Aston University
Birmingham

Week 2:
Regression analysis

Anikó Ekárt

Module CS4730: Machine learning

Learning outcomes

In this unit we studied simple linear regression and multiple linear regression. In this practical, we shall deepen our understanding by applying linear regression on some datasets and interpreting the results.

Instructions

Download the datasets `bp_syst.csv`, `fish.csv` and `ENB2012_data.xlsx`.

We shall build linear models for these datasets and evaluate them.

You are encouraged to use Python Jupyter Notebooks for your work. This is for two main reasons: (1) the fact that you can print out intermediate results and execute code in smaller blocks will help you work through the tasks more efficently and effectively and (2) it is also the format that we shall be asking you to submit your portfolio tasks in.

Task 1 (warm-up)

The first dataset, `bp_syst.csv`, [1] contains systolic blood pressure measurements from for 30 people of different ages. There are 30 rows of data. The data columns include

- the age;
- the systolic blood pressure.

First, create a scatter plot of the dataset, using the `matplotlib` library in `python`. Then apply simple linear regression using the libraries `numpy` and `sklearn` on the full dataset (`sklearn.linear_model.LinearRegression`).

Remember that the equation for simple linear regression is:

$$\hat{y} = \beta_0 + \beta_1 x$$

[1] Helmut Spaeth, Mathematical Algorithms for Linear Regression, Academic Press, 1991, page 304 and D G Kleinbaum and L L Kupper, Applied Regression Analysis and Other Multivariable Methods, Duxbury Press, 1978, page 47.

where β_o is the intercept and β_1 the slope.

A nonzero intercept seems appropriate here, since even a very young person can have a high blood pressure.

Print the coefficients, the mean squared error and the coefficient of determination then plot the regression line and also the original points.

Now split the dataset into 20 points for training and 10 points for testing. Repeat the steps of creating a linear regression model, fitting it to the *t*raining data points, printing coefficients. For testing, print the mean squared error and the coefficient of determination on the *t*esting data. Plot the line and the testing data points.

Compare the results between (a) using the full dataset for training and (b) using $2/3$ of the data for trainig and $1/3$ for testing.

Discussion. *Are the results as you expected? How can you explain them?*

Task 2 (intermediate)

The second dataset, `fish.csv`,[2] is about length of a species of fish. The length of this species of fish is to be represented as a function of the age and water temperature. The fish are kept in tanks at 25, 27, 29 and 31 degrees Celsius. After birth, a test specimen is chosen at random every 14 days and its length measured. There are 44 rows of data. The data columns are:

- the age of the fish,
- the water temperature in degrees Celsius and
- the length of the fish.

First, generate a scatter plot matrix using `pandas.plotting.scatter_matrix` and examine the data visually. Also generate a 3D scatter plot of the data and inspect it visually.

Then apply linear regression on the full dataset, similarly to **Task 1**, with the difference that the model is now based on two input variables x_1 and x_2 and also knowing that a null intercept is desired ($\beta_0 = 0$):

$$\hat{y} = \beta_1 x_1 + \beta_2 x_2$$

Print the coefficients, the mean squared error and the coefficient of determination then try and plot the regression plane on top of the 3D scatter plot.

Discussion. *How good is the result? Would you be satisfied with this model to use on unseen data in the future? Why?*

Task 3 (challenging)

It is great if you got this far. Now let us look at a more complex regression problem. The third dataset, `ENB2012_data.xlsx`,[3] contains 768 data points, with 8 input variables $x_1, \cdots, x_8$. There are two output variables y_1 and y_2. The data were obtained by performing energy analysis using 12 different building shapes. The buildings differ in glazing area, glazing area distribution, orientation, and other parameters. Through simulation, 768 bulding shapes were obtained.

The recorded input and output variables are:

- x_1 Relative Compactness

[2] R J Freund and P D Minton, Regression Methods, Dekker, 1979, page 111 and Helmut Spaeth, Mathematical Algorithms for Linear Regression, Academic Press, 1991, page 305

[3] UCI Machine learning repository, Energy Efficiency Dataset `https://archive.ics.uci.edu/ml/datasets/Energy+efficiency`

- x_2 Surface Area
- x_3 Wall Area
- x_4 Roof Area
- x_5 Overall Height
- x_6 Orientation
- x_7 Glazing Area
- x_8 Glazing Area Distribution
- y_1 Heating Load
- y_2 Cooling Load

Here you can focus on one of the output variables at a time. You may inspect the data in the spreadsheet and decide to build an initial model based on a smaller subset of the input variables. Consider what scatter plots may help you decide and generate these plots.

Apply linear regression on the chosen input and output variables for the full dataset or just $2/3$ of the dataset (keeping $1/3$ for testing), calculate and print out the coefficients, the mean squared error, the mean absolute error and the coefficient of determination. Experiment on different subsets of the provided input variables and compare the results.

If you are curious, try also ridge and lasso (`sklearn.linear_model.Ridge` and `sklearn.linear_model.Lasso`), on the full set of input variables, with a few different values for regularization strength.

Discussion. *What measure would you use for the quality of the model (mean squared error, mean absolute error or coefficient of determination)? What subset of input variables led to your best result? Why? Is this sufficiently accurate? Why? Is your process of obtaining the best result generalizable to other problems?*

You may wish to read about the authors' approach and solution (provided as a pdf file):

A. Tsanas and A. Xifara, 2012. Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools. Energy and Buildings, 49, pp.560-567.