Mark Klobukov

CS 383: Homework 3

Professor Burlick

2/7/2018

Part 1: Theory

Consider the following data:

$$\begin{bmatrix} -2 & 1 \\ -5 & -4 \\ -3 & 1 \\ 0 & 3 \\ -8 & 11 \\ -2 & 5 \\ 1 & 0 \\ 5 & -1 \\ -1 & -3 \\ 6 & 1 \end{bmatrix}$$

Compute coefficients for linear regression with global LSE. Do not separate into training and testing parts.

First step is to standardize the data. Since second column represents the dependent variable, it does not need to be standardized. Only standardize the first column:

$$\mu_1 = -0.9$$

$$\sigma_1 = 4.2282$$

Subtract the mean from each entry and divide by the standard deviation. The resulting matrix looks as follows:

$$\begin{bmatrix} -0.2602 & 1 \\ -0.9697 & -4 \\ -0.4967 & 1 \\ 0.2129 & 3 \\ -1.6792 & 11 \\ -0.2602 & 5 \\ 0.4494 & 0 \\ 1.3954 & -1 \\ -0.0237 & -3 \\ 1.6319 & 1 \end{bmatrix}$$

Add a column of 1's to the data (independent variable) to be able to write the equation in matrix form:

$$\begin{bmatrix} 1 & -0.2602 \\ 1 & -0.9697 \\ 1 & -0.4967 \\ 1 & 0.2129 \\ 1 & -1.6792 \\ 1 & -0.2602 \\ 1 & 0.4494 \\ 1 & 1.3954 \\ 1 & -0.0237 \\ 1 & 1.6319 \end{bmatrix}$$

Now we can find the equation as:

$$g(x) = X\Theta \approx \mathrm{Y}$$

$$g(x) = \begin{bmatrix} 1 & -0.2602 \\ 1 & -0.9697 \\ 1 & -0.4967 \\ 1 & 0.2129 \\ 1 & -1.6792 \\ 1 & -0.2602 \\ 1 & 0.4494 \\ 1 & 1.3954 \\ 1 & -0.0237 \\ 1 & 1.6319 \end{bmatrix} \begin{bmatrix} \Theta_0 \\ \Theta_1 \end{bmatrix} \approx \begin{bmatrix} 1 \\ -4 \\ 1 \\ 3 \\ 11 \\ 5 \\ 0 \\ -1 \\ -3 \\ 1 \end{bmatrix}$$

To find the weights, use least squared error minimization. Take the residual of prediction with each observation, square it, and sum all of the squares together. The following matrix expression can be used to find the error.

$$J(\Theta) = (Y - X\Theta)^T (Y - X\Theta)$$

By differentiating the equation with respect to Θ and setting it to zero (details of derivation are covered in the slides and are not presented here), we get the following equation:

$$\Theta = (X^T X)^{-1} X^T Y$$

This is the closed-form expression for the weights that minimize least squared error. The remaining computations will be shown step-by step.

$$X^T X =$$

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ -0.2602 & -0.9697 & -0.4967 & 0.2129 & -1.6792 & -0.2602 & 0.4494 & 1.3954 & -0.0237 & 1.6319 \end{bmatrix} \begin{bmatrix} 1 & -0.2602 \\ 1 & -0.9697 \\ 1 & -0.4967 \\ 1 & 0.2129 \\ 1 & -1.6792 \\ 1 & -0.2602 \\ 1 & 0.4494 \\ 1 & 1.3954 \\ 1 & -0.0237 \\ 1 & 1.6319 \end{bmatrix}$$

$$= \begin{bmatrix} 10 & -0.0001 \\ -0.0001 & 9.0002 \end{bmatrix}$$

Find the inverse of that:

$$\begin{bmatrix} 10 & -0.0001 \\ -0.0001 & 9.0002 \end{bmatrix}^{-1} = \begin{bmatrix} 0.1000 & 0.0 \\ 0.0 & 0.1111 \end{bmatrix}$$

Multiply this by the transpose of data:

$$(X^T X)^{-1} X^T =$$

$$\begin{bmatrix} 0.1000 & 0.1000 & 0.1000 & 0.1000 & 0.1000 & 0.1000 & 0.1000 & 0.1000 & 0.1000 & 0.1000 \\ -0.0289 & -0.1077 & -0.0552 & 0.0236 & -0.1865 & -0.0289 & 0.0499 & 0.1550 & -0.0026 & 0.1813 \end{bmatrix}$$

