MATH 728D: Machine Learning Lab #6: Linear Regression

John Burkardt

December 5, 2018

What simple linear formula $y = mx + b$ will approximate my data?

A linear model for a set of data (x_i, y_i) is a mathematical formula $y = h(x)$, sometimes called a *hypothesis function*. The hypothesis function is defined by parameters, θ. Unless we know these values in advance, we draw attention to them by writing $y = h(x : \theta)$. Finding the best linear model means searching for the best values of θ, where we state that the best θ minimizes a scalar cost function $E(\theta)$, that we specify.

1 Set up and solve a linear regression problem

Suppose we have m pairs of data (x_i, y_i), and we think there's a linear relationship of the form $y = m * x + b$ that can be a useful approximation. So we need to find the vallues of m and b. In order to apply linear regression techniques, we need to rewrite our problem:

$$y = w * b + x * m$$

where w is always equal to 1. Now we expect this relationship to hold for every pair of data values:

$$\begin{aligned} y_1 &= b\,w_1 + m\,x_1 \\ y_2 &= b\,w_2 + m\,x_2 \\ &\cdots\cdots \\ y_k &= b\,w_k + m\,x_k \end{aligned}$$

and so we have a (rectangular) system of linear equations to solve:

$$\begin{pmatrix} w_1 & x_1 \\ w_2 & x_2 \\ \cdots & \cdots \\ w_k & x_k \end{pmatrix} \begin{pmatrix} b \\ m \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \cdots \\ y_k \end{pmatrix}$$

As we know, the MATLAB backslash operator will always try to return a useful solution, even to rectangular linear systems.

Your task in this exercise is to start with the data from *china_data.txt*, containing 12 data pairs *year* and *income*. We are looking for a relationship of the form

$$income = m * year + b$$

Your task is to set up the linear regression system, solve it, and report the solution. and transform it into a 12×2 matrix A,

Exercise 1:

1. Use `load()` to read the data file *china_data.txt*;
2. Create the system matrix and right hand side;
3. Print the system matrix and right hand side;
4. Solve for the slope m and intercept b and report their values;
5. Scatterplot the data, and include the line $income = m * year + b$;

2 Solve the Ford Escort Problem Again

In the previous lab, we had pairs of data representing mileage and price, and we sought a linear relationship using gradient descent. But since the unknown quantities are in a linear relationship, we can instead use linear regression.

In order to get the gradient descent method to converge, we had to normalize both sets of variables. For our linear regression approach, this will not be necessary. However, we will compute results for both unnormalized and normalized cases.

Exercise 2:

1. Use `lcsvread()` to read the data file *ford_data.csv*, skipping row 1 and column 1;
2. Create the system matrix and right hand side;
3. Solve for the slope m and intercept b and report their values;
4. Scatterplot the data, and include the line $price = m * mileage + b$;
5. Normalize *price* and *mileage* to lie between 0 and 1;
6. Set up and solve the linear system for m and b;
7. Compare your values of m and b for the normalized data to those you got in the lab 5 exercise.

3 Choices for Linear Solvers

Although we can let MATLAB solve our linear system with the backslash operator, there are several ways to solve a rectangular linear system, which can be applied whether or not MATLAB is available.

1. The *Normal equations* $A'Ax = A'b$:

```
x = ( A' * A ) \ ( A' * b );
```

2. The *QR factorization* $A = Q * R$:

```
[ Q, R ] = qr ( A );
x = R \ ( Q' * b );
```

3. The *pseudoinverse* $A^+ = V * D^+ * U'$:

```
Ap = pinv ( A );
x = Ap * b;
```

Exercise 3:

1. read the data file *china_data.txt*;
2. create the system matrix A and right hand side b;
3. compute the solution using:
   (a) the backslash operator;

(b) the normal equations;
(c) the QR factorization;
(d) the pseudoinverse;

How well do your four solutions agree?

4 Linear Regression of Basketball Data

The file *basketball_data.csv* contains a title record, then 30 records of index, height, weight, sponsorship, and age. Let's explore the possibility of a linear relationship of sponsorship based on age. Let sponsorship be y, let w be a variable that is always 1, and x be the age. Our desired formula is

$$y = \theta_1 * w + \theta_2 * x$$

where θ is the vector of parameters we seek, which we can rewrite as a linear system $Ax = b$.

Exercise 4:

1. read the data file *basketball_data.csv*;
2. create the system matrix whose columns are the variable w and age;
3. create the right hand side vector which is sponsorship;
4. solve, using a method of your choice, for the parameters;
5. report the values of the parameters you found;
6. Make a scatterplot of age versus sponsorship, including a plot of your hypothesis function;