ECE368

Lab 1

1) Naive Bayes Classifier for Spam Filtering

a)

$$p_d = \frac{\text{\# of occurrences of word } d \text{ in all spam documents} + 1}{\text{total \# of words in all spam documents} + \text{total \# of distinct words in both spam and ham documents}}$$

Let $D \equiv$ total # of distinct words in both spam & ham documents

$$p_d = \frac{\sum_{i=1}^{N} x_{id} 1(y_i = 1) + 1}{\sum_{i=1}^{N}\sum_{j=1}^{D} x_{ij} 1(y_i = 1) + D} \quad \text{where } 1(y_i = j) = \begin{cases} 1 & y_i = j \\ 0 & y_i \neq j \end{cases}$$

$$q_d = \frac{\text{\# occurrences of word } d \text{ in all ham documents} + 1}{\text{total \# of words in all ham documents} + \text{total \# of distinct words in both spam and ham documents}}$$

$$q_d = \frac{\sum_{i=1}^{N} x_{id} 1(y_i = 0) + 1}{\sum_{i=1}^{N}\sum_{j=1}^{D} x_{ij} 1(y_i = 0) + D}$$

b) See code

2a) $\hat{y} = \underset{y}{\text{argmax}}\, P(y|x) = \underset{y}{\text{argmax}} \frac{P(y)P(x|y)}{P(y)} = \underset{y}{\text{argmax}}\, P(x|y)$ since $P(y)$ is constant, and $P(y=1) = P(y=0) = 0.5$

$$\hat{y} = \underset{y}{\text{argmax}} \frac{(x_1 + x_2 + \dots + x_D)!}{x_1! x_2! \dots x_D!} \prod_{d=1}^{D} P(w_d|y)^{x_d}$$

MAP Rule: $\prod_{d=1}^{D} p_d^{x_d} \underset{\text{ham}}{\overset{\text{spam}}{\gtrless}} \prod_{d=1}^{D} q_d^{x_d}$

b) # Type 1 errors is 2

Type 2 errors is 4

c) We introduce a new parameter r, which is the ratio between the right and left hand sides of rule in 2a. We use it as a critical value for deciding whether email belongs to spam or ham

$$\prod_{d=1}^{D} \left(\frac{p_d}{q_d}\right)^{x_d} \underset{\text{ham}}{\overset{\text{spam}}{\gtrless}} r$$

Note: r was 1 in Questions 1, 2ab

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

1a) From lecture:

$$\mu_m = \frac{\sum_{i=1}^{N} x_i \mathbb{1}(y_i=1)}{\sum_{i=1}^{N} \mathbb{1}(y_i=1)} \qquad \mu_f = \frac{\sum_{i=1}^{N} x_i \mathbb{1}(y_i=2)}{\sum_{i=1}^{N} \mathbb{1}(y_i=2)}$$

