Predicting the melting date of the snow using polynomial regression

1. Introduction

I have planned to make a hiking trip to Koli in April. It would be best to make the trip after snow has melted, as the snow has a huge effect on the equipment needed. To choose optimal timing for the hike, machine learning model was used to predict the melting of the snow. First, in the section 2 the formulation of the problem will be showcased. In section 3, the methods used in the project are displayed. In section 4 the results are discussed. Lastly, in section 5, the conclusions are presented.

2. Problem formulation

This application was modeled as a machine learning problem, where datapoints were individual days. The label of the datapoint was the depth of snow measured in cm and the features were the average temperature and maximum temperature of the day measured in degree Celsius. The model would have to predict the possible melting of the snow using previous weather data combined with weather forecasts for the next seven days, thus alerting about the melting date 7 days in advance.

3. Methods

The data was collected from the Finnish Meteorological Institute's historical weather recordings, which are available for free download from the webpage: https://en.ilmatieteenlaitos.fi/download-observations. The data contained known labels and features introduced earlier (label: depth of snow, features: max and average temperature of the day) and the measurements were from Juuka Niemelä, the closest available weather station near Koli.

Two datasets were acquired, where datapoints were the first 135 days of the year, ranging from the 1st of January to the 15th of May. Dataset 1 was from 2010 and it was used for model training, selection, and validation, while dataset 2 from 2011 was used only for testing the model. The training set and validation sets were chosen randomly from the dataset 1 with 70% of them being for training and 30% for validation. Both polynomial and linear regressions were used to as models, due to the label being continuous numeric value and their easy implementation with excel. Additional features such as maximum temperatures from earlier days and a feature with the depth of snow from 7 days earlier were manufactured. This was chosen to make the model more

accurate, while still preserving some value of the model, as predicting the melting of the snow a week in advance would still be beneficial. Two separate models with polynomial regression and one with linear regression were tested. The loss function used was squared error loss. It was chosen because its suitable due to the label being continuous numeric value (1). Separate training and validation datasets were used within dataset 1 and the resulting training and validation errors for the different models are shown Table 1 below.

Table 1. Training and validation errors of the different models used.

	3rd degree polynomial	2nd degree polynomial	linear
etrain	265,7	210,5	109,4
eval	275,9	224,6	145,2

4 Results

While the training and validation errors were significantly smaller with the linear model it was not suitable for this application. The training and validation errors were not the only categories to decide the best model, as predicting the eventual melting day of the snow was the goal of the model. The linear model was too focused on the previous snow depth that was used as a feature and missed the melting date with exactly seven days. Both polynomial models were much more accurate at predicting the melting date and missed it by only one day. Due to these circumstances the 2nd degree model was chosen as it had smaller errors of the two functional models.

To test the effectiveness of the chosen model was used on previously unused dataset 2. The resulting test error was surprisingly significantly lower with value of 128.5. The date of the melting of the snow was of by 4 days in this test.

5 Conclusions

Three different machine learning models were studied to try to predict the melting day of the snow. The three models used maximum temperatures of and snow depths of the previous days with varying degrees of polynomials. The resulting training and validation errors of the models were quite big, ranging mostly from 0% to 20% error from the actual depth of snow. The final model was chosen with consideration to both validation error and performance on determining the correct date of melting of the snow.

The chosen hypothesis, with polynomials of 2nd degree was chosen and tested with dataset consisting of the first 135 days of the year 2011. Regarding the test error, the model performed better on the training set, with test error of 128.5 compared to validation error of 224.6 on the first dataset. The date of the melting of the snow was off by 4 days on the test set.

The chosen models were probably not best suited for such predictions, as the model would be reliant on the weather forecasts, it could only used to predict snow melt seven days in advance, and it quite frankly did not perform very well. Some kind of time series forecasting machine learning model such as LSTM (2)(Long short-term memory) that takes seasonality in to consideration could have been more suitable for this problem.

6. References

1. [MLBook] A. Jung, "Machine Learning. The Basics", 2021, mlbook.cs.aalto.fi
2. Karevan, Z.; Suykens, J. A. K. Transductive LSTM for time-series prediction: An application to weather forecasting. *Neural Networks* **2020***, 125*, 1-9.