Fundamentals of Data Analytics
Assignment 1

Wesley van der Lee and Mourad el Maouchi

May 2016

1 Introduction

This report concludes our conducted research for the first assignment of the course Fundamentals of Data Analytics (SPM4450). We will present a series of experiments similar to those presented in the section on Prediction Error in Regression by Berthold and Hand [3] and determine whether we can draw the same conclusions.

2 Approach

The goal of data analysis is to gain insight and information from a training sample $T = \{(x_1, y_1), (x_1, y_1), ..., (x_m, y_m)\}$ and determine a linear or polynomial equation which best fits the sample data points, while also not being too sensitive to small changes in the training set. The latter two issues introduce the notion of the bias-variance dilemma. The bias-variance dilemma is the problem of simultaneously minimizing two sources of error that prevent supervised learning algorithms from generalizing beyond their training set:

1. The bias, which describes how accurate a model is across different training sets, should be as low as possible. A high bias may cause the model to underfit.

2. The variance, which describes how sensitive the model is to small changes in the training set, should also be as low as possible. A high variance may cause the model to overfit.

Keeping the bias and variance as low as possible contributes to finding the model with the 'right' complexity. The model with the lowest complexity would be linear and could be derived using linear regression using a least-square approach [4]. If the true relationship of the data points would be of a complexity higher than linear, then this model would always introduce errors, due to it's biasness and hence underfits other data. The bias is then said to be very high. In the other extent, we can also consider a model with a complexity which is too high. A model could overfit the training data, meaning that the model accurately describes the training data, even the noise. A worst case example would be that a training set of n data points would have a corresponding of degree $n - 1$ and hence be very sensitive to small changes. This model has said to have a high variance.

The motivation for finding a model with the right complexity is to predict other (future) points as close as possible to the true relationship of the data points, disregarding the noise as much as possible. To determine whether the model has the right complexity, one could for example switch upon different models of complexity and determine for each model the combination of the bias and variance, which together sum up to the mean squared error (MSE).

The MSE is an approach that assesses the quality of the estimator, which in this case is the regression model. For this MSE there exists a relationship between the bias and variance. From the equation, shown below, it can be rewritten in such a way that it can be expressed in terms of the

variance and the bias squared. MSE reflects both the bias (accuracy) of the estimator, as well as the variance (precision) of the estimator. That means, how much its expected value deviates systematically from the true value (bias) and how much it varies about its expected value due to sampling variability (variance). The approach is used to determine how well the gained estimator actually matches with the data points used.

$$E_T[(f(x) - \hat{f}(x|T))^2] = (f(x) - E_T[\hat{f}(x|T)|)^2 + E_T[(\hat{f}(x|T) - E_T[\hat{f}(x|T)])^2]$$

Models of complexity where can be switched upon can be for example:

- **Linear:** $E(Y) = f_1(x) = \beta_0 + \beta_1 x$
- **Quadratic:** $E(Y) = f_2(x) = \beta_0 + \beta_1 x + \beta_2 x^2$
- **Cubic:** $E(Y) = f_3(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$

3 Set-up

This section describes the applied method in our research. For different types (of polynomial) equations $f_i(x)$, we will generate 1000 training sets of data T_i. The generated data will have some noise, and will thus be subjected to a variance $\sigma_\epsilon^2 = 1$. For each training set $T_{i,j}$ the parameters $\hat{\beta}_k$ can be estimated using linear regression based on a least squares approach. Each data sample $T_{i,j}$ will thus have a set of parameters $\hat{\beta}$ for a i) linear, ii) quadratic or iii) cubic polynomial, which fits the data sample the best.

Once that, for each degree of model, the parameters have been estimated, we are able to calculate the squared bias and the variance, as compounded by E_T. The bias is then the difference between the best prediction $f(x)$, and its average estimate over all possible samples of fixed size. The variance is the expected squared difference between an estimate obtained for a single traning sample and the average estimate obtained over all possible samples [3].

The entire process of sample creation, polynomial regression parameter estimation, calculation of estimates, bias and variance has been automated in a Java in order to avoid redundant work. The script has been published on Github [1]. The script for polynomial regression for parameter estimation has been lend and modified from here [2]. Moreover the results of the polynomial regression have been checked and validated with Excel, so we know that the written code works.

We will conduct this experiment for the following formulae:

- $f_1(x) = 2 + \frac{1}{4}x$
- $f_2(x) = e^x$
- $f_3(x) = \frac{x}{7}4 + sin(x)$

4 Results

This section describes the retrieved results. We devoted one subsection per formula.

