Homework 4[1]

Due by 11:59pm 11/29/2016

Objective

This assignment is an exercise in linear regression. It has two parts. The first part is a single variable exercise. The second is a more complex (multivariate) problem that requires you to do a little more exploration.

Part 1

Write a python program called **lin_regr.py** that takes in two arguments: a filename for training data and a file name for test data. The format of the file is: one example per row. Each example is an (**x**,y) pair where **x** is a vector of features (though in part 1, x is just a scalar) and y is the expected answer. The examples will be given in a .csv format. In the repository, you will find a pair of sample train/test files for part1. Note that for grading purposes, the TA may use a different pair of train/test files.

Your program should do the following:

- Train phase:
  - Read in the training examples from the specified file.
  - Apply gradient descent (see Lecture 17) on the training examples to learn a line that fits through the examples.
  - For every few iterations, print out:
    - The current model: ($w_0 + w_1x_1 + \ldots + w_nx_n$) before this iteration's update
    - The average squared error over the training set using the current line
  - After convergence, print out the parameters of the trained model ($w_0, \ldots w_n$)
- Test phase:
  - Read in the test examples from the specified file.
  - For each test case, apply your learned line on the input x to get a prediction. Compare your prediction against the given answer y (i.e., compute the squared error between them).
  - Your program should print out:
    - The *average* squared error of your trained line on the test cases.

Part 2

Apply your program from Part 1 to a multivariate problem. This dataset (a modified version from what was collected by Cortez and Silva (2008)) is about predicting a group of Portuguese high school students' 3rd year grades based on the following feature set:

---

[1] Shared Google Directory

- student's age (numeric: from 15 to 22)
- mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
- father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
- home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
- weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
- number of past class failures (numeric: n if 1<=n<3, else 4)
- quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
- free time after school (numeric: from 1 - very low to 5 - very high)
- going out with friends (numeric: from 1 - very low to 5 - very high)
- workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
- weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
- current health status (numeric: from 1 - very bad to 5 - very good)
- number of school absences (numeric: from 0 to 93)

The goal is to predict the student's 3rd year grade.

- final grade (numeric: from 0 to 20)

Comments

- Although the raw features about each student is fixed, you do not need to use them "as is" -- you can choose to use fewer features (maybe because you find that some of the features are not very useful on their own) or add more features (e.g., by combining some of the given features in some way). Even for the features that you do decide to keep, you might choose to normalize them so that they all have the same range.
- We do not know whether a linear model is a good choice for this problem. Although you should explore different ways of representing the features, it is not expected that your trained model should predict the answer well. If none of the feature representations you tried seems to be good at the prediction problem, discuss your findings in the report. Try to hypothesize what the problem(s) might be.

Experimental Setup

- Unlike Part1, here, you are given the full dataset. So it's up to you to do some preprocessing to get your train/test files. You should:
    - permute the examples (as given, they are sorted by some features first, I think)
    - Split the dataset set into two parts: 80% training 20% test. (Or, if you know how to setup a 5-fold cross validation, you can do that instead)

What to commit in addition to source code

- A README telling the TA how to run your program, any problems he should be aware of, and, if you used any additional tools or resources, give proper citations for them. If you've discussed the problem with other people, let us know the extent of your collaboration.
- A short Write-Up; it should discuss the following issues:
    - Part 1:
        - How did you choose the value for the learning rate (alpha)?
        - What is your "convergence" criterion? Justify your choice.
    - Part 2:
        - Briefly present your learning rate choice and your convergence criterion for this part.
        - What is the final feature set you decided on?
        - How did you decide to choose these features? What other options have you experimented?
        - How did your trained linear model perform on the test set?

Grading

1. A serious attempt, there is a readme and a cursory write-up.

2. There are some problems with **lin_regr.py** overall, but it seems to work for the univariate case (e.g., finds the right answer for Part 1). There is a readme and a write up.

3. The program **lin_regr.py** works for multivariate problems as well as the univariate case. For Part 2, you've applied **lin_regr.py** to the given feature vector (i.e., "as is"). There is a readme and a write-up. The outcome of Part 2 is discussed and analyzed in the write-up.

.

4. The program **lin_regr.py** works (for both the univariate and the multivariate cases). In Part 2, you've explored at least three different organization of the feature vector. There is a readme and write up. The write up offers some justification for each feature vector configuration that you experimented with. The outcomes of different feature vector configurations are discussed and analyzed in the write-up.

5. Similar to 4, but experimented with more feature vectors in Part 2. Insightful write-up. Has a readme for the grader.

Reference

P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.