UTORid: zhaizixu Student email: shawn.zhai@mail.utoronto.ca

ECE368: Probabilistic Reasoning

Lab 1: Classification with Multinomial and Gaussian Models

Name: Shawn Zhai **Student Number:** 1006979389

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)

$$p_d = \frac{n_d^{SP} + 1}{n^{SP} + D}$$

$$q_d = \frac{n_d^{H} + 1}{n^{H} + D}$$

n_d^{SP}: # of occurance of word d in SPAM word bag
n_d^{H}: # of occurance of word d in HAM word bag
n^{SP}: total # of words in SPAM word bag
n^{H}: total # of words in HAM word bag
D: # of distinct words in both SPAM and HAM word bags

   (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)

$$P[Y|\underline{X}] = \frac{P[\underline{X}|Y]P[Y]}{P[\underline{X}]} = \underbrace{\frac{(X_1+\cdots+X_D)!}{(X_1)!\cdots(X_D)!}}_{\text{constant}} \prod_{d=1}^{D} P(x_d|Y)^{x_d}$$

$P[Y=0] = P[Y=1] = 0.5$
$P[\underline{X}] =$ constant
} Ignore

$$\text{MAP RULE: } \prod_{d=1}^{D} (P_d)^{x_d} \underset{HAM}{\overset{SPAM}{\gtrless}} \prod_{d=1}^{D} (q_d)^{x_d}$$

   (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)

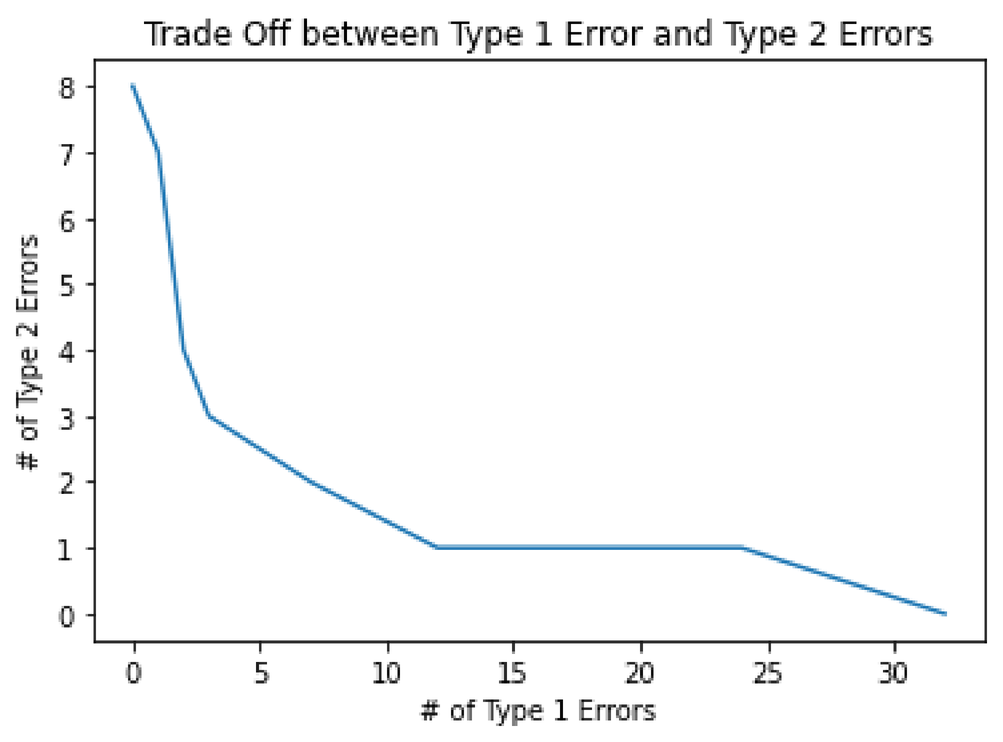
Modified MAP RULE:

$$\frac{P(\underline{X}|Y=1)}{P(\underline{X}|Y=0)} \underset{HAM}{\overset{SPAM}{\gtrless}} \frac{\pi_0}{\pi_1} \Rightarrow \frac{\prod_{d=1}^{D}(P_d)^{x_d}}{\prod_{d=1}^{D}(q_d)^{x_d}} \underset{HAM}{\overset{SPAM}{\gtrless}} k$$

Introduce parameter k equals to the ratio between the prior of two classes

Previously $k = \frac{\pi_0 = 0.5}{\pi_1 = 0.5} = 1$

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)



Trade Off between Type 1 Error and Type 2 Errors
# of Type 2 Errors
# of Type 1 Errors
0
1
2
3
4
5
6
7
8
0
5
10
15
20
25
30

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)

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

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

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

$$\Sigma_m = \frac{1}{\#\text{ of male}} \sum_{i=0}^{N} (\underline{x_i} - \underline{\mu_m})(\underline{x_i} - \underline{\mu_m})^T 1\{y_i = 1\} \qquad \Sigma_f = \frac{1}{\#\text{ of female}} \sum_{i=1}^{N} (\underline{x_i} - \underline{\mu_f})(\underline{x_i} - \underline{\mu_f})^T 1\{y_i = 2\}$$

(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)

