Assignment 03: Linear Regression, Learning Curves and Regularization

CSci 574: Machine Learning

Objectives

- Better understand Linear/Polynomial Regression using standard RMSE cost
- Get experience with what underfit, overfit and good fitting/generalizing models look like.
- Understand basics of using linear curves to tune underfit/overfit models correctly.
- Learn and use regularization to tune overfit models
  - Use both l-2 Ridge regularization and l-1 Lasso regularization
  - Get a feeling for the difference between these and where one or the other might be useful.
- Practice using `scikit-learn` pipelines to create more complex model workflows.
- Experince using grid search in `scikit-learn` to explore a parameter space such as tuning `alpha` parameters for regularization.

Description

In this assignment you have been given a set of data (in `data/assg-03-data.csv`) that has been generated using a secret polynomial function. The polynomial used to generate this data has a degree of at least 2, but no more than 10, and noise (mean 0, standard devaition 0.5) has been added to make it non-trivial to fit a good model and recover the parameters used to generate the data.

You will perform several tasks that will walk you through defining and fitting linear regression models to fit the noisy data you have been given. You will start by creating an obviously underfit model for the data and then a model that should be overfitting the data. You will create a function to display the learning curves that result from training these underfit and overfit models on the data.

From these initial models, you should get a feel for what performance a good performing model should be able to achieve, and what a good generalizing model should look like. You will then use Lasso and Ridge regularization to try and find models that fight the overfitting and obtain good generalization performance.

Overview and Setup

Assignment Tasks

You should read through and perform the tasks described in the given assignment notebook `notebooks/Assg-03-Regression-Learn` In this assignment there are 7 or 8 tasks to complete. For some tasks you need to write and complete a function in the `src/assg_tasks.py` file. These are the functions where you will create and fit the specified model for each of the tasks, and where you will calculate learning curve train and test errors for use in the assignment.

Most tasks have associated tests that will be performed on them. These can be run in the assignment iPython notebook, and/or can be run in VSCode using the test runner. You should ensure that your implementations in the `src/assg_tasks.py` file are passing tests for a task before moving on to the next task.

Assignment Submission

All work will be submitted and evaluated through GitHub classrooms for this assignment. You should work incrementally. Each task should be in a separate commit. So you should end up creating and pushing 7 or 8 commits (or more), at least one for each task, with only the work that complets that task in a commit. You should check the

GitHub Classroom autograder for each commit you push to ensure that your work is passing the expected tests for the assignment.

Additional Information