Coursework of Foundations of Data Science (CMP-5046B)

100241852

May 26 2021

1 Introduction

This assignment aims to implement data science pipelines such as nearest neighbour and least squares approaches to evaluate regression models for real-world problem. This report will explain the results after performing data manipulation using Python libraries and make use of scatter plot and histogram to present the results. This report discusses statistical techniques such as simple regression and multiple regression that uses several independent (explanatory) variables to predict the dependent (target) variable. The results include standardised variables, correlations, performance metrics received from nearest neighbours and least squares in both simple and multiple regression and bootstrapping results.

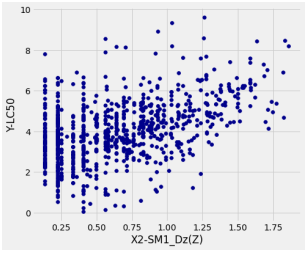
The regression problem for this assignment is based on QSAR model for predicting acute toxicity on fish (fathead minnow) (Cassotti et al., 2015). The independent variables for this problem are the 6 molecular descriptors: MLOGP (molecular properties), CIC0 (information indices), GATS1i (2D autocorrelations), NdssC (atom-type counts), NdsCH ((atom-type counts), SM1_Dz(Z) (2D matrix-based descriptors), which measures the concentration that causes death in fish during the test time-frame. The dependent variable is LC50 [-LOG(mol/L)] which is a response to the independent variable.

2 Description of data

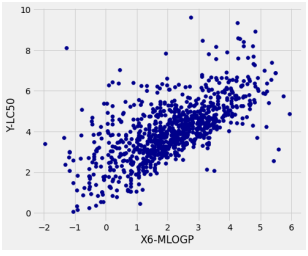
The data contains result of 908 test sets retrieved from UCI Machine Learning site. The figure below displays the first 5 rows of the table. The first six columns are the independent variables x which will be used to predict the value of dependent variable y. The final column contains the actual result of y.

	X1-CIC0	X2-SM1_Dz(Z)	X3-GATS1i	X4-NdsCH	X5-NdssC	X6-MLOGP	Y-LC50
0	3.260	0.829	1.676	0	1	1.453	3.770
1	2.189	0.580	0.863	0	0	1.348	3.115
2	2.125	0.638	0.831	0	0	1.348	3.531
3	3.027	0.331	1.472	1	0	1.807	3.510
4	2.094	0.827	0.860	0	0	1.886	5.390

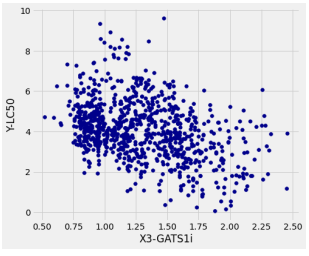
The graph below displays the correlation between a few pair of independent and dependent variables. Graph A and B both shows a strong tendency for a positive correlation. However, graph A shows that the points are more scattered than the points on graph B, this suggests that there is a weaker relationship between x2 and y compared to x6 and y. In contrast, graph C suggests that there is a weakly negative linear relationship between x3 and y.



(a) X2-SM1_Dz(Z)



(b) X6-MLOGP



(c) X3-GATS1i

It is found that the linear association for every pair of independent and dependent variables from descending order are 0.6557147 (x6 and y), 0.44671574 (x2 and y), 0.29496689 (x1 and y), 0.1723897 (x4 and y), 0.17200377 (x5 and y), and -0.37377176 (x3 and y). This supports the initial inspection on the degree of correlation based on the scatter plots.

3 Regression algorithms

This section covers simple and multiple regression with nearest neighbours and least squares. The idea of nearest neighbours is that if two points are near each other in the scatter-plot, then the corresponding measurements are most likely similar. Nearest neighbours is a model that assigns new data point based on the points within the neighbourhood that are most similar to it. The estimated y-value can be examined using the function: *nn_prediction(t, x_col, y_col, x_val, x_dist)*. The parameter t is the data table, *x_col* and *y_col* refers to the column of independent and dependent variable, *x_val* refers to the new data point, and *x_dist* refers to the threshold of distance. The return value is the average of the corresponding y-values for each group as the corresponding prediction.

Least squares is a method to determine the best fit line to data, it involves using the slope-intercept formula: *predicted_y = (slope * x) + intercept*. The regression line aims to predict y-value given values of x however each line typically raises error in estimation. The line of "best" fit tries to obtain the smallest possible overall error among all straight lines. The prediction of y-value is determined applying the slope-intercept formula, using the optimum coefficients.

