Group Projects on EM algorithms

Project 1: Lifestyle Factors and Heart Disease Risk

In this project, you will explore what lifestyle and clinical factors are most predictive of heart disease risk? The data we will use is from the Heart Disease UCI dataset, which contains 14 clinical and lifestyle factors and the stage of heart disease (0 represents no heart diseases and 1-4 represents different stages of heart desises)(see heart.csv and its dictionary).

Logistic regression is a statistical method that is frequently used to determine which factors are associated with the presence of heart diseases. However, as you can observe from the data, some of the factors contain missing data. In fact, Cleverland is the only site that has complete data.

Please propose an E-M algorithm to enable other sites to take part in the analysis. With your E-M algorithm, we can iteratively estimate the missing data and the model parameters. Bootstraping can be used to make the inferences on the estimated model parameters.

Please design and present your E-M algorithm to enable full-data analysis, its subsequent bootstrap-based inferences and compare the results with the analysis using only the data from the Cleveland. Write a report fo your findings.

Project 2: Identifying Subgroups within Type 2 Diabetes Patients

Gaussian Mixture models are commonly used approaches to identify hidden subgroups or patterns, which provide useful insights to follow up analyses. The Pima Indians Diabetes Database includes clinical profiles of 768 subjects from both diabetic and non-diabetic individuals. Clustering those profiles could help group those individuals who might be at different stages of diabetes, and help understand the personalized risk.

The data included 768 subjects with 268 diagnosed with Type 2 Diabetes, and recorded the following features

- Number of times pregnant
- Plasma glucose concentration a 2 hours in an oral glucose tolerance test
- Diastolic blood pressure (mm Hg)
- Triceps skinfold thickness (mm)
- 2-Hour serum insulin (mu U/ml)
- Body mass index (weight in kg/(height in m)^2)
- Diabetes pedigree function
- Age (years)
- Outcome: 1- diebets 0 - not

Please note that the zero's in the variables Glucose, BloodPressure, SkinThickness, Insulin and BMI should be treated as missing data.

Ignoring the class variable, please design and implement an EM algorithm that helps identify distinctive subgroups of features in the data using a Gaussian Mixture Model. Please note that, in your E-M algorithm, the unobserved data include both missing values in those biomarkers and the latent classes.

Once you have identfied your subgroups, please characterize the feature patterns of each subgroup, and examine how the resulting subgroups predict the Type 2 diabetes?

Report your findings.

Project 3: Single-Cell gene expressions

ScRNA-seq is a recent technological breakthrough in biology that measures gene expression levels of thousands of genes at the individual cell level. This technology has significantly advanced our understanding of cell functions.

Cell heterogeneity, the phenomenon of varied gene expression levels in individual cells of the same type, is a complex issue. However, clustering analysis, a powerful tool, can help us navigate this complexity and identify subtypes in cases of cell heterogeneity.

The file ss.csv contains gene expression levels of 558 genes (columns) from 716 cells (rows) of breast cancer tumors.

In this analysis, we will start by applying Principal Component Analysis to our scRNA-seq data. The goal here is to determine the number of principal components needed to capture the variability of gene expressions. To achieve this, we will use the function ‘prcomp’ from the R package ‘stats’.

After identifying the major principal components, you need to fit a Gaussian-Mixture model to their principal component scores. Make sure to use the number of components you selected in Q1. Next, design and implement an EM algorithm to estimate the model and identify the cell clusters. Once you have identified the clusters, explore gene-expression signatures in each of them. Find out which genes are important to differentiate the clusters and report your findings.