Assignment 1

STAT 3006, 2021/22

Due date: Feb 25, 2022 23:59, Friday

Question 1

(25%) Please use the bisection method to find all zero points of the following function,

$$f(x) = x^3 + 6x^2 + \pi x - 12.$$

Question 2

(25%) (Poisson regression) We collected $n = 50$ independent count observations $\{y_i : i = 1, \ldots, n\}$ and their corresponding covariates $\{x_i : i = 1, \ldots, n\}$. Assume the relationship between y_i and x_i (for $i = 1, \ldots, n$) is $y_i \sim Poisson(\lambda_i)$ and $log(\lambda_i) = \alpha + \beta x_i + \gamma x_i^2$. Please

1. write down the likelihood function $L(\alpha, \beta, \gamma | \mathbf{x}, \mathbf{y})$ of the Poisson regression model;
2. derive the Newton method for maximizing $L(\alpha, \beta, \gamma | \mathbf{x}, \mathbf{y})$;
3. implement the Newton method using R to get MLE of (α, β, γ).

The data set $\{(x_i, y_i) : 1 \leq i \leq n\}$ is stored in `PoisRegData.txt`.

Question 3

(20%) (Logistic regression) We collected $n = 50$ independent binary observations $\{y_i : i = 1, \ldots, n\}$ and their corresponding covariates $\{x_i : i = 1, \ldots, n\}$. Assume the relationship between y_i and x_i (for $i = 1, \ldots, n$) is $y_i \sim Bernoulli(p_i)$ and $logit(p_i) = \alpha + \beta x_i$, where $logit(t) = \log \frac{t}{1-t}$. Please

1. write down the likelihood function $L(\alpha, \beta | \mathbf{x}, \mathbf{y})$ of the logistic regression model;
2. derive the Newton method for maximizing $L(\alpha, \beta | \mathbf{x}, \mathbf{y})$;
3. implement the Newton method using R to get MLE of (α, β).

The data set $\{(x_i, y_i) : 1 \leq i \leq n\}$ is stored in `LogitRegData.txt`.

Question 4

(30%) (EM algorithm) The monthly salary of $n = 8000$ employees are drawn from a company. Assume that there are three salary levels, including low income, middle income and high income. We denote the monthly salary of employee i by Y_i, and the salary level of employee i by Z_i. $\{Y_i : 1 \leq i \leq n\}$ are observed, but $\{Z_i : 1 \leq i \leq n\}$ are unknown.

Our model can be formulated as follows. First, for any $k = 1, 2, 3$, $Pr(Z_i = k) = \pi_k$ and $\sum_{k=1}^{3} \pi_k = 1$, where $Z_i = 1$ indicates employee i is low-income, $Z_i = 2$ indicates employee i is middle-income, $Z_i = 3$ indicates employee i is high-income, and $\boldsymbol{\pi}$ can be interpreted as the proportion of employees belonging to each salary level. Second, given $Z_i = k$, Y_i is assumed to be from a normal distribution $N(\mu_k, \sigma_k^2)$. Based on these notations and information, please

1. write down the complete-data likelihood function $L(\pi_1, \pi_2, \mu_1, \mu_2, \mu_3, \sigma_1, \sigma_2, \sigma_3 | \mathbf{Y}, \mathbf{Z})$;
2. derive E step and M step to find MLE of $(\pi_1, \pi_2, \mu_1, \mu_2, \mu_3, \sigma_1, \sigma_2, \sigma_3)$;
3. implement your EM algorithm, give MLE of $(\pi_1, \pi_2, \mu_1, \mu_2, \mu_3, \sigma_1, \sigma_2, \sigma_3)$;
4. distinguish the first 50 employees' salary level.

The data set $\{Y_i : 1 \leq i \leq n\}$ is stored in `SalaryData.txt`.

Requirements

Your answer must contain two parts. The first part is a paper report which includes your derivation and answers for each problem. The second part is a file which includes all your R code to implement your algorithms. Please, by the due date, submit your paper report and your R code file to the blackboard system or TA (1155094474@link.cuhk.edu.hk). You must finish both of the two parts to get a grade. Otherwise, your homework will be regarded as missing. Details of requirements are in the table below.

-	in the paper report	in the R code file
Q1	all zero points	implementing the bisection method
Q2	likelihood function derivation procedure for the Newton algorithm MLE of (α, β)	implementing the Newton method
Q3	likelihood function derivation procedure for the Newton algorithm MLE of (α, β)	implementing the Newton method
Q4	complete data likelihood function L derivation procedure for E step and M step MLE of $(\pi_1, \pi_2, \mu_1, \mu_2, \mu_3, \sigma_1, \sigma_2, \sigma_3)$ The first 50 students' genders you learned	implementing the EM algorithm