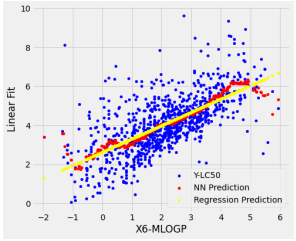
Multiple regression is an extension to simple regression, it involves two or more independent variables. K-nearest neighbours is similar to nearest neighbours but the difference is that it returns top k training table at increasing order of the distances. The distance between two points is calculated using the Euclidean distance formula. The function for k nearest neighbours is: *KNN_model(y_col, training, test, k)*. The parameter *y_col* is the column of the target variable, *training* is the training data set which involves two or more independent variables, *test* is the test data set, and *k* is the chosen k-value. The returned value is the average of the observed y-values in these k training table.

In the case of least squares, cost function is defined to quantify the error between predicted values and expected values. The cost function is specified as following: *multiple_regression_rmse(var1_coef, var2_coef, var3_coef, t, x1_col, x2_col, y_col)*. The parameter *var1_coef* and *var2_coef* is the slope value for first and second explanatory variable, *var3_coef* is the intercept, *t* is the training data, *x1_col* and *x2_col* is the column to the first and second explanatory variable, and *y_col* is the column to the target variable. The return value is the mean squared error (MSE) for the regression model. For simple and multiple regression, the minimize tool is used to search for the optimum coefficients. It tries to return the "best" coefficients by minimizing the mean squared error among all lines.

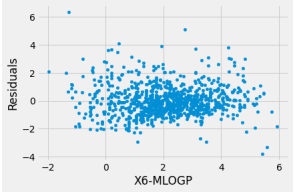
4 Experiments and analysis

4.1 Simple Regression

The figure below displays scatter plots of original data points and prediction lines and residuals for nearest neighbours and least squares from a single independent variable: X6-MLOGP.



(a) Nearest neighbours and least squares



(b) Residuals

The residual plot shows no pattern or flat. This means that the residuals and predictor variable are uncorrelated, hence linear regression is a good alternative to solve the problem. The quality of regression can also be measured by accessing the numerical properties of residuals. Firstly, the correlation between the predictor variable and the residuals should be zero. Here, this value is -1.4e-15(≈ 0). Secondly, the average of the residuals should be zero. The actual result of the average of the residuals is 2.91e-15(≈ 0). Thirdly, the SD of the residuals should be a fraction of the SD of the response variable. This is also supported by the evidence where both values are equal to 1.094.

The figure below shows result for nearest neighbours prediction and least squares. Graph A-C shows nearest neighbours prediction at a few different threshold values, the result shows that the predicted y-value is smaller as the threshold of distance is lower. When comparing the performance of least squares against nearest neighbours, it can be seen that the estimation of y-values given by both least squares and nearest neighbours are nearly similar.

	X1-CIC0	X2-SM1_Dz(Z)	X3-GATS1i	X4-NdsCH	X5-NdssC	X6-MLOGP	Y-LC50	NN Prediction
0	3.260	0.829	1.676	0	1	1.453	3.770	3.417412
1	2.189	0.580	0.863	0	0	1.348	3.115	3.310231
2	2.125	0.638	0.831	0	0	1.348	3.531	3.310231
3	3.027	0.331	1.472	1	0	1.807	3.510	3.942368
4	2.094	0.827	0.860	0	0	1.886	5.390	3.977387

(a) Nearest neighbours at a threshold value of 0.2

	X1-CIC0	X2-SM1_Dz(Z)	X3-GATS1i	X4-NdsCH	X5-NdssC	X6-MLOGP	Y-LC50	NN Prediction
0	3.260	0.829	1.676	0	1	1.453	3.770	3.723888
1	2.189	0.580	0.863	0	0	1.348	3.115	3.669836
2	2.125	0.638	0.831	0	0	1.348	3.531	3.669836
3	3.027	0.331	1.472	1	0	1.807	3.510	3.858971
4	2.094	0.827	0.860	0	0	1.886	5.390	3.910344

(b) Nearest neighbours at a threshold value of 1.5

	X1-CIC0	X2-SM1_Dz(Z)	X3-GATS1i	X4-NdsCH	X5-NdssC	X6-MLOGP	Y-LC50	NN Prediction
0	3.260	0.829	1.676	0	1	1.453	3.770	3.543113
1	2.189	0.580	0.863	0	0	1.348	3.115	3.503766
2	2.125	0.638	0.831	0	0	1.348	3.531	3.503766
3	3.027	0.331	1.472	1	0	1.807	3.510	3.784882
4	2.094	0.827	0.860	0	0	1.886	5.390	3.847195

(c) Nearest neighbours at a threshold value of 0.75

	X1-CIC0	X2-SM1_Dz(Z)	X3-GATS1i	X4-NdsCH	X5-NdssC	X6-MLOGP	Y-LC50	Regression Prediction
0	3.260	0.829	1.676	0	1	1.453	3.770	3.618429
1	2.189	0.580	0.863	0	0	1.348	3.115	3.546795
2	2.125	0.638	0.831	0	0	1.348	3.531	3.546795
3	3.027	0.331	1.472	1	0	1.807	3.510	3.859938
4	2.094	0.827	0.860	0	0	1.886	5.390	3.913834

