Assignment 1

Neel Bhalla

October 2021

1 Intro

All code was done in MATLAB r2021a. Data is stored in files `q1data.mat`, `q2data.mat`, and `q2pdf.mat` for each program to load.

2 Question 1

In order to initialize the data, samples were picked from three possible gaussians: gaussian 01(label 0) with probability 0.325, gaussian 02 (label 0) with probability 0.325, and gaussian 1 (label 1) with probability 0.35. The function `mvnrnd` was utilized to generate the data. Below is a graph of the data with its respective labels.

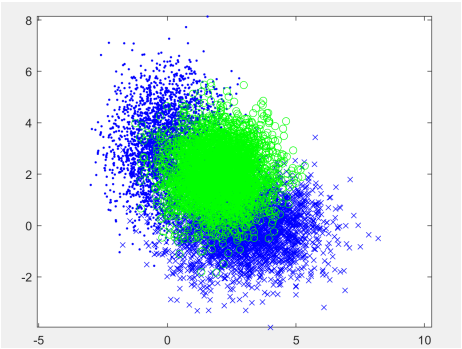


Figure 1: Distribution of data in terms of class labels. Blue is class 0, and Green is class 1.

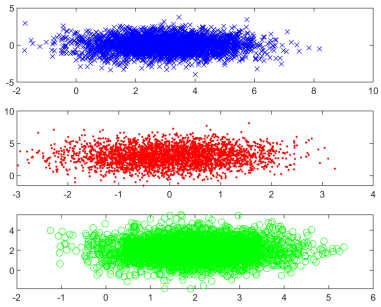


Figure 2: Plot of three gausians. The blue and red distributions formed class 0 (with equal weights), and the green one is class 1 (observe it is centered at 2,2).

2.1 ERM Classification with knowledge of True pdf

Recall the special case ratio test for a 2 class system is:

$$\frac{p(x|L=1)}{P(x|L=0)} \lessgtr \gamma$$

for some threshold γ. When the problem is parametrized by a loss matrix, γ can be expressed as

$$\frac{p(x|L=1)}{P(x|L=0)} \lessgtr \frac{(\lambda_{10}-\lambda_{00})p(L=0)}{(\lambda_{01}-\lambda_{11})P(L=1)}$$

In order to minimize probability of error, we can use the 0-1 loss matrix:

$$\lambda = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

Substituting values into the original equation yields:

$$\frac{p(x|L=1)}{P(x|L=0)} \lessgtr \frac{p(L=0)}{P(L=1)}$$

And since we know the class priors, namely $P(L=0)=0.65$ and $P(L=1)=0.35$:

$$\frac{p(x|L=1)}{P(x|L=0)} \lessgtr \frac{0.65}{0.35} = 1.857$$

2.2 ROC Curve for ERM model

Below is the ROC curve for the 2-class ERM model. When operating at the theoretical optimal, the model performs well, and furthermore the AUC of the model is large indicating a strong model.

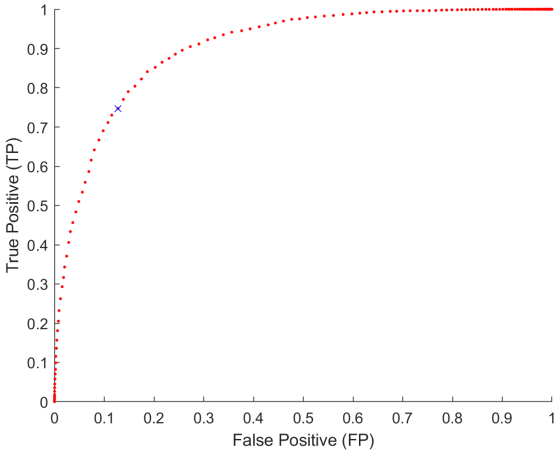


Figure 3: Plot of ROC of the model. Theoretical optimal γ point marked on graph with blue x

.

2.3 Comparison of actual optimal γ to theoretical

Since the dataset is finite, the theoretical optimal not necessarily the actual optimal. That being said, for the current dataset, the acutal optimal is at $\gamma = 2.01$, but depending how the dataset was generated, will fall within 0.1 of the theoretical optimal. Figure 4 depicts the probability of error over different gamma's to visualize the neighborhood.

