Quantitative Foundations Project 1

Linear Feature Engineering

Instructors: Linwei Wang

1 Overview

In this project, you will experiment with linear regression, overfitting, feature "engineering", and basic matrix operations.

You are given a dataset, consisting of 926 examples. Each example has 8 real-valued predictor attributes x ("regressors"), and a single real-valued dependent value y to be predicted. Your goal is to use this data to build a function that will predict values y for *new* data x. Specifically, we have more data that was generated from the same source. We will measure how well your predictor does on this new data.

2 Data

You are given a file, `traindata.txt`. This file contains 926 rows, each with 9 numbers, meaning it constitutes a 926x9 matrix. The first 8 columns contain the data that you will be predicting from, while the last column contain the data that you will predict.

As an example, to read this data in matlab, you might do:

```
traindata = importdata('traindata.txt');
X = traindata(:,1:8);
Y = traindata(:,9);
```

We also provide a file `testinputs.txt`. This contains 103 rows, each with 8 numbers. This is just like above, except that we do not provide the true output value. You might read this in with something like:

```
Xtest = importdata('testinputs.txt');
```

3 Task

Your goal in this project is to provide a text file with 103 numbers in it, consisting of your predictions for each of the 103 test inputs. Specifically, we will measure the mean-squared error of your predictions,

$$\frac{1}{103}\sum_{i=1}^{103}(y_i - y_i^*)^2$$

where y_i is your prediction, and y_i^* is the true output value.

4 Methods

To make these predictions, you should create features. That is, given and input $\mathbf{x}$ (of length 8), you should make a function which creates a (presumably longer) vector $\mathbf{x}'$. Then, you will fit a linear regression model of y to $\mathbf{x}'$.

5 Deliverables

1. Code. Please submit all code for this project, along with a very brief `README.txt` which explains to us how to use it. We should be able to run your code and regenerate your predicted values.
2. Report. Two page maximum, single column. Please be clear and concise. You should discuss the following topics:
   (a) At the top of your report, give two numbers:
      i. The training error $\frac{1}{926}\sum_{i=1}^{926}(y_i - y_i^*)^2$, where y_i is the predicted value, and y_i^* is the data value.
      ii. Your *prediction* for the test error $\frac{1}{103}\sum_{i=1}^{103}(y_i - y_i^*)^2$.
   (b) How did you choose what features to use?
   (c) How did you arrive at your prediction for the test error?
   (d) How did you deal with overfitting?
3. Presentation. Give a 8-minute presentation, describing your experiences, what you tried, what worked, what didn't work, what features you use, and how well you estimate your team did in terms of least-squares error.

6 Grading:

- Content: 50%
- Report: 20%
- Presentation 30% (including Q&A)