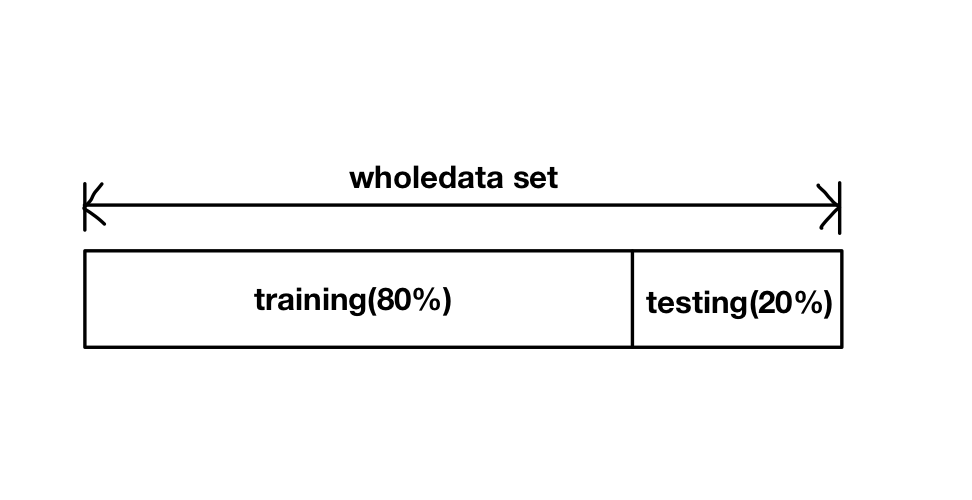
<HW1 Report> 2015310884 박소현

1. Data Reorganization & Plan

train.csv and test.csv should be eliminated header to be parsed. So, I made files **train\_nohead.csv** and **test\_nohead.csv** which have no header. Also, I made a file **wholedata.csv** combining train\_nohead.csv and test\_nohead.csv.

train\_nohead.csv is used for training and building the model.

wholedata.csv is used for predicting and calculating RMSE. I had to use the wholedata set instead of train set because *overfitting problem* can occur when only train set is used. So, test set also should be used to check model works well. In wholedata set, train data and test data ratio is approximately 8:2 because 332 data in train set and 82 data in test set are given.



2. Python Code Explanation

> linear.py

2: import numpy for using float() when parsing

3-4: import LabeledPoint, LinearRegressionWithSGD, LinearRegressionModel,RidgeRegressionWithSGD, RidgeRegressionModel, LassoWithSGD, LassoModel for training and modeling

7: sc = SparkContext() should be added. sc is not defined in example code, which

caused error.

12: LabeledPoint(values[6], values[0:6]) means 7th column is label and 1st~6th columns are features.

