Wave height prediction using Polynomial Regression

CS-C3240 - Machine Learning D

March 28, 2021

1 Introduction

Surf is a sport that depends completely on the weather conditions and especially on the wave height. In some places there are some buoys to measure the height of the wave far from the coast, but for the all the other places where there is no buoy the predictions are usually very bad. That's why this project tries to build a machine learning model to predict the wave height.
In Section 2 the machine learning problem is introduced, then in Section 3 different models for the problem are discussed, later in Section 4 the results obtained with the different models are shown, and finally, Section 5 presents the conclusions of the project.

2 Problem Formulation

For solving this machine learning problem, data points are represented as hours of surf days. For every hour in the dataset, the wind speed (knots), the temperature of the water (^{o}C) and the wave period (s) will be used as features, and the wave height (m) will be used as the quantity of interest (label).
The used dataset represents a data point every three hours in Jaws[1], Hawaii for the year 2020 (2928 data points in total). It was obtained from the Windguru's archive, one of the most used forecast web pages between windsurfers and surfers of the world.

3 Method

Knowing how complex wave formation is and how many variables influence in the height of the wave, the relation between the features and the labels was assumed to be non-linear. Therefore it was decided to implement Polynomial Regression with different maximum degrees.

[1] One of the most famous surf spots in the world. Known for very big and powerful waves.

Polynomial Regression is a special case of multiple linear regression that uses polynomials with maximum degree d as hypotheses maps $h(x)$. In other words, for each degree d we obtain a different model as the space of all polynomials with that maximum degree.
In this project, the performance of 10 models with maximum degree $d = 1, 2, \ldots, 10$ is compared. To evaluate the performance of each model, the mean squared error (MSE) loss $(y - h(x))^2$ is used, where y is the true label and $h(x)$ is the obtained prediction. This loss function was chosen because it works very well with numeric quantity labels as the wave height. Another reason to choose the MSE is that the `sklearn` library has a ready-to-use function easy to implement for Polynomial Regression.
The dataset is first split in two different sets, one for training and validation (80%) and the other for final testing (20%). The first one is then split to train set and validation set using k-fold CV with $k = 10$.

4 Results

After executing a simple code in python several times, it was observed that the best model for the ML problem was not always the same. Thus, it was decided to execute the code 100 times and compute the average error of all the iterations. The obtained results for the training and validation errors are shown in Table 1.

d	1	2	3	4	**5**	6	7	8	9	10
Train e.	0.189	0.168	0.163	0.151	**0.144**	0.140	0.131	0.129	0.122	0.122
Val e.	0.190	0.170	0.167	0.172	**0.159**	0.490	2.440	13.158	23.050	461.075

Table 1: Training and validation errors for Polynomial Regressions with different maximum degrees d.

Observing the results it can be noticed how the performance with $d = 5$ obtained the smallest validation error (0.159), therefore, it was chosen to be the final hypothesis. Note that both, the training and validation errors are very similar with this model (0.015 of difference), hence, it can be assumed that the model is not overfitting the training data.
On the other hand it should be noted how for higher maximum degrees ($d > 5$) the training error decreases, but the validation error increases exponentially, therefore, the higher the maximum degree is, the more the model overfits.

d	1	2	3	4	**5**	6	7	8	9	10
Test e.	0.182	0.163	0.162	0.153	**0.151**	0.162	0.605	1.382	1.344	4.035

Table 2: Test error for Polynomial Regressions with different maximum degrees d.

To asses the performance of the chosen model ($d = 5$) it was also computed a test error with the before mentioned test set. To have a better perspective it was computed for all the models, and the obtained results are showed in Table 2.
Again it can be observed how the best performance (lowest test error) was obtained with the chosen model (test error = 0.151). It should also be noticed that the obtained test error for this and all the models is (curiously) quite smaller than the validation error, specially with higher maximum degrees.

5 Conclusion

During this project ten different models have been compared for predicting the height of a wave. The used models were chosen to be Polynomial Regressions with different maximum degrees $1 \leq d \leq 10$. Then the models were trained and validated with the training and validation sets obtained with k-fold CV. The obtained results showed that the model with $d = 5$ had the best performance with a training error of 0.144 and a validation error of 0.159. This results show that the model is apparently not overfitting. After choosing the model it was also tested with a test set and it obtained a test error of 0.151.
Furthermore, there was found an article by Chih-Chiang Wei [1] where the height of the wave was predicted with data mining techniques where the obtained results are shown in Table 3.

Performance	kNN	LR	M5	MLP	SRV
RMSE (m)	1.034	0.712	0.699	0.691	0.694

Table 3: Average performance of RMSE obtained for five different models. (by: Chih-Chiang Wei)

To be able to compare the results, our obtained MSE error should be transformed to RMSE[2] (Table 4).

d	5
Val RMSE (m)	0.399

Table 4: RMSE obtained with the Polynomial Regression with maximum degree $d = 5$.

Now we can observe how our model performs quite good since the obtained RMSE error for the validation set is lower than the ones obtained by Chih-Chiang Wei. Even though this good results, the model should be tested with data of other surf spots to see how it will perform in other places.

[2] $RMSE = \sqrt{MSE}$

On the other hand it is known that there are a lot of other factors that influence the height of the wave that are not considered in this project such as the direction of the wind, the depth of the water, etc. Therefore there is a possible improvement if this data can be obtained somehow and then applied as additional features. Future models could also try to be more ambitious and try to predict the quality of the wave (in surfing terms) and not only the height.

References

[1] Chih-Chiang Wei. Nearshore wave predictions using data miningtechniques during typhoons: A case study neartaiwan's northeastern coast. *Energies MDPI*, 38(11):23, 11 2017.