Linxin Liu

INFO 7390

April 9th 2021

Rain in Australia Project

When people go outside, they usually make sure that they carry a raincoat or an umbrella if the weather says it will rain outside. In fact, it is unpleasant to a person did not bring a raincoat or an umbrella when this person goes out, and then it start to rain outside. Therefore, knowing the weather ahead will be helpful for people preparing, when they go outside. Thus, we decide to use our knowledge in machine learning and the rain in Australia data set to predict weather in the next day and help people in Australia preparing their belongings for outside.

This rain in Australia project will use K-nearest neighbor(KNN), random forest, and deep neural network(DNN) to predict the weather tomorrow. In this rain in Australia data set, the “RainTomorrow” column determines whether tomorrow will rain or not. In this column, “Yes” or “1” means tomorrow will rain, “No” or “0” means tomorrow will not rain, and “2” means others. Since the “RainTomorrow” columns determines the weather tomorrow, it shall becomes our Y-value and we will use the rest of columns as the X-value and our models to predict it.

As we stated before, we shall use KNN, random forest, and DNN to predict the weather tomorrow. Because knowing the weather ahead will help people get ready, when they go outside. Thus, the goal of this rain in Australia project shall be finding out the best model for predicting weather tomorrow,and getting the correctness of the prediction as high as possible.

To do so, we shall record accuracy score of the prediction of each model in a data frame. As we know, accuracy score is not enough to determine whether a model is the best model for predicting weather tomorrow. Therefore, we shall use confusion matrix to compare the prediction results of KNN and random forest models. As we stated before, one of goals of this project is to getting the accuracy of the prediction as high as possible. To do so, we will use the DNN model and we will apply drop-out layer, hidden drop-out layer, and hidden drop-out layer with weighted constraints to the model. We shall use Receiver Operating Characteristic(ROC) curve to visualize the true positive rate, and record the area under ROC curve for analysis.