## ECE368: Probabilistic Reasoning

## Lab 1: Classification with Multinomial and Gaussian Models

Name:	Mingym Zheng	Student Number: 1903797661
	′ ′ ′ ′ ′ ′ ′ ′ ′ //	

You should hand in: 1) A scanned .pdf version of this sheet with your answers (file size should be under 2 MB); 2) one figure for Question 1.2.(c) and two figures for Question 2.1.(c) in the .pdf format; and 3) two Python files classifier.py and Idaqda.py that contain your code. All these files should be uploaded to Quercus.

## 1 Naïve Bayes Classifier for Spam Filtering

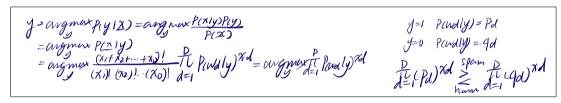
1. (a) Write down the estimators for  $p_d$  and  $q_d$  as functions of the training data  $\{\mathbf{x}_n, y_n\}, n = 1, 2, \dots, N$  using the technique of "Laplace smoothing". (1 **pt**)

Spam: 
$$Pd = \frac{Xnd+1}{Xn_1+\cdots+Xn_N+N} \left\{ \frac{Xn}{N} \right\}$$

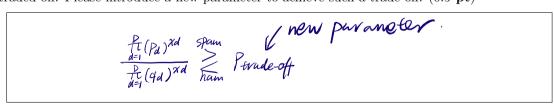
ham:  $Pd = \frac{Xnd+1}{Xn_1+\cdots+Xn_N+N} \left\{ \frac{Xn}{N} \right\}$ 

N: distinct words in spam & ham.

- (b) Complete function learn\_distributions in python file classifier.py based on the expressions. (1 pt)
- 2. (a) Write down the MAP rule to decide whether y=1 or y=0 based on its feature vector  $\mathbf{x}$  for a new email  $\{\mathbf{x},y\}$ . The d-th entry of  $\mathbf{x}$  is denoted by  $x_d$ . Please incorporate  $p_d$  and  $q_d$  in your expression. Please assume that  $\pi=0.5$ . (1 **pt**)



- (b) Complete function classify\_new\_email in classifier.py, and test the classifier on the testing set. The number of Type 1 errors is \_\_\_\_\_\_, and the number of Type 2 errors is \_\_\_\_\_\_. (1 pt)
- (c) Write down the modified decision rule in the classifier such that these two types of error can be traded off. Please introduce a new parameter to achieve such a trade-off. (0.5 **pt**)



Write your code in file classifier.py to implement your modified decision rule. Test it on the testing set and plot a figure to show the trade-off between Type 1 error and Type 2 error. In the figure, the x-axis should be the number of Type 1 errors and the y-axis should be the number of Type 2 errors. Plot at least 10 points corresponding to different pairs of these two types of error in your figure. The two end points of the plot should be: 1) the point with zero Type 1 error; and 2) the point with zero Type 2 error. Please save the figure with name **nbc.pdf**. (1 **pt**)

(d) If we do not use Laplace smoothing and simply use maximum likelihood estimation in the training phase, what will go wrong? What kind of emails such a classifier would fail to classify? (0.5 pt)

For the test files, if a word only shows in one type of email (span/ham) then without Luplace smoothly, he trent the probability of that word show in the other type as o, while its not rigorous enough.

And P. 1 = 1 1 16t depends on 11 sample size)

And Fud = 1 Not deposes on Theory

## 2 Linear/Quadratic Discriminant Analysis for Height/Weight Data

1. (a) Write down the maximum likelihood estimates of the parameters  $\mu_m$ ,  $\mu_f$ ,  $\Sigma$ ,  $\Sigma_m$ , and  $\Sigma_f$  as functions of the training data  $\{\mathbf{x}_n, y_n\}$ , n = 1, 2, ..., N. (1 **pt**)

 $\underline{M} = \frac{2}{4} \underbrace{I}_{1}^{2} \underbrace{M}_{1}^{-1} \underbrace{J}_{1}^{2} \\
\underline{M}_{1}^{-1} \underbrace{J}_{1}^{2} \underbrace{J}_{1}^{2} \underbrace{J}_{1}^{-1} \\
\underline{M}_{1}^{-1} \underbrace{J}_{1}^{-1} \underbrace{J}_{1}^{-1$ 

(b) In the case of LDA, write down the decision boundary as a linear equation of  ${\bf x}$  with parameters  ${\boldsymbol \mu}_m,\,{\boldsymbol \mu}_f,\,$  and  ${\bf \Sigma}.$  Note that we assume  $\pi=0.5.$  (0.5 pt)

Mm Z - = Mm Z - Mm Mg Z - = Mg Z Mf

In the case of QDA, write down the decision boundary as a quadratic equation of  $\mathbf{x}$  with parameters  $\boldsymbol{\mu}_m$ ,  $\boldsymbol{\mu}_f$ ,  $\boldsymbol{\Sigma}_m$ , and  $\boldsymbol{\Sigma}_f$ . Note that we assume  $\pi = 0.5$ . (0.5 **pt**)

- \( \frac{1}{2} \log | \frac{2}{2} m | - \frac{1}{2} (\frac{1}{2} - \frac{1}{2} m) \rightarrow \frac{1}{2} \log | \frac{2}{2} f | - \frac{1}{2} (\frac{1}{2} - \frac{1}{2} f) \rightarrow \frac{1}{2} f \rightarrow \frac{1}{2} - \frac{1}{2} f \rightarrow \frac{1}{2} \rightarrow \frac{1}{2} f \rightarrow \frac{1}{2} \rightarrow \frac{1}{2} f \rightarrow \frac{1}{2} \rightarrow \frac{1}{2} f \rightarrow \frac

- (c) Complete function discrimAnalysis in Idaqda.py to visualize LDA and QDA models and the corresponding decision boundaries. Please name the figures as Ida.pdf, and qda.pdf. (1 pt)
- 2. The misclassification rates are O'//82 for LDA, and O'/O' for QDA. (1 pt)