

Summative Assignment

Artificial Intelligence COMP2261 – Machine Learning 2020/2021

Kramer (1991)

1. Problem framing (10%)

Coronavirus (COVID-19) is spreading fast. Since first reported in December 2019, by mid-January 2021, it has affected more than 95 million people and killed more than 2 million people worldwide. To aid the analysis and inform public health decision making, machine learning models trained on real data can be very useful.

In this paper, we build and compare predictive models using machine learning algorithms and epidemiological data from the COVID-19 outbreak. We explore the dataset and aim to solve the problem of classifying the , with motivations

2. Experimental procedure (35%)

2.1 Data preparation

First, we check for unwanted instances, which include those that are redundant and irrelevant. We do not observe any redundant instances that have appeared during data collection. Irrelevant instances are what are not useful for our specific task, so since we are training a model to predict the number of Covid-19 cases in the future, those instances with no date confirmation are not useful thus are removed.

Next, we check for outliers. Outliers may cause issues, and, since linear regression models are sensitive to outliers, they should be dealt with. We only remove outliers that are [...]; the remaining are informative for our model.

Next, we deal with structural errors. We deal with typos, inconsistent English spelling, inconsistent capitalisation, and abbreviation. In the case of categorical features, we combined those that should be a single category, such as ‘died’ and ‘death’.

Next we deal with missing data. We decide to drop the instances with missing values, because most machine learning algorithms do not accept missing values.

2.2 Model selection

2.3 Model training

2.4 Model testing

2.5 Hyperparameter tuning

2.6 Inference/Prediction

- Clean the dataset.

- Split the dataset into training and test sets.
- Train a logistic regression model.
- Train a polynomial regression model.
- Train a normal regression model.

3. Results (25%)

- Make comparisons between the 3 predictive models
- Provide necessary tables and charts to summarise and support the comparisons.

4. Discussions (20%)

4.1 Chosen models

4.2 Experimental procedure

4.3 Limitations

5. Conclusions and lessons learnt (10%)

- Discuss the results and draw conclusions from your experimentation

References

Mark A Kramer. Nonlinear principal component analysis using autoassociative neural networks. *AIChE journal*, 37(2):233–243, 1991.