Predicting the amount of UVB radiation from air temperature with polynomial regression

Introduction

Ultraviolet radiation is electromagnetic radiation which has a wavelength from 10 nm to 400 nm, just below the wavelength of visible light. Ultraviolet radiation can be categorized into three different types: UVA, UVB and UVC depending on its wavelength. UVB is the main source of sunburns. It can also directly damage the DNA of skin cells thus increasing the risk of developing skin cancer.

In this project I want to find out how the midday air temperature affects the amount of midday UVB radiation. The air temperature, unlike the UVB radiation level, can often be found from various sources: mobile applications, newspapers or subway screens. Understanding the relation between air temperature and UVB exposure is helpful as a heuristic for assessing the need to apply sunburn-preventing measures.

In *Project Formulation*-chapter I will introduce the dataset of the project. It is followed by the *Method*-chapter in which I discuss the model which I've chosen and how I apply it to the dataset. I present the results in the *Results*-chapter and summarize my project in the *Conclusion*-chapter. The references used in this project are listed in *References*.

Project Formulation

Each data point represents a day and consists of a feature: midday air temperature, and a label: the amount of midday UVB radiation. Both measurements are obtained from Kumpula measurement station in Helsinki. The data is collected by the Finnish Meteorological Institute (FMI, 2021). I use data from the year 2019 so in total my data consists of 365 data points. In this project I'll apply polynomial regression to find the relation between the midday air temperature and the midday level of UVB radiation.

Method

We can observe from the scatter plot of the dataset in figure 1 that the relation between the data points appears to be dependant of a polynomial of a small degree. I will use polynomials with degrees { 1, 2, 3, 4, 5, 10, 15, 25 } as my models for the project. I predict that higher degrees will overfit to the data and thus exploring higher degrees of polynomials than 5 will probably not be useful. I will still train the model with them to see their error and to test my hypothesis.

The data points of the days with air temperature greater than 5 degrees Celsius vary a lot. For example some days with air temperature 17°C have almost no UVB radiation as the UVB index is close to 0 whereas some other days with the same air temperature have the UVB index close to 5 which is close to the maximum value reached during the whole year. This could be caused by other meteorological factors, like cloudiness, affecting the level of UVB-radiation more than it affects the air temperature. A possible improvement would be to use multivariate polynomial regression with cloudiness and air temperature to predict the amount of UVB-radiation.

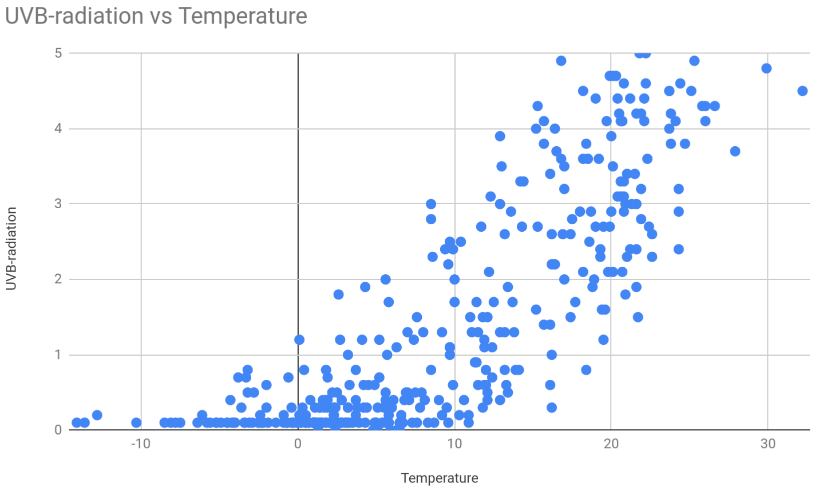

Figure 1 Midday air temperature and UVB plotted in a scatterplot

Another method of improving the prediction could be to cluster the data. The data points in the range [-20, 5] seem to increase less as a function of the air temperature whereas the data points with air temperature > 5 seem to be more greatly affected by the air temperature. From this observation it is be possible to try dividing the data into two clusters $D_{T<=5}$and $D_{T>5}$to improve the quality of the regression.

In this project I will split the data into three sets: the training, the validation and the test set. The data is split into three sets by using single splitting twice. This way the datapoints are divided randomly between the sets. The training set is used to train the model. Then the model is used on the validation set resulting in a validation error. The model that performs the best on the validation set by achieving the lowest validation error is then used on the test set to measure the performance of the model. The sizes of the sets are following: training set: 191, validation set: 64 and test set: 110 data points summing up to the total of 365 data points downloaded from the Finnish Meteorological Institute.

As the loss function I use the squared error loss since the labels are numeric. According to Jung (2021) p.39 the squared error loss is a good first choice for loss function in the context of regression loss with numeric labels. Squared error loss for a labeled data point (x, y) using the hypothesis $h\colon R\ -> \ R$is calculated with formula (1) where x is the feature and y is the label. I use the loss function to calculate the mean squared error (2) for each hypothesis, which I use to compare the performance of the model.

$$L((x,y),h) := (y - h(x))^2 \tag{1}$$

$$1/m * \sum_{i=1}^{m} (y_i - h(x_i))^2 \tag{2}$$

I included the R2-score as it was recommended by a teaching assistant. The R2-score is used as an additional measure of how well the observed outcomes are replicated by the model. According to Scikit-learn (Pedregosa *et al.*, 2011) documentation the R2-score measures the proportion of variance of y that has been explained by the independent variable x in the model. The best R2-score is 1 and the score can be arbitrarily negative.