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

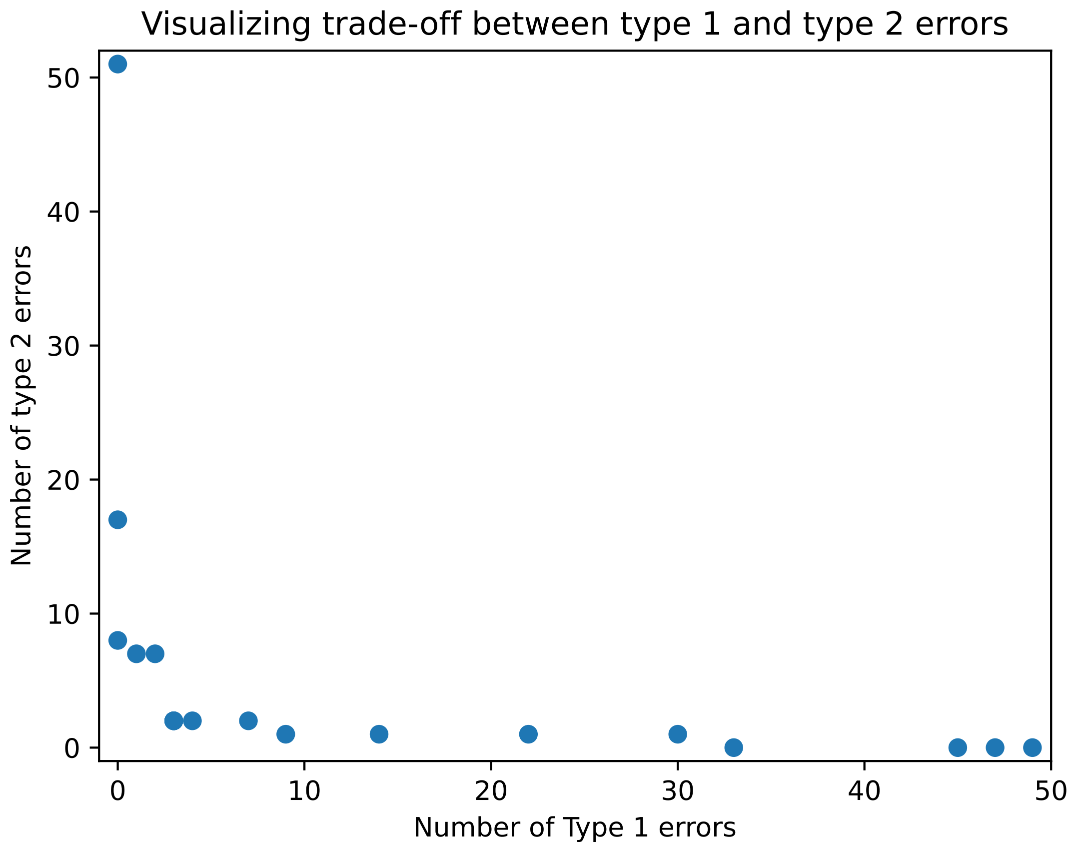
$$\Sigma_m = \frac{\sum_{i=1}^{N} (x_i-\mu_m)(x_i-\mu_m)^T \mathbb{1}(y_i=1)}{\sum_{i=1}^{N} \mathbb{1}(y_i=1)} \qquad \Sigma_f = \frac{\sum_{i=1}^{N} (x_i-\mu_f)(x_i-\mu_f)^T \mathbb{1}(y_i=2)}{\sum_{i=1}^{N} \mathbb{1}(y_i=2)}$$

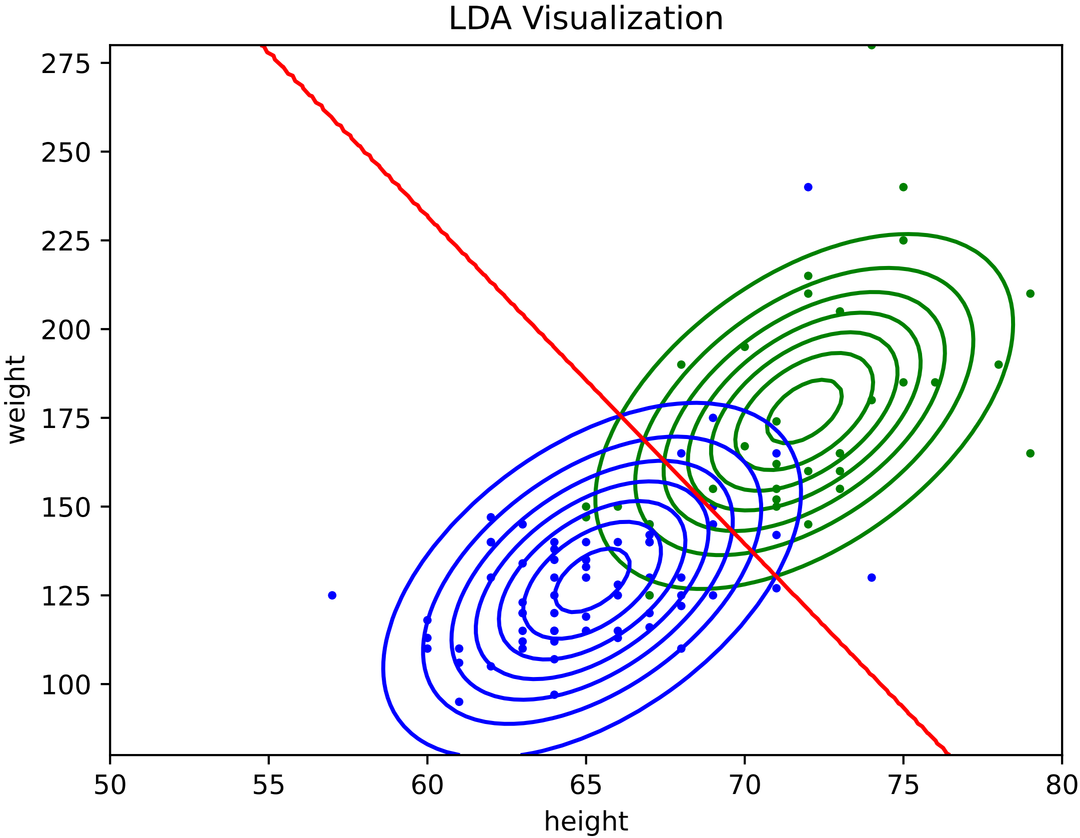
b) From lecture:

