Weight-of-evidence 2.0 with shrinkage and spline-binning

Jakob Raymaekers

Department of Mathematics, KU Leuven, Leuven, Belgium

Department of Mathematics, University of Antwerp, Antwerp, Belgium

Wouter Verbeke

Faculty of Economics and Business, KU Leuven, Leuven, Belgium

Tim Verdonck[1]

Department of Mathematics, University of Antwerp, Antwerp, Belgium

Department of Mathematics, KU Leuven, Leuven, Belgium

Abstract

In many practical applications, such as fraud detection, credit risk modeling or medical decision making, classification models for assigning instances to a predefined set of classes are required to be both precise as well as interpretable. Linear modeling methods such as logistic regression are often adopted, since they offer an acceptable balance between precision and interpretability. Linear methods, however, are not well equipped to handle categorical predictors with high-cardinality or to exploit non-linear relations in the data. As a solution, data preprocessing methods such as weight-of-evidence are typically used for transforming the predictors. The binning procedure that underlies the weight-of-evidence approach, however, has been little researched and typically relies on ad-hoc or expert driven procedures. The objective in this paper, therefore, is to propose a formalized, data-driven and powerful method.

To this end, we explore the discretization of continuous variables through the binning of spline functions, which allows for capturing non-linear effects in

[1]Corresponding author. E-mail: tim.verdonck@uantwerpen.be

the predictor variables and yields highly interpretable predictors taking only a small number of discrete values. Moreover, we extend upon the weight-of-evidence approach and propose to estimate the proportions using shrinkage estimators. Together, this offers an improved ability to exploit both non-linear and categorical predictors for achieving increased classification precision, while maintaining interpretability of the resulting model and decreasing the risk of overfitting.

We present the results of a series of experiments in a fraud detection setting, which illustrate the effectiveness of the presented approach. We facilitate reproduction of the presented results and adoption of the proposed approaches by providing both the dataset and the code for implementing the experiments and the presented approach.

Keywords: Classification, Transformation, Preprocessing/Feature engineering, Interpretability, Regression analysis, Fraud detection

---

Contents

1 Introduction **4**

2 Literature review (background methodology) **5**

3 Methodology **8**

3.1 (Local) shrinkage of WOE 8

3.2 Discretized splines for non-linear effects 10

3.3 Model building 13

4 Empirical results **16**

4.1 Fraud detection data 16

4.2 Experimental design 16

4.2.1 Experiment 1 17

4.2.2 Experiment 2 18

4.2.3 Experiment 3 19

4.3 Results 20

4.3.1 Experiment 1 20

4.3.2 Experiment 2 23

4.3.3 Experiment 3 24

4.4 Discussion 28

5 Conclusion and further research **28**

1. Introduction

Classification is a well-studied machine learning task which concerns the assignment of instances to a set of outcomes. Classification models support the optimization of managerial decision-making across a variety of operational business processes. For instance, fraud detection models classify instances, such as transactions or claims, as fraudulent or non-fraudulent. This allows to efficiently and effectively allocate limited inspection capacity by selecting the most suspicious cases for investigation by a human fraud analyst [1].

A broad variety of powerful classification methods for learning classification models from data have been proposed in literature, such as neural networks, support vector machines and ensemble methods. These methods are intrinsically able to model highly complex data structures and have been shown to outperform simpler classification methods, such as logistic regression and single decision trees [2, 3]. In industry, however, simple logistic regression remains to date among the most frequently used approaches for developing classification models across various fields of application. This may be explained by users or regulation, e.g., the Basel regulation, requiring the resulting model to be both interpretable [4] and powerful, i.e., accurately classifying instances. Logistic regression is broadly perceived as offering the best balance between both objectives. Other possible explanations for the popularity of logistic regression are the broad expertise and experience in using logistic regression that exists in industry, but also a follow-the-herd behavior and some degree of inertia and resistance to change may explain its enduring popularity. Moreover, these complex analytical techniques sometimes only provide marginal performance gains on structured, tabular data sets as frequently encountered in common classification tasks such as credit scoring and healthcare analytics [2, 5].

Compared to learning classification models, relatively few studies focus on preprocessing data, i.e., transforming the data prior to learning a classification model, and on developing approaches for optimally preparing the data so as to maximize predictive power, out-of-sample performance, or, importantly,

to improve the interpretability of the resulting model. Specifically, we identify a lack of approaches that allow to optimally transform non-linear patterns and categorical variables with high cardinality for incorporation within linear models, so as to achieve an interpretable yet powerful classifier [6]. Currently, the weight-of-evidence (WOE) approach appears to be frequently used to this end as it offers a good balance between interpretability and predictive power, complementary with and similar to logistic regression [7, 8]. In case of high cardinality, however, WOE may lead to overfitting. Moreover, WOE does not have an integrated binning approach for optimally merging categories or discretizing continuous predictors.

In this article, we present an integrated WOE-based approach for optimally transforming predictor variables, which mainly improves upon the existing WOE approach in case of non-linear predictor variables (continuous or ordinal) as well as categorical predictor variables with high-cardinality, with the aim to maximize both predictive power and interpretability of, although not limited to, logistic regression models. The presented approach is experimentally evaluated, providing an illustration of the use and an indication of the merits of the proposed approach. An open source implementation of the method is provided in digital annex to this paper, so as to facilitate peer researchers to reproduce and verify the presented results, and practitioners to adopt the method for practical use. This paper is structured as follows. In the following section, we present the standard methodology that uses logistic regression and weight-of-evidence, upon which we expand in Section 3. In Section 4 we present experimental results in a fraud detection case, and in Section 5 we conclude and present directions for future research.

