Machine Learning – Final Report

Volker Strobel

June 8, 2016

1 Introduction

Predicting the challenge is to predict, whether a person earns over EUR 40k a year.

2 Analysis

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

The used loss 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

3 Methods

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.

4 Missing data values

Missing data values are a common problem in machine learning problems. 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.

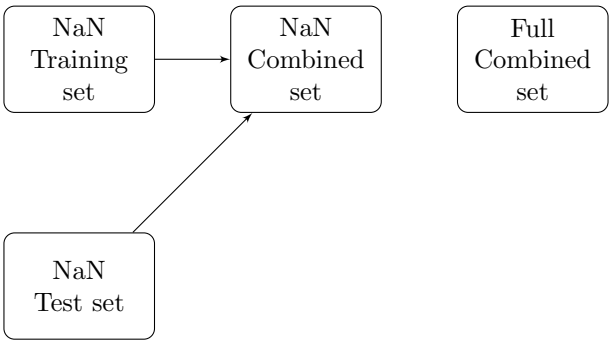

Figure 1: The pipeline