Machine Learning – Final Report

Volker Strobel

July 23, 2016

This reports presents the techniques and results of a classification problem involving missing data as well as a mixture of categorical and continuous data. The problem of missing data points is being addressed by Multiple Imputation Chained Equations. The classification is conducted using the AdaBoost classifier. The method has been locally evaluated using cross-validation and remotely on a hold-out test-set using the Kaggle platform.

1 The Challenge

The challenge at hand is to predict, whether a person earns over EUR 40k a year based on a mixture of both qualitative and quantiative variables (Table 1).

Therefore, this competition involves three main challenges:

- Number of missing data points
- Mixture of categorical and continuous variables
- Classification of the output variable

We will address each of them in turn.

Section 2 briefly compares different methods for missing data. In Section 3, we analyze and visualize the structure of the data, to set the stage for feature-extraction-to-classification pipeline. In Section 3, the used method—Adaboost classification—is described in detail. Section 4 describes the results obtained during cross-validation and on Kaggle. In Section 5,

2 Background

Missing data values are a common problem in statistics. The failure of sensors, or the conscious loss due to anonymity impede the machine learning accuracy. While the *imputation* of these missing data is still a open problem, several method have been put forth. For the competition, a maximum-likelihood (expectation-maximization) method has been used.

3 Analysis

In order to motivate classifier, design and technique choices, we start with an in-depth analysis and visualization of the given data sets.

Table 1 shows the discrete and continuous features. The percentage of missing values is given in the next table.

The used loss function is the 1/0 loss.

Categorical	Continuous
work class	age
education	number of years of education
marital status	income from investment sources
occupation	losses from investment sources
relationship	working hours per week
race	
sex	
native country	

Table 1: Overview of the used features

4 Methods

4.1 Missing Data Values

A first analysis shows that on average ca. 19.6 % of the data values are missing, with ca. 18.8 % some variables and 23.1 % for workclass and occupation.
For the possibly best performance, we would like to show that the data is missing completely at random (MCAR). Therefore, we conduct Little's MCAR test **little1988test** to analyze the interaction structure of the variables.

4.2 Dummy Variables

Seven of the twelve measured variables are qualitative (workclass, education, marital status, occupation, relationship, sex, native country)—that is, they are measured only at the nominal level. Since measurements on the nominal level do not allow for a particular ordering, a proxy method has to be used. Therefore, we define dummy variables which take the value 0 or 1 depending on if a certain category is present. Defining N dummy variables for the N different possible values of a categorical feature allows us to capture the full information in the orginal unmodified dataset and use qualitative data in a straight-forward manner in regression models that are usually based on decision boundaries or linear relationships.

After transforming the dataset, xx features are obtained, which largely increases the size of the dataset.

5 Feature-extraction-to-classification Pipeline

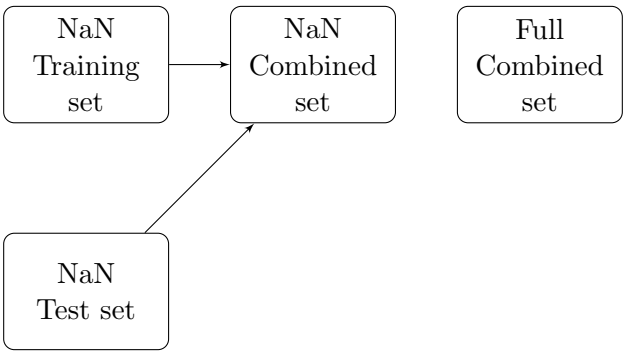

Figure 1: The pipeline

6 Results

7 Discussion

8 Conclusion