Project 1: Multivariate Linear Regression

Due Oct 5, 2019

In this project you will apply *multivariate* (or *multiple*) *linear regression* to predict the gas milage of automobiles from their other attributes.

General Guidelines:

1. The document "Workflow" (in the Files section on Canvas) suggests a workflow for all of your machine learning projects. It might be updated from time to time.
2. We advise using python or Matlab (or its free equivalent, Octave) for the projects. C++ and Java are also acceptable, but if you want to use another language, check with us first.
3. You are expected to implement your own ML programs. You are not expected to implement other software, such as matrix manipulations (e.g., multiplication, pseudo-inverse), plotting, and data wrangling (e.g., format conversion, data frames, pandas). When in doubt, ask us.
4. We will post links to useful resources (e.g., software libraries) on the course website. If you know of additional resources, please email us or share them on Piazza so that we can add them to the website.
5. For each project you will be expected to write a report in which you describe your observations about the data, any manipulation or standardization that you applied to the data, and the results of your ML experiments. You will be graded on the quality of your report, so make sure it is well organized and well written. We won't try to make sense out of badly written reports!
6. Credit will be assigned as follows:
   - Project Description (10):
     - (5) Purpose of the project (what method are you analyzing?)
     - (5) Include database that's being tested
   - Pre-processing Steps (10):
     - (10) Explain format of initial data as well as any changes made to it (ex: separation into training and test data, adding placeholders for missing data, etc.)
   - Solution Description (20):
     - (10) Include required equations
     - (10) Explain all variables used in the equations
   - Analysis (50):
     - (25) Provide results in legible format
     - (25) Explain results in detail (not general summation)
   - Discussion (10):
     - (10) General summation of results and report
7. Submit your project by zipping a folder containing your report (in pdf format) and your software, which you should upload on Canvas. The software should be runnable by us, should we wish to do so. We will not accept emailed projects (barring extraordinary circumstances).
8. These are individual (not group) projects unless we state otherwise.

Specifics for Project 1:

1. In the folder for Project 1 on Canvas, you will find two files, `auto-mpg.data`, which is the data file, and `auto-mpg.names`, which is the dictionary describing the attributes. The first attribute is the mpg (miles per gallon) value, which your regression should predict from the other seven numeric attributes together.
2. As explained in General Guidelines #3, you are expected to implement your own multivariate linear regression and not use existing toolboxes (e.g., `sklearn`).
3. Compare the performance of your program with and without data standardization.
4. Extra credit: See if you can improve the performance of your algorithm through polynomial regression.
5. Based on your experiments, can you draw any conclusions about which attribute(s) are most determinative of gas milage? Include these in your report.