Implement Regression with Gradient Descent

Yueqi Li
yueqi.li@uky.edu

March 6, 2022

I. Introduction

A regression model provides a function that describes the relationship between one or more independent variables and a dependent, or target variable. A regression analysis is the basis for many types of prediction and for determining the effects on target variables.Regression methods continue to be an area of active research.

Regression methods continue to be an area of active research, and also commonly used in industry, such as leveraged across an organization to determine the degree to which particular independent variables are influencing dependent variables. Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal, but are also a popular tool in machine learning.

II. Methods

In this project, we are given three data sets. We use liner regression model to predict the wine quality for wine quality data set, which contain eleven feature and a wine quality as a label. For synthetic data sets, we use polynomial regression model to describe it. We also implement Ridge Regression, which involve L2 norm regularization for the synthetic data sets.

i. Linear Regression

For linear regression,we want to use $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{11} x_{11}$ to describe data, where y is wine quality; x is features, β_0 is the intercept, and β_1 to $\beta_1 1$ is the coefficient or weight of the corresponding feature.

In our project, we can view the feature as 11x1560 matrix, and weight $t\theta$ as a 12x1 vector. We implement a design matrix, which we denote as X. In other word, we add a column 1 in front of feature matrix. Thus,the computation will be much easier. we denote weights as θ, wine quality vector as Y, learning rate as α.

First, we implement the hypothesis function:

$$h_\theta(x) = \theta_0 + \sum_{j=1}^{n} \theta_j x_j = X\theta \tag{1}$$

Then we implement the loss function:

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)} - y^{(i)}))^2 = \frac{1}{2m} \sum_{i=1}^{m} (X\theta - Y)^2 \tag{2}$$

We can take derivative of $J(\theta)$ respect to θ to get our gradient

$$\frac{\partial J}{\partial \theta} = \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)} - y^{(i)})) \cdot x^{(i)} = \frac{1}{m}(X^T X\theta - X^T Y) \tag{3}$$

Now, we can start to implement gradient decent:
First, we randomly initialize θ by using uniform methods from -1 to 1.

Since equation (3) calculate weight, based on it, we use following function to update θ where

$$\theta = \theta - \alpha \cdot \frac{1}{m}(X^T X\theta - X^T Y) \tag{4}$$

In terms of evaluation, we will use Mean Square Error:

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)} - y^{(i)}))^2 = \frac{1}{m}\sum_{i=1}^{m}(X\theta - Y)^2 \tag{5}$$

When we arrive certain amount of steps, we will stop training.The linear model will give us a set of parameter, that we can use for future wine quality prediction. We will also have a mean square error to evaluate the training error.

ii. *Polynomial Regression*

For Polynomial regression,we want to use $y = \theta_0 + \theta_1 x_1 + \theta_2 x^2 + \cdots + \theta_n x^n$ to describe data, where y is label; x is features, β_0 is the intercept, and θ_1 to θ_n is the coefficient or weight of the corresponding x polynomial.

In order to use polynomial regression, we need to expand the synthetics data set by taking power of x and add it as a feature of the data.

Then we can apply same methods as we mentioned in linear regression.

After training, we will have same number of parameter as number of polynomial, then we can use it to describe data, as well as Mean Square Error to describe the training error.

III. Result

In the linear regression, which is wine quality dataset, we use 1e-5 as learning rate, and we stop at 10000000 steps, we arrive 0.4332347 mean square error, and we have parameter as following:

θ_0	θ_1	θ_2	θ_3	θ_4	θ_5
-9.32631714e-2	4.63891272e-2	-9.21517069e-1	3.42736529e-2	-4.29593279e-4	5.19908644e-1
θ_6	θ_7	θ_8	θ_9	θ_{10}	θ_{11}
4.67079740e-3	-2.76123961e-3	1.21035357e+0	2.41734938e-1	6.77708553e-1	3.23866568e-1

For the synthetic data sets, we use polynomial 2, 3, 5 to describe the data, we use 0.001 as learning rate, and we stop at 100000 steps, we arrive mean square errors and parameters as following tables

<table>
<tr><th colspan="4">Polynomial Regression Mean Square Error</th></tr>
<tr><td></td><td>Polynomial 2</td><td>Polynomial 3</td><td>Polynomial 5</td></tr>
<tr><td>synthetic-1</td><td>30.405389915964605</td><td>8.938301388620976</td><td>8.282870958898158</td></tr>
<tr><td>synthetic-2</td><td>0.3276429479455291</td><td>0.32761960364296244</td><td>0.3019140483644252</td></tr>
</table>

