**Study the effect of training set noise over different regression algorithms and settings**

Alevizopoulou Sofia, AM: 2022201704002; Avgeros Giannis, AM: 2022201704003

Tsiatsios Georgios, AM: 2022201704023

Athens 2018, MSC Data Science, Applied Data Science

# **Abstract**

It is very common machine learning techniques to deal with noisy data, which may affect the accuracy of the resulting data models. For that reason, effective dealing with noise is a key aspect in machine learning to obtain reliable models from data. In this assignment, we address this issue by comparing how the noise affect linear, Lasso and Ridge regression on the final outcomes of 3 different datasets. We will compare the final outcomes from the original datasets with the outcomes from the noisy datasets. [ADD RESULTS OF REGRESSORS]

## **Keywords**

Attribute noise, noisy data, performance of regression algorithms, statistical test, noise effects

# 1. Introduction

Machine learning techniques and more specifically supervised learning techniques, are applied to extract novel and interesting information from data collected out of real-world problems. An important characteristic of these datasets is that the data frequently contains noise. Noisy data may be the consequence of human error due to mistakes in the translation phase or due errors while collecting them. Noisy data, may produce bias to the learning process, and as a result, is more difficult for learning algorithms to form accurate models from them. Developing learning techniques that effectively deal with noisy data is a key aspect in machine learning.

        In this assignment we provide a comparison of the effect of attribute noise and class noise on three regression models (a) Linear Regression, (b)Lasso Regression, (c) Ridge Regression. We are going to use 3 different datasets from different domains, each of them describes a different regression problem. The first dataset is the *Boston Housing* dataset thatconcerns housing values in suburbs of Boston. [4] The second is *air Qualit*y dataset that contains instances of hourly averaged responses from metal oxide chemical sensors embedded in an Air Quality Chemical Multisensor Device. [5] And the third dataset, *nba players* generated by us and contains info about the performance of nba players. We used these datasets to generate the “noisy” version of them. Finally, having the clean and the noisy version of each dataset, we compared the effect of noise on the models created (9 different models, for each dataset 2 different kind of noise).

Training a linear regression model say we try to find out coefficients for the linear function that best describe the input variables by choosing a function to help us measure the error, this function called cost function. In regression problems, we need evaluation metrics designed for comparing continuous values. Mean Square Error (MSE) is the most commonly used regression loss function. MSE is the sum of squared distances between our target variable and predicted values. We would like to minimize the MSE function as much as possible, so the prediction will be as close as possible to the ground truth by updating the coefficients.

The results enable us to highlight the characteristics of the different approaches. The whole analysis is developed from the following null (initial) hypothesis.

*Initial hypothesis:* Given the characteristics of each learning technique, we initially propose that Linear Regression is LESS/MORE  ---POIO EINAI TO NULL HYPOTHESIS MAS- THA TO GRAPSOUME GENIKA H SUGKEKRIMENA GIA KATHE REGRESSOR.”

The structure of the present paper is as follows: in Sect. [2](#_2._Related_work) summarization related work in noise analysis. In Sect. [3](#_3._Regression_algorithms) presentation of the three Regression algorithms and the learning techniques. Sect. [4](#_4._Design_of) description of the data sets, and explanation of the methodology used for the noise generation. Sect. [5](#_.5._Results_of) presents and analyzes the results of the different experiments.  Sect.[6](#_6._Hypothesis_testing) describe and run hypothesis tests based on null hypothesis.  Sect. [7](#_7._Conclusions) summarization the overall conclusions.

# 2.Related work

        Understanding the impact of noisy data in the performance of machine learning algorithms is a key issue for improving algorithms reliability. Below will be described some techniques of noise generation. In current researches, some machine learning techniques are considered more “robust” to noise, errors and missing values than others.

[DO WE NEED MORE ??]

## 2.1.   Noise Generation

Noise generation can be described in different ways. First, it is very important where the noise is introduced. Noise can be introduced in the input attributes or/and the target feature of the data (feature because it is a regression problem). Second, the distribution that noise follows, for example, normal or Gaussian. Third, the generated noise values can be relative to min, max, std (standard deviation) of each variable or to the variable value itself. Noisy training data will impact on the outcomes of the final level.