Finally, multiply the matrix above by the dependent variable column:

$$(X^T X)^{-1} X^T Y = \begin{bmatrix} -0.00000408 \\ -0.408245 \end{bmatrix} \begin{bmatrix} 1 \\ -4 \\ 1 \\ 3 \\ 11 \\ 5 \\ 0 \\ -1 \\ -3 \\ 1 \end{bmatrix} = \begin{bmatrix} 1.4000 \\ -1.7449 \end{bmatrix}$$

So the answer is: $\Theta = \begin{bmatrix} \Theta_0 \\ \Theta_1 \end{bmatrix} = \begin{bmatrix} 1.4000 \\ -1.7449 \end{bmatrix}$

With this model, RMSE = 3.7024, according to the formula:

$$(\text{RMSE}): \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2}.$$

2) Closed-form Linear Regression

Theta for the dataset:

```
[[ 3267.43333333]
 [ 1109.30363764]
 [ -198.34795734]]
```

Predicted values for test data vs. actual values in test data and RMSE:

```
Predicted =  [[ 3024.85782476]]  Actual =  3920.0
Predicted =  [[ 2309.52545335]]  Actual =  1215.0
Predicted =  [[ 1559.41569719]]  Actual =  590.0
Predicted =  [[ 2388.79859589]]  Actual =  2140.0
Predicted =  [[ 4092.99719534]]  Actual =  3935.0
Predicted =  [[ 2093.48538248]]  Actual =  1305.0
Predicted =  [[ 2490.78813946]]  Actual =  2140.0
Predicted =  [[ 5149.07558222]]  Actual =  4600.0
Predicted =  [[ 4513.01635335]]  Actual =  4535.0
Predicted =  [[ 5047.08603865]]  Actual =  4570.0
Predicted =  [[ 3444.87698276]]  Actual =  3030.0
Predicted =  [[ 4398.96582607]]  Actual =  3257.0
Predicted =  [[ 3138.90835204]]  Actual =  2600.0
Predicted =  [[ 2207.53590977]]  Actual =  620.0
RMSE =  783.177472477
```

RMSE: 783.1775

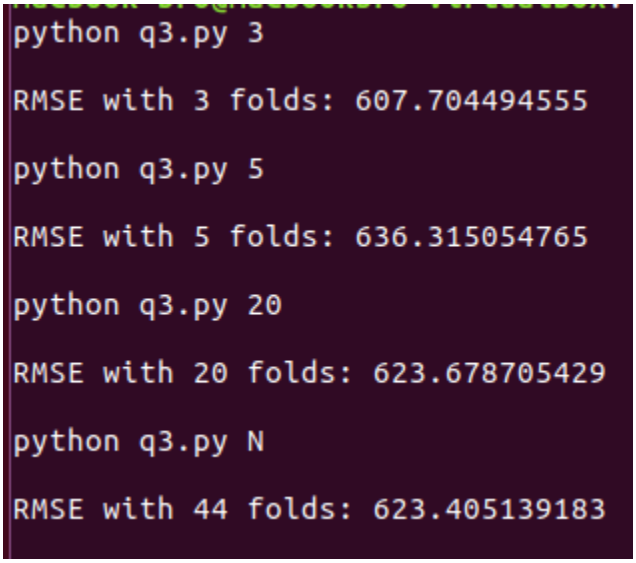
Model: $g(X) = \Theta_0 + \Theta_1 X_1 + \Theta_2 X_2 = 3267.433 + 1109.3036 \times X_1 - 198.348 \times X_2$

where $X_1 = age\ of\ the\ fish, X_2 = temperature\ of\ water$

Or in matrix form:

$$g(X) = [1 \quad X_1 \quad X_2] \begin{bmatrix} \Theta_0 \\ \Theta_1 \\ \Theta_2 \end{bmatrix} = [1 \quad X_1 \quad X_2] \begin{bmatrix} 3267.433 \\ 1109.3036 \\ -198.348 \end{bmatrix}$$

3) S-Folds cross validation (execute `make` to replicate the results below)



Number of Folds	**RMSE**
3	607.704
5	636.315
20	632.679
N (44 in this case)	623.4051

Table 1: RMSE values for different numbers of folds