FYS-STK 4155 project 1

”When you are solving a problem, don’t worry. Now, after you have solved the problem, then that’s the time to worry.” - Richard Feynman

Introduction

The amount of data being produced in today’s society is ever growing within all scientific areas. Making predictions and estimations with the use of data has become so big that this branch of scientific approach has been granted a name of it’s own. This buzzword is known as data science. The use of data science is applicable in many fields from physics, biotechnology to engineering of self-driving cars to name but a few. Interpreting these data play an important role in the end result. Not being aware about of the pitfalls may cause crucial and misleading results.

In Japan earthquakes tend to occur frequently as do tidal waves, also known as tsunamis. These natural occurrences are therefore necessary to take into account when building constructions. In Stacey (2015) regression analysis was applied and attempted to predict the structural and safety measures needed for a power plant. Looking at historical data dating back at 1600 CE it was concluded that the appropriate model was a polynomial rather than a linear one. This resulted in the construction not being able to withstand a earthquake of magnitude 9 registered in march 2011. The reason for this was that the polynomial model had estimated the building to withstand earthquake of a magnitude up to 8.6. The outcome of this had large impact on Japan’s economy.

Within the field of data science is Machine Learning (ML) a word which is usually mentioned when talking about the applications of data science. The goal of ML is to make certain predictions given a set of data. When talking about ML two types of approaches are common. In some cases the output of the model is known, in other cases not. These two types are called supervised and unsupervised learning, respectively. Here we will be looking at the supervised learning type. In supervised learning the goal is to predict a model given a set of training data and make predictions on a set of test data. Supervised learning can be grouped into regression and classifications problems. From Brownlee (2016) is the distinction: ”A classification problem is when the output variable is a category, such as “red” or “blue” or “disease” and “no disease” and ”a regression problem is when the output variable is a real value, such as “dollars” or “weight”.”

In this project, the aim is to do a polynomial fit on a model using several regression types given two sets of data. Data drawn randomly and a real data set. A set of quantities will be derived in order to see what the benefits of each method is, also to make the fit as good as possible. These quantities are the mean squared error(MSE), variance, R2 score and bias. The bias and variance are related to how good the fit of the model is i.e do we need to fit the model with a higher degree of polynomial or lower. This is known as under and over-fitting, respectively. The R2 score is a statistical measure explaining the correlation between variables. Depending on the field in which one studies does the R2 score not necessarily give one wanted value over all studies. In this case, how close the data are to the fitted model will be the R2 score. The value ranges from 1 being good fit and 0 being bad fit. Regression methods will include ordinary least squares(OLS), ridge and lasso. The function we will be looking at is the Franke function which is a two-dimensional function. For the second model have we been awarded terrain data over Norway where the same implementations of regression will be made. This will hopefully grant some clarification in what machine learning constitutes. The remaining of the report is arranged as follows: II method, III results, IV discussion and V conclusion.

References

[1] Stacey Brian,*The failure of predictive models*, 2015

[2] Mehta Pankaj, Wang Ching-Hao, Day Alexandre G.R and Richardson Clint, A high-bias, *low-variance introduction to Machine Learning for physicists*, 2018

[3] Frost Jim, *Regression Analysis: How Do I Interpret R-squared and Assess the Goodness-of-Fit?*, 2013

[4] Brownlee Jason, *Supervised and Unsupervised Machine Learning Algorithms*, 2016