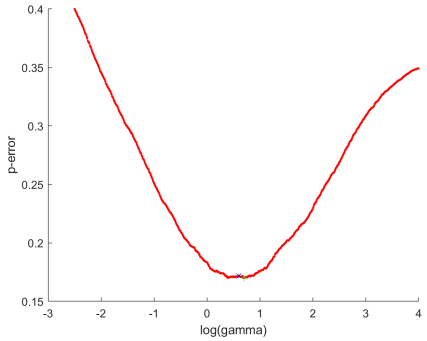


Figure 4: Plot of p-error wrt $\log \gamma$. Theoretical optimal denoted by blue x, and actual optimal denoted by green plus.

.

2.4 LDA classifier

The next approach was done by utilizing an LDA based classifier. The classifier takes the form:

$$y = w^T x \lessgtr \tau$$

For some trained vector w, and threshold τ. Ideally the projection vector w would result in the two class' points to be as apart as possible, with τ being the best threshold for w.

In order to estimate w, sample means and covariances were estimated from the samples generated. Since class 0 is the mixture of two gaussians, its mean and covariance are stretched to explain both distributions. Next the scatter within, and scatter between matrices were computed where.

$$S_B = (m_1 - m_0)(m_1 - m_0)^T$$

$$S_W = \Sigma_0 + \Sigma_1$$

Next the eigenvector corresponding with the largest eigenvalue of the two matrices was found and treated as w. Finally all data was projected onto w, and values of τ were iterated to find the minimum p-error threshold.

The largest eigenvalue was 0.2179 with an eigenvector of $\langle 0.7259, 0.6878 \rangle$.

2.5 ROC

The LDA based classifier did not perform well. This is partially due to the fact that class 0 was estimated as a gaussian, where it is in fact a mixture of 2, but also the fact that class 1 distribution sits between the two distributions that compose class 0. A linear boundary will never perform well (Refer to figure 1), a nonlinear boundary is required. The ROC is shown on Figure 5.

Looking at the graph of γ vs p-error, it is clear that the model struggled to find a threshold given w. It was able to find a threshold of $\tau = 3.35$ which brough p-error to 0.3447, which is barely below a threshold of $\tau \to \inf$, aka, classifying everything with label 0. LDA performed much worse than ERM, which was able to achieve a p-error of 0.16 which is half of LDA's p-error. Figure 6 depicts the graph of p-error vs τ.

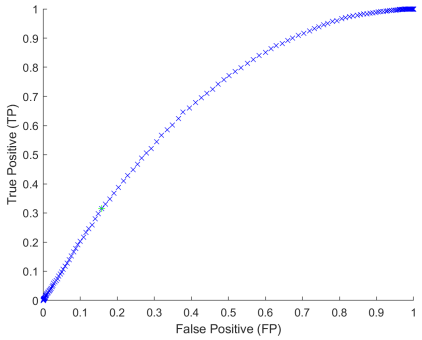


Figure 5: ROC for LDA classifier. The AUC is not very large (much less than ERM classifier), indicating a weak / mismatched model. The green plus indicates the threshold that minimizes p-error

.

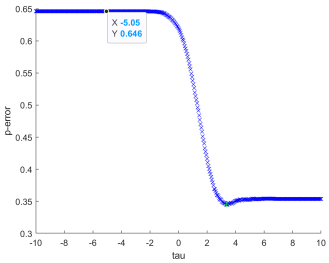
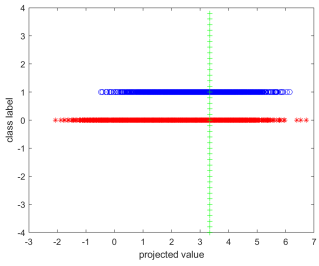
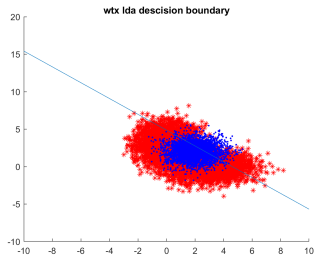


Figure 6: τ vs p-error. It is clear that the optimal value, indicated by a green plus, performs barely better than classifying everything as class 0

.



(a) Projected LDA threshold visualization. Data is split by classes on different horizonta lines, and the threshold is indicated as a green vertical line.



(b) Optimal LDA descision boundary on data. All data below line is classified as 0(red), and all above as 1(blue)

Figure 7: LDA descision boundaries (projected and superimposed on dataset)

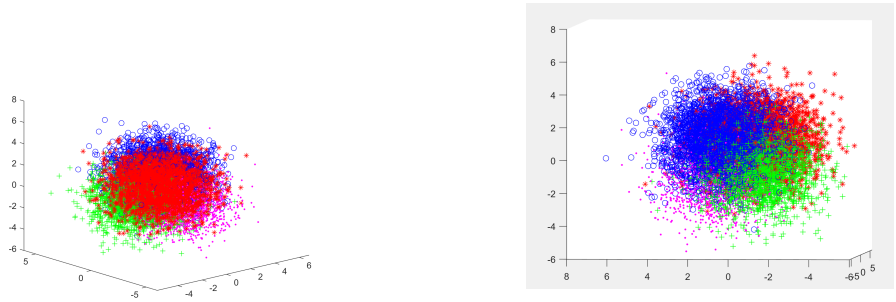


Figure 8: Distribution of points. Green and magenta are merged to form 1 class label.

