Homework 1

2-AIN-150, Winter 2023

Deadline: 9.10.2023, 23:59, via Google classroom

Before you start solving the homework, please read the general instruction at the end of the document. Submitted solutions should be your own. Do not copy and do not tried to find solution in literature or over the internet.

Model for data with few nonzero attributes

Consider linear regression, with m attributes and n training samples, but where most elements of matrix X are zeroes and only few elements in each rows are nonzero.

In these situation we usually go for stochastic gradient descent solution ($X^T X$ matrix would be too big). Your task is to implement stochastic gradient descent with better time complexity than $O(mn)$ (or better than $O(m)$ for one training sample). Then use this to calculate linear regression coefficients.

In other words: Given n, m, X, $\vec{y}$, where X is almost everywhere zero, train linear regression effieciently.

There is one training input in supplemented package. Your solution should have mean quadratic error approximatelly 1.05. (if your error is around 2330 then you are missing one crutial thing).

Your program can use any libraries for basic math, matrix operation, matrix inversion, solving systems of linear equations, and calcualating numerical or symbolic derivatives. You are forbidden to use scikit-learn library and its equavalents (basically anything solving the homework in one liner).

Python instructions There is `template.py` in package. You should fill out the function `fit(n, m, X, y)`. Program can be runned using `python3 template.py <input file>`.

Instructions for other languages Fill out similar function as in Python. Input file contains numbers n, m on the first line. Then it contains n lines with training samples. Each sample is on one line and it has following format: First it has number k – the number of nonzero elements. Then it contains k pairs p, v where p is position of nonzero element and v is its value. The last number is y – expected output for the sample.

General instructions

You should submit homework via Google classroom.

Ideally submit your homeworks in Python (fill out required functionality from assignment). You can use different language if you really want, but you need also to add auxiliary functionality like reading input and output. But your solution should be runnable under Linux using easily accessible open source software.