

MLT - Assignment 3

Ayush Sekhari (12185)

March 11, 2015

1 Part I and II

Part I and II were hand solved and are attached with the report. Please scroll down to view them.

2 Part III

2.1 Part III-a

2.1.1 Generating Data

I used inbuilt `runif`, `rnorm` functions to generate the data. 600 points for each class were used. 400 of them were used to train and 100, 100 for validation and test respectively. The metadata of the data is :

1. Training Data : 1200
2. Validation Data : 300
3. Test Data : 300
4. Priors = $[1/3, 1/3, 1/3]$

2.1.2 Calculation of the Initial Model

Each class is to be represented as a mixture of three gaussians. I used **K-means clustering** on training data for each class with $k = 3$. So each class data is clustered into 3 groups. These give 3 cluster means and co-variance matrices for each class. These lead to 9 mean vectors and 9 covariance matrices. I also estimate initial priors for each cluster as the cluster group strength.

2.1.3 Estimation of Co-Variance Matrix

I took the pooled co-variance matrix as the weighted average of the cluster co-variance matrices. The weights used were (priors - 1) in order to ensure unbiased estimator of the co-variance. This is kept constant for the problem.

Approach	CV error	Test-Set error
k-means {training error = 12.333}	16.666	15.666
Stochastic Variation	16.333	15.333
Varying all of the 21 components simultaneously with $c = \{-0.5, 0, 0.5\}$	14	17.333
Varying all of the 21 components simultaneously with $c = \{-0.1, 0, 0.1\}$	15.666	16.333

Table 1: Observed Best errors while searching the space around means using various approaches

2.1.4 Generic Approach:

I have an initial model with 9 mean vectors and a SIGMA matrix. I vary the mean vectors locally using various approaches, recompute the priors each time and then use the mixture model to estimate the error on the validation set. The one with the best is used and the errors are reported on the test set. Performance and search results using various search approaches are listed in the table below.

2.2 Part III-b – LDA and QDA

1. LDA gives the error of 0.136 on the test set
2. QDA gives the error of 0.153 on the test set
3. **LDA performs the best**