Regression and K-fold Cross Validation

ABSTRACT

PURPOSE: Evaluate root mean squared errors from mutltiple Linear Regression models. Use k-fold Cross Validation to determine the reliability of linear regression performance on a modest dataset.

METHODS: Comparing different linear regressions, construct k-fold cross validation process, and evaluate results on cross validation.

RESULTS: A feature within a dataset that is best explained by other variables. Errors from testing and training for 5-fold Cross Validation for the variable with the best modeled variable.

CONCLUSIONS: Using cross validation technique, linear regression performance can be evaluated across a limited dataset.

Word count: 88

INTRODUCTION

The objective is to construct a cross validation approach and evaluate the performance of linear regression. In this study, linear regression model is the subject of the evaluation.

Linear Regression estimate a weight vector for a design matric that is "tall and thin", and a data vector.[7] This is a problem of pattern recognition from sparse data. The assumption made on these data si that with each set of data has the form of (a,c) where a is an independent value and a is a dependent value. The relationship between these data pairs has a model parameter w that will learn from the pattern in the dataset. [5] Residual Error is the difference between the dependent value and the model value. The goal of the modeling process is to minimize the residual error, by a function of model parameter w. A popular method of evaluating linear regression is by measuring the sum of the squares of the individuals residaul errors. The solution to this is to use the normal equation

$$[A^T A]\vec{w} = A^T\vec{c}.$$ [5]

In machine learning, the problem is described in linear regression as $X\vec{w} \approx \vec{y}.$ [7] If the model includes an intercept coefficient, each observation would then be defined as $\vec{t_i}^T = [a_i\ 1]$. [5] The intercept term is the expected mean value of the function when all independent values are 0, accounting for bias that parameters do not cover. [8] In order to optimize the equation, RMS error is used as a common way to measure the error between the model and the data. This is because Euclidean norm of an error vector can be expected to increase with the number of entires in the error vector. The RMS error of equation can be written as

$$RMS(X, \vec{y}; \vec{w}) = \sqrt{\frac{[X\vec{w} - \bar{y}]^T [X\vec{w} - \vec{y}]}{m}}$$ [6]

Cross validation is a statistical approach to compare a selected model for a given predictive problem. In Machine Learning, training is the process of finding parameters or coefficients for the

model in training, while testing is the process of evaluation the model performs using a score.[6] With a medium sized dataset, a common method of approach is k-fold corss validation. The process is commonly performed with 5 folds. The advantage of k-fold cross validation is that it is robust when dealing with multiple statistical outliers in the dataset. The disadvantage of k-fold cross validation is that only k folds are generated during the evaluation, causing sparse sampling and high variance during testing process. [6]

The testable hypothesis for this study is to implement linear regression in MATLAB on a dataset and then access the accuracy of training and testing using k-fold cross validation method.

METHODS

From the original data matrix, this work produces a cross validation on the linear regression task for the most predictive data column.

First given task was to determine the variable of best regression from the dataset. The goal is to determine the data column that is best explained by other data column.

Preprocessing of data require dealing with loading data, manage missing data, and standardize data. To load the data matrix, csvread built-in function is used to extract the data without including the first row with column titles. Then, the missing values (-200's) are then replaces as NaN in the data matrix. In the case of time interval data collecton for air quality, missing values are associated with Missing Completely At Random (MCAR) or Missing At Random (MAR), the reasons for its absence are external and not related to the value of the observation.[1] With MAR, the general approach is to use generative method to fill out missing values based on completed data points. With MCAR, the safe approach is to remove the missing data because the result will be unbiased and reliable, with a trade off that the test won't be as powerful.[2] For the current dataset, an attempt to remove all rows with missing values will result in only 827 entries left. With the current dataset, a combination of both method is required to retain a usable amount of data while avoid bias from filling out missing values. Thus, only rows with more than 2/3 data are missing will be removed from the dataset, leaving 8991 rows left from the initial 9357 rows. Then, the leftover missing values in the data set will be determined using the built-in fillmissing function with 'linear' method. When filling out missing value using fillmissing built-in function, method 'linear' is chosen. This method works cohesively with slow changing data where there are a lot of sample points.[3] This approach completes the dataset using linear interpolation of neighboring, non-missing value. Normalization is then applied to the data matrix. Standardize comes in use when orginal data has varying scale. Furthermore, linear algorithm makes assumptions about data having a Gaussian distribution. [6]

Linear Regression is performed multiple times on the dataset, using each column as the dependent vector and the rest as independent observations. This is done using a for loop going through all data columns. For every new iteration will update a Xmat matrix and an yvec vector accordingly to the current column that is being tested on. When data is standardized, no intercept term is added to the matrix. This is because standardized data will have no intercept term since the process of creating a zero-mean vector for the final column of the data matrix will result in a zero vector.[4] Then, a vector uval is used to store results from using linsolve on Xmat and yvec. The RMS result from uval is then stored accordingly in rmsvars to its column. Variable lowndx is the decided by taking on the index of the lowest rms value within the array.