3 Question 2

In order to find 4 equidistant points, the vertices of a tetrahedron utilized. The Gaussian means were assigned from (1, 1, 1), (-1, -1, 1), (-1, 1, -1), and (1, -1, -1). Since the distances between the points are $2\sqrt{2}$, the intended $\sigma = \sqrt{2}$, and var $= \sigma^2 = 2$. The covariance matrix becomes $2I_3$ as a result.

The gaussians centered around (-1, 1, -1) and (1, -1, -1) are merged together with weights $\frac{1}{2}$ each to form the third class label.

For generation, samples were draw from label 0 with $\frac{3}{10}$, label 1 with $\frac{3}{10}$, and label 2 with $\frac{4}{10}$. Since label 2 was composed of 2 gaussians, they were each sampled with probability $\frac{1}{5}$. Plot of points is on Figure 8.

3.1 ERM 3-Class classifier

Recall that an ERM classifier aims to reduce the expected risk, by computing $R(L = i|x)$ for class-labels i and picking the smallest one. In order to calculate the Risk vector (where each entry is the risk assosciated with deciding label l for sample x), is:

$$R = \Lambda p(L|x)$$

where $P(L|x)$ is a column vector representing $P(L = i|x)$ $\forall i$, and Λ is a 3x3 matrix with $\Lambda_{i,j}$ is the penalty associated with deciding i when the true label is j.

In order to compute $p(L = i|x)$, we utilize bayes' rule:

$$p(L = i|x) = \frac{p(x|L = i)p(L = i)}{p(x)}$$

where $p(x|L = i)$ is a pdf lookup operation, $p(L = i)$ is a class prior, and $p(x) = \sum_{i=1}^{c} p(x|L = i)p(L = i)$

Putting it all together, we get

$$P(L|x) = \frac{1}{p(x)} \begin{bmatrix} p(x|L=1)p(L=1) \\ p(x|L=2)p(L=2) \\ \cdots \\ p(x|L=c)p(L=c) \end{bmatrix}$$

And the risk vector formula becomes

$$R = \frac{1}{p(x)} \Lambda \begin{bmatrix} p(x|L=1)p(L=1) \\ p(x|L=2)p(L=2) \\ \cdots \\ p(x|L=c)p(L=c) \end{bmatrix}$$

with

$$d = \underset{i \in \{1,2,\ldots c\}}{\text{argmin}} \; R_i$$

3.2 Minimizing P(error) with 0-1 loss

In this experiment, we utilize

$$\Lambda = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$$

which penalizes all mistakes equally, and gives 0 penalty for correct answers.

Testing the classifier on Λ with our pre-defined dataset yielded:

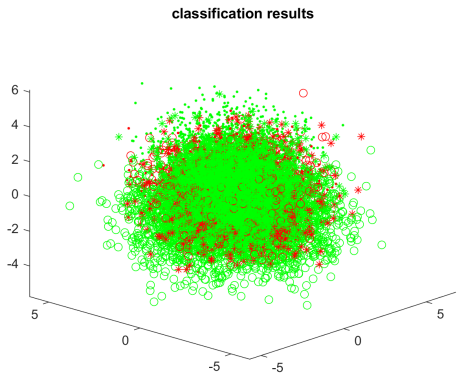


Figure 9: Classification results on 0-1 error. Accuracy was 0.7313 with a p-error and expected loss of 0.2687. Dots are label 1, stars are label 2, and circles are label 3.

Looking at the confusion matrix,

$$\begin{bmatrix} 0.7135 & 0.1106 & 0.1209 \\ 0.1058 & 0.7116 & 0.1197 \\ 0.1807 & 0.1778 & 0.7594 \end{bmatrix}$$

It is clear that mistakes were made with simmilar probabilities, likely caused by the large amount of overlap between the different data-points.

