Design Assignment 1: Multiple Linear Regression

Bryan Guner, James Martinez, and Olivia Shanley
Larry Pearlstein, Ph. D.
Artificial Neural Networks - ELC 470-01
The College of New Jersey
Department of Electrical and Computer Engineering (ECE)
03/29/2017

Description:

The data set used in this neural network is on baseball players’ pitching statistics over the past century and a half. This includes the number of games played; hits; walks by batters; home runs; errors; wins, losses, etc. We used this data to predict the number of hits that a pitcher gives up during a season. The input to this network was the hits given up by a pitcher during a season and the other pitching statistics. The output is the predicted number of hits that the pitcher would give up during the season. This data set was one of the many provided to the class from github user adilmoujahid’s repository “Baseball Mathematics” (https://github.com/adilmoujahid/Blog/blob/master/content/2014/Baseball%20Mathematics.md). The expected result of this analysis is a set of weights that can be used to predict the number of hits the pitcher will give during the season.

Results:

W	**L**	**G**	**GS**	**CG**	**SHO**	**SV**	**IPouts**	**H**	**ER**	**HR**
0.0447	-0.0606	0.0516	0.3228	-0.3996	-0.5676	-0.0943	0.2413	0.7338	-0.529	-0.3603
BB	**SO**	**BAOpp**	**ERA**	**IBB**	**WP**	**HBP**	**BK**	**BFP**	**GF**	**R**
-0.1715	0.248	-0.0313	0.2496	-0.2433	-0.2224	-0.31	0.0027	0.0759	0.3763	0.257

Table 1. Optimal Weights

The error for the training data set and the test data set were calculated using the mean square error equation: . The training data error was calculated to be 58.914577555412670. The test data error was calculated to be 15.491855286948436.

Discussion:

The neural network performed better than expected with the data points near the regression line. The regression line has a slope of 1 and a y-intercept of 0. The points closer to the regression line have less error. This is because the closer the predicted value and the actual value are to each other than the point would be closer to the regression line. The highest weights were assigned with games started, games finished and runs with games finished being higher. This means that the likelihood of the pitcher giving up a hit is more correlated to if that pitcher finished the game rather than started it. The error for the test data set is significantly lower than the training data set error so it is believed that the model is a good fit.

Matlab Code:

```
%Open the csv file
M=csvread('C:/Users/James/Documents/elc470/Pitching.csv',1,5);
% The total hits given up are in the 9th column
train_hits = M(1:34066,9); %training
test_hits=M(34067:42583,9);
% Collect all other numerical data into the array "not_hits"
train_not_hits = [M(1:34066,1:8) M(1:34066,10:23)]; %training 80% of data
test_not_hits=[M(34067:42583,1:8) M(34067:42583,10:23)]; %training 20% of data

% Name the train array 'X', to conform to the common naming for regression
X=train_not_hits;
% Name the test array 'Y'
Y=test_not_hits;
% Compute the covariance matrix
R=X'*X;
% Compute the cross correlation vector
P=X'*train_hits;
% Use the Normal Equation to solve for the optimum weights
w = inv(R)*P;

%calculate train error
train_e=train_hits-(X*w);
train_error= (1/34066)*(train_e')*train_e;
%calculate test error
test_e=test_hits-(Y*w);
```

```
test_error= (1/34066)*(test_e')*test_e;

% Plot the value predicted from the train data vs. true value
figure(1)
scatter(train_hits, X*w, 1)
hline=refline(1,0);
hline.Color = 'red';
ylabel('\fontsize{16}Predicted Hits Given Up by Pitcher in Each Season');
xlabel('\fontsize{16}Actual Hits Given Up by Pitcher in Each Season');
title('\fontsize{20}Predicting Hits From Other Pitching Stats (Training Data)');

% Plot the value predicted from the test data vs. true value
figure(2)
scatter(test_hits, Y*w, 1)
hline=refline(1,0);
hline.Color = 'green';
ylabel('\fontsize{16}Predicted Hits Given Up by Pitcher in Each Season');
xlabel('\fontsize{16}Actual Hits Given Up by Pitcher in Each Season');
title('\fontsize{20}Predicting Hits From Other Pitching Stats (Test Data)');
```