Group Assignment 1

Incremental Learning
on Large Data

Learning objective:

Estimating a feed-forward neural network on data that is larger than what can be held in a typical computer's RAM.

Data:

All data is contained in *pricing.csv*. The data is copyrighted and confidential.

One line in the data represents a product selling on the company's e-commerce website. If a product goes out-of-stock and then becomes in-stock again a new line is created in the data for that product

- sku: stock keeping unit
- price: the price of the product on the website
- quantity: total quantity sold
- order: identifies the n^{th} time the product was in-stock and selling on the website
- duration: how long the product appeared on the site before going out-of-stock
- category: product category

Note 1: All categorical variables are integer encoded (not necessarily consecutively). All numeric variables are divided by a constant.

Variables

Input variables

- sku
- price
- order
- duration
- category

Response variable:

- quantity

Deliverables:

- A feed forward neural network that predicts quantity sold. The neural network should have three hidden layers, with sigmoid activation. The model should be learned incrementally: read in one record, update the model, repeat. Among other things, the better the model in terms of R^2 on pricing_test.csv set the higher your grade:

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

- Do not use pricing_test.csv for training or tuning
- A learning curve: a plot with on the x-axis the number of instances learned, and on the y-axis the moving average of the MSE.
- A plot showing variable importances
- Multiple partial dependence plots
- Information on RAM usage
- Information on training time

Format:

Code (.py file) and Presentation (pdf or PowerPoint)