Chapter 1

Typhoon Prediction: A Supervised Machine Learning Approach

The enormous devastation caused by super typhoon Haiyan (2013) motivates us to consider whether data from past typhoons hold a key to predicting the grade ("category") of an emerging tropical storm from early best track data (lat, lon, windspeed and pressure in the first 18 hours). In particular, we describe a machine learning (ML) approach to classification of typhoon grade and regression of maximum wind speed. We train and validate our model and make 'hindcasts' using Japan Meteorological Agency (JMA) best-track data.

1.1. Introduction

In November 2013, Typhoon Haiyan struck the Philippines with sustained wind speeds of nearly 200 miles per hour, claiming over 6,000 lives, displacing 4 million people, damaging over a million houses, and resulting in a request for over 750 million dollars in humanitarian aid. Devastation caused by “super typhoons" like Haiyan makes the on-going development of early warning systems an absolute necessity to provide vulnerable communities sufficient preparation time to mitigate damage to both life and property. This motivates us to apply Machine Learning algorithms to predict the typhoon grade (category model), and maximum wind speed (regression model).

1.2. Overview of Supervised Machine Learning

Supervised Machine learning (ML) refers to a set of algorithms which utilize a training data set to develop a model which takes one or more input variables called *predictors* and outputs the value of a variable called a *response*. We will use Japan Meteorological Agency (JMA) data for 2009-2014 as training data:

- Predictor Variables: Lat, Lon, WS and Pressure at times $t = 0, 6, 12, 18$ hours.
- Response Variables: Typhoon grade (3,4, or 5) and maximum sustained wind speed.

In the next section we describe a number of different ML algorithms such as decision tree, k-nearest neighbor, ensemble. In this chapter, we do not consider *unsupervised* ML algorithms where the response variables are not known in advance, and hence not specified in the training data set.

When the response is a discrete set, such as is the case for typhoon grade, we use a *categorical* ML algorithm. When the response is continuous, such as maximum wind speed, we use a *regression* algorithm. In both cases, the accuracy of the prediction is our primary performance measure.

In the case of categorical models, one way to report accuracy is via a *confusion matrix*. Figure 1.1 gives an example of the latter, for a model which on the training data set accurately classifies 13 grade, 13 grade 4, and 47 grade 5 typhoons. In addition, specific information about the classification errors is given. For example, 8 typhoons predicted as grade 5 were actually grade 4, and 9 typhoons predicted as grade 4 were actually grade 5.

For a regression model, a scatter plot which compares the actual to the predicted values, with the difference between them called residuals is a standard visual representation of the error. The root mean squared error (RMSE) is calculated as the square root of the mean squared residuals. Figure 1.2 shows such a scatterplot, with in particular shows fairly good wind speed prediction of Grade 5 violent typhoon (maximum wind speeds of at least 105 knots/hr) as well as grade 3 tropical storms, but has greater difficulty in achieving accurate wind speed prediction for the typhoons in the mid-range (category 4).

Once a model has been developed using the training data set, it should be validated using a *validation* data set. In our case, we use JMA data for 2015 as our validation set. If the model does not perform well on the validation data, we must revise our model,

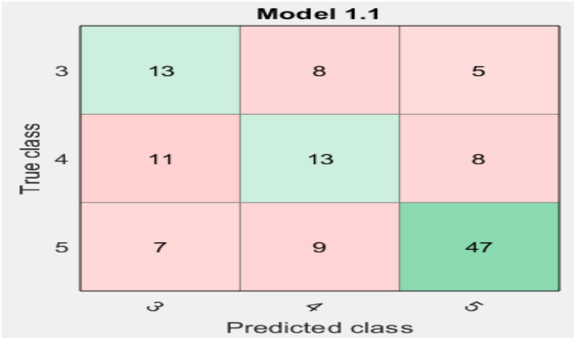

Figure 1.1. Confusion matrix for a typhoon grade Fine Tree categorical ML model.