2. Literature review (background methodology)

Consider a model with binary response Y and p continuous predictors $\boldsymbol{X} = (X_1, \ldots, X_p)$. The goal is to model the conditional mean $p_{\boldsymbol{x}} = E(Y|\boldsymbol{X} = \boldsymbol{x}) = P(Y = 1|\boldsymbol{X} = \boldsymbol{x})$. The classical logistic regression model assumes a linear

relationship between the predictor variables and the log-odds of the event $Y = 1$. More specifically, we have

$$\log\left(\frac{p_{\boldsymbol{x}}}{1-p_{\boldsymbol{x}}}\right) = \beta_0 + \sum_{i=1}^{p} \beta_i x_i = \beta_0 + \boldsymbol{\beta}\boldsymbol{x}$$

where β_0 denotes an intercept, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ denotes a vector of model parameters. This model can be reformulated in terms of probabilities as

$$P(Y = 1|\boldsymbol{X} = \boldsymbol{x}) = \frac{1}{1 + e^{-(\beta_0 + \boldsymbol{\beta}\boldsymbol{x})}}.$$

The classical logistic regression model serves as a very popular benchmark for many binary classification tasks due to its ease of computation, high interpretability and solid performance. However, it also has several shortcomings, two of which we want to focus our attention on:

1. categorical variables with many categories
2. continuous variables with a non-linear effect on the log-odds

Categorical variables are often one-hot encoded (also known as "dummy encoding"), after which they can be included in the model as numerical variables. This has the drawback that a categorical variable with N categories leads to $N - 1$ variables. If N is large, this leads to a lot of variability in the estimation process and usually many insignificant predictors. One way to avoid this problem, is by converting the categorical variable into a continuous one by using a weight-of-evidence transformation. The weight-of-evidence (WOE) transformation of a categorical predictor is commonly defined as follows. Suppose we have a category j with N_j elements. Denote with P_j the number of true cases in our category and F_j the number of false cases in our category. Additionally, let P be the total number of true cases in the data and F the total number of false cases in the data. The WOE value of category j is then given by:

$$\log\left(\frac{P_j/P}{F_j/F}\right). \tag{1}$$

The WOE transformation usually provides an elegant solution, but since it is based on the estimation of a proportion, its variance can be high when there

are categories with few observations, which is common for categorical variables with high cardinality.

Continuous variables are modeled by logistic regression as having a linear effect on the log-odds of the response. While this is often a reasonable idea, there can be variables which do not satisfy this assumption. This happens unexpectedly, but sometimes by design. Suppose we want to predict whether a transaction is fraudulent based on a single predictor x_t which characterizes the time of a transaction in terms of hours (i.e. taking values in $[0, 24)$). Now suppose we make the reasonable assumption that the influence of the time on the probability of a transaction being fraudulent is roughly continuous and we interpret x_t close to 24 as x_t close to 0. Then we would have $P(Y = 1|x_t = 0) = \lim_{T \to 24^-} P(Y = 1|x_t = T)$. In terms of log-odds, this would imply that $\beta_0 = \lim_{T \to 24^-} \beta_0 + \beta_1 T = \beta_0 + 24\beta_1$, which clearly can only be satisfied when $\beta_1 = 0$. In other words, under the assumptions above, the only relationship that can be fit is a constant one, which is not of much interest. This example illustrates that some variables will display non-linear relationships with the response by design.

One way to incorporate non-linear effects of continuous predictors on the log-odds is to use the generalized additive model (GAM, [9]) for logistic regression:

$$\log\left(\frac{p_{\boldsymbol{x}}}{1 - p_{\boldsymbol{x}}}\right) = \beta_0 + \sum_{i=1}^{p} f_i(x_i) \tag{2}$$

where $f_1, \ldots, f_p$ are arbitrary smooth functions of the predictor variables $x_1, \ldots, x_j$. The model in 2 is very flexible but this flexibility comes at a price. As the functions f_i can be arbitrary smooth functions of the predictors, they can display rather unusual patterns. These make the model harder to interpret and hence lesser used in practical situations such as fraud detection where the predictions resulting from the model may have to be explained. In order to improve the interpretability of the model, [10] proposed a data-driven way of binning the fitted functions f_i into a limited number of categories. Afterwards, a classical logistic regression can be fit to the binned variable. This strategy allows for

capturing non-linear effects, while greatly improving the interpretability of the model.

3. Methodology

In the following, we describe our proposal to address the issues described in the previous section. The underlying goal is develop a powerful predictive model while maintaining interpretability, by allowing to incorporate non-linear effects within a logistic regression model and by improving upon traditional WOE binning

3.1. (Local) shrinkage of WOE

Our starting point for the treatment of categorical variables is the WOE transformation that transforms a categorical variable into continuous values. In order to introduce our shrinkage estimator of the WOE values, we will first rewrite the definition of Equation (1) in a different but equivalent form. More specifically, for a given categorical variable we assign the empirical log-odds to each bin, i.e. each element in a given category j gets assigned the value

$$\widehat{\mathrm{WOE}}_j = \log\left(\frac{\hat{p}_j}{1-\hat{p}_j}\right), \tag{3}$$

where $\hat{p}_j$ denotes the proportion of successes (e.g. fraudulent transactions) in category j. The equivalence with the earlier definition in Equation (1) can be seen as follows. With the notation introduced before, we have that $\mathrm{WOE}_j = \log\left(\frac{p_j}{1-p_j}\right) = \log\left(\frac{P_j/N_j}{F_j/N_j}\right) = \log\left(\frac{P_j}{F_j}\right) = \log\left(\frac{P_j/P}{F_j/F}\right) + \log\left(\frac{P}{F}\right)$. Therefore, both values only differ by a constant, which typically does not play a role in most statistical or machine learning models. As an example, the constant will disappear in the intercept for generalized linear models (GLMs, [11]. It is worth noting that sometimes, categories with $\hat{p} = 0$ or $\hat{p} = 1$ can occur, and that they lead to undefined WOE values. In those cases, we can slightly adjust the WOE by introducing a small offset c with $0 < c < 1$ and replace $\hat{p} = 0$

by $\hat{p} = \frac{c}{n_j}$ and $\hat{p} = 1$ by $\hat{p} = 1 - \frac{c}{n_j}$. Note that this offset disappears as the number of observations in the category becomes large (i.e. when $n_j \to \infty$). We use $c = 0.01$ by default. In practice, categories are often merged to avoid this boundary case, but this merging introduces a certain level of arbitrariness. In particular, it raises the question as to whether all possible combinations of categories should be considered as possible merging candidates. Additionally, it does not use the performance or quality of the final model in evaluating which merges are most interesting. We thus prefer working with a small offset, after which we can deal with the WOE values in a rigorous way.

