Ensemble distance-based regression and classification for large sets of mixed-type data

Amparo Baíllo (Univ. Autónoma de Madrid)

Aurea Grané (Univ. Carlos III de Madrid)

Mathematics 2021, 9, 2247

- **1.** Introduction
- **2.** Distance-based linear and logistic regression
- **3.** Ensemble techniques
- **4.** Aggregation and DB-LM on two real data sets
- **5.** Bagging and DB-GLM: a simulation study

New Bridges between Mathematics and Data Science
10 Nov. 2021, Valladolid (Spain)

1. Introduction

Nowadays the nature of data is of mixed type: quantitative and qualitative variables, textual and functional data, manifolds, ...

We focus on data mixing qualitative and quantitative variables.

Many classical multivariate analysis techniques just deal with quantitative data. But statistical problems involving multivariate mixed data can also be solved via an adequate inter-object metric.

Some statistical techniques, such as certain clustering procedures, multidimensional scaling (MDS) or the distance-based linear model (DB-LM), use only the matrix of pairwise distances as an input to perform the procedure.

When the sample size n is very large, the dimension $n \times n$ of the matrix of pairwise distances between sample individuals can be prohibitively large to carry out DB regression.

Ensemble methods, such as bagging or stacking have dealt flexibly with high-dimensional data sets in regression and classification (Bühlmann 2003).

Big data sets frequently have an inhomogeneous structure due to, e.g., the presence of outliers or mixed populations. Bühlmann and Meinshausen (2015) proposed maximin aggregation (magging) to adapt the subsampling and aggregation techniques in linear regression to large samples of heterogeneous observations.

Aim: To check if subsampling and aggregation strategies circunvent the computational problems posed by a large sample size in distance-based linear and logistic regression.

2. Distance-based linear and logistic regression

Distance-based (DB) linear regression is a prediction procedure that can be applied to qualitative or mixed regressors.

It was introduced by Cuadras (1989) and extended to the framework of functional data by Boj *et al.* (2010) and to the case of generalized linear models by Boj *et al.* (2016).

When all the predictors are quantitative and the metric is the Euclidean one, then the DB linear regression is equivalent to the classical least-squares linear regression (Cuadras and Arenas 1990).

For each sample individual we observe the value of a response y and a "vector" $\mathbf{z}$ of predictive variables, which can be qualitative. The sample values are $\mathbf{y} = (y_1, \ldots, y_n)'$ and $\mathbf{z}_1, \ldots, \mathbf{z}_n$.

Let $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)'$ be the weight vector, satisfying $\omega_i \in (0, 1)$ and $\sum_{i=1}^{n} \omega_i = 1$, and define $\mathbf{D}_{\boldsymbol{\omega}} = \text{diag}(\boldsymbol{\omega})$.

We define a distance or dissimilarity δ between any two values of $\mathbf{z}$. Denote by $\boldsymbol{\Delta} = (\delta^2(\mathbf{z}_i, \mathbf{z}_j))_{i,j=1}^{n}$ the matrix of squared distances.

The inner-products or Gram matrix is $\mathbf{G}_{\boldsymbol{\omega}} = -\frac{1}{2}\mathbf{J}_{\boldsymbol{\omega}}\,\boldsymbol{\Delta}\,\mathbf{J}'_{\boldsymbol{\omega}}$, where $\mathbf{J}_{\boldsymbol{\omega}} := \mathbf{I} - \mathbf{1}\boldsymbol{\omega}'$ is the $\boldsymbol{\omega}$-centering matrix, $\mathbf{I}$ is the $n \times n$ identity matrix and $\mathbf{1}$ is an $n \times 1$ vector of 1's.

If an $n \times k$ matrix $\mathbf{X}_\omega$ satisfies $\mathbf{G}_\omega = \mathbf{X}_\omega \mathbf{X}_\omega'$, then $\mathbf{X}_\omega$ is a k-dimensional Euclidean configuration of $\boldsymbol{\Delta}$. This happens if and only if $\mathbf{G}_\omega$ is a positive semidefinite matrix.

The rows of $\mathbf{X}_\omega$ are n observations of a latent regressor in $\mathbb{R}^k$ such that the pairwise Euclidean distance between them match the pairwise dissimilarities between the z's.

The key idea in DB regression is to substitute the original "matrix" of mixed-type predictors $\mathbf{Z}$ by $\mathbf{X}_\omega$.

The advantage of DB regression is that the predicted response depends solely on the distance matrix (Boj *et al.* 2010) and not on the (implicit) Euclidean configuration.

To check that DB-LM does not depend on the choice of $\mathbf{X}_\omega$, we display the expression of the predicted response $\hat{y}$.

