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

Name: RUI ZENG	**Student Number:** 1003979091

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 $\{\mathbf{x}_n, y_n\}, n = 1, 2, \ldots, N$ using the technique of "Laplace smoothing". (1 **pt**)

$$pd = \frac{\text{\# of occurence of wd in spam bag} + 1}{\text{total \# of words in spam bag} + \text{total \# of distinct words in spam and ham}}$$

$$qd = \frac{\text{\# of occurence of wd in ham bag} + 1}{\text{total \# of words in ham bag} + \text{total \# of distinct words in spam and ham}}$$

   (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**)

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

$$\because P(y=1) = P(y=0) = 0.5$$

$$\therefore y = \underset{y}{\text{argmax}}\ P\{x|y\} = \text{argmax} \frac{(x_1 + \cdots x_D)!}{x_1! \cdots x_D!} \prod_{d=1}^{D} P(x_d|y)^{x_d}$$

$$\prod_{d=1}^{D} pd^{x_d} \underset{ham}{\overset{spam}{\gtrless}} \prod_{d=1}^{D} qd^{x_d}$$

   (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 $\boxed{4}$. (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**)

introduce a ratio parameter r

$$\frac{\prod_{d=1}^{D} pd^{x_d} \cdot \underbrace{\pi}_{0.5}}{\prod_{d=1}^{D} qd^{x_d} \cdot \underbrace{(1-\pi)}_{0.5}} \underset{ham}{\overset{spam}{\gtrless}} r$$

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**)

Male

$$\mu_m = \frac{1}{\#\text{ of male}} \sum_{i=1}^{N} 1\{y_i = 1\}\, x_i$$

$$\Sigma_m = \frac{1}{\#\text{ of male}} \sum_{i=1}^{N} (x_i - \mu_m)(x_i - \mu_m)^T 1\{y_i = 1\}$$

Female

$$\mu_f = \frac{1}{\#\text{ of female}} \sum_{i=1}^{N} 1\{y_i = 2\}\, x_i$$

$$\Sigma_f = \frac{1}{\#\text{ of female}} \sum_{i=1}^{N} (x_i - \mu_f)(x_i - \mu_f)^T 1\{y_i = 2\}$$

$$\Sigma = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu_m)(x_i - \mu_m)^T 1\{y_i = 1\} + (x_i - \mu_f)(x_i - \mu_f)^T 1\{y_i = 2\}$$

(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**)

$$\mu_m^T \Sigma^{-1} \underline{x} - \frac{1}{2} \mu_m^T \Sigma^{-1} \mu_m = \mu_f^T \Sigma^{-1} \underline{x} - \frac{1}{2} \mu_f^T \Sigma^{-1} \mu_f$$

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**)

$$-\frac{1}{2}\log|\Sigma_m| - \frac{1}{2}(x - \mu_m)^T \Sigma_m^{-1} (x - \mu_m)$$
$$= -\frac{1}{2}\log|\Sigma_f| - \frac{1}{2}(x - \mu_f)^T \Sigma_f^{-1} (x - \mu_f)$$

(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 0.11818 for LDA, and 0.10909 for QDA. (1 **pt**)