Topics on High-Dimensional Data Analytics
ISYE 8803 - Spring 2019
Homework 4
Due 02/27/2019 11:59pm

Problem 1. Convex sets and convex functions (10 points)
Prove that the following sets are convex.

a. The set $\{x : ||x||_2 \leq 1\}$, where $||x||_2 = \sqrt{\sum_{i=1}^n x_i^2}$ is the Euclidean norm.

b. The set $\{x : Ax = b\}$, where $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$ are fixed.

Prove that the following functions are convex.

a. Exponential: $f : \mathbb{R} \to \mathbb{R}$, $f(x) = e^{ax}$ for any $a \in \mathbb{R}$.

b. Negative logarithm: $f : \{x : x > 0\} \to \mathbb{R}$, $f(x) = -\log x$.

c. Non-negative weighted sum of convex functions: $f(x) = \sum_{i=1}^k \omega_i f_i(x)$, where $f_1, \cdots, f_k$ are convex functions and $\omega_1, \cdots, \omega_k$ are non-negative numbers.

Problem 2. Medical image estimation (60 points)
In the setting of emission tomography, let x_i represent the number of photons emitted by source i, $i = 1, \cdots, n$. Suppose x_i has Poisson distribution with

$$P(x_i = k) = \frac{e^{-\mu_i}\mu_i^k}{k!}$$

with unknown mean μ_i, $i = 1, \cdots, n$.
We consider an experiment design to determine the means μ_i. The experiment involves m detectors. If event i occurs, it is detected by detector j with probability p_{ji} ($p_{ji} > 0$ and $\sum_{j=1}^m p_{ji} \leq 1$). The total number of events recorded by detector j is denoted by y_j,

$$y_j = \sum_{i=1}^n y_{ji}, \quad j = 1, \cdots, m.$$

The likelihood function for estimating the means μ_i, based on observed values of y_j, $j = 1, \cdots, m$ is:

$$L(\mu_1, \cdots, \mu_n) = \prod_{j=1}^m \frac{\exp\left(-\sum_{i=1}^n p_{ji}\mu_i\right)\left(\sum_{i=1}^n p_{ji}\mu_i\right)^{y_j}}{y_j!}.$$

We used the following two facts:

- The variables y_{ji} have Poisson distribution with means $p_{ji}\mu_i$.
- The sum of n independent Poisson variables with means $\lambda_1, \cdots, \lambda_n$ has a Poisson distribution with mean $\lambda_1 + \cdots + \lambda_n$.

Follow the following steps.

a. (5 points) Write down the log-likelihood for estimating the means μ_i, based on observed values of $y_j,\ j = 1, \cdots, m$.

b. (5 points) Write down the corresponding maximum likelihood formulation.

c. (20 points) Derive the gradient and the Hessian of the log-likelihood function.

d. (20 points) In the file emission.mat, you will find the observed values of y_j, $j = 1, \cdots, 5$, and the probabilities p_{ji}, $j = 1, \cdots, 5$ and $i = 1, \cdots, 3$. Use the following optimization methods to estimate the values of μ_i, $i = 1, \cdots, 3$. Please submit your codes. Plot the value of the log-likelihood function versus the number of iterations.

   - Gradient descent
   - Accelerated gradient descent
   - Stochastic gradient descent
   - Newton's method

e. (10 points) Compare the previous methods in terms of prediction error (the true values of μ_i, $i = 1, \cdots, 3$ can be found on the file true.mat) and number of iterations. Which method has a better performance?

Problem 3. Leak estimation in transmission systems (30 points)
Leakage of the transmission fluid or oil in power train systems can cause engine overheating and permanent damage. Therefore, it is crucial to run a leak test to inspect for any possible porosity in the casting parts. For this purpose, a non-linear regression model is developed. The goal is to model the leak flow profiles as a function of both the leak testing time and the part temperature. The model is the following:

$$F_{ijk} = (\beta_{0i} + \beta_{1i}T_{ik})(1 - \exp(-2\beta_{2i}t_{ijk})) + \epsilon_{ijk}$$

where F_{ijk} is the leak flow measured for part i at time t_{ijk} with temperatures T_{ik} ($i = 1, \cdots, m;\ j = 1, \cdots, n_i;\ k = 1, \cdots, K_i$). T_{ik} is the kth temperature at which part i is tested, t_{ijk} represents the testing time for recording the jth measurement of the leak flow profile for part i tested at its kth temperature, and ϵ_{ijk} are independent and identically distributed random noise variables with $\epsilon_{ijk} \sim N(0, \sigma_\epsilon^2)$.

In this problem, your goal is to estimate the parameters $\beta_{0i},\ \beta_{1i},\ \beta_{2i}$, for every part i. The data can be found on the file leak.mat. Leak is a matrix where the first column is the part, the second column is time, the third column is temperature, and the fourth column is the corresponding leak value. Additionally, b contains the true values of the parameters. For these purpose, you will minimize the following loss function:

$$L(\beta_{0i},\ \beta_{1i},\ \beta_{2i}) = \sum_{j=1}^{J}\sum_{k=1}^{K} ||F_{ijk} - (\beta_{0i} + \beta_{1i}T_{ik})(1 - \exp(-2\beta_{2i}t_{ijk}))||_2^2$$

Follow the following steps:

a. (5 points) The function can be decomposed as

$$L(\beta_{0i},\ \beta_{1i},\ \beta_{2i}) = g(\beta_{0i},\ \beta_{1i},\ \beta_{2i})^T g(\beta_{0i},\ \beta_{1i},\ \beta_{2i})$$

Define the function $g(\cdot)$.

b. (10 points) Derive the Jacobian matrix of $g(\cdot)$.

c. (5 points) Present the gradient and the Hessian of the loss function as a function of the Jacobian matrix.

d. (10 points) Use Gauss-Newton algorithm to estimate the values of $\beta_{0i},\ \beta_{1i},\ \beta_{2i},\ i = 1, \cdots, 5$. Please submit your codes. Don't forget to report your answers in the PDF file. Additionally, compute the root mean square error for each part.