CS436/CS580L Homework 2

Fall 2018, Xi Peng

Instruction:

(a) Release: Oct 25, 2018.
(b) Due: Nov 6, 2018.

You need to upload a report to answer all the questions. The source codes should be directly pasted into the report as the appendix.

Question 1.

In this problem you will replicate and expand upon the prototype data analysis problem in lecture slides "L5&6_Linear Regression_Gradiant Descent_Polynomial Regression 2.pdf". In this case someone has generated data by sampling the sinusoidal function $f(x) = \sin(2\pi x), x \in [0,1]$. However, pretend not to know this fact and try to fit different models to discover the relationship between x and y.

1. Generate a dataset $\mathcal{D}$ consisting of $N = 200$ by evaluating the function $f(x)$ at N uniformly spaced points in x and adding iid white Gaussian noise with deviation $\sigma = 0.1$. I.e., $x_i = \frac{i-1}{N-1}$ $and\ y_i = f(x_i) + \epsilon_i, \epsilon_i \sim \mathcal{N}(0, \sigma^2), i = 1, \cdots, N, and\ \mathcal{D}_N = \{(x_{i,} y_i)\}_{i=1}^{N}$. (Plot all the points and the sinusoidal function in the same figure; 10 points.)
2. Repeat the polynomial fitting experiment by selecting a random subset of $\mathcal{D}$ of the appropriate size to train the model. Report estimate coefficient (model parameters) θ, MSEs, both as a function of the chosen training set size N_{train} and the regularization λ. (Training set size changes from 10, 50, 100, 150, 200; Choose at least 5 different λ, including $\lambda = 0$; Report the results as requested; 30 points)
   1) As we explained in the class, you need to plot 5 different figures (horizontal axis is the number of the training data points, vertical axis is the MSE).
   2) Each figure corresponds to a λ.
   3) Suppose you will try K different polynomials, you will have K curves in each of the 5 figures.
   4) You can decide K, a possible combination could be $\{0, 1, 2, 3, 9\}$.
   5) Totally you will train 5x5xK different models. Report ALL the model parameters in 5 different tables. Each table corresponds to a figure.
3. Select a test dataset $\mathcal{D}_{test}$ as $N_{test} = 199$ uniformly spaced points $x_{i,test} = x_{i,train} + \frac{1}{2\cdot(N_{train}-1)}, i = 1, \cdots, N_{train} - 1, y_{i,test} = f(x_{i,test})$, and plot the comparison of training and test MSEs for the fitted models as a function of N_{train} and λ. Choose at lease 10 different values of both parameters for these graphs.
   (Pick out 10 trained models trained in 2 to test; Report the results as requested; 30 points)
   1) You will plot 5 different figures similar to Q2.
   2) For every figure, you need to pick out at least 2 out of K models trained in Q2 to plot the MSE curve.
   3) The testing set has a fixed number (199) of data points. No matter how many training data points are used to train the model, all the testing data points will be tested.
   4) You can put the corresponding MSE curves on the training data in the same figure.
4. Report your analysis and conclusion.

(For instance, which one has the best performance and why? Which one has the worst performance and why? How about the balance of variance and bias? 10 points)

Question 2.

In this problem you will implement gradient descent algorithm from scratch. Use MATLAB build-in dataset for binary classification problem by running “load ionosphere”. You need to shuffle the dataset first. Use the first 301 data points as the training set and the following 50 data points as the testing set. Suppose we choose logistic regression as the classification model.

1. Implement gradient descent algorithm without regularization. (10 points)
2. Implement stochastic gradient descent algorithm without regularization. (10 points)
3. Implement gradient descent algorithm WITH regularization. (10 points)

Notice:

a) Use functions as demonstrated in the class, e.g., LR_GD(), LR_SGD(), LR_Test(), Sigmoid(), so some functions can be shared in the three questions.
b) For all the three questions, you need to try at least 3 different learning rates, or design the learning rate as a function of iterations. Plot the training and testing errors with respect of iterations. What do you observe by using different learning rate? What is the best learning rate? Explain how to find it?
c) For the third question, you also need to try at least 3 different weights for the regularization term. Plot the training and testing errors with respect of iterations. What do you observe by using different weights? What is the best weight? Explain how to find it?