Udacity Machine Learning Nanodegree
Capstone Proposal

Carsten Krüger

May 13, 2018

Domain Background

Machine learning is used in a wide variety of fields today. Luca Talenti et al. [1] for example used a classification model to predict the severity criteria in imported malaria. In this project, machine learning will be used to build a model that can decide based on the role information of an employee whether that employee shall have access to a specific resource [2].

An employee that has to use a computer in order to fulfill their tasks needs access to certain areas of software programs or access rights to execute actions such as read, write or delete a document. While working, employees may encounter that they don't have a concrete access right required to perform the task at hand. In those situations a supervisor or an administrator has to grant them access. The process of discovering that a certain access right is missing and removing that obstacle is both time-consuming and costly. A model that can predict which access rights are needed based on the current role of an employee is therefore relevant.

Problem Statement

The problem stems from the *Amazon.com Employee Access Challenge Kaggle Competition* [2] and is there described as follows:

> *"The objective of this competition is to build a model, learned using historical data, that will determine an employee's access needs, such that manual access transactions (grants and revokes) are minimized as the employee's attributes change over time. The model will take an employee's role information and a resource code and will return whether or not access should be granted."*

Datasets and Inputs

As well as the problem the data stems from the *Amazon.com Employee Access Challenge Kaggle Competition* [2]. It is real historical data that was collected in 2010 and 2011 by Amazon. The dataset consists of approx. 32000 rows. Each row contains eight columns that provide role information for an employee, and an ID specifying a resource. The eight columns consist of the manager of the current employee, two role grouping categories, a role department description, a role business title, a role family description, an extended role family description and a role code, which is unique to each role. Each of these columns contains numbers which are IDs. Therefore it is expected that a vectorization of the columns is necessary and that the dimensionality of the dataset will increase tremendously. Each row has an action, labeled either 0 or 1. Where 1 means that the resource was approved for the employee and 0 means that it was not. Those class labels are clearly balanced. In more than 94 % of the cases the access was granted where only in nearly 6 % of the cases the access was denied.

The dataset is appropriate for the problem at hand because it contains all the required information to train a classification model that determines whether an employee should be granted access to a given resource based on their current role.

Solution Statement

This is clearly a supervised learning problem because the dataset is labeled. The goal is to decide whether access to a resource shall be granted or denied for an employee in a given role. This implies the use of a binary classification model. For each row in the labeled dataset it will be possible to determine whether the model predicted the correct class. The solution will be clearly quantifiable and measurable. For example by dividing the number of correctly classified entries by the total number of entries in the dataset one can calculate the accuracy of the model.

Benchmark Model

An out of the box logistic regression model will be used as the benchmark for this project, because the model is fast, simple to implement and to interpret and should give far better results than random guessing for the problem at hand.

Because the problems stems from a Kaggle competition, as a secondary benchmark, the result of the final solution will be compared to the result of the solution of the team that won the competition. The submissions to the competition were judged on the *area under the ROC curve (auc)* metric. Therefore this metric will be used to compare the results. The winning team got an *auc* value of 0.92360, which is an excellent result and the author will strive to come close to it.

Evaluation Metrics

The area under the ROC curve is proposed as a metric to be used to quantify the performance of the model because it was used in the herein before mentioned Kaggle competition and it can be used for binary classification problems.

The metric is derived by first constructing the ROC curve and then calculating the area under that curve.

> *"The ROC curve is created by plotting the true positive rate against the false positive rate at various threshold settings"* [3],

where the threshold is a value between 0 and 1 that determines how sure the model needs to be in order to classify a data entry as positive (access granted in the problem at hand). For example if the threshold was 0.7 the model would have to have calculated a probability of at least 70 % to classify a data entry as positive.

Project Design

First of all the data needs to be explored. For each feature it is useful to examine the percentage of missing values, to detect outliers and to find out the type of the feature (binary, categorical, continuous etc.). In order to gain further inside the data will be visualized. Any two features could be drawn in a scatter plot were positive results (access granted) are marked in e.g. green and negative results (access denied) are marked in e.g. red. A scatter plot would also help to visualize whether any two features are correlated. If the exploration step found outliers they will be removed. Missing data will be filled in or the whole row will be dropped.

The dataset will be split into a training and testing set. Many different models will be trained using standard parameters. The plan is to try logistic regression, Naive Bayes, decision trees, SVM, random forrest, AdaBoost, XGBoost and LightGBM as models. The models will be quantified based on the *auc* metric and based on computation time. Other metrics which are applicable for classification problems, such as *F-Scores*, *confusion matrix* or *precison* and *recall* might also be taken into consideration. It will be checked whether the models are overfitting or underfitting. Because the dataset is rather small, K-fold cross-validation will be used. This ensures that the training set doesn't have to be reduced any further to create a separate validation set. A *ROC* curve will be plotted for each model.

After the models have been trained, if applicable, transformations such as regularization will be applied. Features that were observed to don't provide any insides will be disregarded. A principle component analysis might be used in order to discover the features with the highest influence to the outcome or to justify the feature removal. The performance of the updated models will be compared to the unmodified ones to see whether the modifications are helping or not.

The best three models will be further examined and tweaked in order to improve the performance. Different sets of hyper-parameters will be tried out. It is planned to use grid search or random search to find the best hyperparameters for each model. The tweaked models will be evaluated on the test set and compared based on the *auc* performance metric. If it is possible to make a late submission to the Kaggle competition and Kaggle returns the value of the *auc*, this result will be compared with the benchmark. Otherwise the result on the test set will be used for this comparison.

References

[1] L1 logistic regression as a feature selection step for training stable classification trees for the prediction of severity criteria in imported malaria

Luca Talenti, Margaux Luck, Anastasia Yartseva, Nicolas Argy, Sandrine Houz, Cecilia Damon

arXiv:1511.06663 [cs.LG]

[2] Amazon.com – Employee Access Challenge

Predict an employee's access needs, given his/her job role.

https://www.kaggle.com/c/amazon-employee-access-challenge,

[3] Receiver operating characteristic

https://en.wikipedia.org/wiki/Receiver_operating_characteristic