Second given task was to determine the reliability of the chosen chemical variable as a proxy for other chemical variables in the given data using 5-fold cross validation.

Loading and processing the data is the same as the previous task, including managing missing data however no standardization for this task. Without standardization being applied to the data an intercept term is added since a zero mean is not guaranteed with the original dataset. Furthermore, adding an intercept term can reduce bias in the linear regression model.

Performing cross validation on the data matrix requires random fold selection for all the rows in the data matrix. Within mykfold function, the built-in function randperm creates a random permutation of index from 1 to the number of rows in the data matrix. This row vector is then being transposed to get a column vector. The value within the vector is then being replaced by its modulo with k value. This creates a vectors of randomized fold assignment (from 1 to k) for all the rows in the dataframe. Within a for loop iterating through k times validation, every round will create a new testing set with all the row in that fold and everything else will be left to training set. Since the dataset for this task is unstandardized, an intercept term is added to the training and testing set to decrease any bias. After k times, the rmstrain and rmstest accumulate all the rms values for all the cross-training attempts.

This program was tested on "air-quality.csv" dataset from the University of California at Irvine. The data has previously been processed by the instructor to remove 4 variables: date of readings, time of readings, relative humidity of readings, and absolute humidity of readings.

In order to evaluate the result from this study, a variance of the rmstrain and rmstest is recorded. After 5-fold cross validation, the variance for rmstrain is recorded around 0.2591 and variance for rmstest is 1.0552. This ensure that the cross-validation process produces error within a similar range and the model was accurate.

RESULTS

Table 1. RMS errors for choosing the Best Modeled Chemical in standardized data. Each column represents the RMS for its according variable column.

Index	Chemical Name	RMS errors
1	CO(GT)	0.4175
2	PT08.S1(CO) Tin oxide	0.3489
3	NMHC(GT) Non Metanic HydroCarbons	0.9335
4	C6H6(GT Benzene	0.1550
5	PT08.S2(NMHC) Titania	0.1333
6	NOx(GT)	0.4385
7	PT08.S3(NOx) Tungsten oxide	0.4835
8	NO2(GT)	0.5194
9	PT08.S4(NO2) Tungsten oxide	0.3603
10	PT08.S5(O3) Indium oxide	0.3407
11	Temperature	0.6202

Table 2. Values of RMS errors with 5- fold cross validation on the variable best explained by other variable with index 5 – Titania (PT08.S2(NMHC)).

Fold	1	2	3	4	5
rmstrain	35.2765	35.7330	35.5546	35.6259	35.5534
rmstest	36.7704	34.9121	35.6543	35.3102	35.6216

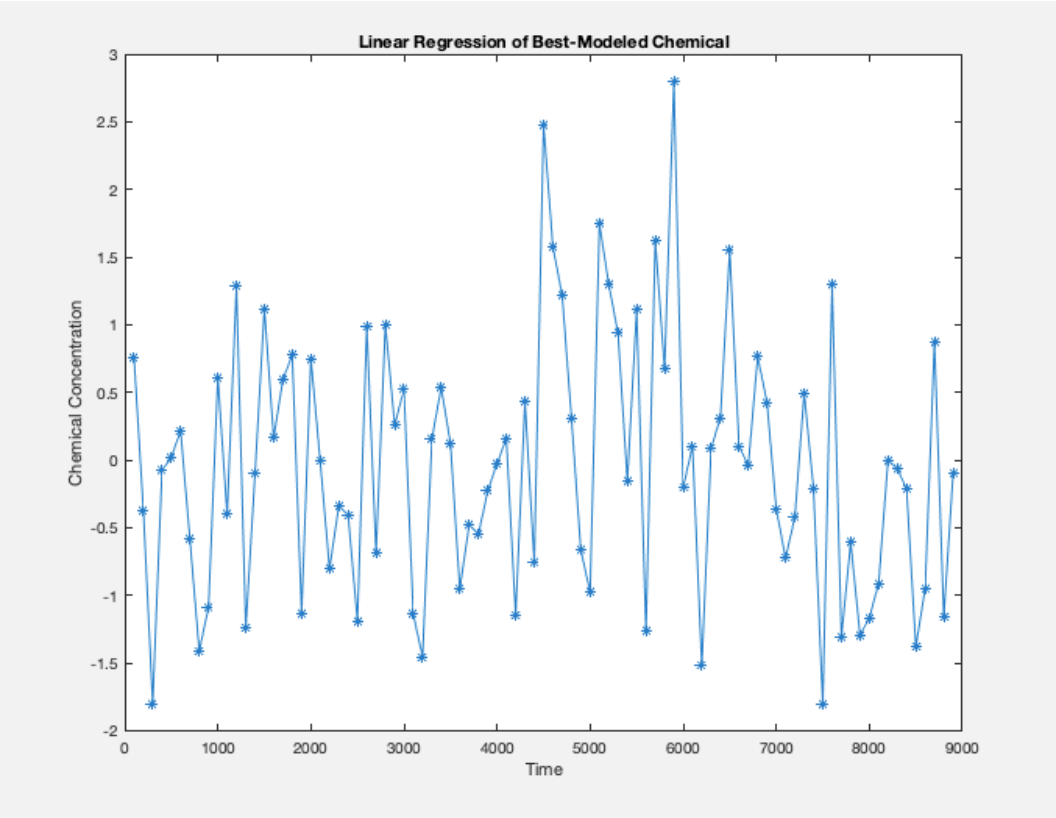

Figure 1. Plot of Linear Regression using the lowest RMS value variable as the dependent data. The plot is normalized with a center 0 and standard deviation 1.

