ECE368: Probabilistic Reasoning

Lab 1: Classification with Multinomial and Gaussian Models

Name: Pratyush Menon **Student Number:** 1004282661

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**)

$$p_d = \frac{\#\text{spam emails containing } w_d + 1}{\#\text{words in spam emails} + D} = \frac{\left(\sum_{n=1}^{N} I(w_d \in \underline{x_n}) \cdot y_n\right) + 1}{\left(\sum_{d=1}^{D} I(w_d \in \underline{x_n}, y_n = 1)\right) + D}$$

$$q_d = \frac{\#\text{ham emails containing } w_d + 1}{\#\text{words in ham emails} + D} = \frac{\left(\sum_{n=1}^{N} I(w_d \in \underline{x_n}) \cdot |y_n - 1|\right) + 1}{\sum_{d=1}^{D} I(w_d \in \underline{x_n}, y_n = 0) + D}$$

   (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**)

$$\hat{y}_{MAP} = \arg\max_{y} P_{y|\underline{x}_n}(y|\underline{x}_n)$$

$$\Rightarrow \log(\pi) + \sum_{d=1}^{D} x_d \log p_d \underset{\text{HAM}}{\overset{\text{SPAM}}{\gtrless}} \sum_{d=1}^{D} x_d \log q_d + \log(1-\pi)$$

$$\Rightarrow \sum_{d=1}^{D} x_d \log p_d \underset{\text{HAM}}{\overset{\text{SPAM}}{\gtrless}} \sum_{d=1}^{D} x_d \log q_d$$

   (b) Complete function classify_new_email in classifier.py, and test the classifier on the testing set. The number of Type 1 errors is 5, and the number of Type 2 errors is 3. (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**)

$$\frac{P(y=1|\underline{x})}{P(y=0|\underline{x})} \underset{\text{HAM}}{\overset{\text{SPAM}}{\gtrless}} \zeta \quad \Rightarrow \quad \log P(y=1|\underline{x}) - \log P(y=0|\underline{x}) \underset{\text{HAM}}{\overset{\text{SPAM}}{\gtrless}} \log \zeta$$

$$\sum_{d=1}^{D} x_d \log p_d - \sum_{d=1}^{D} x_d \log q_d \underset{\text{HAM}}{\overset{\text{SPAM}}{\gtrless}} \log \zeta$$

   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**)

The classifier will fail to classify emails with words that it hasn't seen before, as the probability of the email being in either SPAM or HAM would be 0 for both, as the probability of the word being in SPAM or HAM would be 0.

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**)

$$\hat{\underline{u}}_M = \frac{1}{N_M} \sum_{i=1}^{N} \underline{x}_i \cdot I(y_i = M)$$

$N_M \Rightarrow$ number of samples with $y_i = M$.

$$\hat{\underline{u}}_F = \frac{1}{N_F} \sum_{i=1}^{N} \underline{x}_i \cdot I(y_i = F)$$

$N_F \Rightarrow$ number of samples with $y_i = F$.

$$\hat{\Sigma}_M = \frac{1}{N_M} \sum_{i=1}^{N} (\underline{x}_i - \hat{\underline{u}}_M)(\underline{x}_i - \hat{\underline{u}}_M)^T \cdot I(y_i = M)$$

$$\hat{\Sigma}_F = \frac{1}{N_F} \sum_{i=1}^{N} (\underline{x}_i - \hat{\underline{u}}_F)(\underline{x}_i - \hat{\underline{u}}_F)^T \cdot I(y_i = F)$$

$$\hat{\Sigma} = \frac{1}{N} \left(N_M \hat{\Sigma}_M + N_F \hat{\Sigma}_F \right)$$

(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**)

$$\underline{u}_M^T \Sigma^{-1} \underline{x} - \underline{u}_F^T \Sigma^{-1} \underline{u}_F = 0.5\, \underline{u}_M^T \Sigma^{-1} \underline{u}_M - 0.5\, \underline{u}_F^T \Sigma^{-1} \underline{u}_F$$

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**)

$$-\frac{1}{2} \underline{x}^T \Sigma_M^{-1} \underline{x} + \underline{x}^T \Sigma_M^{-1} \underline{u}_M - \frac{1}{2} \underline{u}_M^T \Sigma_M^{-1} \underline{u}_M - \frac{1}{2} \log|\Sigma_M| + \log(0.5)$$

$$=$$

$$-\frac{1}{2} \underline{x}^T \Sigma_F^{-1} \underline{x} + \underline{x}^T \Sigma_F^{-1} \underline{u}_F - \frac{1}{2} \underline{u}_F^T \Sigma_F^{-1} \underline{u}_F - \frac{1}{2} \log|\Sigma_F| + \log(0.5)$$

(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 $\boxed{0.1}$ for LDA, and $\boxed{0.127}$ for QDA. (1 **pt**)