4.1 Formula: $f_1(x) = 2 + \frac{1}{4}x$

The formula, $f_1(x)$, is obviously a linear formula or in other words, a 1^{st} degree polynomial. We would therefore assume that this type of formula would have the lowest mean squared error using linear regression. The motivation for starting off with such an easy example, is to get the script up and running and to quickly test if we can get the same results. We generated 1000 data samples

according to $Y_i \sim N(\mu = 2 + \frac{1}{4}x_i, \sigma_\epsilon^2 = 1)$ on a domain of $[0, 10]$. Based on this, $1^{st}, 2^{nd}$ and 3^{rd} degree polynomials are created and tested against the data samples by calculating the MSE for each one. Together with the eventual mean squared error, the squared bias and variance are depicted in the table below. As can be seen, the MSE for the linear regression model is the lowest and thus fits $f_1(x)$ the best. This is a logical result, because the function used is also linear.

	Squared bias			Variance			MSE		
m	10	100	1000	10	100	1000	10	100	1000
Linear	0.206	0.037	0.001	0.124	0.152	0.157	0.330	0.187	0.158
Quadratic	0.212	0.038	0.001	0.222	0.236	0.237	0.434	0.274	0.238
Cubic	0.288	0.039	0.002	0.267	0.302	0.319	0.555	0.340	0.321

Table 1: MSE Results $f_1(x)$

One can also see that when considering different numbers of data samples, the mean squared error changes accordingly. It becomes smaller as more training data sets are used. This follows logically from the literature, in such a way that the mean squared error will be lower as the average estimator function given the training data set will be more accurate.

4.2 Formula: $f_2(x) = e^x$

In calculus, e^x can be written as the infinite series $1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + ... + \frac{x^n}{n!}$ for n nearing infinity. The formula $f_2(x)$ can therefore best be described using an infinite-degree polynomial. Since such an exploration would be infeasable to show with our current set of experiments, we will investigate the mean squared error if we would also calculate the bias and variance for 4^{th} and 5^{th} degree polynomials. The same approach as for $f_1(x)$ is used to determine the squared bias, variance and MSE. The results are shown below:

	Squared bias			Variance			Mean square error		
m	10	100	1000	10	100	1000	10	100	1000
Linear	2.352E8	2.352E8	2.352E8	1.960E7	1.960E7	1.960E7	2.548E8	2.548E8	2.548E8
Quadratic	8.104E7	8.104E7	8.104E7	0.675E7	0.675E7	0.675E7	8.780E7	8.780E7	8.780E7
Cubic	1.871E7	1.871E7	1.871E7	0.156E7	0.156E7	0.156E7	2.027E7	2.027E7	2.027E7
4-degree	0.293E7	0.293E7	0.293E7	0.244E7	0.244E7	0.244E7	0.317E7	0.317E7	0.317E7
5-degree	0.031E7	0.031E7	0.031E7	0.003E7	0.003E7	0.003E7	0.033E7	0.033E7	0.033E7
6-degree	0.002E7	0.002E7	0.002E7	0.002E7	0.0002E7	0.0002E7	0.0002E7	0.002E7	0.002E7

Table 2: MSE Results $f_2(x)$

As can be seen from Table 2, the higher the degree of the polynomial regression model, the lower the MSE value is. The squared bias, variance and even the MSE get lower with every degree. This was expected, since $f_2(x)$ is best described as an infinite-degree polynomial. It is only logical that the MSE decreases with the amount of degrees getting higher. Therefore, the higher the degree of the polynomial regression model, the better it describes the function $f_2(x)$.

Another observation we can make from the retrieved data is that the squared bias and varias non-significantly differ throughout a different number of data samples. This is probably related to the corresponding range of the actual function.

4.3 Formula: $f_3(x) = \frac{x}{4} + sin(x)$

Another point of influence we want to investigate is the influence of the data domain. Depending on the size of a domain, a variable $sin(x)$ can be best estimated with a linear model in the case of a large

domain, a quadratic model in the case of a domain $[0, 2\pi(\sim 3)]$, a cubic model in the case of a domain $[0, 3\pi(\sim 6)]$, etc. To thus also emphasize this importance, we will also show this with an example. Since the importance is not about showing the behaviour of underfitting, we will disregard the 10 and 100 number of data samples.

Using a least squares approach based on all polynomial regression models, we constructed the estimator which gives on average for all values in a certain domain, the best estimation for $f_3(x) = \frac{x}{4} + sin(x)$. This again is done for a linear, quadratic and cubic model, and the results are shown below in table 4.3.

