MACHINE LEARNING METHODS FOR CLASSIFICATION PROBLEMS*

Heiko Groenitz

Philipps University of Marburg, Germany
School of Business and Economics, Marburg, Germany
e-mail: groenitz@staff.uni-marburg.de

ŚLĄSKI
PRZEGLĄD
STATYSTYCZNY
Nr 18(24)

ISSN 1644-6739
e-ISSN 2449-9765

Quote as: Groenitz, H. (2020). Machine learning methods for classification problems. *Śląski Przegląd Statystyczny, 18*(24).

DOI: 10.15611/sps.2020.18.14

1. Introduction

Machine learning means the application of computer algorithms onto a dataset to discover structure. The term ‘machine’ indicates that a computer (i.e. machine) is usually needed to conduct the algorithms (large datasets, lots of calculations). The term ‘learning’ indicates that one would like to formulate some system from the data. The discovered structure is intended to be applied beneficially in the future.

In this contribution, the author focused on classification problems with predefined classes, taking a dataset with n statistical units. Each unit belongs to one of k classes. Let Y denote the class. In addition to class Y, we observe a vector of further characteristics $X = (X1, ..., X\_p)$. The structure of interest is function fallowing good predictions of Y based on the input characteristics X, i.e. requiring that $f(X\_1, ..., X\_p) = Y$ often holds.

Thus the found function f can be employed to predict class Y for new statistical units based on known values for the input variables X_1, ..., X_p.

In practice, classification problems often occur. For example, let us consider a bank that offers loans. The classes here may be ‘correct repayment’ and ‘problems with repayment’. Typical input characteristics are income, savings, real estate, duration of employment contract, further loans, age, and family status. The bank is interested in a prediction of

* 25th Scientific Statistical Seminar „Marburg-Wrocław”. Gollhofen, 23-26 September 2019. Extended abstract.

the repayment behaviour based on the input variables. Such a prediction function can be applied to decide new credit applications.

Many classification algorithms exist. There are approaches with a long tradition (e.g. discriminant analysis, logistic regression) and methods that have become increasingly popular (e.g. support vector machines, random forests). The aim of this paper was to review some classification methods, especially some newer techniques, and demonstrate the procedures in a real-data study with credit data.

2. Classification methods

This section outlines several classification methods, throughout restricts the descriptions to two classes: $Y = 1$ and $Y = 2$.

In discriminant analysis (e.g. [Jobson 1992, Chapter 8.2]), normal distribution of the input variables X_1, ..., X_p is assumed in each class. The parameters are estimated by the given data. The prediction is the class with the highest estimated probability given the outcomes of the input variables.

Logistic regression (e.g. [Pathak 2014, Chapter 7.2.2]) applies the logistic distribution function to model the chance of class $Y = 1$ given the explanatory variables X. The unknown parameter is specified via maximum likelihood estimation. The prediction is the category with the highest estimated chance given X.

In nearest neighbour classification (e.g. [Pathak 2014, Chapter 7.3.1]), the assignment rule for a unit is as follows: first, calculate the distance of the unit to any unit in the dataset with respect to X_1, ..., X_p. Second, identify K's nearest neighbours. Then, determine the distribution of the class variable Y among those neighbours. Finally, assign the class with the largest frequency (see: Figure 1).

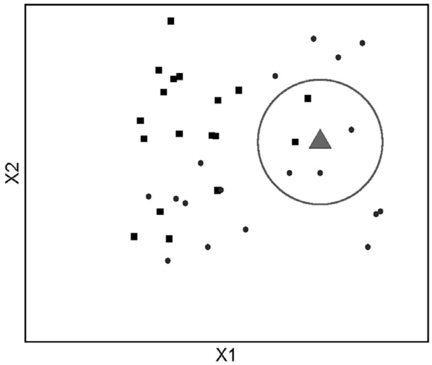

Fig. 1. Nearest neighbours classification: Class $Y = 1$ is represented by points, and $Y = 2$ by rectangles. The unit to classify is illustrated by a triangle

Source: own elaboration.