3.3 Experimenting with other loss functions

3.3.1 10 loss

The next experiment was to work with situations where penalties for classifying datapoints of label 3 is increased. For example,

$$\Lambda = \begin{bmatrix} 0 & 1 & 10 \\ 1 & 0 & 10 \\ 1 & 1 & 0 \end{bmatrix}$$

penalizes deciding 1 or 2 where the truth is 3 with penalty 10, 10 times more than all other mistakes. As a reuslt, the ERM classifier will over-classify points as label 3 when somewhat uncertain, in hopes to reduce the expected loss.

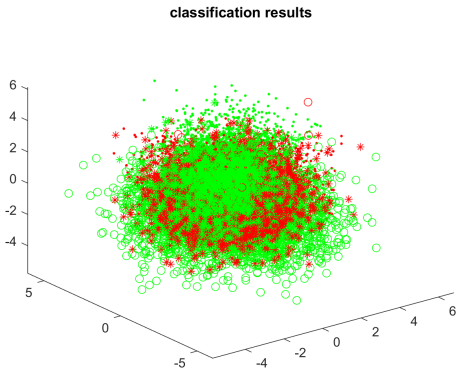


Figure 10: Classification results on 10 error. Accuracy was 0.5653 with an expected loss of 0.5084. Dots are label 1, stars are label 2, and circles are label 3.

While the accuracy is much lower at 0.5653, it should be noted if the confusion matrix from 0-1 error was applied to 10-error Λ, the expected loss would be 1.13, which is nearly twice the expected loss by the current experiment at 0.5084. This demonstrates that accuracy is not the primary objective function, and classifiers with relatively low accuracy, can still have optimal expected loss given a non-uniform loss function.

The confusion matrix,

$$\begin{bmatrix} 0.2852 & 0.0291 & 0.0092 \\ 0.0256 & 0.2931 & 0.0112 \\ 0.6892 & 0.6778 & 0.9795 \end{bmatrix}$$

affirms our suspicions that the classifier is over-classifying samples as label 3, in hopes to not incur the large penalties from $\Lambda_{1,3}$ and $\Lambda_{2,3}$. Classes' 1 and 2 were correctly identitfied with probability 0.28 and 0.29, whereas Class 3 was correctly identified with probability 0.9795. It is reasonable to believe that the remaining 2% of samples from Class 3 may be significant outliers nearly $\geq 2\sigma$ away from the gaussian mixture for class 3.

3.3.2 100 loss

The final experiment was for making penalties for deciding 1 and 2 where truth is 3 as 100. This is extreme, but applicable in cases where classifying 3 could be catastrophic (like maybe missing a category cancer cells that progress extremely fast). The loss matrix becomes

$$\Lambda = \begin{bmatrix} 0 & 1 & 100 \\ 1 & 0 & 100 \\ 1 & 1 & 0 \end{bmatrix}$$

Looking at the confusion matrix,

$$\begin{bmatrix} 0.0316 & 0.0003 & 0.0005 \\ 0.0013 & 0.0297 & 0 \\ 0.9671 & 0.9699 & 0.9995 \end{bmatrix}$$

shows that the classifier decided label 3 for almost everything unless it was **extremely certain** that it was from another label. Since $\Lambda_{2,3}$ was so high, it made 0 mistakes and made only 5 mistakes deciding label 1 with truth 3. Given the loss matrix, the confusion matrix produced makes sense, as while classes 1 and 2 were only correctly labeled with probability 0.03, class 3 was correctly predicted with probability 0.9995.

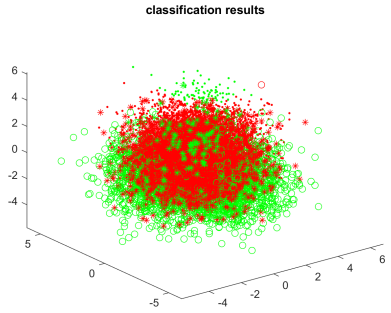


Figure 11: Classification results on 100 error. Accuracy was 0.4182 with an expected loss of 0.6016. Dots are label 1, stars are label 2, and circles are label 3.

4 Appendix

All the figures are included in the `Images/` folder wihin the zip, but also can be produced in MATLAB (especially for 3d plots that couldnt be captured as nicely. The data generation files will generate new data, but not save it in order to preserve replicable results in the other parts of the assignment.

file	purpose	Figures created
`q1datagen.m`	make data	1 2
`q1aroc.m`	make roc	3 4
`q1b.m`	lda and visuals	5 6 7a 7b
`q2datagen.m`	make data	8
`q2a.m`	0-1 loss experiment	9 and confusion matrix
`q2b.m`	10 loss experiment	10 and confusion matrix
`q2c.m`	100 loss experiment	11 and confusion matrix

Thanks,
Neel.