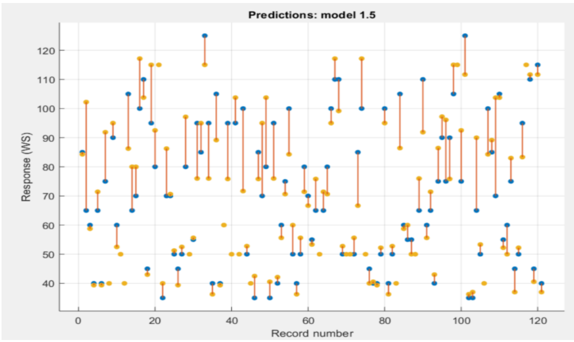

Figure 1.2. Fine Tree ML regression model RMSE: 11.0

returning to the training stage. If the model does perform well on the validation set, we are ready to test it on new data. For a quasi-operational ("hind-casting") model, our "new" data is JMA data for 2016.

Algorithms

Predicting the JMA grade of an emerging tropical storm based on its initial 18 hour bext track data is an example of a standard supervised learning problem ([2]). A training data set has the form $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$ where $\mathbf{x}_* = (YEAR, MONTH, DAY, LAT1, LON1, PRES1, WS1, L$ specifies the date and first three best track points of an emerging tropical storm (recorded windspeed of 35 knots) and $y_* \in \{3, 4, 5\}$ is the highest typhoon grade achieved by the storm. A learning algorithm outputs a function $f(\mathbf{x})$ called a classifier, which given values for the predictor variable $\mathbf{x}$, outputs a value for the response variable $y = f(\mathbf{x})$ (typhoon grade). An ensemble method combines responses of two or more classifiers to improve the prediction accuracy of individual classifiers.

There are a number of basic supervised ML algorithms [1] such as

- Decision Tree
- Support Vector Machines (SVM)
- K nearest neighbors
- Ensemble

1.3. Exercises

1. What is the overall percentage of correct typhoon grade prediction based on the information in an ensemble model ML confusion matrix shown in Figure 1.3? How well does the model predict grade 5 typhoons (give percentages of correct, false positives and false negatives).)

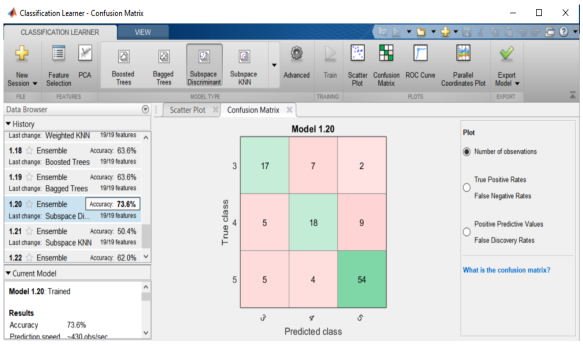

Figure 1.3. Confusion matrix for a MATLAB subspace discriminant ensemble classification algorithm.

1.4. Computer Project

Project website: https://www.overleaf.com/project/5c057323a87843644f4893a6

MATLAB has a Statistics and Machine Learning Toolbox which has apps to build both categorical and regression model. The main steps in creating these ML models are:

- Train MATLAB's suite of ML models using a specified dataset.
- Choose one of the models with high accuracy, and export a MATLAB function based on your selected trained model.
- Apply the exported model to a validation dataset. If satisfied, continue to the next step. If not, repeat the previous steps.
- Test the model on a new data set and record the error.

1. Using the same training and validation data sets discussed earlier in this chapter, use MATLAB to develop an categorical ML model for the various level of Grade 5 typhoons, and a regression model for the lat, lon and time where the maximum wind speed is first attained on the best track.

Project Team: Michaela Flitsch, Mark Nussbaum, Zach Oslund, Matthew Rueger.

References

[1] Bonnardot, G., 8 Machine Learning Algorithms Explained in Human Language. Available at https://www.datakeen.co/en/8-machine-learning-algorithms-explained-in-human-language/

[2] Dietterich, T. Ensemble Methods in Machine Learning. Available at http://web.engr.oregonstate.edu/ tgd/publications/mcs-ensembles.pdf

[3] Japan Meteorological Agency, Best track archives. Available at http://www.jma.go.jp/ jma/jma-eng/ jma-center/ rsmc-hp-pub-eg/trackarchives.html.