For categories with a small number of observations n_j, the estimation of p_j (and the corresponding WOE_j) has a high variance, often yielding unreliable estimates. This is more likely to occur in categorical variables with many levels. In order to address this issue, we consider shrinkage estimation of the proportion of successes in each category j. The shrinkage estimator of a proportion is given by [12]:

$$\tilde{p}_j = (1 - b_j)\hat{p}_j + b_j\hat{p}$$

where $\hat{p}$ denotes the proportion of successes calculated over all possible values of j (i.e. over all categories). We thus effectively shrink the proportion of successes towards the sample mean. The shrinkage coefficient b_j determines the amount of shrinkage: $b_j = 0$ corresponds with no shrinkage, whereas $b_j = 1$ corresponds with taking the population proportion. The value of b_j is chosen to minimize the expected mean squared error over all estimated proportions, given by $\text{EMSE} = E_s[E_j[(\tilde{p}_j - p_j)^2|p_j]$. The minimum is given by (provided $n_j/n < 0.5$):

$$b_j^* = \frac{v_j(1 - n_j/n)}{v_j(1 - 2n_j/n) + v + \sigma^2}$$

where $v = \text{var}(\hat{p})$ is the sampling variance of $\hat{p}$, v_j denotes the sampling variance of $\hat{p}_j$ and σ^2 equals the between-area variance (i.e. $\text{var}_j(p_j)$). By plugging in the shrinkage estimator in the WOE calculation, we obtain the shrinkage estimator of the WOE values:

$$\widehat{\text{SWOE}}_j = \log\left(\frac{\tilde{p}_j}{1 - \tilde{p}_j}\right)$$

for each category j. In the rest of the paper, we will denote the WOE transformation based on shrinkage estimation of the proportions with SWOE($\cdot$).

In addition to the global shrinkage described above, which shrinks proportions towards the overall proportion in the data, we consider shrinking the proportions locally. More specifically, we cluster the WOE values using weighted k-means [13, 14], where the weights are taken inversely proportional to the sampling variability of the WOE values. Note that by the CLT and delta method, it holds that $\sqrt{n}(g(\hat{p}) - g(p)) \xrightarrow{D} N\left(0, \frac{1}{p(1-p)}\right)$, where $g(t) = \log\left(\frac{t}{1-t}\right)$. The asymptotic variance of the WOE estimates is thus $1/(np(1-p))$. Denoting with $z_1, \ldots, z_n$ the thus solve the optimization problem given by

$$\hat{B}_1, \ldots, \hat{B}_K = \underset{B_1,\ldots,B_K}{\arg\min} \sum_{k=1}^{K} \sum_{i \in B_k} w_i (z_i - \bar{z}_k)^2$$

where $w_i \sim n_{j_i} \hat{p}_{j_i}(1 - \hat{p}_{j_i})$ where j_i is the category of the original observation x_i. Note that these weights are small for categories with very few observations, which makes it more likely that these categories are put in the same cluster with other categories. Clustering the WOE values induces local shrinkage, since WOE values that are close together tend to end up in the same cluster and will receive an WOE value that is a weighted average of the WOE values in the cluster. In addition to achieving less variability in the estimation of the WOE values, we also obtain a natural "fusing" of the categorical variable which joins similar categories into a categorical variable with fewer categories. This allows for easier interpretation and visualization of the effect of the categorical variable. In the rest of the paper, we will denote the WOE transformation based on shrinkage estimation of the proportions with CWOE($\cdot$).

3.2. Discretized splines for non-linear effects

In order to allow for non-linear effects of the predictor variables on the log-odds, we revisit the approach of [10] and start from the generalized additive model (GAM) of Equation 2. After fitting GAM, the goal is to discretize the fitted spline functions into a limited number of bins. These can then be used

as input to a classical logistic regression. As such, we can capture non-linear effects, while greatly improving the interpretability of the model.

We differ in our approach in three main ways. First, we unite different binning types in one framework consisting of "constrained" and "unconstrained" binning. Both types have the same elegant objective function, with an additional constraint in the former case, which can be optimized exactly and efficiently. Second, we avoid the use of evolutionary trees for constrained binning, which are typically slower to compute and do not guarantee a global optimum of the objective function. Finally, our framework allows for a natural inclusion of weights in both types of binning, which are typically chosen inversely proportional to the variance of the estimated spline function at the observed value. This strategy avoids creating too many bins in regions of the spline function which are only supported by few observations.

Depending on the nature of the predictor variable, a different type of binning may be desirable. We distinguish two cases:

1. **Unconstrained** binning: the value of the original feature does not play a role in the binning.

2. **Constrained** binning: the value of the original feature puts a constraint on the binning we are looking for.

Let's consider an example. Suppose x_j is a variable characterizing the age of a person making a transaction. After fitting model 2, we have obtained a smooth function $f_j(x_j)$ which linearly influences the log-odds. Suppose we want to create bins for this transformed variable. If we would apply unconstrained binning, the binning of $f_j(X_j)$ would be independent of the value of X_j. This means that the resulting bins may combine different age groups. We could have a bin of ages $\{0-20, 80+\}$ and another bin of ages $\{21-79\}$. While this may be fine in some situations, there may also be situations where the binning is required to be contiguous in x_j, so as to facilitate a user in interpreting or explaining the model. This means that the categories cannot "jump" over ages.

An example of such a binning would be $\{0-50\}$ and $\{50+\}$.

Unconstrained binning is arguably the easiest problem. Given a predictor $\boldsymbol{x} = x_1, \ldots, x_n$ where $i = 1, \ldots, n$ ranges over the observations, consider the transformed values $z_i = f(x_i)$. We want to find K disjoint bins $\hat{B}_1, \ldots, \hat{B}_K$ for the original observations $x_1, \ldots, x_n$ such that within each bin, the corresponding values of z_i are roughly homogeneous. This is a univariate clustering problem for which many approaches have been proposed. We propose to optimize the weighted k-means objective function:

$$\hat{B}_1, \ldots, \hat{B}_K = \underset{B_1,\ldots,B_K}{\arg\min} \sum_{k=1}^{K} \sum_{i \in B_k} w_i (z_i - \bar{z}_k)^2$$

where $w_i \geq 0$ are weights such that $\sum_{i=1}^{n} w_i = n$. We choose the weights inversely proportional to the variance of the fitted function at point x_i. Once we have obtained the bins $\hat{B}_1, \ldots, \hat{B}_K$, we can transform the original predictor $\boldsymbol{x} = x_1, \ldots, x_n$ to $\bar{z}_{k_1}, \ldots, \bar{z}_{k_n}$ where k_i denotes the cluster to which observation $i = 1, \ldots, n$ was assigned. Alternatively, we can include the predictor as a categorical variable with the categories equal to the cluster memberships. We choose not to do this to avoid the creation of many dummy variables.

The weighted k-means clustering problems can be solved exactly in $\mathcal{O}(n \log(n))$ time using dynamic programming. Finally, note that the k-means approach with all weights equal to 1 is equivalent to Fisher's natural breaks algorithm [15] used in [10]. The issue of choosing the number of cluster K is a challenge in cluster analysis and a multitude of heuristic approaches exist. Among the more popular ones are the gap statistic [16] and the silhouette coefficient [17]. While these can be used in our setting, it may be more appropriate to adopt a hyperparameter tuning approach and determine the value of K by evaluating the quality of the resulting logistic regression model, since this is our primary concern. We will address this issue in Section 3.3.