Figure 2 shows the basic idea of a support vector machine (SVM). We look for a strip that separates the groups and has maximum width. The decision border is the middle of the strip, and around the border there is a margin without observations. The points on the boundary of the strip are the support vectors. Removing such a point results in a new decision border.

A strip that divides the classes exactly often does not exist. Therefore we allow some points within the margin or misclassified points (see Figure 3). We can then speak of a soft margin SVM instead of a hard margin SVM. Here the strip is determined in a way that it has preferably a large width and few points within the margin or on the wrong side of it.

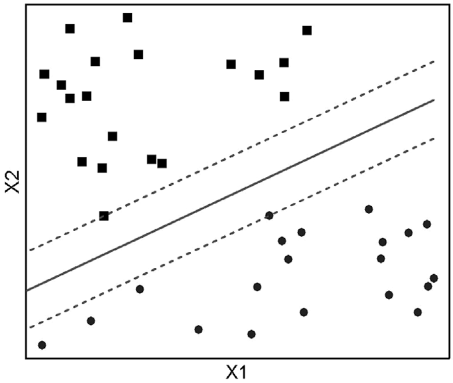

Fig. 2. The desired strip without observations in support vector machines

Source: own elaboration.

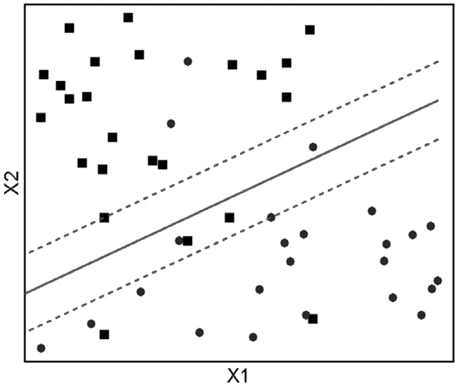

Fig. 3. Soft margin support vector machine

Source: own elaboration.

A further extension are non-linear SVMs. They are motivated by the fact that linear decision borders sometimes are not appropriate. As for linear SVMs, both hard and soft margins are possible for non-linear SVMs (cf. Figures 4 and 5). Further information on SVMs can be found in Cortes and Vapnik [1995], Moguerza and Munoz [2006], and Hamel [2009].

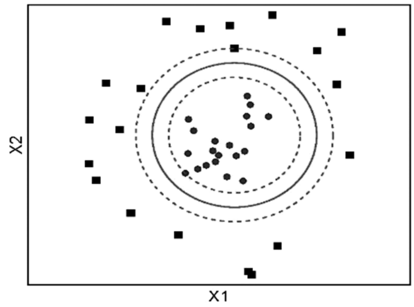

Fig. 4. Non-linear support vector machine with soft margin

Source: own elaboration.

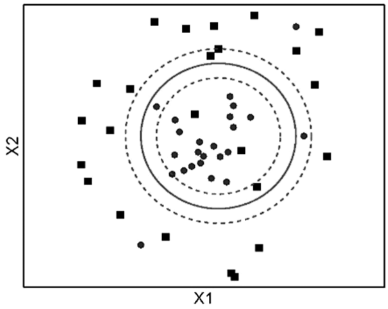

Fig. 5. Non-linear support vector machine with hard margin

Source: own elaboration.

The next method used was the classification tree, see e.g. Breiman et al. [1984] and Lantz [2015, Chapter 5]. Successively the observations were split into smaller groups. The goal was to obtain homogeneous groups with respect to class *Y*. Each partition was based on some input variable *X_i*. To make a prediction for a unit, one should detect the terminal point of the tree ('leaf') the unit belongs to and assign the most frequent class from this leaf. An example of a classification tree is given in Figure 6.

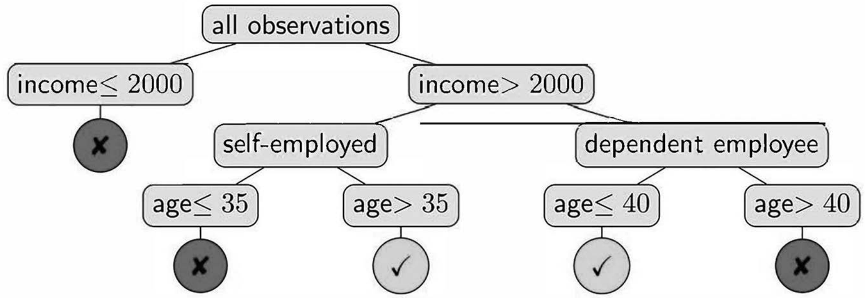

Fig. 6. Possible classification tree for repayment of credits. In two leaves, correct repayment predominates, while in three leaves, default occurred more often than correct repayment

Source: own elaboration.

