Data mining HW1: Report

Malik DAHMANI

April 16, 2024

Contents

1 Question 1: How do you select features for your model input, and what preprocessing did you perform?

For data pre-processing, I used Pandas' interpolation function to replace invalid values with interpolated values. Where certain values could not be interpolated, I filled in the gaps by taking the average of the neighbouring valid values. This approach preserves as much information as possible while guaranteeing the validity of the data.

Finally, to process the temporal data, I used the window-sliding technique to efficiently account for trends and variations over time.

In order to select the features of my model, I followed several steps. First, I performed a correlation test to assess the linear relationships between the different variables. Next, I removed features with very high correlation coefficients between them to avoid redundancy in the data. I also removed a feature with a very low variance, as this indicates that it provides little discriminative information for prediction.

To conclude, after having carried out all these pre-processing and feature selection steps, I finally proceeded to Ridge regression. Regarding the choice of the lambda regularisation parameter for the ridge regression, I chose the one that maximised the performance score of the model. This ensures a better fit to the data while avoiding overfitting.

2 Question 2: Compare the impact of different amounts of training data on the PM2.5 prediction accuracy. Visualize the results and explain them.

To compare the impact of different amounts of training data on PM2.5 prediction accuracy, I decided to change the number of months in the training data.

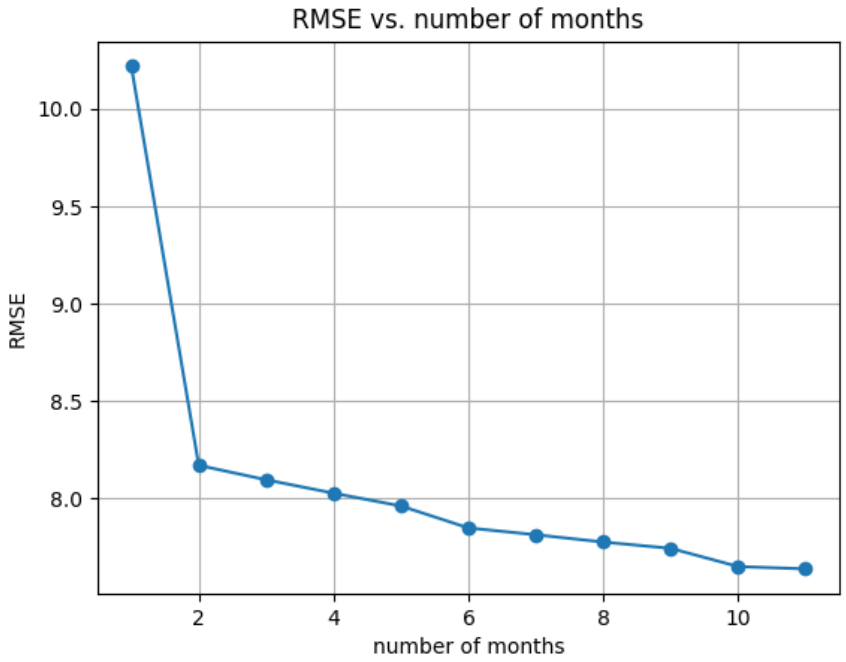

Figure 1: Figure showing evolution of RMSE depending on number of months

We notice that between the first and second month, the RMSE drops sharply, which is explained by the fact that the training set is small. Then, as the size of the dataset increases (from two month to three months and more), the model has access to more information to learn and generalise from the data. With a sufficient amount of data, the model becomes more stable and is better able to capture general trends in the data. This leads to more consistent performance and less change in RMSE as the size of the

dataset increases.

3 Question 3: Discuss the impact of regularization on PM2.5 prediction accuracy.

We note that regularization has an impact on the model's RMSE on the training set, but regularization is essential for better generalization of the model and to have better predictions. It therefore offers several advantages:

- Prevent model overfitting: Without regularization, models can become over-adapted to training data, leading to over-fitting and poor performance when confronted with new data.
- Improve the model's ability to generalize: By restricting the model's freedom during training, regularization encourages it to learn more general patterns and avoid memorizing the noise present in the training data.
- Reduce model variance: By limiting model complexity, it helps stabilize predictions, making the model less sensitive to minor variations in the training data.