Machine Learning Lab 3

Linear Regression

Linear regression is a statistical model that examines relationship between two or more variables and a dependent variable and independent variables. This means that when one (or more) independent variable increases (or decreases) the dependent variable increases (or decreases) too. A linear relationship between the variables can be both positive as well as negative.

The dataset

The dataset used to perform this experiment is the wine quality dataset, it is a combination of data on two types of wine variants, namely red wine and white wine, of the portuguese "Vinho Verde" wine. The dataset contains information on the parameters for fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol.

Experiment

The algorithms I have used in this experiment are :

* Linear Regression
* Linear Regression with Gradient Descent
* RANSAC (RANdom SAmple Consensus) algorithm, it is a non-deteministic iterative algorithm for the robust elimination of parameters from a subset of inliers from the complete data set.

For performing this experiment I started with calculating the correlations between the different features in the dataset and then went forward with fitting a regression model between the features that had high correlation between them. By high correlation I meant more than 0.7 or less than - 0.7

I used linear regression with gradient descent to find the correlation between residual sugar and the density of the wine, and achieved a mean square error of 0 over the training data.

I used a RANSAC regressor to calculate the plot the regression line between total sulfur dioxide and density of the wine.

Finally I applied a simple linear regression over the above two features to evaluate the performance of the regression model. The model got a MSE error of 0 over both training and test set, this is clearly because of the less number of training samples for the model to train upon leading the model to overfit very easily. However the r^2 error was about 0.3 on the training data and 0.22 on the test data. Finally I experimented upon the data with using polynomial features and modelling non linear relationships. The results were better in the non-linear case with a r^2 error of 0.281 in linear, 0.283 in quadratic and 0.292 in the case of cubic relationship.

The code and plots can be found in the accompanying jupyter notebook.