(d) Least squares

4.2 Multiple Regression

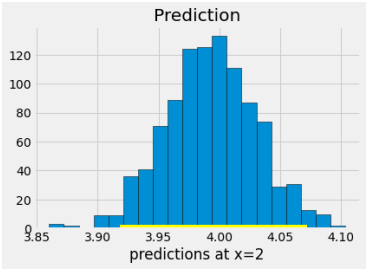
Table below shows the prediction accuracy of nearest neighbour at different value of k and least squares using different combinations of model inputs:

Model Input	Nearest Neighbour				Least Squares
	k=1	k=3	k=5	k=11	
x2 and x6	0.167	0.590	0.725	0.861	1.018
x1 and x6	0.162	0.671	0.854	1.012	1.093
x1 and x2	0.284	0.718	0.934	1.110	1.179

The result shows that the prediction accuracy for nearest neighbours is affected by tuning the parameter k and using different combinations of model inputs. It has been observed that for all model inputs, the mean squared error of estimation is greater when larger k-value is specified. For all cases, as a large k is specified, the errors for k-nearest neighbour is nearly similar to the least squares prediction. In contrast, k-nearest neighbour with a low k-value provides a lower value of errors, this indicates a better fit. Overall, the combination of variable x2 and x6 gives a lower MSE value compared to the other two model inputs, this suggests that the molecular component CIC0 and SM1_Dz(Z) is a better measurement to predict the acute toxicity on fish.

4.3 Bootstrapping

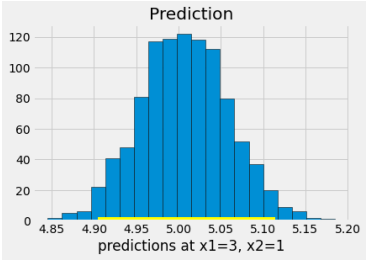
With bootstrapping sample is drawn without replacement, each time the original data is re-sampled, a different predicted values for given value of x is returned. Bootstrapping estimates the properties of confidence intervals by measuring the data from a distribution. A confidence interval indicates that the prediction will likely to fall within a certain region. The figure below shows the histogram of 1000 bootstrap predictions for least squares in simple regression where x=2 using 96% confidence level, the independent variable chosen here is X6-MLOGP. The empirical histogram is symmetric or bell-shaped, this means that the bootstrap percentile method works well for estimating confidence intervals. Furthermore, having a large sample size allows us to visualise a good probability distribution which result in decent approximations. By taking 96% confidence level, the prediction interval ranges from 2th percentile to the 98th percentile of the predictions. The interval ranges from about 3.918 to 4.071. Using the original data, the predicted y-value is 3.992 which is roughly the centre of the interval. This range is narrow which suggests that the uncertainty in the estimation of y is statistically small. This is because the x-value is close to the centre of the distribution (mean=2.11). In general, instances are closer to each other near the mean of x and as a result the confidence interval is narrow.



The table below shows confidence interval obtained for predictions at a different value of x for simple regression of least squares using the same independent variable X6-MLOGP. It shows that the confidence interval are wider with x=-1 (0.478) and x=5 (0.396) compared to the x-values that are closer to the mean. It is assumed that instances are farther apart from each other at x=1 or x=5 compared to x=2 or x=3 and therefore the predictions varies as well.

x=-1	x=2	x=3	x=5
0.478	0.154	0.173	0.396

The figure below shows the histogram of 1000 bootstrap predictions for least squares in multiple regression at x1=3 and x2=1 where x1 represents X6-MLOGP, and x2 represents X2-SM1_Dz(Z). The interval ranges from about 4.904 to 5.114, the predicted y-value based on the original table data is 5.001 which lies within the interval range. The confidence interval above using a single variable MLOGP is slightly narrower than combining MLOGP and SM1_Dz(Z). This might suggests that using one explanatory variable is enough to measure the concentration that causes death in fish.



5 Conclusion

This assignment aims to implement simple and multiple regression model with nearest neighbour and least squares approaches to evaluate acute toxicity on fish based on the data retrieved from UCI Machine Learning site. Several explanatory variables have been explored to predict acute toxicity on fish. Among all explanatory variables, it is found that there is a stronger relationship between MLOGP (molecular properties) with LC50, a concentration that causes death in fish. The properties of residuals in regression analysis show that linear regression is a suitable method to solve the problem. Experiments were carried out to find the prediction accuracies for nearest neighbour and least squares. It is concluded that there are many factors to predict Y-LC50 this can be in the approach measurement or the combination of model inputs.

I am satisfied with the outcome of the report, however if I would have done it differently, I would allocate more time to prepare the evaluation report. There is a part which I missed and that is testing the prediction accuracies for simple regression. This is because I did not pay close attention to the question.

References

Cassotti, M., Ballabio, D., Todeschini, R., and Consonni, V. (2015). Qsar fish toxicity data set.