Foundations of Data Mining (2IMM20)
Homework Assignment 2 (HW2)
Linear Regression

Decebal Mocanu

Release date: October 3, 2018
Deadline: October 17, 2018
Feedback: October 24, 2018

Preamble:

This homework can be solved using pen-and-paper, or your preferred programming language (e.g. Python, MATLAB). Please submit your solutions as a report in PDF format through Canvas, in groups of two (recommended) or three (tolerated). State clearly in the report the group number and the group members.

Problem:

In this homework you are requested to perform linear regression on the data collected from a Combined Cycle Power Plant. The data can be found in Canvas in the file “HW2data.xlsx”. It contains two columns. The first column represents the *Temperature* (the input data - X) on the Celsius scale and the second column represents the *Net hourly electrical energy output* (the output data – Y) in MW. Each row represents a data point. It is not necessary for solving this homework, but if you would like to know more about these type of data you can look over [1].

I. Linear regression (10 points). Please solve the following subproblems and give clear explanations for your solutions:

1) *Data acquisition* (1 point). Each group has to perform linear regression on 5 data points (not on all data points from the file). Assuming that your group number is *n* then the data points corresponding to your group starts with row number *(n-1)*5+2* and ends with the row number *n*5+1* (e.g. Group 1 has the data from row 2 until row 6, Group 2 has the data from row 7 until row 11, and so on). Put your data in a table in the report.

2) *Data transformation* (1 point). Perform a transformation of your acquired data. Please feel free to use any type of transformation you prefer (e.g. Min-Max scaling, normalize data to zero mean and unit standard deviation). Report the formula that you used for the transformation and the transformed data in the report. Justify your choice.

3) *Linear Regression with Least Squares* (2 points). Perform linear regression on your data using least squares (closed form solution). Print the obtained regression parameters. Plot the regression line and the data points.

4) *Linear Regression with Gradient Descent - loss function* (1 point). Choose a suitable loss function to perform linear regression with gradient descent. Motivate your choice. Perform the partial derivatives of the loss function with respect to each of the regression parameters.

5) *Linear Regression with Gradient Descent - first iteration* (1 point). Start from a random choice of the regression parameters (print the choice in the report) and perform the first iteration of gradient descent on your data. Report the chosen learning rate and the computed values. Plot the regression line and the data points.

6) *Linear Regression with Gradient Descent - second iteration* (1 point). Perform the second iteration of gradient descent. Report the computed values. Plot the regression line and the data points.

7) *Linear Regression with Gradient Descent - third iteration* (1 point). Perform the third iteration of gradient descent. Report the computed values. Plot the regression line and the data points.

8) *Linear Regression with Gradient Descent - last iteration* (1 point). Continue performing gradient descent for an arbitrary number of iterations until the regression line fits well the data. Plot the values of the loss function for all iterations. Report the regression parameters and plot the regression line that you obtained after the last iteration.

9) *Discussion* (1 point). Compare the performance of Linear Regression with Least Squares and Linear Regression with Gradient Descent. Discuss the similarities and the differences between them.

II. Peer Review paragraph (0 points)

Finally, each group member must write a single paragraph outlining their opinion on the work distribution within the group. Did every group member contribute equally? Did you split up tasks in a fair manner, or jointly worked through the exercises. Do you think that some members of your group deserve a different grade from others?

References:

[1] Pınar Tüfekci, *Prediction of full load electrical power output of a base load operated combined cycle power plant using machine learning methods*, International Journal of Electrical Power & Energy Systems, Volume 60, September 2014, Pages 126-140, ISSN 0142-0615

Success!