Machine Learning Project Report

Lena Loye, Anas Yusuf

December 2021

1 Introduction

In this competition we get measurements from different weather stations in Switzerland. The goal was to predict whether there is some precipitation (rain, snow etc.) on the next day in Pully. This problem is a typical binary classification task. We first need to explore our data set, and then we will approach this problem using both linear and non-linear methods. To compare the different approaches throughout the project we mainly used the AUC (area under curve).

2 Process and results

2.1 Exploration

A preliminary exploration of the data allow us to get a first ensemble view of the main characteristics of the data set. To predict if there will have precipitation on the next day, we were provided 529 predictors and 3176 observations in our training data. These predictors are numerical (Float64) and the y (*precipitation_nextday*) is a boolean. We also observed that among the 3176 observations, 1477 have missing values. We will replace them using the *FillImputer()* function. We also observe that in both training and test data, some predictors have a null variance, respectively 1 and 5 for each data. That means they can not bring useful information to our predictions and we will remove them before constructing our models. All predictors are provided in our test data. We also see that there are more false responses. To summarize we observed the need to transform missing values and remove predictors with null variance.

2.2 Linear Methods

At the beginning of the notebook we prepare our training data set by filling in the missing values with some standard values and standardize both training and test data sets. We decided to use the Lasso regularization for Logistic Regression in order to decrease the test error. We cross-validated on different lambda parameters (the strength of the regularization) to choose the one with the best performance on our model. We used the AUC measure to compare

the different models. We first obtain a lambda of approximately 4.3 using an interval from $1e^{-2}$ to 10. Then we decided to reduce the interval and find the best model with lambda equal to 4.00775. We posted it on Kaggle and obtain a result of approximately 0.956.

2.3 Non-Linear Methods

We first prepare our data sets as in the Linear Methods notebook.

2.3.1 K-Nearest Neighbor

We used the *KNNClassifier()* method. We cross-validated on different K parameters (1 to 50) to choose the one with the best performance on our model. We also used the AUC measure to compare the different models. We obtain a K equal to 22. Then we posted it on Kaggle but obtained a result of approximately 0.944, lower than the result obtained with the Logistic Regression.

2.3.2 Trees

We used the *RandomForestClassifier()* method. We cross-validated on different *n_trees* parameters on the interval 100 to 500 to choose the one with the best performance on our model. We also used the AUC measure to compare the different models. We first obtain a *n_trees* equal to 254. We decided to reduce the interval and found the best model at $n\_trees = 251$. We posted it on Kaggle but obtained a result of approximately 0.943, lower than the result obtained with the Logistic Regression and with KNN.

2.3.3 Neural Networks

We used the *NeuralNetworkClassifier()* method. We first tried to build a two-layer network with 50 neurons on each layer. We post it on Kaggle and obtained approximately 0.93. Then we tried to build a full-connected three-layer network with 128 nodes in the hidden layer by using the MLJFlux builder. We posted it on Kaggle and obtained a result of 0.956.

3 Conclusions

To summarize, we tried to predict whether there is some precipitation (rain, snow etc.) on the next day in Pully based on measurements from different weather stations in Switzerland using linear and non-linear machine learning methods. The best methods were the Lasso regularized regression and the full-connected three-layer network. Both are efficient and quick methods. Ultimately, we can not expect to perfectly predict the precipitation in Pully using measurements from other places.