We have implemented 1 python script which generates 2 noisy datasets – all features are noise, target feature is noise - based on an input dataset. The process is described below.

### 2.1.1. Attribute noise

This kind of noise adds noisy data at a specific feature of the train dataset. We don’t modify test data. First, we define the percentage of noise which can be between 0%-100%, let that be the Fc variable. After using this percentage, we find out how many random variables will be generated by multiplying the percentage of noise with the number of instances, so we have N random values. Now, for each variable of this attribute, we generate N random numbers Rv with a Gaussian distribution and within a range between the max and min of the corresponding variable. Then we generate another random numbers Rc with a uniform distribution within the range 1 to the number of instances, which indicates the instance whose data value should be overwritten by the generated noisy value. [1]

Specifically, we assign create a dictionary of lists. Each element of the dictionary is of the type {Feature name: list}, where each list contains N (number of instances) elements. Each element is assigned a probability, which represents the numerical probability of the record’s feature value to be altered by adding noise. We achieve by that to introduce noise per feature.

For example, if we assign a probability of 0.1% to add noise, that means that values produced out of the gaussian distribution have a probability (given by the beforementioned dictionary of lists) that must be below that threshold in order for the noise value to be added into the records feature value. Adding value per feature (and not per whole record) simulates at best the real case scenario of a non-simulated dataset.

### 2.1.2. Target feature noise

This kind of noise adds noisy data at the target feature of the train dataset. We don’t modify test data at this case also. The process is the same as before, as we seek to generate noise values for one feature instead of the multiple features we did on the previous feature case.

## 2.2 Stratified split for test train dataset

We wanted to create two different datasets for the implementation of noise effect. The first dataset would be the train one and the second one would be the test one. The test dataset will not contain any noise as stated previously. For the train dataset, we create two different versions of it, one with noise for the attribute and one with noise for the class dataset.

What is important here is the fact that we needed to perform the split in a stratified way. Thus, we wanted for the test and the train distribution to be as similar as possible. Due to the fact that for regression purposes, sciikit learning was not able to perform the split, we decided to make it manually, implementing it as explained below:

1. We sort the dataset based on the target feature
2. We split it the sorted dataset to lists of 10 items.
3. The proportion of split to test and train is the number of items we select out of the 10 items. We decided to split the dataset based on 80 % train – 20% test, so the number of elements taken from the sorted 10 items would be 2 items. Those 2 items would be added to the test dataset, whereas the remaining 8 are part of the training dataset.

As we see by that, we force with the way we create the datasets that the distribution of the train dataset would be preserved and be present also to the test dataset.

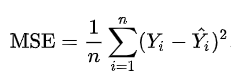
# 3. Regression algorithms

## 3.1    Regression techniques

Regression analysis is a form of predictive modelling technique which investigates the relationship between a dependent and independent variable. This technique is used for forecasting, time series modelling and finding the curve / line to the data points, in such a manner that the differences between the distances of data points from the curve or line is minimized.

        One of the first steps in a regression analysis is to determine if multicollinearity is a problem. Multicollinearity, or collinearity, is the existence of near-linear relationships among the independent variables. Multicollinearity causes two basic types of problems. The coefficient estimates can swing rapidly based on which other independent variables are in the model, so coefficients become very sensitive to small changes in the model. Multicollinearity reduces the precision of the estimate coefficients, which makes the statistical power of the regression model weak.

In regression problems, we need evaluation metrics designed for comparing continuous values. We would like to minimize the MSE (loss function) for a specific data-point as much as possible so the prediction will be as close as possible to the ground truth. This is done by learning the parameters, executing an iterative process that updates coefficients at every step and reduce the loss function as much as possible until the minimum point of the loss function has been reached.



Where n: number of instances, Y^: predicted values, Y: observed values

## 3.2.   Linear regression

Linear Regression establishes a relationship between dependent variable (Y) and one or more independent variables (X) using a best fit straight line. The dependent variable is continuous, independent variable(s) can be [continuous or discrete](https://en.wikipedia.org/wiki/Continuous_and_discrete_variables), and the regression line is linear.

