

ECE368: Probabilistic Reasoning

Lab 1: Classification with Multinomial and Gaussian Models

Name: Shadman Kaif

Student Number: 1005303137

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 `ldaqa.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)

Let D be the total # of distinct words in spam and ham.

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

- (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)

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

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

$$y = \underset{y}{\operatorname{argmax}} P(\mathbf{x}|y) = \underset{y}{\operatorname{argmax}} \frac{(x_1 + x_2 + \dots + x_D)!}{x_1! x_2! \dots x_D!} \frac{0}{d=1} \prod_{d=1}^D P(x_d|y)^{x_d}$$

$$\frac{0}{d=1} \prod_{d=1}^D p_d^{x_d} \sum_{\text{spam}} \frac{0}{d=1} \prod_{d=1}^D q_d^{x_d}$$

- (b) Complete function `classify_new_email` in `classifier.py`, and test the classifier on the testing set. The number of Type 1 errors is 2, and the number of Type 2 errors is 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 new parameter, r , which is the ratio

$$\frac{\prod_{d=1}^D (p_d)^{x_d}}{\prod_{d=1}^D (q_d)^{x_d}} \sum_{\text{spam}} r$$

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 μ_m , μ_f , Σ , Σ_m , and Σ_f as functions of the training data $\{x_n, y_n\}, n = 1, 2, \dots, N$. (1 pt)

$$\begin{aligned}\mu_m &= \frac{1}{\# \text{ of males}} \sum_{i=1}^N 1\{y_i = 1\} x_i \\ \mu_f &= \frac{1}{\# \text{ of females}} \sum_{i=1}^N 1\{y_i = 2\} x_i \\ \Sigma_m &= \frac{1}{\# \text{ of males}} \sum_{i=1}^N (x_i - \mu_m)(x_i - \mu_m)^T 1\{y_i = 1\} \\ \Sigma_f &= \frac{1}{\# \text{ of females}} \sum_{i=1}^N (x_i - \mu_f)(x_i - \mu_f)^T 1\{y_i = 2\} \\ \Sigma &= \frac{1}{N} \left(\sum_{i=1}^N (x_i - \mu_m)(x_i - \mu_m)^T 1\{y_i = 1\} + (x_i - \mu_f)(x_i - \mu_f)^T 1\{y_i = 2\} \right)\end{aligned}$$

- (b) In the case of LDA, write down the decision boundary as a linear equation of x with parameters μ_m , μ_f , and Σ . Note that we assume $\pi = 0.5$. (0.5 pt)

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

In the case of QDA, write down the decision boundary as a quadratic equation of x with parameters μ_m , μ_f , Σ_m , and Σ_f . Note that we assume $\pi = 0.5$. (0.5 pt)

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

- (c) Complete function `discrimAnalysis` in `lda_qda.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)