Multivariate Data Analysis. Home Assignment 1

Xijia Liu

Part I: Theoretical problems

Solve the following theoretical exercises: 2.32, 3.18 and 4.16 in Applied Multivariate Statistical Analysis (AMSA). **Note**: there are (at least) two versions of AMSA, 6th edition, with the same ISBN-number. In the most recent one they have changed the numbering of the chapters: Ch 2 became Ch3 and Ch3 became Ch 2. The numbering of the exercises in this assignment refers to the older version. The easiest way to know what exercises to do is: In the chapter (2 or 3) with 42 exercises, do Number 32. In the chapter (2 or 3) with 20 exercises, do Number 18. The chosen exercise in Ch 4 has the same number in both versions.

Part II: Gaussian model, Gaussian discriminant analysis, GMM, and EM algorithm in R.

Task 1: Do you know pseudo-random numbers? Basically, pseudo-random numbers are a sequence of numbers generated from an initial number by a certain algorithm. They can simulate the behavior of a random sample evenly distributed between 0 and 1. Once the pseudo-uniform random numbers are ready, different algorithms can be applied to them to generate random numbers from a specific distribution.

The task is to implement the Box and Muller's algorithm as a function that can generate random numbers from a Gaussian distribution with **arbitrary means and covariance matrix**.

You need to use the knowledge from chapter 4 to explain your function. Once the function is implemented, you need to apply it to generate 1000 observations from a two-dimensional Gaussian distribution with mean vector $\boldsymbol{\mu} = (3, 1)^\top$, standard deviations are $\sigma_1 = 1$, $\sigma_2 = 2$, and correlation $\rho = -0.6$. Make a scatter plot for your random sample.

Box and Muller's Algorithm:

1. Independently generate u_1 and u_2 from uniform distribution, $U(0, 1)$.
2. Set $x = \sqrt{-2\log(u_1)}\cos(2\pi u_2)$ and $y = \sqrt{-2\log(u_1)}\sin(2\pi u_2)$. Then x and y are independent normal distributed with mean 0 and variance 1

Can you theoretically proof the random numbers generated by this algorithm do follow Gaussian distribution? (Optional[1])

Task 2: In this task, we apply Gaussian discriminant analysis to a real data set. The data set contains breast cancer diagnostic results and 30 features created from medical images of breast mass from 569 patients. For more information about the data, please check the 'help.txt' document. We want to build up a classifier to predict the diagnostic results given the values of those 30 features. The original data set is split into two parts. 469 observations are randomly drawn as training set and saved in 'train.txt'. The rest of the data is saved in 'test.txt' for testing.

Train Gaussian discriminant model on the training set. Then apply the classifiers on the testing set to predict the diagnostic for each case. Given the testing results, calculate the accuracy of your classifier. It is allowed to apply the build-in function 'lda' in R to solve this problem. For this option, you can practice on how to read the help document in R and learn a function by yourself. Another option is to solve this problem by your own code. It is also a good practice!

[1] It means this question will not be recorded in the total score.

Task 3 GMM and estimation with EM algorithm has been implemented in a useful package 'mixtools' in R. Please install this package first.

Task 3.1: 'mixtools' provides a function, 'rmvnorm' by which data can be simulated from multivariate normal distribution with arbitrary mean vector and covariance matrix. Please try to learn this function by typing '?rmvnorm', and generate 1000 realizations from a two dimension Gaussian distribution with mean vector $\boldsymbol{\mu} = (2, 3)^\top$, variances are $\sigma_1^2 = 1$, $\sigma_2^2 = 4$, and correlation $\rho = 0.7$. Make a scatter plot for your random sample.

Task 3.2: Please simulate 1000 realizations from the following GMM:

- The latent (label) variable z_i belongs to Bernoulli distribution with parameter $p = 0.6$ for $i = 1, ..., 1000$
- The conditional distribution given the value of latent variable:

  $$\mathbf{X}_i | z_i = 1 \sim \mathcal{N}_2(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) \text{ and } \mathbf{X}_i | z_i = 0 \sim \mathcal{N}_2(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$$

  where $\boldsymbol{\mu}_1 = (2, 3)^\top$ and $\boldsymbol{\mu}_2 = (3, 2)^\top$. For $\boldsymbol{\Sigma}_1$, the standard deviations are 0.2 and 0.6, the correlation is 0.5. For $\boldsymbol{\Sigma}_2$, the standard deviations are 0.4 and 0.3, the correlation is 0.5

Task 3.3: Please read the help document of function 'mvnormalmixEM', then fit a GMM on your simulated data and compare the estimation results of parameters with the true values.

Task 4 In this task, we will apply the EM algorithm to obtain the maximum likelihood estimation (MLE) of the parameters of a mixture of the binary distribution. First, apply the following R code to generate the random observations with random seed 2021.

```
n = 50
N = 30
```

```
p = 0.6
p_a = 0.4
p_b = 0.7
z = rbinom(n, 1, p)
x = rbinom(n, N, p_a)
x1 = rbinom(n, N, p_b)
x[z!=1] = x1[z!=1]
x
rm(z, x1)
```

This data-generating process can be viewed as simulating the following experiment. There are two kinds of coins, A and B, with different probabilities of getting head in a black box. Every time, we randomly draw a coin and flip it 30 times and record the number of heads. Repeat this procedure 50 times. Write an R function to apply the EM algorithm to get the MLE of the proportion of coins A, the probabilities of getting head. You may also do some experiments to investigate how do the parameters n and N affect your estimation results (the last question is optional).