CS-443 Machine Learning Project 1: J-D-S Team

Julie Camille Rosalie Giunta
274957

Samuel Chassot
270955

Daniel Filipe Nunes Silva
275197

2019 October

1 Introduction

The goal of this project is to apply machine learning methods learned in class on a real dataset. We take a strong interest in testing a lot of techniques and comparing their results. This comparison encourage us to tweak hyperparameters and check their effectiveness using cross-validation.

We do not use least_squares_SGD because we consider that it would provide us results really close to other methods we already us. Finally, we assess the following methods.

- least_squares
- least_squares_GD
- ridge_regression
- logistic_regression
- reg_logistic_regression

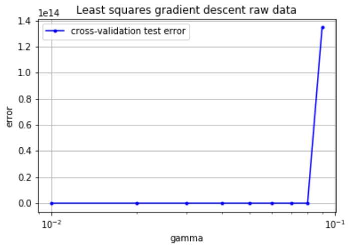
2 Least squares

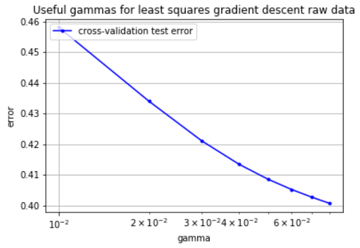
Samuel

3 Least squares gradient descent

To test the efficiency of *least_squares_GD* on our dataset, which we standardize, we use cross-validation with five sets to train the hyperparameter γ. In figure 3, you can observe how fast the cross-validation test error is growing when you change gamma from 0.08 to 0.09 but in order to have a more useful representation of the progression of the error, you can take a look at figure 3, which omits the error corresponding to $\gamma = 0.09$. For the initial weights chosen, they are first all initialized to 0.5. We then modify the initial weights it to be closer to the final weights we got. Surprisingly, 0.4 gives a better accuracy result on AICrowd (69.7%) than 0.0 (69.3%) which is closer to the final weights in general and ouputs a smaller loss. For the number of iterations, 200 and 1000 converge almost to the same loss so to be more efficient, we choose 200. Our best submission with *least_squares_GD* has an accuracy of 69.7% and has, as hyperparameters,

- max_iters = 200
- k_fold = 5
- initial_weights = np.array([0.4 for i in range(tX_stdrzed.shape[1])])
- gamma = 0.08


Least squares gradient descent raw data


Useful gammas for least squares gradient descent raw data

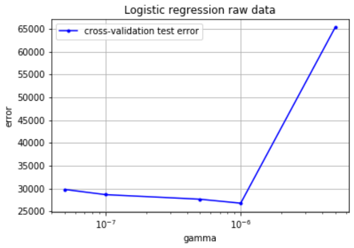
4 Ridge regression

Daniel

5 Logistic regression

We also used cross-validation to test the efficiency of *logistic_regression* on our standardized dataset. We use five sets to set the hyperparameter γ. For the initial weights chosen, they are all initialized to 0.5 since $0.0, 0.1, ..., 0.6$ do not change the loss a lot but worsen the accuracy on AICrowd. Our best submission with *logistic_regression* had an accuracy of 73.9% and had

- max_iters = 1000
- k_fold = 5
- initial_weights = np.array([0.5 for i in range(tX_stdrzed.shape[1])])
- gamma = 1e-06


Logistic regression raw data

6 Regularized logistic regression

Samuel