Udacity Machine Learning Nanodegree
Capstone Project

Carsten Krüger

June 13, 2018

1 Definition

1.1 Project Overview

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.

1.2 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."

This is a supervised learning problem because the dataset is labeled. Anticipated solution:

1. Explore data in order to gain insights.
2. Train many different binary classification models using standard parameters.
3. Apply transformations or regularizations.
4. Compare plain models and transformed models.
5. Pick the three best models based on the *auc* [3] metric.
6. Tweak the hyper-parameters for each of the chosen models in order to improve their performance.
7. Evaluate the tweaked models on the test set.

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