Problem 6: Air Goodness Index (25 points) [File Upload]

By Pedro Paredes

*Read through the **Problem Statement** and **Grading** sections of this problem fully.*

Problem Statement

The Princeton Computer Science Department has decided to monitor the air quality around campus, and post regular updates on their website (which will be generated using a large language model). To do this, the department acquired a special sensor that is capable to collect D parameters related to air quality, with the ultimate goal of using this to determine the *air goodness index* (or AGI). To calibrate this sensor, the department rented a much more powerful one from the Physics Department that can determine the value of AGI with small error.

The CS Department then placed both sensors at N random locations and collected data in each one. Formally, we represent the readings from the weak sensor in a $N \times D$ matrix $\boldsymbol{X}$, and the AGI readings of the strong sensor in a N dimensional vector $\boldsymbol{y}$. Thankfully, it is known that the D parameters the sensor collects are linearly correlated with the value of AGI. Formally, there is some vector $\boldsymbol{\beta}$ in D dimensions such that given a vector in D dimensions $\boldsymbol{x}$ we have that $\text{AGI} = \boldsymbol{x}^T\boldsymbol{\beta} = \sum_{i=1}^{D} x_i\beta_i$.

You can assume that the values of $\boldsymbol{X}$ were all drawn uniformly over $[-10, 10]$ (because the collection locations were random). The value $\boldsymbol{y}_i$ is equal to $\boldsymbol{x}_i^T\boldsymbol{\beta} + \varepsilon_i$, where is the ith row of $\boldsymbol{X}$ and ε_i is a small error that is distributed as a standard Gaussian distribution. You also know that all values of $\boldsymbol{\beta}$ are in the range $[-10, 10]$.

Your task is to use the readings of $\boldsymbol{X}$ and $\boldsymbol{y}$ to find $\boldsymbol{\beta}$.

However, life sometimes isn't perfect, and you were given the data you realized that some of the data got corrupted. In particular, it looks like a δ fraction of all the strong sensor readings got corrupted, and got replaced by a uniformly random value in the range of the original readings. Formally, this means that a δ fraction of the values in $\boldsymbol{y}$ were replaced with a new values that is uniformly random in $[-M, M]$, for some parameter M that you determined.

Task summary: Determine the value of $\boldsymbol{\beta}$ from $\boldsymbol{X}$, $\boldsymbol{y}$, δ and M.

Grading

You will be given 5 groups of data sets (i.e. values of $\boldsymbol{X}$ and $\boldsymbol{y}$) each worth 5 points. Each group contains 10 data sets with the same values of N, D, δ and M. For each data set, you should determine the vector $\boldsymbol{\beta}$. Your solution is successful on a test case if the mean absolute difference (average of the absolute value of the coordinate-wise differences) between each coefficient of your predicted $\hat{\boldsymbol{\beta}}$ and the real $\boldsymbol{\beta}$ should be at most 0.1. You will be awarded the 5 points for a group if your solution is successful on all data sets.

Here is a summary of the parameter for each group of data sets.

0. $N = 100$, $D = 10$, $\delta = 0$, $M = 0$.
1. $N = 1000$, $D = 40$, $\delta = 0.1$, $M = 3000$.
2. $N = 1000$, $D = 40$, $\delta = 0.1$, $M = 630$.
3. $N = 50000$, $D = 100$, $\delta = 0.2$, $M = 1000$.
4. $N = 50000$, $D = 100$, $\delta = 0.5$, $M = 1000$.

The format of the input and output is specified on the next section.

Input Format

You can find all the the test cases in the `agi_data.zip` file listed on the problems section of the website. This zip file contains 50 files, each one follows the format `agi_G_T.txt`, where `G` is a digit between 0 and 4, indicating the group that file belongs to, and `T` is a digit between 0 and 9 indicating which data set it corresponds to. Each one of this files starts with a line with four space separated numbers, N, D, δ and M (you could deduce this from the group, but we included it in the header of the file to help you). Then follow N lines, each containing D space separated floating point numbers. This represents the matrix $\boldsymbol{X}$. Then follows one line with N space separated floating point numbers, representing $\boldsymbol{y}$.

You should submit a zip file called `XXXXX_p6.zip` containing files following the format `XXXXX_G_T.txt`, where `XXXXX` should be your team ID, `G` is a digit between 0 and 4, indicating the group this file belongs to, and `T` is a digit between 0 and 9 indicating which data set it corresponds to. Each one of these should contain D space separated numbers in one line. If some file doesn't follow this format then the corresponding solution is considered incorrect.

IF you don't have the solution to one of the groups you don't have to include it in the zip file submission.