Summative Assignment
Artificial Intelligence COMP2261 –
Machine Learning 2020/2021

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 predicting patient outcome as either "died" or "discharged" from hospital. The motivation is that the proposed prediction model may be helpful for the quick triage of patients without having to wait for the results of additional tests such as laboratory or radiologic studies, during a pandemic when limited medical resources must be wisely allocated without hesitation.

2. Experimental procedure (35%)

The experimental procedure was to follow the machine learning workflow, which consists of 7 stages: problem framing, data preparation, model selection, model training, model testing, hyperparameter tuning, and inference/prediction.

2.1 Data preparation

We used the dataset by Xu et al. (2020) which is publicly available on GitHub[1].

First, we removed irrelevant instances. These are instances that are not useful for the task. Since we are training models to predict patient outcome, irrelevant instances had 'null' as the outcome. We also corrected structural errors, such as inconsistent capitalisation. We made one category for 'died', 'Died', 'dead', or similar, and another for 'discharged', 'Discharged', 'recovered', or similar. Instances which did not fall under either one of these two categories were dropped.

Second, we kept attributes which we thought could be useful to the models. The attributes kept were 'age', 'sex', 'country', 'latitude', 'longitude', 'date_onset_symptoms', 'date_confirmation', 'symptoms', 'chronic_disease_binary', and 'travel_history_binary'. After this, we cleaned up the data by replacing 'date_onset_symptoms' and 'symptoms' with a new attribute, 'symptoms_binary', then dropping any remaining instances with missing values.

[1] github.com/beoutbreakprepared/nCoV2019/

Third, we transformed numeric data. We perform data binning on the ‘age’ attribute. The ranges were ‘0-14’, ‘15-34’, ‘35-59’, ‘60-79’, and ‘80+’. These were chosen because the ‘age’ attribute was in fact already categorical, and the latter 4 ranges were the 4 most frequent categories. We also perform standardisation on the ‘latitude’ and ‘longitude’ attributes. (remove country)

Fourth, we transformed categorical data.

2.1.1 Data sampling

We notice we have an imbalanced data set, so we employ down sampling and up-weighting during model training.

2.1.2 Data splitting

We split the dataset into training set and test set. We keep them separate, as we don’t want the model to memorise. Before splitting, we randomise the dataset as we don’t want the order of the instances, which is irrelevant, to affect the model training process. We make our test set meet two conditions: it is large enough to yield statistically meaningful results, and it is representative of the dataset as a while.

To prevent overfitting, we produce a validation set. We train the moddel on the training set, then evaluate the model on the validation set and use those results to tweak the odel iteratively. We leave the test set separate to only confirm results of the model that does best on the validation set. This creates fewer exposures to the test set.

We employ k-fold cross validation to reduce the chance of overfitting, assess how well the model performs on previously unseen data, and resampling producedure to tes models on a limited data sample

2.1.3 Data transformation

Before feature transformation we explore and clean up data and visualise data in graphs and charts.

We perform data transformation for data compatibiltiy and better model performance.

Numeric data. We perform binning on age. We perform feature scaling because we will be using SVM, kNN, PCA, clustering (not necessary for logistic regression, decision tree).

categorical data. We encode categorical data in roder to be able to fit and evaluaet models. We use one-hot encoding

2.2 Model selection

We chose SVM, kNN, and logistic reg.

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

Bo Xu, Bernardo Gutierrez, Sumiko Mekaru, Kara Sewalk, Lauren Goodwin, Alyssa Loskill, Emily Cohn, Yulin Hswen, Sarah C. Hill, Maria M Cobo, Alexander Zarebski, Sabrina Li, Chieh-Hsi Wu, Erin Hulland, Julia Morgan, Lin Wang, Katelynn O'Brien, Samuel V. Scarpino, John S. Brownstein, Oliver G. Pybus, David M. Pigott, and Moritz U. G. Kraemer. Epidemiological data from the COVID-19 outbreak, real-time case information. *Scientific Data*, 7(106), 2020. doi: doi.org/10.1038/s41597-020-0448-0.