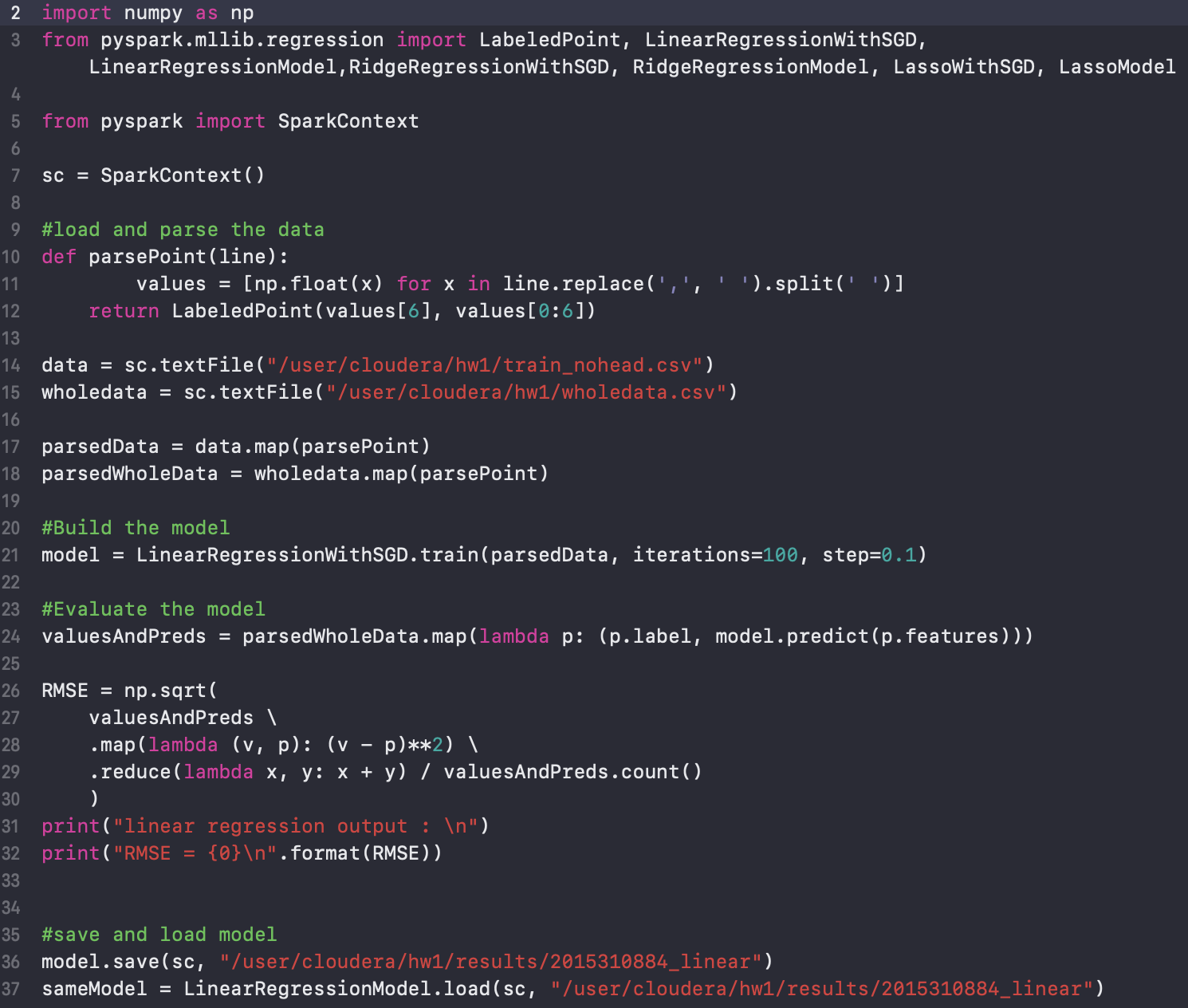
14-18: load and parse the data

21: build the model

24: predict on wholedata using the model

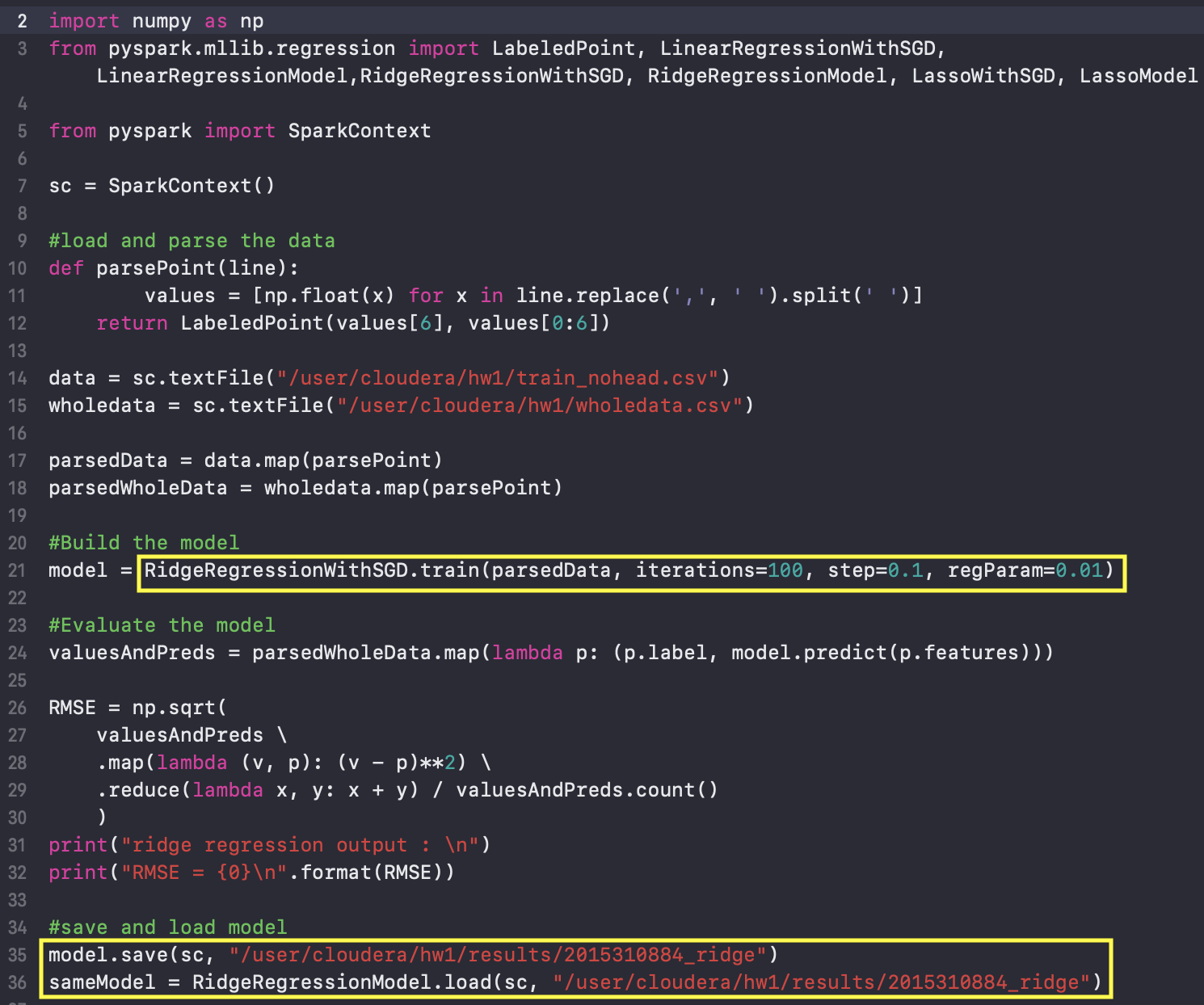
26-30: calculate RMSE

36-37: save and load model

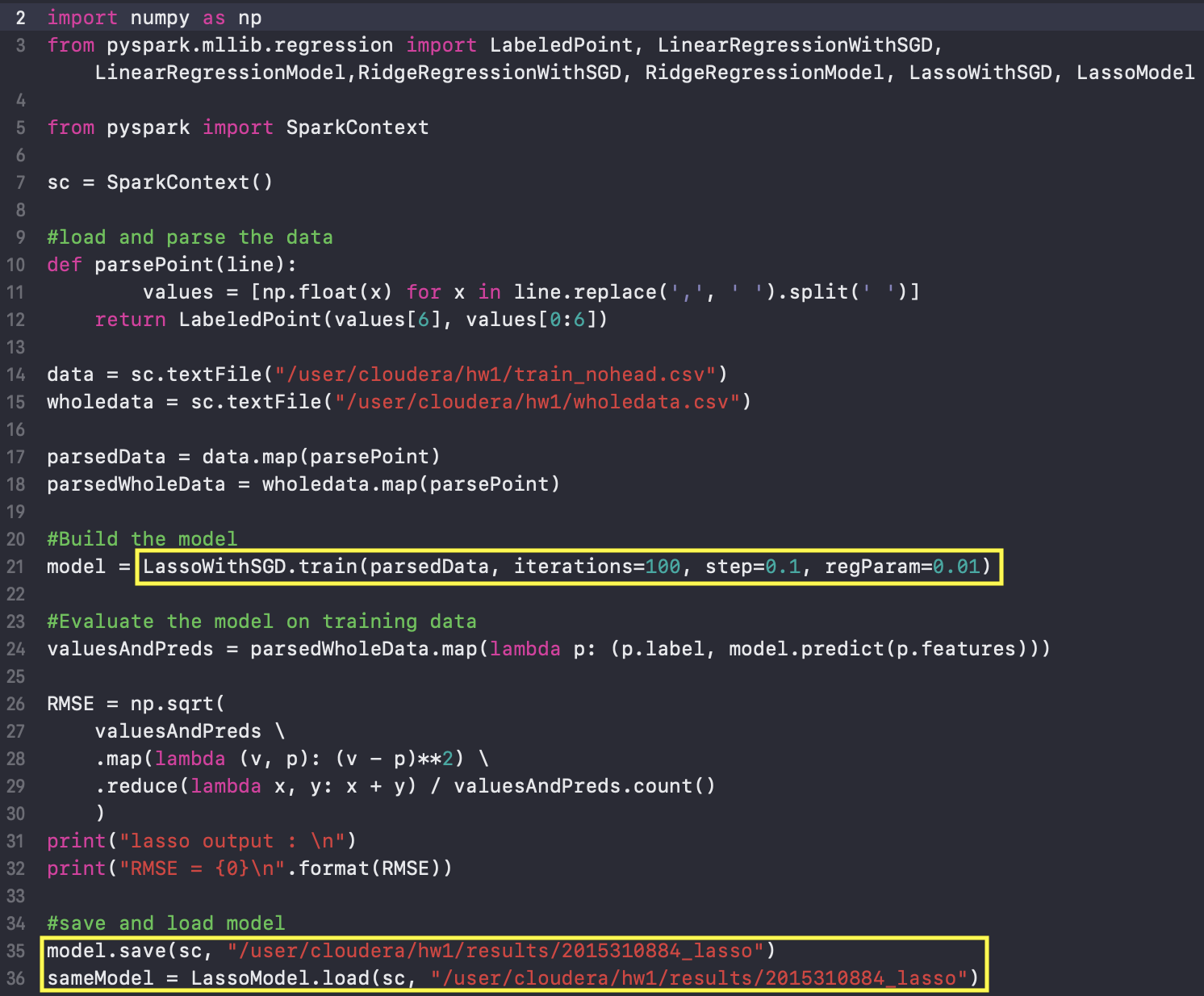


ridge.py and lasso.py are slightly different from above linear.py code in the part of building model and saving model.

>ridge.py



>lasso.py



3. Results

For each algorithm, the results come out.

Smaller value means more accurate. Three results show similar values.