The random forest (e.g. [Breiman 2001]) is an extension of the classification tree and utilizes an ensemble of many trees. Each tree is generated taking the following two principles into account: first, draw a random sample from the data (with replacement, same sample size *n*). We speak of a bootstrap sample and base the tree on this sample. Second, for any split, only a subset of input variables is available. The subset is selected randomly. For a prediction of class *Y*, one makes a classification with each tree, and the final classification comes from a majority vote.

3. Real-Data application with credit data

The authors now demonstrate the methods from Section 2 in a case study with a dataset of bank loans. The dataset was provided by the UCI Machine Learning Repository, see the following link:

https://archive.ics.uci.edu/ml/index.php.

The authors last viewed the website on November 22, 2019. The repository operator is the University of California, Irvine. The name of the dataset is "Statlog (German Credit Data)". For this study the program R was applied.

The dataset comprises 1000 loans in Germany. The dependent variable *Y* describes whether a loan was repaid correctly or not; 700 loans duly repaid.

There are 20 input characteristics, for example: duration of credit, designated use, credit amount, available savings, current duration of employment, ratio repayment rate/income, real estate in existence, age, rental flat or home ownership.

To measure the performance of an algorithm, the cross-validated error rate was computed, splitting dataset *D* into *N* = 10 parts D_1, ..., D_N. In step *i* (*i* = 1, ..., *N*), part D_i is the test dataset and the other parts are the training data. The classification rule is derived from the training data. The test data were used to compute the error rate corresponding to the classification rule.

In addition to the cross-validated error rate, the authors also computed the cross-validated monetary loss corresponding to an algorithm. The following assumptions were made: if target variable Y indicates that the repayment was not correct, assume that the repayment stopped half-way through the duration, taking an interest rate of 5% per year.

For this study, the authors also addressed the selection of method parameters and the variable selection. The method parameters were, for example, the number of neighbours for the nearest neighbour approach, the degree of softness of the margin in SVMs, or the number of trees for random forests.

Variable selection means deciding which input variables were finally included into the classification rule.

Let us assume the method parameters were fixed, then apply sequential backward selection to select the final set of input characteristics. Here, input variables were removed step by step. In each step, we removed the variable that led to the smallest reduction in power where the power was measured by cross-validated error rates or losses. A 'bath-tub effect' typically occurs (see Figure 7). Finally, the study involved those input variables that resulted in the largest power.

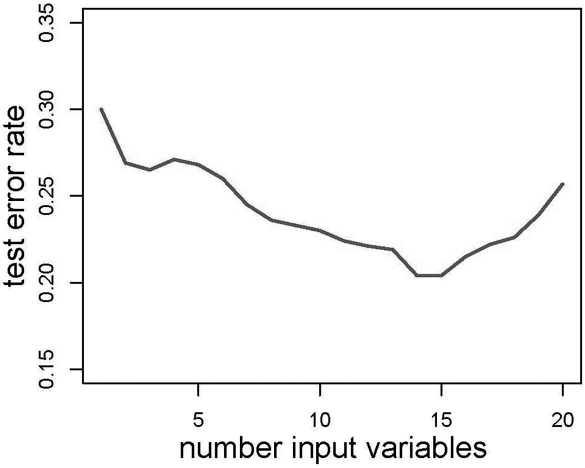

Fig. 7. 'Bath-tub effect' for a SVM. Here, 14 input variables should be included

Source: own elaboration.

To adjust the method parameters, one proceeds as follows: first, choose a starting parameter. Second, detect the corresponding optimal set of input variables. Third, record the power for this parameter and the optimal input variables. Steps 2 and 3 are repeated for other parameter values. Eventually, the parameter with the maximum power is chosen.

