Week 5 Experiments: Linear Regression and Logistic Regression

Introduction to NLP (Sep. 2021)

Student Name: Student ID:

Lectured by: Shangsong Liang
Mohamed bin Zayed University of Artificial Intelligence
Do NOT Distributed This Document and the Associated Datasets

1 Experiment One: Linear Regression

In this experiment, you will investigate multivariate linear regression using Gradient Descent and Stochastic Gradient Descent. [1] You will also examine the relationship between the cost function, the convergence of gradient descent, overfitting problem, and the learning rate.

Download the file "dataForTrainingLinear.txt" in the attached files called "experiment 5" . This is a training dataset of apartment prices in Masdar city, where there are 50 training instances, one line per one instance, formatted in three columns separated with each other by a whitespace. The data in the first and the second columns are sizes of the apartments in square meters and the distances to the MBZUAI university in kilo-meters, respectively, while the data in the third are the corresponding prices in billion Dirhams. Please build a multivariate linear regression model with the training instances by script in any programming languages to predict the prices of the apartments. For evaluation purpose, please also download the file "dataForTestingLinear.txt" (the same format as that in the file of training data) in the same folder.

(a) How many parameters do you use to tune this linear regression model? Please use Gradient Descent to obtain the optimal parameters. Before you train the model, please set the number of iterations to be 1500000, the learning rate to 0.00015, the initial values of all the parameters to 0.0. During training, at every 100000 iterations, i.e., 100000 , 200000,···, 1500000, report the current training error and the testing error in a figure (you can draw it by hands or by any software). What can you find in the plots? Please analyze the plots.

(b) Now, you change the learning rate to a number of different values, for instance, to 0.0002 (you may also change the number of iterations as well) and then train the model again. What can you find? Please conclude your findings.

(c) Now, we turn to use other optimization methods to get the optimal parameters. Can you use Stochastic Gradient Descent to get the optimal parameters? Plots the training error and the testing error at each K-step iterations (the size of K is set by yourself). Can you analyze the plots and make comparisons to those findings in Exercise 1?

2 Experiment Two: Logistic Regression

You will implement a logistic regression classifier and apply it to a two-class classification problem. To get started, download the two datasets, "dataForTrainingLogistic.txt" and "dataForTestingLogistic.txt" from the folder called "Homework 3". In both of these two datasets, each instance is put per line with the first to the six columns being the features of the instance and the last column being the ground-truth

[1] To see what are "Gradient Descent" and "Stochastic Gradient Descent", please refer to the slides.

label of the category (either "1" or "0") that the instance should be classified into. Each column per line is separated by a whitespace.

(a) In logistic regression, our goal is to learn a set of parameters by maximizing the conditional log likelihood of the data. Assuming you are given a dataset with n training examples and p features, write down a formula for the conditional log likelihood of the training data in terms of the the class labels $y^{(i)}$, the features $x_1^{(i)}, \ldots, x_p^{(i)}$, and the parameters $w_0, w_1, \ldots, w_p$, where the superscript (i) denotes the sample index. This will be your objective function for gradient ascent.

(b) Compute the partial derivative of the objective function with respect to w_0 and with respect to an arbitrary w_j, i.e. derive $\partial f / \partial w_0$ and $\partial f / \partial w_j$, where f is the objective that you provided above. Please show all derivatives can be written in a finite sum form.

(c) Train your logistic regression classifier on the data provided in the training dataset "dataForTrainingLogistic.txt". How do you design and train your logistic regression classifier? What are your optimal estimated parameters in your logistic regression classifier? Use your estimated parameters to calculate predicted labels for the data in the testing dataset "dataForTestingLogistic.txt" (Do not use the label information (the last column in the file) for testing).

(d) Report the number of misclassified examples in the testing dataset.

(e) Plot the value of the objective function on each iteration of stochastic gradient ascent, with the iteration number on the horizontal axis and the objective value on the vertical axis. Make sure to include axis labels and a title for your plot. Report the number of iterations that are required for the algorithm to converge.

(f) Next, you will evaluate how the training and test error change as the training set size increases. For each value of k in the set $\{10, 20, 30, \ldots, 380, 390, 400\}$, first choose a random subset of the training data of size k. Then re-train your logistic regression classifier using the k random subset of the training data you just chose, and use the estimated parameters to calculate the number of misclassified examples on both the current training set (k random instances) and on the original test set "dataForTestingLogistic.txt". Finally, generate a plot with two lines: in blue, plot the value of the training error against k, and in red, plot the value of the test error against k, where the error should be on the vertical axis and training set size should be on the horizontal axis. Make sure to include a legend in your plot to label the two lines. Describe what happens to the training and test error as the training set size increases, and provide an explanation for why this behavior occurs.