<table>
<tr><th colspan="7">Polynomial Regression Parameter for synthetic-1</th></tr>
<tr><td></td><td>θ_0</td><td>θ_1</td><td>θ_2</td><td>θ_3</td><td>θ_4</td><td>θ_5</td></tr>
<tr><td>Polynomial 2</td><td>-4.4205321</td><td>1.90026024</td><td>0.6768219</td><td>-</td><td>-</td><td>-</td></tr>
<tr><td>Polynomial 3</td><td>-3.95660592</td><td>11.14927217</td><td>0.87702711</td><td>-3.8329229</td><td>-</td><td>-</td></tr>
<tr><td>Polynomial 5</td><td>-3.02093059</td><td>10.82368801</td><td>-1.62334179</td><td>-3.38242625</td><td>0.73638042</td><td>-0.12239048</td></tr>
<tr><td colspan="7">Polynomial Regression Parameter for synthetic-2</td></tr>
<tr><td>Polynomial 2</td><td>0.37024944</td><td>-0.04780947</td><td>-0.17884706</td><td>-</td><td>-</td><td>-</td></tr>
<tr><td>Polynomial 3</td><td>0.36947691</td><td>-0.05770504</td><td>-0.17811546</td><td>0.00455444</td><td>-</td><td>-</td></tr>
<tr><td>Polynomial 5</td><td>0.48433072</td><td>-0.38658959</td><td>-0.52815878</td><td>0.33797196</td><td>0.10995532</td><td>-0.0684146</td></tr>
</table>

IV. Bonus In order to prevent over fitting, we implement the Ridge Regression, which involve the L2 norm regularization in the loss function and gradient decent. The hypothesis function is same as before, $h_\theta(x) = \theta_0 + \sum_{j=1}^{n}\theta_j x_j = X\theta$

Then we implement the loss function for Ridge Regression:

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)} - y^{(i)}))^2 + \frac{1}{2}\lambda||\theta||_2^2 = \frac{1}{2m}\sum_{i=1}^{m}(X\theta - Y)^2 + \frac{1}{2}\lambda||\theta||_2^2 \tag{6}$$

We can take derivative of $J(\theta)$ respect to θ to get our gradient

$$\frac{\partial J}{\partial \theta} = \frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)} - y^{(i)})) \cdot x^{(i)} + \lambda\theta_i = \frac{1}{m}(X^TX\theta - X^TY) + \lambda\theta \tag{7}$$

Now, we can start to implement gradient decent:
First, we randomly initialize θ by using uniform methods from -1 to 1.
Since equation (7) calculate weight, based on it, we use following function to update θ where

$$\theta = \theta - \alpha \cdot \frac{1}{m}(X^TX\theta - X^TY) + \lambda\theta \tag{8}$$

In terms of evaluation, we will use Mean Square Error as before:

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)} - y^{(i)}))^2 = \frac{1}{m}\sum_{i=1}^{m}(X\theta - Y)^2$$

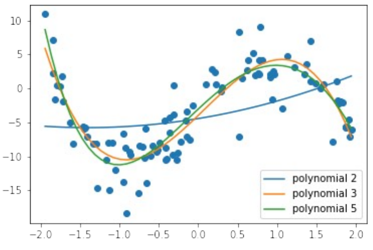
Then we use $\alpha = 0.01$, $\lambda = 0.05$, and 100000 steps as stop point, we have Mean square errors and parameter as follow:

Ridge Regression Mean Square Error			
	Polynomial 2	Polynomial 3	Polynomial 5
synthetic-1	30.50155083924403	10.136090166120733	9.60086849090402
synthetic-2	0.3284697151811236	0.3284480743902003	0.31041904238497076

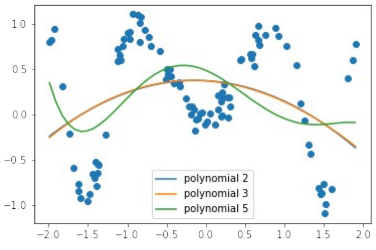
Ridge Regression Parameter for synthetic-1						
	θ_0	θ_1	θ_2	θ_3	θ_4	θ_5
Polynomial 2	-3.98149947	1.85528501	0.48925374	-	-	-
Polynomial 3	-3.68583936	8.84887521	0.68431916	-3.01320725	-	-
Polynomial 5	-3.14720065	6.79101706	-1.28104163	0.40847235	0.65195103	-0.89057433
Ridge Regression Parameter for synthetic-2						
Polynomial 2	0.33103623	-0.04446019	-0.15718136	-	-	-
Polynomial 3	0.33057632	-0.05090691	-0.15667458	0.00311149	-	-
Polynomial 5	0.3786813	-0.16152214	-0.32537145	0.13887873	0.05553962	-0.03067927

V. Visualization

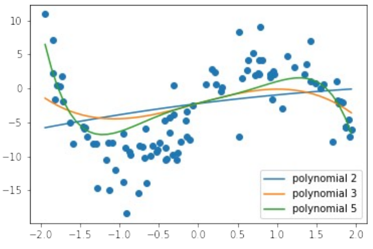
We can visualize Polynomial Regression and Ridge Regression as following graphs:



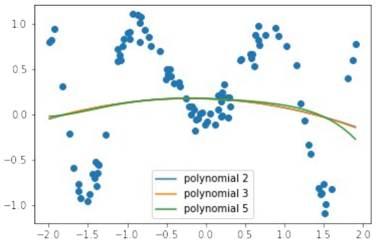
Synthetic-1 Polynomial Regression



Synthetic-2 Polynomial Regression



Synthetic-1 Ridge Regression



Synthetic-2 Ridge Regression