DISCUSSION

The results demonstate proper validation of a linear regression model using a 5-fold cross validation. This is reflected in the values of RMS errors for training and testing folds with minor differences. A few choices are made to the original dataset that affect the results. These changes will be the main focus of the discussion section.

Preprocessing methods for the first task implemented are managing missing data, filling out missing values, and normalize data. Through previous trial and error, mishandling missing data can lead to changes in the results. Processing missing data and not normalize them result in lowndx of column 1. This is the column with the lowest value range compare to other columns in the original dataset. Thus, properly preprocess data avoid this bias caused by unit ranges. A similar observation happens within mykfold function where if column 1 is being tested for cross validation. Within mykfold function, unstandardized data are being used to run cross validation with no normalization, thus running cross validation results in significantly lower RMS values for average of 0.5996(rms train) and 0.60076(rms train) compared to the RMS values for lowndx column 5 with 35.5516(rmstrain) and

35.6168(rmstest). Thus, preprocessing of data can greatly alter the result achieved at the end of study through dealing with bias and variance within data.

With the choice to standardize data in the first task, this normalizes all data column with different value ranges. This choice results in RMS values within the range of 0 to 1. On a test run with unstandardized value, the range of RMS varies from 0.6 to 205, causing change in the lowndx to change to 1 instead of 5 with standardized values. This is because column 1 has the smalles range of value compared to any other column in the orginal dataset. This is the inherent bias that leads to column 1 having lower RMS value with unstandardized data.

With the choice to add an intercept term for the dataset, the result is varied when data is normalized and when it is not. Within the first task, when data is standardized, whether an intercept vector is added to the matrix or not, the RMS values stay in the same range. However, for the second task with unstandardized value, adding an intercept term significantly reduce the rmstrain and rmstest from average value 49.8116 down to 35.1884.

Cross validation successfully validates the result of linear regression model. Preprocessing data to avoid bias and variance is requried to get an accurate prediction from linear regression model.

REFERENCES

[1] Arroyo Á, Herrero Á, Tricio V, Corchado E, Woźniak M. Neural Models for Imputation of Missing Ozone Data in Air-Quality Datasets [Internet]. Complexity. Hindawi; 2018 [cited 2021Feb26]. Available from: https://www.hindawi.com/journals/complexity/2018/7238015/

[2] How to Deal with Missing Data [Internet]. Master's in Data Science. [cited 2021Feb26]. Available from: https://www.mastersindatascience.org/learning/how-to-deal-with-missing-data/#:~:text=Data may be missing due,the value of the observation.

[3] 7.2. Data Interpolation [Internet]. 7.2. Data Interpolation - Applied Data Analysis and Tools. [cited 2021Feb26]. Available from: http://faculty.salina.k-state.edu/tim/DAT/numeric/interp.html

[4] Ellis Rnady E. Class 12: Patterns – Linear Regression [unpublished lecture notes]. CISC 271: Linear Data Analysis, Queen's University; lecture given 2021 Febuary. Available from https://onq.queensu.ca/content/enforced/505646-CISC271W21/notes/Class12.pdf

[5] Ellis Rnady E. Class 13: Cross-Validation of Linear Regression [unpublished lecture notes]. CISC 271: Linear Data Analysis, Queen's University; lecture given 2021 Febuary. Available from https://onq.queensu.ca/content/enforced/505646-CISC271W21/notes/Class13.pdf

[6] Towards AI Team. How, when, and why should you normalize / standardize / rescale your data? [Internet]. Towardsai.net. Towards AI — The Best of Tech, Science, and Engineering; 2019 [cited 2021 Feb 26]. Available from: https://towardsai.net/p/data-science/how-when-and-why-should-you-normalize-standardize-rescale-your-data-3f083def38ff

[7] Ellis Randy E. Winter 2021 Assignment #2: Regression and Cross Validation, CISC 271: Linear Data Analysis, Queen's University; 2021 Febuary. Available from https://onq.queensu.ca/content/enforced/505646-CISC271W21/homework/A2.pdf

[8] Grace-Martin K. Interpreting the intercept in a regression model [Internet]. Theanalysisfactor.com. 2009 [cited 2021 Feb 26]. Available from: https://www.theanalysisfactor.com/interpreting-the-intercept-in-a-regression-model/