Its equation is *Y = a+b1X1+b2X2+b3X3+…*

where Y: the response, a: the intercept, b: model coefficients, Xn (the nth feature)

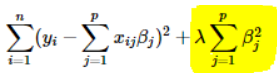
This equation can be used to predict the value of target variable based on given predictor variable(s). We would like to minimize the MSE (loss function) for a specific data-point as much as possible. Linear Regression is very sensitive to Outliers. Having noisy data can influence on the computation of MSE and as a result the regression line. Prediction errors can occur due to two sub components or combination of them. First component is the bias and the second is the variance (multicollinearity problem). [2]

## 3.3.   Ridge Regression

Ridge regression adds “squared magnitude” of coefficient as penalty term to the loss function, it is known as L2 regularization. It is another regression method and used when the data suffers from multicollinearity, independent variables are highly correlated. In multicollinearity, even though the least squares estimates are unbiased, their variances are large. Large variances mean that the observed value is far from the true value. Adding a degree of bias, standard errors are reduced. Ridge regression solves the multicollinearity problem through [shrinkage parameter](https://en.wikipedia.org/wiki/Shrinkage_estimator) λ (lambda). Ridge regression is a well-known approach to dealing with noisy data. [4]

Its equation is Y=a+b\*X+ε

where Y: the response, a: the intercept, b: model coefficients, Xn (the nth feature), ε: the error. Error term is the value needed to correct the prediction error between the observed and predicted value. The highlighted part below represents L2 regularization element. [6]



If *lambda* is very large then it will add too much weight and it will lead to under-fitting, if *lambda* equals to zero then we have linear regression. We can understand that it’s important how *lambda* is chosen. Minimizing the sum of the squares of the vertical deviations from each data point to the line enforce β coefficients to be lower but it does not enforce them to be zero.

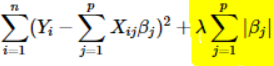
## 3.4.   Lasso Regression

Lasso Regression (Least Absolute Shrinkage and Selection Operator) adds “*absolute value of magnitude*” of coefficient as penalty term to the loss function, its known as L1 regularization. It penalizes the absolute size of the regression coefficients, reduces the variability and improving the accuracy of linear regression models.

Lasso regression differs from ridge regression in a way that it uses absolute values in the penalty function, instead of squares. This leads to penalizing values which causes some of the parameter estimates to turn out exactly zero. Larger the penalty applied, further the estimates get shrunk towards absolute zero. Therefore, you might end up with fewer features included in the model than you started with, which is a huge advantage. Lasso regressor is not equipped to deal with noisy or missing data. [2]

Its equation is Y=a+b\*X+ε

where Y: the response, a: the intercept, b: model coefficients, Xn (the nth feature), ε: the error. Error term is the value needed to correct the prediction error between the observed and predicted value. The highlighted part below represents L1 regularization element. [6]



If *lambda* is a large value will make coefficients zero hence it will under-fit, if *lambda* equals to zero then we have linear regression. The key difference between Lasso and Ridge techniques is that Lasso shrinks the less important feature’s coefficient to zero thus, removing some feature altogether. So, this is effective for feature selection in case we have a very big number of features.

# 4. Design of the dataset

In our experiments we will use the 3 different datasets. The two of them: boston-houses, air-quality are available on web and the third is generated by us: nba-players. The nba-players dataset designed to have no noise. The other datasets as they are based on real data (data descriptions of real objects or events), may contain a certain amount of background noise, erroneous values, and so on. All the data sets have as target numerical attributes. Forced noise will be added only at the train data. Before adding noise, we have splitted the dataset into train and test.

## 4.1. Boston houses

The Boston dataset contains information collected by the U.S Census Service concerning housing in the area of Boston Mass. The dataset contains a total of 506 instances and 14 attributes. The 14 features give several information for the specific area, eg: the feature CRIM describes the crime rate by town. At this case, we try to find the best regression model at which the median value of houses will be predicted.  There is a readme file available at which there are detailed information about the dataset. [7]

As said before, this is a real dataset so for sure it contains noisy data. Because we don’t know exactly the percentage of noise we generated another two datasets based on this at which we will include 5% noise. The first generated dataset contains noise data at the target feature -MEDV (median value of houses) - whereas the other contains noise data at all features of the dataset except the target one. For example if the dataset contains 100 instances and 3 features (one of them is the target), 5% noise on the target feature will overwrite target value on 5 instances, whereas 5% noise on features in general will changes 5 instances at both features. The generation of noise has been analyzed on Sect. [2.1.](#_2.1.__)