We now turn to the problem of constrained binning. Consider again the transformed variable $z_i = f(x_i)$. In contrast to the unconstrained binning, the

value of x_i now influences the clustering of the values z_i. Suppose without loss of generality that the values of x_i are ordered in the relevant order (e.g. it contains the observed ages in ascending order). We are now interested in K bins $B_1, \ldots, B_K$ which each contain disjoint subsets of $x_1, \ldots, x_n$ such that if $x_i, x_j \in B_k$ for certain $i < j \in \{1, \ldots, n\}$, then $x_l \in B_k$ for all $i \leq l \leq j$. Of course we still want the bins to contain homogeneous values for the corresponding transformed values z_i. This problem is equivalent to fitting a step function to the set of bivariate points (x_i, z_i), i.e. we look for a piecewise-constant approximation of z_i within clusters of x_i. This problem has been considered in many areas including function approximation, time series analysis and cluster analysis. In the same spirit as the weighted K-means approach, we propose to optimize the weighted K-segments objective function:

$$\hat{B}_1, \ldots, \hat{B}_K = \underset{B_1,\ldots,B_K}{\arg\min} \sum_{k=1}^{K} \sum_{i \in B_k} w_i (z_i - \bar{z}_k)^2$$

which is the exact same objective as the weighted k-means problem, with the added constraint that the bins need to be contiguous. The weights $w_i \geq 0$ are again chosen to be inversely proportional to the variance of the fitted function at point x_i.

The k-segments clustering can be found exactly in $\mathcal{O}(n^2)$ time using dynamic programming, but an approximate $\mathcal{O}(n \log(n))$ algorithm exists [18]. Alternatively, one could use (evolutionary) regression trees to bin the z_i values. However, they are typically slower to compute and do not guarantee a global optimum of the objective function.

3.3. Model building

We now discuss how to incorporate the new techniques in building a model for logistic regression. For each continuous effect which is discretized into a step function, there is one tuning parameter in the form of the number of bins used. For the categorical data, the global shrinkage estimation of the WOE values does not have additional tuning parameters, but when using clustering to achieve local shrinkage, there is again the number of clusters as a tuning parameter.

Ideally, one would optimize a performance criterion of choice over all possible combinations of the tuning parameters, but this evidently becomes computationally cumbersome when there are multiple non-linear continuous variables and clustered categorical variables.

We propose to simplify the problem as follows. A simple AIC criterion for univariate k-means clustering is [19] AIC = WCSS + $2k$, where k equals the number of clusters and WCSS denotes the within-cluster sums of squares. One could use this formula to select the number of clusters for each clustering problem, but this would not take into account the performance of the final model. Therefore, we adapt this criterion by introducing a parameter that balances the strength of the fit with the number of clusters:

$$\text{WCSS} + \lambda k. \tag{4}$$

As λ increases, we encourage the algorithm to use fewer bins or clusters. When there are only continuous variables which need to be preprocessed, we propose to use two tuning parameters, λ_c and λ_{uc} for the constrained and unconstrained effects respectively. Note that it is necessary to distinguish between these two effects, since the constrained problem will have a naturally higher WCSS. For tuning the model, we thus use the procedure outlined in Table 1.

In the end, we choose the combination of tuning parameters yielding the lowest AIC value. This procedure can be used in combination with the shrinkage estimation of the WOE values, since the latter procedure does not have any tuning parameter. If the WOE values need to be clustered as well, there is one additional tuning parameter λ_{cat}. In that case, this parameter is optimized first, as the nature of an effect (linear vs. non-linear) may change after clustering the WOE values. For each value of λ_{cat} we thus execute the procedure outlined in Table 2, after which we can tune the remaining parameters using the previous procedure in Table 1. The value of λ_{cat} yielding the lowest AIC of the resulting GAM is retained, after which we can proceed with tuning λ_c and λ_{uc}. Note that by using a validation set approach, the evaluation of the resulting GLM can also be done using out-of-sample criteria, such as a measure of prediction

Tuning of λ_c and λ_{uc}	
Step 1	Bin the unconstrained non-linear continuous effects, using the number of bins k which yields the lowers value of the objective in Equation 4 with $\lambda = \lambda_{uc}$.
Step 2	Bin the unconstrained non-linear continuous effects, using the number of bins k which yields the lowers value of the objective in Equation 4 with $\lambda = \lambda_c$.
Step 3	Fit a glm using the binned effects, possibly including other variables.
Step 4	Evaluate the glm model using the AIC criterion.

Table 1: Tuning strategy for the clustering of the WOE of categorical variables.

Tuning of λ_{cat}	
Step 1	For each categorical variable, find the number of clusters associated with the value of $\lambda = \lambda_{cat}$.
Step 2	Cluster the WOE values of each of the categorical variables using the appropriate number of clusters found in the previous step.
Step 3	Train the GAM using the splines for the continuous variables, and the clustered WOE values for the categorical variables.
Step 4	Evaluate the GAM using the AIC criterion.

Table 2: Tuning strategy for the clustering of the WOE of categorical variables.

accuracy. This requires more data to be available and more computation time, but prevents better against overfitting of the training data. The parameters λ_{co} and λ_{ca} yielding the lowest out-of-sample prediction errors are then retained, and the final model is fit using these values.

4. Empirical results

4.1. Fraud detection data

This data set is on fraud detection in credit card transactions done on the East Coast of the USA. It consists of 6669 observations of 5 variables: `amount`, `age`, risk `category` (previously assigned by the bank), `country` and `time`. The response is a binary variable indicating fraudulent transactions, of which there are 73 in this dataset (i.e. roughly 1 %). Table 4.1 presents an overview of the variables in the dataset.

Variable name	description
`amount`	transaction amount (USD)
`age`	age of the person executing the transaction
`category`	risk category of the transaction (low-medium-high)
`country`	transaction destination (43 countries)
`time`	time of transaction (0-24h)

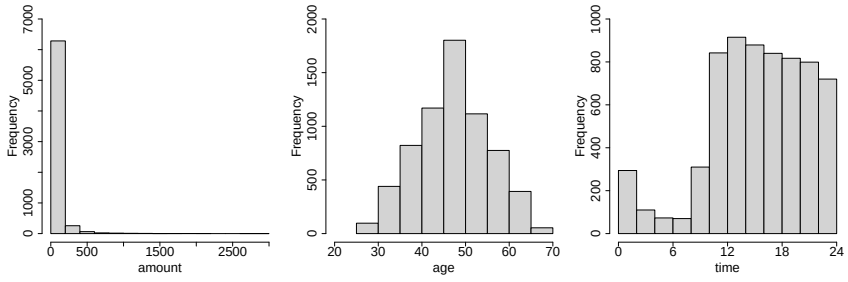


Figure 1: Histograms of the continuous variables in the credit card fraud data. The `amount` variable is heavily right skewed, and the `time` variable shows few transactions between 1 and 7 a.m..

