ECE368: Probabilistic Reasoning
Lab 1: Classification with Multinomial and Gaussian Models

Name: Jia Ming Huang **Student Number:** 1003893245.

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 ldaqda.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**)

$$P_d = \frac{x_{nd,1} + 1}{x_{n1,1} + \dots + x_{nN,1} + N}, \quad q_d = \frac{x_{nd,0} + 1}{x_{n1,0} + \dots + x_{nN,0} + N},$$

P_d : spam
q_d : ham
N: total distinct words in P_d, q_d.

(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**)

$$y_{MAP} = \text{argmax}_y\, P(y|x) = \log(\pi) + \sum_{d=1}^{D} x_d \log P_d \underset{\text{ham}}{\overset{\text{spam}}{\gtrless}} \sum_{d=1}^{D} x_d \log q_d + \log(1-\pi).$$

$$= \text{argmax}_y \frac{P(x|y)P(y)}{P(x)} \uparrow$$

$$= \text{argmax}_y\, P(x|y). \quad P(y), P(x) \text{ constant.}$$

For $y=1$, $P(w_d|y) = P_d$ Note: $\pi = 1 - \pi$ *.
$y=0$, $P(w_d|y) = q_d$.

(b) Complete function classify_new_email in classifier.py, and test the classifier on the testing set. The number of Type 1 errors is $\boxed{2}$, and the number of Type 2 errors is $\boxed{5}$. (1.5 **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**)

$$\frac{\sum_{d=1}^{D} x_d \log P_d}{\sum_{d=1}^{D} x_d \log q_d} \underset{\text{ham}}{\overset{\text{spam}}{\gtrless}} T$$

T: new trade-off parameter.

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**)

2 Linear/Quadratic Discriminant Analysis for Height/Weight Data

1. (a) Write down the maximum likelihood estimates of the parameters $\boldsymbol{\mu}_m$, $\boldsymbol{\mu}_f$, $\boldsymbol{\Sigma}$, $\boldsymbol{\Sigma}_m$, and $\boldsymbol{\Sigma}_f$ as functions of the training data $\{\mathbf{x}_n, y_n\}, n = 1, 2, \ldots, N$. (1 **pt**)

$$\mu_m = \frac{1}{N_m} \sum_{n=1}^{N} x_n \mathbb{I}(y_n=1) \qquad \Sigma_m = \frac{1}{N_m} \sum_{n=1}^{N} \left[(x_n-\mu_m)(x_n-\mu_m)^T\right] \cdot \mathbb{I}(y_n=1)$$

$$\mu_f = \frac{1}{N_f} \sum_{n=1}^{N} x_n \cdot \mathbb{I}(y_n=2) \qquad \Sigma_f = \frac{1}{N_f} \sum_{n=1}^{N} \left[(x_n-\mu_f)(x_n-\mu_f)^T\right] \cdot \mathbb{I}(y_n=2)$$

$$N_m = \sum_{n=1}^{N} \mathbb{I}(y_n=1) \qquad \Sigma = \frac{1}{N}\left(N_m \cdot \Sigma_m + N_f \Sigma_f\right).$$

$$N_f = \sum_{n=1}^{N} \mathbb{I}(y_n=2)$$

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

$$x^T \Sigma^{-1} \mu_m - \frac{1}{2}\mu_m^T \Sigma^{-1} \mu_m + \log(\pi) \underset{y=2}{\overset{y=1}{\lessgtr}} x^T \Sigma^{-1} \mu_f - \frac{1}{2}\mu_f^T \Sigma^{-1} \mu_f \log(1-\pi)$$

Note: $\log \pi = \log(1-\pi)$ for $\pi = 0.5$.

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**)

$$-\frac{1}{2}(x-\mu_m)^T \Sigma_m^{-1} (x-\mu_m) - \frac{1}{2}\log|\Sigma_m| + \log(\pi) \underset{y=2}{\overset{y=1}{\lessgtr}} -\frac{1}{2}(x-\mu_f)^T \Sigma_f^{-1}(x-\mu_f) - \frac{1}{2}\log|\Sigma_f| + \log(1-\pi)$$

(c) Complete function discrimAnalysis in ldaqda.py to visualize LDA and QDA models and the corresponding decision boundaries. Please name the figures as lda.pdf, and qda.pdf. (1 **pt**)

2. The misclassification rates are $\boxed{11.82\%}$ for LDA, and $\boxed{10.91\%}$ for QDA. (1 **pt**)