MATH 728D: Machine Learning Lab #7: Multilinear Regression

John Burkardt

December 5, 2018

What simple linear formula $y = ax_1 + bx_2 + cx_3 + d$ will approximate my data?

An output quantity y might depend on the value of several independent inputs. For instance, a realtor might suggest the selling price for a house based on a list of factors that include age, square footage, number of bathrooms. A doctor might choose the dosage of a medicine based on weight, age, blood pressure, and severity of symptoms. We can still seek a simple linear model, but now there will simply be more factors to keep track of, and parameters to determine.

If we represent our linear model as $y = A * x$, then it is important to realize that the matrix A represents our known data, and the unknown vector x represents the linear coefficients that multiply our input data to estimate the output.

1 Factors in Medical Insurance Charges

The file *insurance_data.txt* includes 1338 records about medical insurance. Each record lists, for a given person, 7 factors: age, sex ("male"=0, "female"=1), BMI, children, smoker ("no"=0, "yes"=1), region ("NE"=1, "NW"=2, "SW"=3, "SE"=4), total medical charges ($). We seek a linear formula that can predict the charges based on the numerical values of age, sex, BMI, number of children, and smoker .

Our data array A will have six columns, with the first being the vector of 1's, followed by columns for age, sex, BMI, children, smoker. Our model will have the form $y = A * x$ where x represents the unknown coefficients in the linear relationship. We estimate the coefficients x using the backslash operator.

Exercise 1:

1. read the data from *insurance_data.txt* using `load()`;
2. extract columns 1, 2, 3, 4, 5, and 7, storing them in vectors named age, sex, bmi, children, smoker, and charges;
3. build a matrix `A` with a column of 1's, then age, sex, bmi, children, smoker;
4. Seek a solution of A*x=charges;
5. print the x parameters;
6. compare the actual charges to the charges predicted by your model for the patient with data

```
age=38,sex=1,bmi=19.3,children=0,smoker=1,charges=15820.699
```

2 Analyzing a Subgroup

Column 6 of the data gives the patient's region as 1="northeast", 2="northwest", 3="southwest" or 4="southeast". Repeat Exercise 1, but only using data for people in the southwest region.

Exercise 2:

1. read the data from *insurance_data.txt* using `load()`;
2. use MATLAB's `find()` command to index the southwest records.
3. copy only the southwest data items into vectors named age, sex, bmi, children, smoker, and charges;
4. build a matrix `A` with a column of 1's, then age, sex, bmi, children, smoker;
5. Seek a solution of A*x=charges;
6. print the x parameters;
7. compare the actual charges to the charges predicted by your model for the patient with data

```
age=38,sex=1,bmi=19.3,children=0,smoker=1,charges=15820.699
```

3 Training and Testing

In machine learning, a common technique is to build a model with part of the data, and then test how well the model predicts the behavior of the remaining set of data.

We can do a simple version of this training and testing sequence by compute our parameters on the first 1000 sets of data, then comparing the model to the actual data for the last 338 sets

Exercise 3:

1. read the data from *insurance_data.txt* using `load()`;
2. copy only first 1000 data items into vectors named age1, sex1, bmi1, children1, smoker1, and charges1;
3. build a matrix `A1` with a column of 1's, then age1, sex1, bmi1, children1, smoker1;
4. Seek a solution of A1*x=charges1;
5. print the x parameters;
6. Compute the residual norm: $r1 = ||charges1 - A1 * x||$;
7. Print the average residual norm (divide r1 by 1000);
8. Now copy data items 1001:1338 into vectors named age2, sex2, bmi2, children2, smoker2, and charges2;
9. build a matrix `A2` with a column of 1's, then age2, sex2, bmi2, children2, smoker2;
10. Compute the residual norm: $r2 = ||charges2 - A2 * x||$;
11. Print the average residual norm (divide r2 by 338);
12. Why do we expect $r1 < r2$?

4 Estimating Importance of Factors

When we compute a linear relationship, the size of the parameters can tell us something about the relative importance of the corresponding factors. However, since the factors may have different units and ranges, it is typical to normalize the data items. In our problem, the data items of age, bmi, and children can be normalized. For instance, to normalize age:

$$age = \frac{age - \min(age)}{\max(age) - \min(age)}$$

If our data ranges between 0 and 1, then when we find an approximate linear relationship between the data and the output (in this case, the charges), then large coefficients correspond to more important factors.

Exercise 4:

1. Repeat Exercise 1, after normalizing age, bmi, and children;
2. Print the coefficients x;
3. Indicate which data item seems to be most important in determining the charges;