CS 397 | Machine Learning Implementation Techniques

Lab 1 | Regression

Goal

The goal of this lab is to implement a supervised learning regression model (both linear and logistic) to predict the outcome in a course based on the hours the student has studied.

Description

For this lab the only needed tool is an Excel spreadsheet. A template with the input data is provided along with some tables to fill out.

Data

The data that can be found on the spreadsheet contains the data of the hours a student invested studying for a course and the final grade for that course.

Hours	0.5	0.75	1	1.25	1.5	1.75	1.75	2	2.25	2.5	2.75	3	3.25	3.5	4	4.25	4.5	4.75	5	5.5
Grade	35	45	35	50	64	68	72	55	78	61	75	53	81	65	85	90	92	79	95	89

Linear Regression

First sheet on the Excel file is meant for the student to implement a linear regression that will fine the line that best predicts the course grade based on the input (hours studied). In order to get that line some parameters are provided:

- Hypothesis: Since we are looking for a line we will have two θ values to learn, the slope (θ_1) and the intercept (θ_0). The equation we are trying to fit to the data is $h_\theta(x) = \theta_0 + \theta_1 x$.

- Learning rate: the learning rate is set to 0.1 by default, but that can be modified once the algorithm has been implemented.

Apart from the parameters there is a table that needs to be filled out with the actual regression iterations. That table has the following three columns and computations that needs to be completed:

- PREDICTION#: Value of the hypothesis for each input in the training set for this iterations.
- DIFFERENCE: Difference (subtraction) between the predicted output and the target output.
- DIFFERENCE SQUARED: The difference previously computed squared.
- COST (LEAST SQUARES): Cost of the prediction for that specific iteration.
- COST DERIVATIVES: Each θ will compute its corresponding cost derivative so that the slope and the intercept can be updated for next iterations.

Once the first two iterations are computed, extend it to 20 iterations and graph the input dataset and the last prediction line.

Logistic Regression

The goal of this regression is similar to the one linear regression, but the input data will need to be transformed to get better input for a logistic regression.

Otherwise the main two factors to start learning are provided (hypothesis and learning rate), so the only part that differs is the values computed in each iteration:

- EQ. OUTPUT: The output of each input after applying the boundary equation for this iteration.

- PREDICTION#: Value of the hypothesis for each input in the training set for this iteration.
- ERROR: Result of the error function for each input.
- DIFFERENCE: Difference (subtraction) between the predicted output and the target output.
- COST: Cost of the prediction for that specific iteration.
- COST DERIVATIVES: Each θ will compute its corresponding cost derivative so that the slope and the intercept can be updated for next iterations.

Once the first two iterations are computed, extend it to 20 iterations and graph the input dataset and the last prediction line.

Excel functions

The following Excel functions will be helpful in order to complete the lab:

- SUM(A2:A10): Adds the values in cells A2:10.
- SUMPRODUCT(C2:C10,D2:D5): Sum of the products of corresponding ranges or arrays. The default operation is multiplication, but addition, subtraction, and division are also possible.
- LOG(A2): Logarithm of a number (A2) to the base you specify (10 by default).
- IF(A2 > 30, 10, 5): Logical comparisons between a value and what you expect (A2 > 30). The first result is if your comparison is True (10 in example), the second if your comparison is False (5 in the example).

Submission

The final .xlsx file needs to be submitted with the following name **cs397_login_lab1.xlsx**. Deadline for the submission is the **04/16/2021**. The grade will just be a **pass** or a **not pass**.