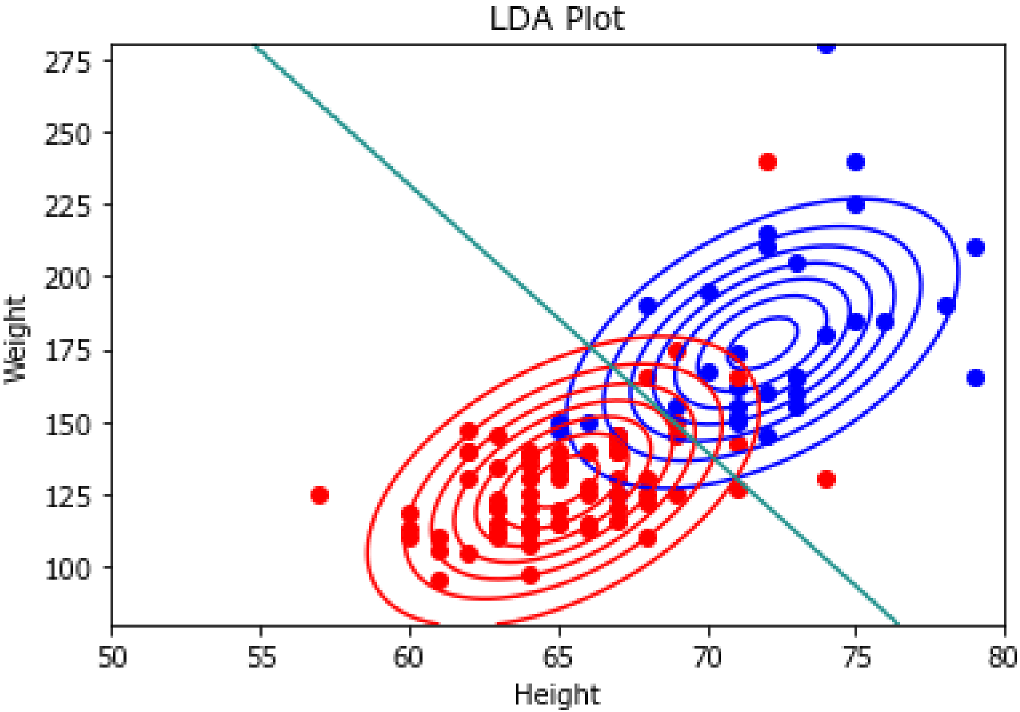
$$-\frac{1}{2} \underline{\mu_m}^T \Sigma^{-1} \underline{\mu_m} + \underline{\mu_m}^T \Sigma^{-1} \underline{x} = -\frac{1}{2} \underline{\mu_f}^T \Sigma^{-1} \underline{\mu_f} + \underline{\mu_f}^T \Sigma^{-1} \underline{x}$$

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)

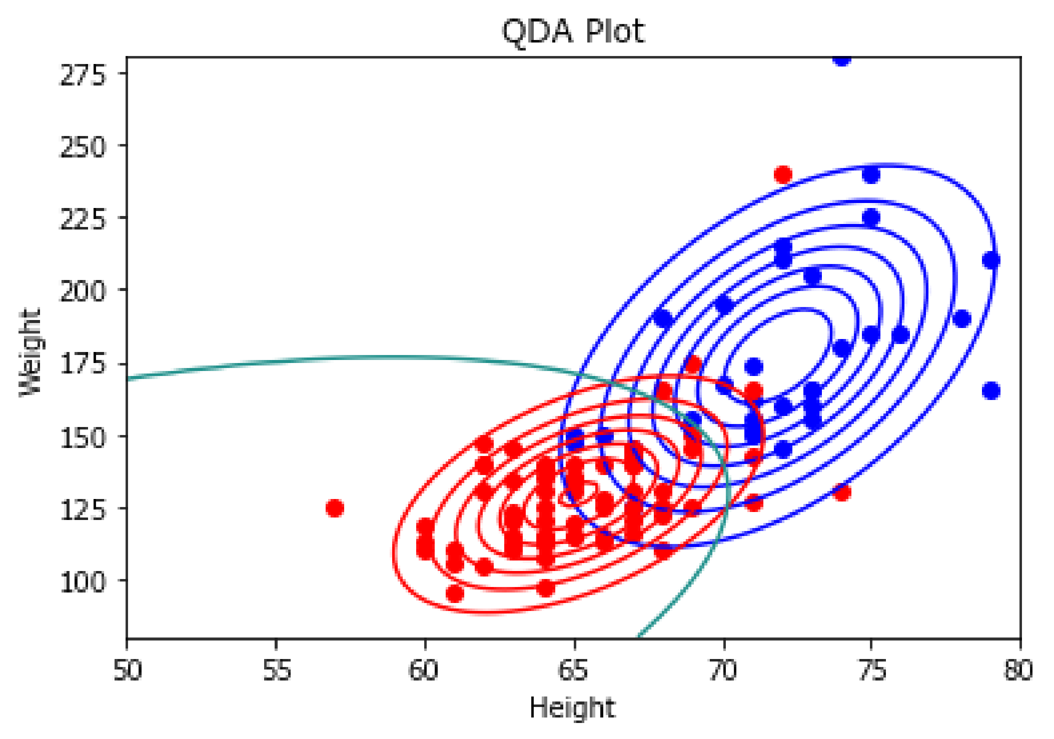
$$\underline{x}^T (\Sigma_m^{-1} - \Sigma_f^{-1}) \underline{x} + 2 (\Sigma_f^{-1} \underline{\mu_f} - \Sigma_m^{-1} \underline{\mu_m})^T \underline{x} + (\underline{\mu_m}^T \Sigma_m^{-1} \mu_m - \mu_f^T \Sigma_f^{-1} \mu_f) + \log \frac{|\Sigma_m|}{|\Sigma_f|} = 0$$

(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 $\boxed{0.1182}$ for LDA, and $\boxed{0.1091}$ for QDA. (1 pt)



LDA Plot
Weight
275
250
225
200
175
150
125
100
50
55
60
65
70
75
80
Height


QDA Plot
Weight
275
250
225
200
175
150
125
100
50
55
60
65
70
75
80
Height