Polynomial degree	1	2	3	4	5	10	15	25
Training error	0.6494089012	0.562979046	0.5414901124	0.5343412778	0.5295050807	0.525626238	0.5341712023	0.8400245777
Validation error	0.7343725579	0.6261864993	0.6539815906	**0.5979425611**	0.8406222188	123.4398314	15409.25714	4979298.529
Training R2	0.7135293772	0.7516557632	0.7611350731	0.7642886041	0.7664219724	0.7681330277	0.7643636287	0.6294440014
Validation R2	0.6518465623	0.7031357177	0.6899585415	**0.7165256845**	0.6014754199	-57.5207075	-7304.264595	-2360599.054

Table 1 The results of polynomial regression

Results

From the table 1 we can observe that the smallest validation error is achieved with polynomial regression with maximum degree of 4. The smallest validation error, which is bolded in the table, is approximately 0.598. The greatest R2-score of approximately 0.717 is also achieved by the 4th degree polynomial. When using the trained model of 4th degree polynomial on the test set we achieve a mean squared error of approximately 0.607 which is not significantly greater than the validation error. The R2-score for the test set is 0.734. From these results we can conclude that it is possible to use air temperature to predict the amount of UVB radiation.

From table 1 we can observe that the validation error increases significantly with polynomial degrees 10, 15 and 25. This is due to overfitting, which is visualized in scatterplot of the model and the whole dataset in Figure 2.

Contrary to expectations the training error also increases slightly when using polynomials of degrees 15 and 25. Theoretically using higher degree polynomials should reduce the training error. In this case the slight increase of training error can be due to some limitation of the computing models, for example the inaccuracies that are inherent with the floating point number representation. Being multiplied many times, these small errors can compound and accumulate to greater inaccuracy in the model.

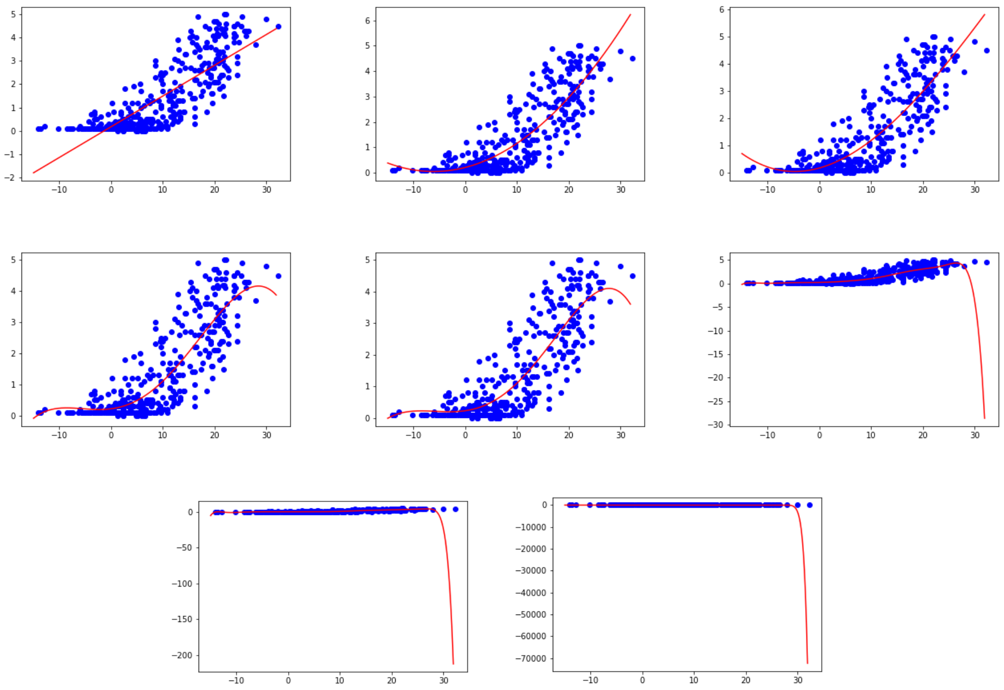

Figure 2 the model and the dataset visualized in a scatterplot

Conclusion

In this project I applied polynomial regression to predict the amount of UVB radiation from air temperature. I used a dataset consisting of 365 datapoints sourced by the Finnish Meteorological Institute. Each data point consisted of midday air temperature and midday UVB intensity measured in Kumpula, Helsinki. The data points correspond to days of the year 2019.

The results of this project show that it is possible to predict the amount of UVB radiation from the air temperature with polynomial regression. The best results were achieved by fitting the data with a 4th-degree polynomial. In the experiments I used polynomials of degrees { 1, 2, 3, 4, 5, 10, 15, 25 }. To train, choose and test the models I divided the data into training, validation and testing sets of sizes 191, 64 and 110.

The possible improvements that can be conducted to possibly improve the quality of the prediction are clustering the data and taking cloudiness into account. As discussed in the *Method*-chapter, I assume that cloudiness contributes to the variance of amounts of UVB on days with equal air temperature. This can be confirmed by using multivariate regression and including the level of cloudiness as a feature. Clustering the data can improve the polynomial model’s ability to fit the data as many of the data points are located between temperatures -10 and 5 having a low UVB index whereas the points with temperature greater than 5 seem to depend more greatly on the temperature.

References

FMI, 2021 “Download Observations”, Accessed [25.3.2021]
Available:https://en.ilmatieteenlaitos.fi/download-observations

Jung, A., 2021, “Machine Learning: The Basics”

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, É 2011, "Scikit-learn: Machine Learning in Python", *Journal of Machine Learning Research,* vol. 12, no. 85, pp. 2825-2830.