Applied Machine Learning Lab: Unit 1, Lab 2

Date: 21st Jan 2015

Time: 3:30 pm to 5:30 pm IST

Problem 1: Price Predictor for Mobile Phones

The goal of this lab experiment is to help the student learn how to build and use linear regression based classifier for practical problems.

You are provided with a dataset that has samples of key specifications of mobile phones along with price of the phone, which is the output we would like to predict using our classifier (target attribute). We will use the following features from the dataset:

- Number of cores (1, 2, 4, 8)
- Clock (in GHz)
- Display Size (in inches)
- Internal memory (in GB)
- RAM (in GB)
- Primary_camera_mp (in mega pixels)
- Price (in Indian Rupees)

The dataset file is self-explanatory. In this experiment, you will build a linear regression based machine learning system using normal equation method taught in the class. The model should be developed to predict the price of the phone as a real valued number, given the input vector of features as explained above.

The high level steps to do this are as follows:

1. Read the dataset from the input file (which is in CSV format) and set up the X_{train} matrix and Y_{train} vector. Also set up X_{test} matrix and Y_{test} vector for testing. Use 80% of the given dataset examples for training and 20% for testing.
2. Compute the pseudo inverse using the numpy library available in Python. The function to be used is: numpy.linalg.pinv

   See: http://docs.scipy.org/doc/numpy/reference/generated/numpy.linalg.pinv.html
3. Determine the model parameter matrix theta as the product of pseudo inverse of X and Y vector
4. Use the model generated as above to perform prediction. Let the predicted values for the test dataset be $Y_{predicted}$. Compute the out of sample error as explained in the class and report the results.

Deliverables:

1. Source code of all your py modules
2. A report that describes or mentions:
    a. Performance numbers and your comments on the same. Did you get a reasonable performance (say > 70%) or bad? What are the possible reasons behind the results you obtained?
    b. If you want to improve on these results what would you do?
    c. If you are to do this using a decision tree, how would you go about?
    d. What would be the pros and cons between these two approaches?

Please zip all deliverables as above and mail to course.aml.2015@gmail.com

Best wishes from your faculty ☺