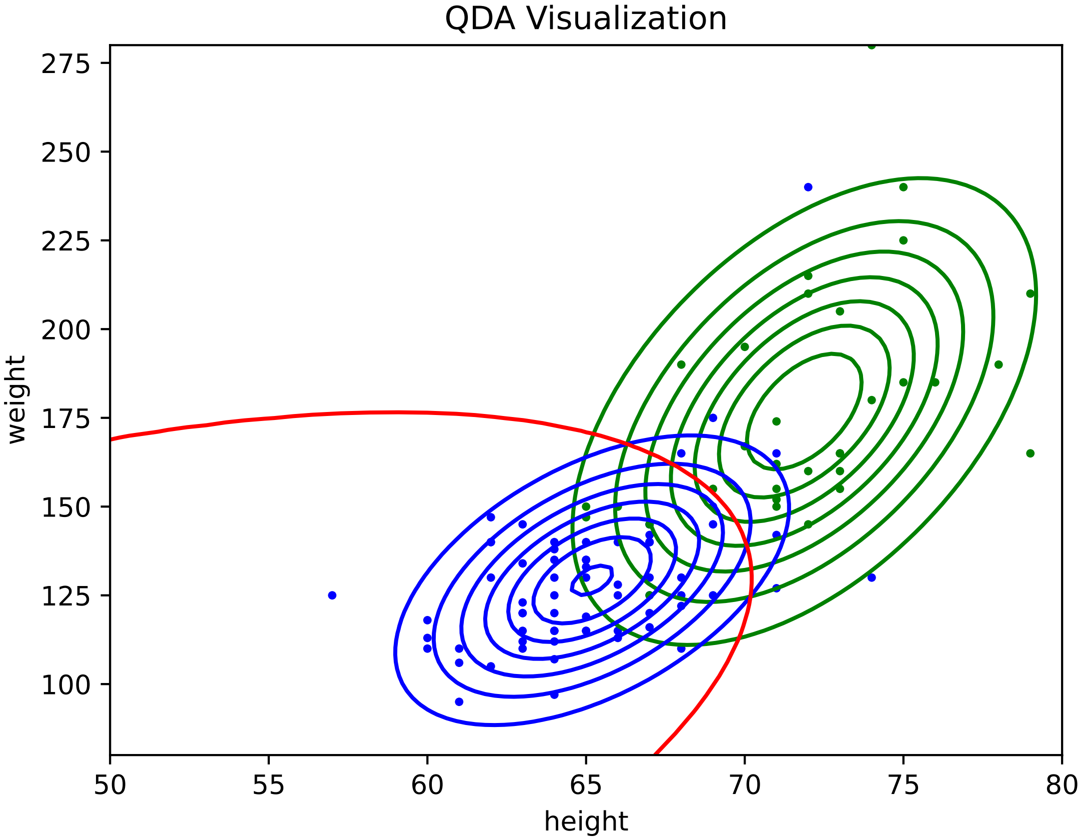
LDA: $\mu_m^T \Sigma^{-1} x - \frac{1}{2}\mu_m^T \Sigma^{-1} \mu_m = \mu_f^T \Sigma^{-1} x - \frac{1}{2}\mu_f^T \Sigma^{-1} \mu_f$

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

2) The misclassification rates are 0.118 for LDA and 0.109 for QDA


Visualizing trade-off between type 1 and type 2 errors
Number of type 2 errors
Number of Type 1 errors
0
10
20
30
40
50


LDA Visualization
weight
height
275
250
225
200
175
150
125
100
50
55
60
65
70
75
80


QDA Visualization
weight
height
275
250
225
200
175
150
125
100
50
55
60
65
70
75
80