Spring 2020
CS 397 | Machine Learning Implementation Techniques
Assignment 1 | Regression

Topics
The assignment will cover implementing a regression model, including:

1. Support different settings (Features and Exponents) in order to determine the best predictive model for the data.
2. Computing the model's "cost"
3. Adjustable learning rate.
4. Data mean normalization.
5. Adjust features and exponents to get the best possible prediction.

Goal
The goal of this assignment is to implement a supervised learning regression model, which will be used to predict the prices of a house per square meter.

Description

For this assignment three main code resources are provided:

Regression.h

This file contains the public interface of the class that need to be implemented:

- `Constructor:` Regression constructor that where all the data to execute the learning algorithm is provided.
- `Predict:` Given a single or multiple input datapoint(s), computes the value(s) that the regression would predict for that input. Two version are given for convenience, so that testing a full dataset can happen in one single call. The input data given to predict may not be part of the training dataset.
- `Cost:` Given an output (probably resulted from `Predict`) and a target value or real results of a dataset, it will compute the cost of the algorithm. This value will indicate how well the algorithm is learning and should most likely decrease the returned value after each iteration.
- `Iteration:` Function where the actual learning happens. It will adjust the theta value of each feature based on gradient descent for the model to fit the training data provided in the constructor. It will return false when the update is smaller than the specified one.

The non-public interface can (and should) be extended by the student in order to complete the assignment.

DatasetCreator.h/.cpp

These files should not be modified by the student. They contain a helper functions to easily generate datasets for the various tests.

RegressionTests.cpp

A sequence of tests using gtest.lib that test different features of the Regression. The first part of the assignment consists on passing those tests without modifying them. During the second part the student will fill the empty `HousePriceDeduction` and `LifeExpectancyData` tests to evaluate the implemented algorithm and they will create the best possible prediction for the RealStatePrices.csv and LifeExpectancyData.csv dataset. Only the code for this test should be modified on this file.

Assignment Submission

The submission should include 4 files: **Regression.h/.cpp**, **RegressionTests.cpp** and **README.txt**. The first three files will include the code with the student´s implementation and the tests (including the ones that were provided). The README.txt file will have the following content:

- Define the features of your model
  - What exponent value to use for each feature? What are the theta values for each feature?
  - Write down the function the model uses to predict. Explain how you interpret the function you got.
- Learning process
  - How many iterations does it need to converge?
  - Why are you using the learning rate you are using?

- What if you divide the dataset into training and test sets? What is the result in the test set? Overfitting? Underfitting?

Grading Rubrics

Regression implementation: 70%

- Features and exponents: 10%
- Cost function: 15%
- Learning rate: 10%
- Iteration: 20%
- Normalization: 15%

Best prediction models: 30%

- Explanation of prediction: 20%
- Prediction model: 10%