	Domain: [0,25]	Domain: [0,6]	Domain: [0,3]
Linear	$y = 0.23x + 0.22$	$y = 0.03x + 0.60$	$y = 0.3x + 0.38$
	bias: 3.13; variance: 0.48	bias: 1.8; variance: 0.46	bias: 0.65; variance 0.5
Quadratic	$y = -3.06x^2 + 0.23x + 0.18$	$y = -0.03x^2 + 0.23x + 0.44$	$y = -0.38x^2 + 1.45x$
	bias: 3.17; variance: 0.52	bias: 1.73; variance:0.58	bias:0.002; variance: 0.57
Cubic	$y = 0.005x^3 + 0.02x^2 + 0.07x + 0.5$	$y = 0.09x^3 - 0.82x^2 + 1.97x - 0.08$	$y = -0.01x^3 - 0.33x^2 + 1.4x$
	bias: 2.52; variance: 0.48	bias: 0.04; variance:0.49	bias: 0.002; variance: 0.759

Table 3: Estimators for $f_3(x)$

From the retrieved data we can immediately validate the earlier stated statement. On the large domain, the linear function $y_{linear} = 0.23x + 0.22$ outperformes the quadratic function in terms of it's mean squared error. This is mainly due to the fact that on a domain of $[0, \infty]$ the best estimator would be $y = 0.25x$ as the sinus function itself hovers around 0. Surprisingly on the domain of $[0, 25]$ the cubic model has a lower bias compared to the linear model, probably because a cubic model can better scope with the edge created by the $\frac{x}{4}$ factor.

On a somewhat smaller domain, the domain $[0, 6]$ the sinus function would behave just like a third degree polynomial, with two rising edges and one falling edge. Note that the cubic model better fits the data samples retrieved from the $[0, 6]$ domain than compared to the $[0, 25]$ domain, as the bias becomes 63 times smaller. This all acts like expected. On the last and smallest domain analyzed, the domain $[0, 3]$, we expected that the quadratic function would best fit the data sample, as only on this domain, the data looks parabolical distributed. Considering only the quadratic and the cubic polynomials, they have an equal bias but differ hugely in variance. The cubic polynomial has a higher variance because it just is too complex for the parabolically distributed data, for which a quadratic function would suffice. Again, this confirms our hypothesis that a quadratic model would suffice on this interval, but it all depended on the chosen domain to show certain characteristics of the true function.

5 Conclusion

In this report, we conducted a research in an attempt to draw the same conclusions as done in the section on Prediction Error in Regression by Berthold and Hand [3]. Based on three formulas defined, and the regression models created for them, we calculated the squared bias, variance and eventually the mean squared error. In these experiments the true formulae are known and henceforth the bias-component could be truly calculated. For the first linear formula we conclude that the linear regression model indeed fits it best, as for this model the mean squared error was the lowest. For the second formula, it is concluded that a higher degree polynomial describes the function $f_2(x)$ best. Since the true form of $f_2(x)$ can be calculated according to an infinite-degree polynome, the error would only be less by applying additional complexity. Too show this, we even shown this fact for 4, 5, and 6-degree polynomes and saw the mean squared error only decreasing. We can henceforth conclude that the accuracy will increase as complexity will be added. In our opinion this will not hold forever, due to the static and non-decreasing variance $\sigma_\epsilon^2 = 1$. There will be a point at which an increase in complexity

will not provide additional accuracy, because the precision is within the variance boundary. It is interesting to investigate by using the bias-variance decomposition at what degree of complexity such occurrences will occur, but this is however not within the scope of this assignment. Thirdly we saw that a given domain shows characteristics of a true function, and based on the chosen domain, one can find different models of different complexity which fits the data sample the best. This has been researched with an example based on a sinus function.

A general conclusion can be made from these results. In our case it is that we gained the same results as done by Berthold and Hand. This confirms their approach and the fact that regression models, of different orders, can be derived from sample data and the best can be concluded by using the Mean Squared Error approach. A model of a higher complexity can lead towards a higher error, because of the variance due to the notion of overfitting the data. This becomes visible when higher degree polynomials are compared to models with a lower complexity. While determining the variance, the number of data samples also come into play, as the more data samples are used, the more robust an estimator will be to certain changes, and will thus experience a conveniently lower variance. A model should on the other hand also be not of a too low complexity, as it will then become biased. As we have seen, this tradeoff, better known as the bias-variance dilemma, can be investigated through trial and error over different regression polynomes with different complexity. The investigation is only trivial though if there is enough data available.

References

[1] `https://github.com/wesleyvanderlee/FDA`.

[2] `https://gist.github.com/sanity/234253`.

[3] Michael Berthold and David J Hand. *Intelligent Data Analysis, An Introduction*. Springer Science & Business Media, 2007.

[4] John Neter, Michael H Kutner, Christopher J Nachtsheim, and William Wasserman. *Applied linear statistical models*, volume 4. Irwin Chicago, 1996.