4.2. Experimental design

In order to avoid interference of the effects caused by the treatment of continuous and discrete variables, we set up two experiments.

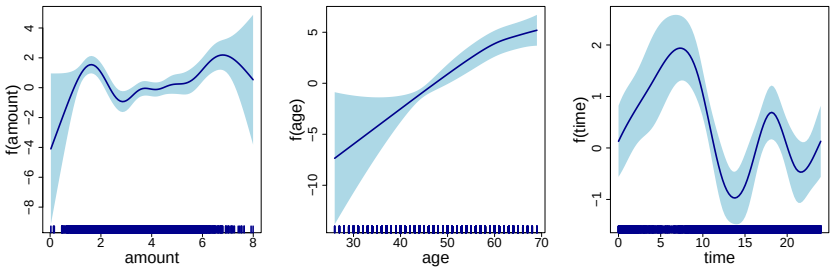


Figure 2: Result of a univariate GAM fit to the continuous predictors. The fitted splines suggest non-linear effects for the `amount` and `time` variables on the log-odds.

4.2.1. Experiment 1

In the first experiment, we use the continuous variables to predict the fraudulent transactions. In order to quickly scan for the variables which may have a potential non-linear effect with the response, we fit a univariate GAM on each of the continuous predictors. Figure 2 shows the results, indicating that the `amount` and `time` variables are likely to influence the log-odds of fraud in a non-linear way. Note that the `time` variable is a typical example of an inherent non-linear effect, as discussed in Section 3.

We randomly split up the data in a training and test set of equal size, keeping the percentage of fraudulent cases equal to the population percentage. Denoting with p the probability of fraud, we train the GAM:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \texttt{age} + f_1(\texttt{amount}) + f_2(\texttt{time}) \tag{5}$$

where f_1 is a thin-plate regression spline and f_2 is a cyclic cubic regression spline, which captures the periodic nature of the time-effect.

In the second step, the continuous effects $f_1(\texttt{amount})$ and $f_2(\texttt{time})$ are discretized (i.e. approximated by step functions) using the strategy of section 3.3 to obtain `f(amount)` and `f(time)`. The `amount` variable is discretized using constrained binning, whereas we use unconstrained binning for the `time` variable.

Finally, a classical logistic regression is fit to the transformed variables:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \texttt{age} + \beta_2 \text{f}_1\texttt{(amount)} + \beta_3 \text{f}_2\texttt{(time)}$$

The results are evaluated based on different criteria. In addition to the AIC on the training set, we also compare the AUC, a weighted brier-score and the H-measure on the test set. The AUC is the well-known area under the receiver operating curve (also equivalent to a linearly transformed Gini coefficient). The classical brier-score is the mean squared error between the predicted probabilities and observed response, i.e. $\frac{1}{n}\sum_{i=1}^{n}(\hat{p}_i - y_i)^2$. This measure is clearly inadequate for imbalanced classification tasks, as it gives equal importance to each individual prediction. We therefore use weights inversely proportional to the prior probabilities: $\text{wbrier} = \frac{1}{n}\sum_{i=1}^{n} w_i(\hat{p}_i - y_i)^2$ where $w_i = \frac{1}{\pi_0} I_{y_i=0} + \frac{1}{\pi_1} I_{y_i=1}$. Note that these weights make the predictions of all fraudulent cases together as important as those of all the regular transactions. The H-measure is a more recent alternative to the AUC which avoids the dependence on the classifier and is therefore more reliable. It requires the severity ratio as an input, for which we take recommended ratio of the class priors (π_1/π_0), see [20, 21] for details.

4.2.2. Experiment 2

For the second experiment, we concentrate on the two categorical variables in the dataset, `category` and `country`. We are interested in training the GLM:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 g(\texttt{category}) + \beta_2 g(\texttt{country})$$

where g can be either the classical WOE transformation (WOE($\cdot$)), the shrinkage WOE transformation (SWOE($\cdot$)), or the clustered WOE (CWOE($\cdot$)).

The effect of shrinkage estimation of WOE will be most pronounced in situations where the variances of the estimated proportions are rather large, i.e. in smaller datasets. We thus need a careful experimental setup to illustrate the potential advantages of shrinkage estimation for WOE. We proceed as follows. We randomly split up the data in five parts while keeping the proportion of fraudulent transactions constant in each split. Each split is then used as a

training dataset and validated on the observations not in the split. In other words, we use 5-fold cross validation where the smaller dataset is used for training. This procedure enforces the use of small training datasets, in which the effect of shrinkage estimation will be most pronounced. We also repeat this experiment with 10 splits.

The performance of the method is again evaluated using the AIC, AUC, weighted brier score and H-measure.

4.2.3. Experiment 3

In the final experiment, we combine both techniques on a 50-50 training-test split of the data as in experiment 1. The goal is to evaluate the different combinations of the treatment of categorical and continuous variables. We use the same performance measures as in the previous experiments. For the combination of discretized splines with the shrinkage estimation of the WOE values, we first convert the catgorical variables into continuous ones use the shrinkage estimators. Then, we proceed as in Experiment 1, with the difference that the GAM now includes the transformed categorical variables:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \texttt{age} + \beta_2\,\text{SWOE}(\texttt{category}) + \beta_3\,\text{SWOE}(\texttt{country}) + f_1(\texttt{amount}) + f_2(\texttt{time})$$

where f_1 is a thin-plate regression spline and f_2 is a cyclic cubic regression spline, which captures the periodic nature of the time-effect.

For the combination of the clustered WOE values with the discretized splines, we follow the strategy outlined in Section 3.3. We thus first optimize the number of clusters for each of the categorical variables using the approach in Table 2. Afterwards, we proceed with as in Experiment 1 but now with the GAM:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \texttt{age} + \beta_2\,\text{CWOE}(\texttt{category}) + \beta_3\,\text{CWOE}(\texttt{country}) + f_1(\texttt{amount}) + f_2(\texttt{time})$$

For the evaluation, we use the same performance measures as in the previous two experiments: the AIC, AUC, weighted brier score and H-measure.

