CS725 Fundamentals of ML Assignment

9th September 2021

Gourav Tilwankar (203070051)
Saketh VNS (20307R001)
Milind Dushyant Patil (203170020)
Samay Pritam Singh (20307R002)
B Vasant (203070085)
Instructor: Prof Dr Preethi Jyothi, Department of Computer Science and Engineering, IIT Bombay

---

Within submission/report.pdf, report your best MSE losses on the development set obtained using the analytical solution and the gradient descent solution.

RMSE Losses

	Analytical Losses	Gradient Descent Losses (Best)
Development Set Losses	198.3960229253915	203.95386631925086
Training Set Losses	162.0483098088209	170.36674574729219

MSE Losses

	Analytical Losses	Gradient Descent Losses (Best)
Development Set Losses	39360.981912612464	41597.17958657085
Training Set Losses	26259.6547118956	29024.8280565225

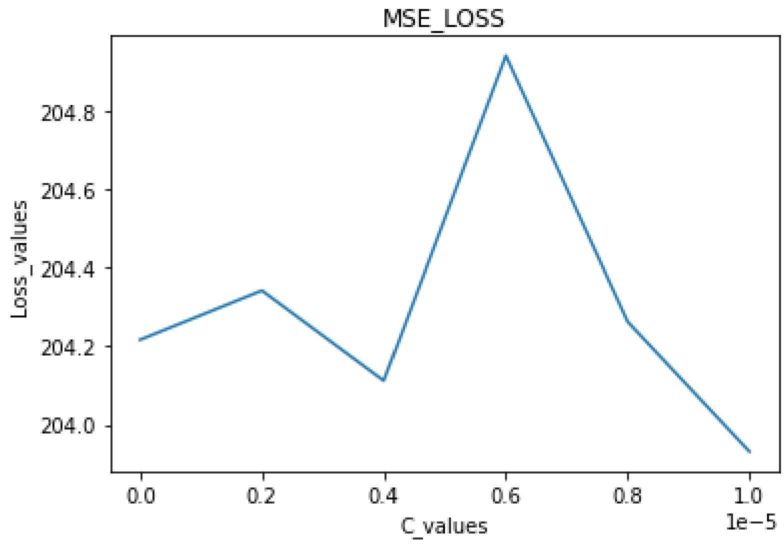
What stopping criterion did you use for convergence in gradient descent? Write this down in submission/report.pdf

We used a cap on the number of iterations as the stopping criterion, max_steps was set to 150000

Within submission/report.pdf, report MSE losses on dev.set instances with and without the use of early stopping.

	With early stopping	Without early stopping
Dev_loss	-	203.93115294994587

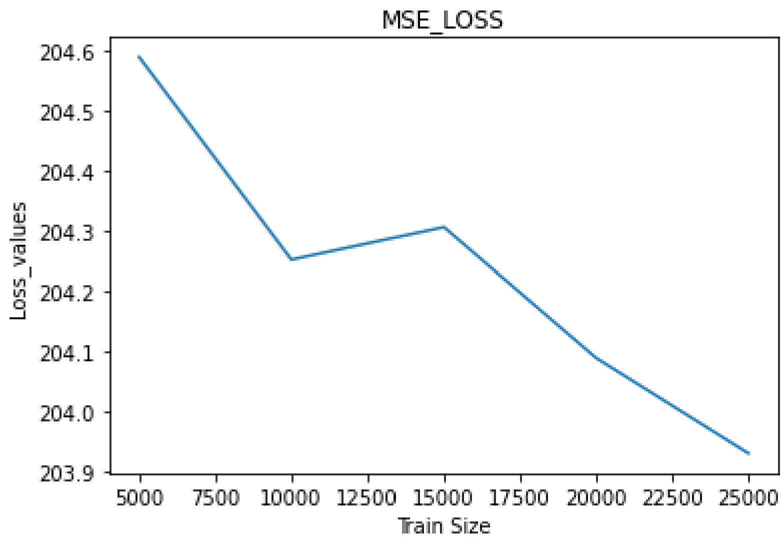
Setting C to 0 in do_gradient_descent yields the unregularized least squares solution. Plot MSE on the dev.set instances for different values of C (including C=0). Include this plot within submission/report.pdf.



Implement two different basis functions that will be applied to your input features with the L2-regularized model and optimized using gradient descent. Add your implementations of the basis functions within get_features. Describe your choice of basis functions in submission/report.pdf and report the MSE on the development set samples using both basis functions.

Basis functions are not completely implemented, but functions are created in LR.py

Create subsets of the training data containing the first 5000, 10000, 15000, 20000 instances in train.csv. Create a plot where the X-axis is the size of the training set (5000, 10000, 15000, 20000, full) and the Y-axis is the mean-square error on the dev.csv instances. A placeholder function plot_trainsize_losses has been defined for you to fill out for this question. You can use existing libraries (like matplotlib) to create this plot and add it to submission/report.pdf.



Of the fourteen features, which would you consider the most important and the least important for the frp prediction? Describe how you identified these two features in submission/report.pdf.

These are the model coefficients:

```
bias 16.571033181507342
latitude -14.428438506739482
longitude 5.098272399348363
brightness 122.15925210661204
scan 33.06030794353354
track 32.15601615612771
acq_time 13.483018892914123
satellite 0.8486946385489659
confidence 0.8537389373035555
bright_t31 37.86447696592324
daynight 30.080951283894656
month -0.73578107297859
year -0.31138437391649243
day -0.7339327076716736
```

Here, brightness has the highest weight, therefore we conclude that it is the most important feature for our ‘frp’ predictions.
Since, we have excluded two features namely, instrument and version du to their unique and single values, we conclude that it’s the least important feature

Describe your main innovations in submission/report.pdf.

-