## 4.2. Air-Quality

The air quality dataset contains 9358 instances of hourly averaged responses from an array of 5 metal oxide chemical sensors embedded in an Air Quality Chemical Multisensor Device. The device was located on the field in a significantly polluted area, at road level,within an Italian city. Data were recorded from March 2004 to February 2005 (one year) representing the longest freely available recordings of on field deployed air quality chemical sensor devices responses. At this case, we try to find the best regression model at which the predict the amount of C6H6(GT) will be predicted.  There is a readme file available at which there are detailed information about the dataset. [8]

As said before, this is a real dataset so for sure it contains noisy data. Because we don’t know exactly the percentage of noise we generated another two datasets based on this at which we will include 5% noise on target feature and 5% noise at all features of the dataset. The first generated dataset contains noise data at the target class - C6H6(GT) - whereas the other contains noise data at all features of the dataset except the target one. The generation of noise has been analyzed on Sect. [2.1.](#_2.1.__)

## 4.3. Nba-players

The nba-players dataset contains information about the statistics of nba-players and made by us. More specifically each feature is a different metric of performance. GFor example 3-point, 2-ponit etc. The information has been selected from web. The target feature is the salary of the player. At this case, we try to find the best regression model at which the predict of salary will be predicted.  There is a readme file available at which there are detailed information about the dataset. There is a readme file available at which there are detailed information about the dataset. [9]

As said before, this is a dataset generated by us there are no noisy data. We generated another two datasets based on this at which we will include 5% noise. The first generated dataset contains noise data at the target class - SALARY - whereas the other contains noise data at all features of the dataset except the target one. The generation of noise has been analyzed on Sect. [2.1.](#_2.1.__)

# 5.Results of the experiments

results for all datasets all algorithms

# 6. Hypothesis testing

# 7. Conclusions

# References

1. A study of the effect of different types of noise on the precision of supervised learning techniques - David F. Nettleton · Albert Orriols-Puig · Albert Fornells, <https://link.springer.com/article/10.1007/s10462-010-9156-z>
2. Noisy and Missing Data Regression: Distribution-Oblivious Support Recovery - Yudong Chen, Constantine Caramanis, http://proceedings.mlr.press/v28/chen13d.pdf
3. The Truth About Linear Regression 36-350, Data Mining 21 October 2009,  <https://www.stat.cmu.edu/~cshalizi/350/lectures/17/lecture-17.pdf>
4. Scattered Data Approximation of Noisy Data via Iterated Moving Least Squares - Gregory E. Fasshauer and Jack G. Zhang, <https://pdfs.semanticscholar.org/9b7d/891601e006b85a3f49ef432f35524aa2a328.pdf>
5. Machine Learning Algorithms, a study of noise sensitivity – Elias Kalapanidas, Nikolaos Avouris, Marian Craciun, Daniel Niagu, <http://delab.csd.auth.gr/bci1/Balkan/356kalapanidas.pdf>
6. A robust hybrid of lasso and ridge regression Art B. Owen Stanford University October 2006, http://statweb.stanford.edu/~owen/reports/hhu.pdf

#### Datasets

1. <https://github.com/rupakc/UCI-Data-Analysis/tree/master/Boston%20Housing%20Dataset/Boston%20Housing>
2. <http://archive.ics.uci.edu/ml/datasets/air+quality>
3. https://github.com/JoHNNyB92/applied/tree/master/nba\_players

#### Useful Links

1. <https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/>
2. <https://codingstartups.com/practical-machine-learning-ridge-regression-vs-lasso/>
3. <https://sci2s.ugr.es/noisydata>
4. <https://www.kellogg.northwestern.edu/faculty/dranove/htm/dranove/coursepages/Mgmt%20469/noisy-variables.pdf>
5. [**https://www.theanalysisfactor.com/assessing-the-fit-of-regression-models/**](https://www.theanalysisfactor.com/assessing-the-fit-of-regression-models/)