4.3. Results

4.3.1. Experiment 1

The initial fit of the GAM of Equation (5) yields the estimates $\hat{\beta}_0 = -19.518$ and $\hat{\beta}_1 = 0.268$ in addition to the spline functions f_1 and f_2 shown in Figure 3. The fitted amount effect suggests that extreme amounts (both large and small) are more likely to be fraudulent. The time effect suggests that transactions in the morning and late afternoon are more likely to be fraudulent, whereas transactions in the early afternoon and early evening are less likely to be fraudulent. The fitted GAM has an AIC of 284.495. For the out-of-sample measures, we have an AUC of 0.919, a weighted brier score of 0.407 and a H-measure of 0.604. This is a reasonable performance, and we will compare it to the final model and classical GLM later.

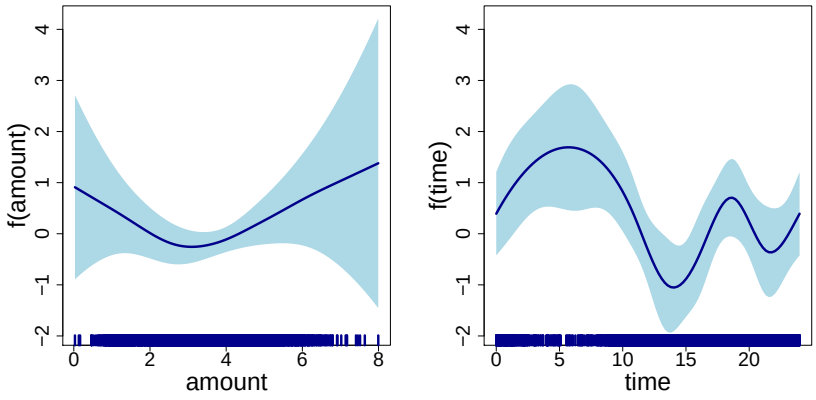


Figure 3: The estimated spline functions of the initial GAM fit.

We now discretize the fitted spline functions. We choose a maximum of $k = 10$ bins and use the selection strategy detailed in Section 3.3. This yields 7 bins for the constrained `amount` binning, and 6 bins for the unconstrained binning of the `time` variable. Figure 4 shows the original and binned effects of both variables. In the left panel, we see the amount variable which discretized in a step function with 7 steps. Note that the first and last steps span a rather large interval of transaction amount. The reason is that the observations in

these regions are fewer in numbers and the variance of the estimated spline is much larger. Therefore, due to the weighting strategy with weights inversely proportional to the variances, we obtain larger bins at the extremes of the spline. The right panel shows the time variable, which we have wrapped around a circle in a clock plot for the purpose of presentation. This plot visually illustrates the time windows in which transactions are more likely to be fraudulent. Note that an effect like this could never be estimated using classical logistic regression.

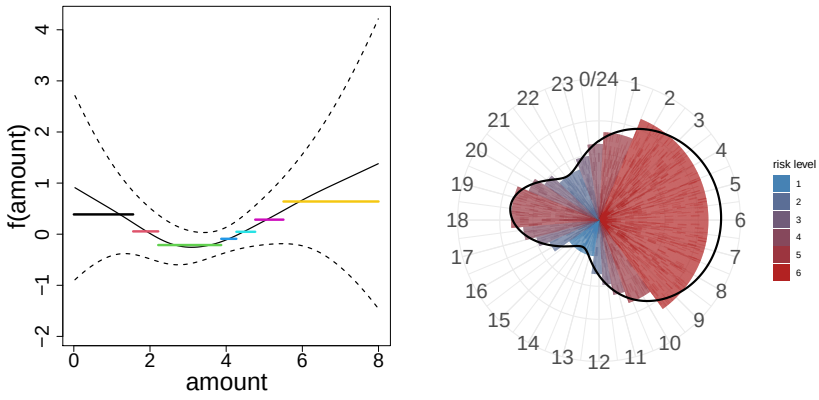


Figure 4: The estimated spline functions of the initial GAM fit.

We now evaluate the performance of the obtained model using the various performance measures discussed above. The final GLM fit on the discretized splines and the original `age` variable has an AIC of 286.429. This is slightly above the AIC of the full GAM, but it is clear that the difference is rather small. Furthermore, the tables turn when considering out-of-sample performance. The proposed method yields an AUC of 0.925, a weighted brier score of 0.396 and a H-measure of 0.624. All of these are in fact better than the corresponding performance measures of the full GAM fit. This can be explained by the fact that the classical GAM may slightly overfit the training data. By discretizing the resulting spline functions, we gain robustness against this overfitting. Table 3 shows the comparison of the performances. We have additionally added the results of the classical GLM. We see that the enhanced GLM outperforms the

classical GLM on all levels. The most significant difference is found in the H-measure, with an increase of almost 15 %.

Method	AIC	AUC	wbrier	H-measure	maxMCC	AUCPR
`classical glm`	293.656	0.896	0.438	0.549	0.226	0.101
`classical GAM`	**284.495**	0.919	0.407	0.604	0.296	0.175
`enhanced glm`	286.429	**0.925**	**0.396**	**0.624**	0.288	0.190
`xgboost`	NA	0.891	0.363	0.567	**0.360**	**0.241**

Table 3: Comparison of the different models trained on the continuous predictors of the fraud detection data. The enhanced glm outperforms the other methods in out-of-sample evaluation, whereas the classical GAM has a slightly lower AIC.

4.3.2. Experiment 2

The results of experiment 2 are summarized in tables 4 and 5 for respectively 5 and 10 splits. The conclusions for both tables are similar. In terms of out-of-sample performance, the shrinkage estimation of the WOE values yields a clear improvement over the baseline of just using the standard WOE values. The clustering of the WOE values shows a slight in-sample advantage in terms of AIC, but this does not seem to translate directly to an improved out-of-sample performance. The main benefit of the latter approach is thus that is improves interpretability by reducing the number of categories in the categorical variable.

Method	AIC	AUC	wbrier	H-measure
`woe`	112.593	0.733	0.442	0.283
`shrinkage woe`	113.258	**0.744**	**0.437**	**0.302**
`clustered woe`	**112.569**	0.732	0.443	0.278

Table 4: Comparison of the different strategies for treating categorical variables using 5 splits in the data. The WOE values based on shrinkage estimation generally outperform the other options in out-of-sample evaluation, whereas the glm using clustered WOE has a slightly lower AIC.

Method	AIC	AUC	wbrier	H-measure
`woe`	53.713	0.714	0.451	0.248
`shrinkage woe`	54.227	**0.726**	**0.449**	**0.256**
`clustered woe`	**53.672**	0.714	0.451	0.243

Table 5: Comparison of the different strategies for treating categorical variables using 10 splits in the data. The WOE values based on shrinkage estimation generally outperform the other options in out-of-sample evaluation, whereas the glm using clustered WOE has a slightly lower AIC.

