CS460G Programming Assignment 2 Report
Caleb Geyer

Part 1: Linear Regression with Gradiant Descent

In the first part of this assignment, we are using the gradiant descent technique on a set of data to predict the quality of wine given a list of 11 parameters:

```
features = ['fixed acidity', 'volatile acidity', 'citric acid', 'residual sugar',
'chlorides', 'free sulfur dioxide', 'total sulfur dioxide', 'density', 'pH',
'sulphates', 'alcohol']
```

We create a θ-value for each feature as well as an additional value. Using the normalized feature values as well as a chosen α-value (alpha), we can update the weights every iteration to lower the mean-squared error and get a better predicting function for the quality of the wine:

```
 # Update theta values
for j in range(len(X[0])):
  thetas[j] = thetas[j] - alpha * (1/m) * summation_regress(j, X, Y, thetas_copy)
```

Update θ values:

$$\theta_i=\theta_i-\alpha*(\frac{1}{m})\times\sum_{i=1}^{m}(h_\theta(x_i)-y_i)x_i$$

Hypothesis function:

$$h_\theta(x)=\theta_0+\sum_{i=1}^{n}\theta_i x_i$$

Results

Given each θ-value initialized to 0, an α-value of 0.001, and 1,000 iterations, the resulting mean-squared error recorded is as follows:

$$MSE(X,Y)=1.0620$$

And the final θ-values (weights) are as follows:

$$\theta_0=2.1386,\theta_1=1.0223,\theta_2=0.9228,\theta_3=0.9355,\theta_4=0.6677,\theta_5=0.6905,\theta_6=0.8174,\theta_7=0.7015,\theta_8=1.2536$$
$$\theta_9=1.2407,\theta_{10}=0.8267,\theta_{11}=1.0628$$

Additional Implementation Information

The implemented program uses a full-batch gradiant descent. As stated before, the α-value chosen was 0.001 and the feature values were normalized between 0 and 1.

Part 2: Polynomial Regression Using Basis Expansion

In the second part of the assignment, we are using polynomial regression and basis expansion to create a predicting function for two sets of synthetic data:

```
synthetic_data_files = ["data/synthetic-1.csv", "data/synthetic-2.csv"]
```

As before the data is normalized, but before that it is expanded using basis expansion.

Basis Expansion

Given an integer representing the order of the predicting polynomial function, we can create new features with data before the data is normalized. We will use the following hypothesis function during regression training:

$$h_\theta(x)=\theta_0+\theta_1 x+\theta_2 x^2+\ldots+\theta_n x^n$$

So given an order of 2, for example, we will create a new feature column in our data which will be equal the the X-values raised to the exponent 2, and so on for higher orders.

Once the data has been expanded, each column is normalized between 0 and 1.

Now that we have expanded the data, the steps for polynomial regression are the same as those from gradiant descent.

Results

Given each θ-value initialized to 0, an α-value of 0.05, and 10,000 iterations, the resulting mean-squared error and weights recorded are as follows:

synthetic-1.csv at order 2:

$$MSE(X,Y)=30.4053$$
$$\theta_0=-8.1140,\theta_1=7.3910,\theta_2=2.5582$$

synthetic-1.csv at order 3:

$$MSE(X,Y)=9.1575$$
$$\theta_0=1.5833,\theta_1=39.5670,\theta_2=3.2706,\theta_3=-50.7228$$

synthetic-1.csv at order 5:

$$MSE(X,Y)=8.5777$$
$$\theta_0=4.7397,\theta_1=35.8824,\theta_2=-3.4606,\theta_3=-22.6517,\theta_4=7.9792,\theta_5=-29.4471$$

synthetic-2.csv at order 2:

$$MSE(X,Y)=0.3276$$
$$\theta_0=0.4653,\theta_1=-0.1861,\theta_2=-0.7078$$

synthetic-2.csv at order 3:

$$MSE(X,Y)=0.3276$$
$$\theta_0=0.4436,\theta_1=-0.2366,\theta_2=-0.7040,\theta_3=0.0872$$

synthetic-2.csv at order 5:

$$MSE(X,Y)=0.3085$$
$$\theta_0=0.4915,\theta_1=-0.3890,\theta_2=-1.8544,\theta_3=0.5025,\theta_4=1.4715,\theta_5=-0.1818$$

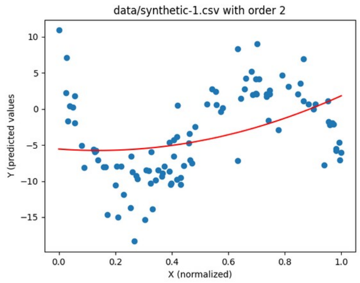
Additional Implementation Information

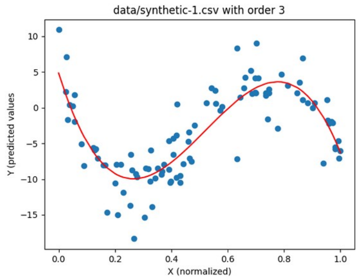
As in Part 1, the implemented program uses a full-batch gradiant descent. As stated before, the α-value chosen was 0.05 and the feature values were normalized between 0 and 1.

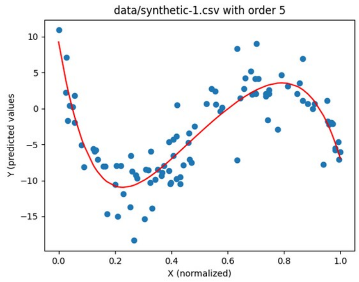
Part 3: Plot your regression lines

In each plot, the blue dots represent the datapoints and the red line represents the regression line. The corresponding dataset and order are labeled at the top:

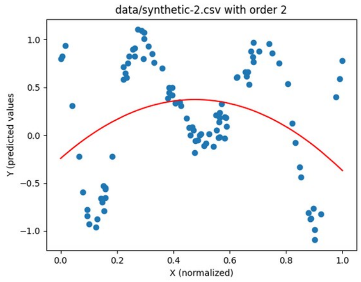
synthetic-1.csv

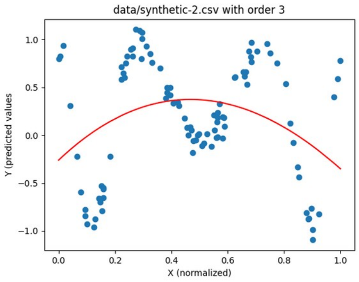


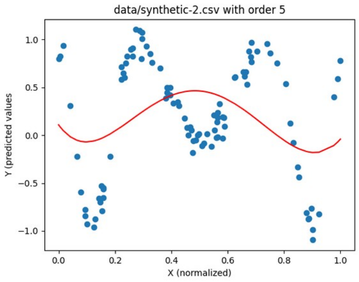




synthetic-2.csv


data/synthetic-2.csv with order 2
Y (predicted values
X (normalized)


data/synthetic-2.csv with order 3
Y (predicted values
X (normalized)


data/synthetic-2.csv with order 5
Y (predicted values
X (normalized)