The results with respect to error rates are illustrated in Figure 8. For comparison, note that the trivial assignment that always predicts correct repayment had an expected error rate of 30%. It was recognized that the SVM performed best, and the second place went to the random forest. It turned out that adjusting the method parameters is important. Moreover, the finally selected input variables depend on the classification method and the most important explanatory variable varies from method to method.

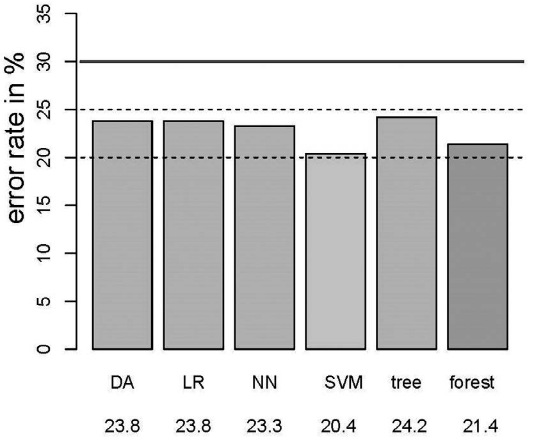

Fig. 8. Cross-validated error rate for discriminant analysis (DA), logistic regression (LR), nearest neighbour method (NN), support vector machine (SVM), classification tree (tree), and random forest (forest)

Source: own elaboration.

Figure 9 shows the findings for monetary losses. Again, first place goes to the SVM, and the random forest in the second place. It was found again that it is important to adjust the method parameters and conduct a variable selection.

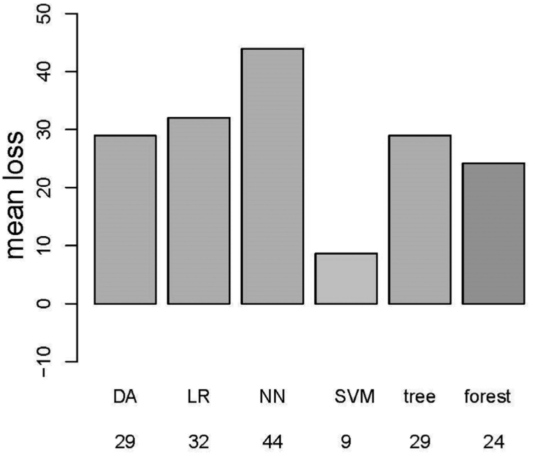

Fig. 9. Cross-validated monetary losses for discriminant analysis (DA), logistic regression (LR), nearest neighbour method (NN), support vector machine (SVM), classification tree (tree), and random forest (forest)

Source: own elaboration.

4. Discussion

The paper reviewed some machine learning methods for classification, and provided a basic impression, especially regarding the support vector machine and the random forest. The classification techniques were applied in the study on credit data, showing that the support vector machine and the random forest outperformed the other methods. Naturally the results cannot be generalized arbitrarily, however the potential was shown. To sum up, it can be said that increasing computer power enables the application of new computer-intensive methods. Thus the pool of methods becomes larger giving more options to improve classification power.

References

Breiman L. (2001), *Random forests*. Machine Learning Journal, 45, pp. 5-32.

Breiman L., Friedman J.H., Olshen R.A., Stone C.J. (1984), *Classification and regression trees*, Chapman & Hall/CRC, Boca Raton.

Cortes C., Vapnik V.N. (1995), *Support-vector networks*, Machine Learning, 20, pp. 273--297.

Hamel L. (2009), *Knowledge Discovery with Support Vector Machines*. John Wiley & Sons, Hoboken.

Jobson J.D. (1992), *Applied Multivariate Data Analysis – Volume II: Categorical and Multivariate Methods*, Springer, New York.

Lantz B. (2015), *Machine Learning with R. Packt Publishing*, Birmingham.

Moguerza J.M., Munoz A. (2006), Support vector machines with applications, Statistical Science, 21, pp. 322-336.

Pathak M.A. (2014), *Beginning Data Science with R*. Springer, Cham.

Keywords: algorithm, credit analysis, pattern recognition, random forest, support vector machine.