4.3.3. Experiment 3

In the final experiment we compare the different combinations of our proposed preprocessing techniques. The results of this comparison are presented in Table 6. Several interesting conclusions can be made from these results. First of all, we see that the classical GLM is vastly outperformed by any of the other methods. This is mainly due to the inclusion of 42 dummy variables for the categorical variable `country`. Second, we can see that the shrinkage estimation of the WOE values outperforms the classical WOE, whether or not the continuous effects are estimated using discretized splines. The clustered WOE values do not significantly outperform the classical WOE values, and their main benefit thus lies in the fact that the final model is more interpretable, since it enforces a natural reduction of the number of categories in the categorical variables. Finally, we see that the discretized spline approach always improves upon the model using the original continuous variables.

WOE	sWOE	cWOE	enhanced	AIC	AUC	wbrier	H	maxMCC	AUCPR
☐	☐	☐	☐	285.264	0.831	0.366	0.520	0.322	0.212
☑	☐	☐	☐	226.609	0.925	0.352	0.596	0.358	0.257
☐	☑	☐	☐	226.179	0.928	0.354	0.615	0.356	0.249
☐	☐	☑	☐	225.405	0.924	0.357	0.589	0.346	0.264
☑	☐	☐	☑	216.578	0.941	**0.335**	0.638	0.379	0.253
☐	☑	☐	☑	**216.185**	**0.943**	**0.336**	**0.653**	0.387	0.250
☐	☐	☑	☑	219.345	0.936	**0.336**	0.627	0.362	**0.277**
☐	☐	☐	xgb	NA	0.905	0.347	0.637	0.343	0.254
☑	☐	☐	xgb	NA	0.918	0.364	0.611	**0.404**	0.274

Table 6: Evaluation of the combined strategies. Shrinkage estimation of the WOE values in combination with the discretized splines outperforms the other models. Clustered WOE values in combination with discretized splines is the second best performing model.

For illustrative purposes, we further analyze the model using clustered WOE

values and discretized splines. The clustering of the categorical variables yields an optimal tuning parameter of $\lambda_{\text{cat}} = e^{-7}$. This parameter enforces a clustering of the `country` variable into 12 bins (down from 42 categories), whereas the `category` variable is left untouched with its original 3 categories. Figure 5 shows the binned `country` variable, with 12 different levels. It turns out that transactions going to Europe are generally connected to lower probabilities of fraud, with the exception of receivers in Greece and the UK to a lesser extent. The highest risk is associated with national transactions and those to Canada and Mexico. International transactions to Australia, China, South Africa and Chili have a neutral risk level.

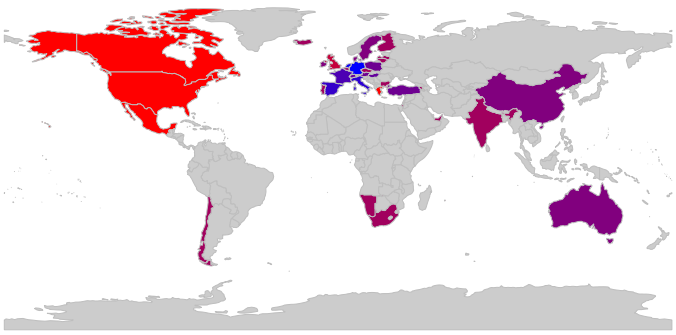


Figure 5: The `country` variable, reduced to 12 categories instead of the original 42.

The GAM fit with the optimal value of λ_{cat} no longer displays a non-linear effect for the `amount` variable as was the case in experiment 1. This means that the inclusion of the categorical variables resolves the non-linearity issue for this variable and we can treat it as a linear effect. The `time` variable however does still display a non-linear relationship with the response, as shown in Figure 6.

Discretizing the continuous effect `time` variable yields 3 bins. The result of the latter discretization step is shown in Figure 7. It is clear that the transactions made in the morning or early evening are more likely to be fraud than the transactions around noon or late in the evening. The coefficients of the

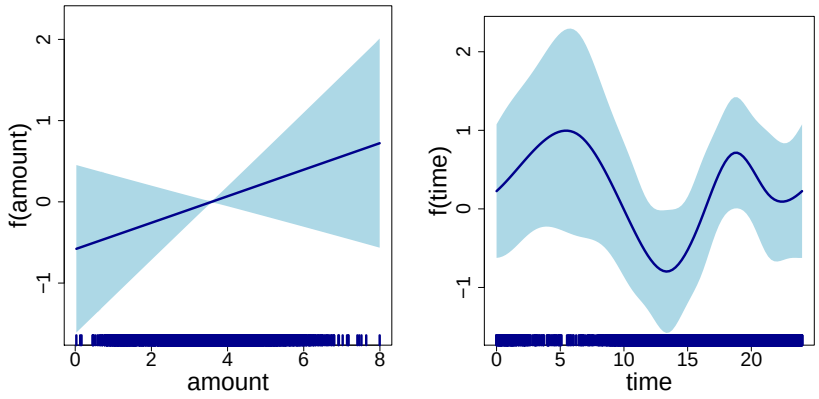


Figure 6: The estimated spline functions of the initial GAM fit when all variables are included in the model. The `amount` variable no longer displays a non-linear effect on the response variable, as was the case for the model with only continuous variables.

final model are presented in Table 7, which suggests that all predictors have a significant contribution to the model.

	Estimate	P-value
(Intercept)	-12.55	0.00
`amount`	0.19	0.19
`age`	0.27	0.00
CWOE(`category`)	0.64	0.01
CWOE(`country`)	0.90	0.00
`f(time)`	1.88	0.00

Table 7: Coefficients of the final model

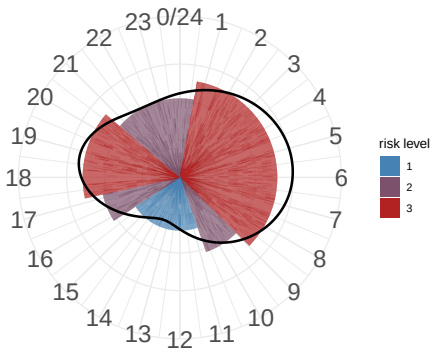


Figure 7: The discretized time effect in the final model.

