Machine Learning – Final Report

Volker Strobel

July 30, 2016

Abstract

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 given problem is to predict, whether a person earns over EUR 40k a year (Table 1). The competition involves three main challenges:

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

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 the feature-extraction-to-classification pipeline. In Section 4, the used method—Adaboost classification—is described in detail. Section 6 describes the results obtained during cross-validation and on Kaggle. In Section 7, the results are discussed.

2 Background

Missing data values are a common problem in statistics. The failure of sensors, or the concealment of data impede machine learning accuracy. However, methods have been put forth for the imputation of missing data points, such as case deletion or single mean imputation **??**. However, such simple methods often discard useful information and are not favorable if a large part of the data is missing.

3 Analysis

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

Table 1 shows the discrete and continuous features.

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

Figure 1: Scattermatrix of the given dataset.

4 Methods

4.1 Missing Data Values

A first analysis shows that in total 19.6 % of the data values are missing, with 23.1 % for the variables workclass and occupation, and 18.8 % for the remaining variables.

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 Little 1988 to analyze the interaction structure of the variables.

Since the relative frequency of missing data points is rather large, simple methods, like mean imputation are likely to give a suboptimal result. Therefore, data points were imputed busing a technique called Multivariate Imputation by Chained Equations (MICE) Buuren and Groothuis-Oudshoorn 2011 using the `R` package `mice` (Version 2.25). The algorithm uses multiple imputations, which allows for incorporating the statistical uncertainty in the imputed values. It is a flexible approch that can aldo handle continuous and categorical variables. The general idea is to model the conditional probability of each missing variable given the remaining variables.

The algorithm works as follows Azur et al. 2011:

1. In the beginning, a first simple imputation is performed. To this end, all missing data points are replaced by random sampling with replacement from the observed datapoints.

2. One variable is selected at random, for example "occupation". An "intermediate" regression model is build, using the remaining variables as predictors and occupation as target value. The chosen model is dependent on the target value. I used a logistic regression for binary data, a polytomous regression model for categorical data, and predictive mean matching for numerical data. The missing values in occupation are replaced by the predictions of the model

3. The previous step is performed for each variable with missing data. The models that are build use the imputed and existing values for the training of the model. If all variables were predicted, one cycle is completed.

4. Several cycles are performed to stabilize the imputation results. I used $c = 5$ cycles.

5. The entire procedure is performed multiple times to yield multiple imputed datasets. Due to the random factors, the imputated values will be different, while the non-missing data entries will be the same. I used $m = 5$ imputations.

The MICE algorithm was used on a combined dataset, by combining the training and the testset to one large dataset. This should increase the quality of the impuated values since more training examples can be used for building the intermediate models (Step 2 in the algorithm outline).

4.2 Pooling

Since the multiple imputation methods yields multiple datasets ($m = 5$), the predictions have to be aggregated. For this, majority voting was used based on the predictions (0 or 1) of the single dataset.

4.3 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 that indicate if a certain category is present. Defining N dummy variables for the N different possible values of a categorical feature allows for capturing 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, 104 variables were obtained, which largely increases the size of the dataset.

4.4 Cross-Validation

To evaluate submissions and determine the rank of participants, the Kaggle challenge used the 1/0 loss. While the score in the challenge is the best validation of the methods employed, only one submission could be submitted per participant and day. In order to evaluate the used methods more frequently, local 5-fold cross-validation was implemented. The cross validation was performed per dataset of the imputation result, that is, no cross-dataset cross-validation was performed. Therefore, in the case of five imputations, 25 cross-validation results were obtained.

The cross-validation achieved similar results to the public Kaggle evaluation (Table 2).

Method	Local Score	Global Score
AdaBoost with 1 imputation	0.83886	0.84205
Random Forest with five imputations	0.82726	0.85056
SVM with five imputations	0.84413	0.85494

Table 2: Comparing local and public scores

4.5 The Classifier

As can be seen in Section 4.4, different classifiers were tested for the given problem. The best performing one was a support vector machine (SVM). For running the algorithm `Python 2.7.11` was used with the package `scikit-learn` in Version 0.17.0.

Support vector machines build a $D-1$-dimensional hyperplane that tries to maximize the margin between the two classes. The support vectors are the samples on the margin for the positive class and the negative class. The maximum-margin hyperplane is the hyperplane "right in the middle" between the support vectors.

The classification function is $h(x) = g(w^T x + b)$, with w being the coefficient vector and b the intercept term. The function $g(z)$ is defined as $g(z) = 1$, if $z \geq 0$ and $g(z) = -1$ otherwise. The goal is to find w, b such that the resulting decision boundary maximizes the margin.

4.5.1 The kernel

To extend SVMs to non-linear classification, the *kernel trick* can be used.

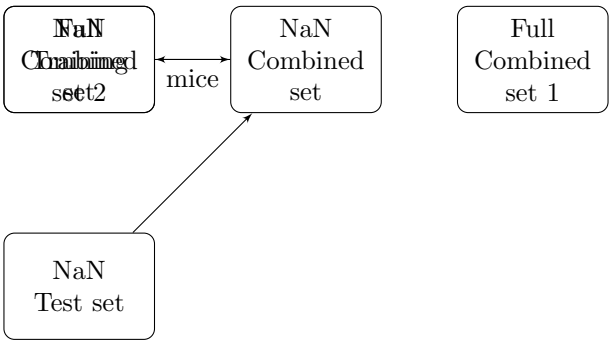

Figure 2: The pipeline

5 Feature-extraction-to-classification Pipeline

6 Results

In total, I made XX submissions during the competition. The best submission resulted in an accuracy of XX, which corresponds to position XX out of XX participants.

Using a support vector machine, random forest regression, and AdaBoost classifier resulted in very similar results.

7 Discussion

In summary, the main challenge lay in handling the missing data values. Afterward, any standard regression technique could be used. The use of different regression techniques had no substantial influence on the achieved accuracy.

My code for this competition can be found at:
`https://github.com/Pold87/ml-final-ass`

References

Azur, Melissa J et al. (2011). "Multiple imputation by chained equations: what is it and how does it work?" In: *International journal of methods in psychiatric research* 20.1, pp. 40–49.

Buuren, Stef and Karin Groothuis-Oudshoorn (2011). "mice: Multivariate imputation by chained equations in R". In: *Journal of statistical software* 45.3.

Little, Roderick JA (1988). "A test of missing completely at random for multivariate data with missing values". In: *Journal of the American Statistical Association* 83.404, pp. 1198–1202.