer Science Department
California State University Channel Islands

COMP 478 - Homework 2

Deadline: 03/23/2022, 11:59 am

1. (85 points) In this problem, we will write a python code to for a naive Bayes classifier to detect spam/ham SMS. This classifier assumes features are independent given the label:

$$p(x_{1:K}|y) = \prod_{i=1}^{k} p(xi|y)$$

   (a) *Step 1*: Download the SMS spam collection data set from `https://archive.ics.uci.edu/ml/datasets/sms+spam+collection`. The test dataset includes indices: 0, 10, 20, 30, 40, 50, 60, 70, 80, 90, ...(the multiples of 10) and the rest of the data points will be your training dataset.

   (b) *Step 2*: For each word in the training dataset, extract: $p(word|ham)$ and $p(word|spam)$. Each SMS message is a sequence of words (case-insensitive). Ignore punctuation. To avoid overfitting, use **additive smoothing** to smooth $p(word|ham)$ and $p(word|spam)$. e.g.:

$$\text{P(word | ham)} = \frac{count(word,ham) + \propto}{count(ham) + N\propto}$$

   Lets use $\alpha = 0.2$ and $N = 20000$.

   (Hint: use CountVectorizer from sklearn.feature-extraction.text for feature extraction)

   (c) *Step 3*: Calculate the testing accuracy, confusion matrix, precision, recall, and F-score for your classifier.

2. (15 points) We want to train a binary Logistic Regression classifier for the given training dataset:

   d_1 = {X:(0, 1), p($Y = 0|X$) = 0.3}

   d_2 = {X:(1, 0), p($Y = 0|X$) = 0.7}

   d_3 = {X:(0, 0), p($Y = 0|X$) = 0.5}

   Find the decision boundary for this classifier. What is the prediction for the given test data:

   d_1 = {X:(-1, 0)}