4.4. Discussion

The results of the experiments above lead us to several conclusions. First of all, when it comes to the estimation of WOE values, estimating the proportions using the shrinkage estimator seems to improve the out-of-sample performance of the model. Secondly, clustering the WOE values does not generally yield a substantial improvement over the regular WOE values but has the advantage of fusing the categorical variable into a variable with fewer categories, which improves the interpretability of the model. Finally, the use of discretized splines on the continuous variables significantly improves the out-of-sample performance of the model. Additionally, one could argue that this also leads to improved interpretability, as the continuous variable is reduced to a select number of discrete values. Note that the advantage of using the discretized splines may not be significant if there are no important non-linear effects in the set predictor variables.

5. Conclusion and further research

We have proposed and studied two advanced techniques for preprocessing the data for logistic regression. The first considers the treatment of WOE values, which we propose to estimate using shrinkage estimators for the proportions. Alternatively, the original WOE values can be clustered for improved interpretability. Secondly, we have studied the discretization of continuous variables through the binning of spline functions. It allows for capturing non-linear effects in the predictor variables and yields highly interpretable predictors taking only a small number of discrete values.

Through three different experiments on the fraud detection dataset, we have illustrated the advantages of using the advanced preprocessing techniques. In particular, the out-of-sample performance was improved using the discretized spline treatment of the continuous variables. Additionally, the WOE values based on shrinkage estimation of the proportions also increased the out-of-

sample performance. The clustering of WOE values showed improved interpretability, but no clear improvement in predictive performance.

Further research could treat the combination of the two strategies for categorical variables, by using the classical WOE values as an input to a GAM. This combined method would be able to capture non-linear effects of the WOE values on the response. However, due to the nature of WOE in logistic regression (which implies a linear effect of WOE on the response), it is not clear that this would yield an improvement over the current method. Another line of research could investigate a more precise approximation of the spline functions in the GAM. For example, one could use a piecewise linear approximation instead of a step function, which would still be easy to interpret but more flexible to work with. Finally, the shrinkage estimation of the proportions could be combined with the clustering, i.e. one could first compute WOE values based on shrinkage estimation and then cluster the resulting values in a number of bins.

Acknowledgements

The authors gratefully acknowledge the financial support from the BNP Paribas Fortis Research Chair in Fraud Analytics at KU Leuven and the Internal Funds KU Leuven under grant C16/15/068. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

[1] B. Baesens, V. Van Vlasselaer, W. Verbeke, Fraud analytics using descriptive, predictive, and social network techniques: a guide to data science for fraud detection, John Wiley & Sons, 2015.

[2] S. Lessmann, B. Baesens, H.-V. Seow, L. C. Thomas, Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research, European Journal of Operational Research 247 (1) (2015) 124–136.

[3] H.-Y. Shi, K.-T. Lee, H.-H. Lee, W.-H. Ho, D.-P. Sun, J.-J. Wang, C.-C. Chiu, Comparison of artificial neural network and logistic regression models for predicting in-hospital mortality after primary liver cancer surgery, PloS one 7 (4) (2012) e35781.

[4] D. Martens, J. Vanthienen, W. Verbeke, B. Baesens, Performance of classification models from a user perspective, Decision Support Systems 51 (4) (2011) 782–793.

[5] A. Rajkomar, E. Oren, K. Chen, A. M. Dai, N. Hajaj, M. Hardt, P. J. Liu, X. Liu, J. Marcus, M. Sun, et al., Scalable and accurate deep learning with electronic health records, NPJ Digital Medicine 1 (1) (2018) 18.

[6] J. Moeyersoms, D. Martens, Including high-cardinality attributes in predictive models: A case study in churn prediction in the energy sector, Decision support systems 72 (2015) 72–81.

[7] E. P. Smith, I. Lipkovich, K. Ye, Weight-of-evidence (woe): quantitative estimation of probability of impairment for individual and multiple lines of evidence, Human and Ecological Risk Assessment 8 (7) (2002) 1585–1596.

[8] R. Anderson, The credit scoring toolkit: theory and practice for retail credit risk management and decision automation, Oxford University Press, 2007.

[9] T. Hastie, R. Tibshirani, Generalized additive models: some applications, Journal of the American Statistical Association 82 (398) (1987) 371–386.

[10] R. Henckaerts, K. Antonio, M. Clijsters, R. Verbelen, A data driven binning strategy for the construction of insurance tariff classes, Scandinavian Actuarial Journal 2018 (8) (2018) 681–705. arXiv:https://doi.org/10.1080/03461238.2018.1429300, doi:10.1080/03461238.2018.1429300.
URL https://doi.org/10.1080/03461238.2018.1429300

[11] J. A. Nelder, R. W. Wedderburn, Generalized linear models, Journal of the Royal Statistical Society: Series A (General) 135 (3) (1972) 370–384.

[12] N. T. Longford, Multivariate shrinkage estimation of small area means and proportions, Journal of the Royal Statistical Society: Series A (Statistics in Society) 162 (2) (1999) 227–245.

[13] J. MacQueen, et al., Some methods for classification and analysis of multivariate observations, 1967.

[14] S. Lloyd, Least squares quantization in pcm, IEEE transactions on information theory 28 (2) (1982) 129–137.

[15] W. D. Fisher, On grouping for maximum homogeneity, Journal of the American Statistical Association 53 (284) (1958) 789–798.
URL http://www.jstor.org/stable/2281952

[16] R. Tibshirani, G. Walther, T. Hastie, Estimating the number of clusters in a data set via the gap statistic, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63 (2) (2001) 411–423. arXiv:https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/1467-9868.00293, doi:10.1111/1467-9868.00293.
URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/1467-9868.00293

[17] P. J. Rousseeuw, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, Journal of Computational and Applied Mathematics 20 (1987) 53–65. doi:https://doi.org/10.1016/0377-0427(87)90125-7.

URL http://www.sciencedirect.com/science/article/pii/0377042787901257

[18] H. Wang, M. Song, Ckmeans. 1d. dp: optimal k-means clustering in one dimension by dynamic programming, The R journal 3 (2) (2011) 29.

[19] S. A. Ramsey, S. L. Klemm, D. E. Zak, K. A. Kennedy, V. Thorsson, B. Li, M. Gilchrist, E. S. Gold, C. D. Johnson, V. Litvak, et al., Uncovering a macrophage transcriptional program by integrating evidence from motif scanning and expression dynamics, PLoS Comput Biol 4 (3) (2008) e1000021.

[20] D. J. Hand, Measuring classifier performance: a coherent alternative to the area under the roc curve, Machine learning 77 (1) (2009) 103–123.

[21] D. J. Hand, Evaluating diagnostic tests: the area under the roc curve and the balance of errors, Statistics in medicine 29 (14) (2010) 1502–1510.