For a new individual with predictors $\mathbf{z}_{n+1}$, the fitted response given by DB-LM (using the interpolation formula in Gower 1968) is

$$\hat{y}_{n+1} = \bar{y}_\omega + \frac{1}{2}(\mathbf{g}_\omega - \boldsymbol{\delta}_{n+1})' \left(\mathbf{D}_\omega^{1/2}\, \mathbf{F}_\omega^{+}\, \mathbf{D}_\omega^{1/2}\right) (\mathbf{y} - \bar{y}_\omega\, \mathbf{1}),$$

where $\mathbf{F}_\omega^{+}$ is the Moore-Penrose pseudo-inverse of

$$\mathbf{F}_\omega = \mathbf{D}_\omega^{1/2}\, \mathbf{G}_\omega\, \mathbf{D}_\omega^{1/2},$$

$\mathbf{g}_\omega$ is the vector containing the diagonal of $\mathbf{G}_\omega$ and $\boldsymbol{\delta}_{n+1}$ is the column vector of squared distances from $\mathbf{z}_{n+1}$ to $\mathbf{z}_1, \ldots, \mathbf{z}_n$.

To extend the DB-LM to the generalized linear model (GLM) setting, Boj *et al.* (2016) substitute the linear regressions of the IWLS algorithm for GLM fitting, by the corresponding DB-LMs.

DB logistic regression can be used in classification problems with mixed-type data.

As the distance measure between mixed predictors $\mathbf{z}_i$ and $\mathbf{z}_j$ we use $\delta^2(\mathbf{z}_i, \mathbf{z}_j) = 1 - s_{ij}$, where s_{ij} is the similarity coefficient of Gower (1971) given by

$$s_{ij} = \frac{\sum_{h=1}^{p_1} (1 - |z_{ih} - z_{jh}|/R_h) + a + \alpha}{p_1 + (p_2 - d) + p_3}, \quad 0 \leq s_{ij} \leq 1,$$

p_1 is the number of quantitative predictors, a and d are the number of coincidences $(1,1)$ and $(0,0)$ for the p_2 binary variables, α is the number of coincidences in the p_3 multi-state qualitative variables and R_h is the range of the h-th quantitative variable.

Some data sets have such a large sample size n that the decomposition of the Gram matrix cannot be performed, or even the computation and storage of the squared-distances matrix are unmanageable. A simple solution is to use ensemble techniques.

We use subsampling (to generate several predictions for the response of a predictor input) and aggregation (that combines the predictions into a single output).

Let $\mathcal{G}_1, \ldots, \mathcal{G}_G$, with $\mathcal{G}_g \subset \{1, \ldots, n\}$, denote the subsample indices. The subsamples can overlap. For each subsample $\mathcal{G}_g$ we predict the value of the response in a new individual, $\hat{y}_{n+1;g}$. These ensemble predictions $\hat{y}_{n+1;1}, \ldots, \hat{y}_{n+1;G}$ can be aggregated to a single predicted response $\hat{y}_{n+1;\text{aggr}}$ in different ways.

Bagging (Breiman 1996a)

In regression the predicted response is the average (with equal weights) of the ensemble predictions

$$\hat{y}^B_{n+1} := \sum_{g=1}^{G} v_g \, \hat{y}_{n+1;g},$$

where $v_g = 1/G$, for all $g = 1, \ldots, G$.

In classification with two populations, the response Y takes only two values (0 or 1). The bagging classifier takes the vote of the majority of the ensemble classifications $\hat{y}_{i;1}, \ldots, \hat{y}_{i;G}$:

$$\hat{y}^B_{n+1} = \arg\max_{k=0,1} \sum_{g=1}^{G} \mathbb{1}_{\{\hat{y}_{i;g}=k\}}.$$

Stacking (Breiman 1996b and Wolpert (1992))

The predicted response is a weighted average of the subsample predictions

$$\hat{y}^S_{n+1} := \sum_{g=1}^{G} v_g \, \hat{y}_{n+1;g},$$

where

$$\mathbf{v} = (v_1, \ldots, v_G)' = \arg\min_{\mathbf{v} \in V} \left\| \mathbf{y} - \sum_{g=1}^{G} v_g \, \hat{\mathbf{y}}(g) \right\|_2 .$$

The space V of possible weight vectors $\mathbf{v}$ can be

$$\begin{array}{rcll} V_c & = & \{\mathbf{v} : \min_g v_g \geq 0, \sum_{g=1}^{G} v_g = 1\} & \text{(convex constraint)} \\ V_r & = & \{\mathbf{v} : \|\mathbf{v}\|_2 \leq s\} \quad \text{for some } s > 0 & \text{(ridge constraint).} \end{array}$$

The use of stacking and DB-GLM for classification is not computationally feasible.

Magging (Bühlmann and Meinshausen 2015)

The predicted response is a weighted average of the subsample predictions

$$\hat{y}_{n+1}^{M} := \sum_{g=1}^{G} v_g \, \hat{y}_{n+1;g},$$

with weights such that

$$\mathbf{v} = \arg\min_{\mathbf{v} \in V_c} \left\| \sum_{g=1}^{G} v_g \, \hat{\mathbf{y}}(g) \right\|_2 .$$

Magging has not yet been extended to the discrimination framework.

Aggregation and DB-LM on two real data sets

The number G of subsamples can be chosen via cross-validation. Here we compare the performance of the ensemble procedure for different choices of a fixed G and different subsample sizes, m. Each subsample is formed by sampling m individuals from $\{1, \ldots, n\}$ without replacement.

The analysis is a Monte Carlo (MC) experiment, with 500 runs. In each MC run we randomly separate the original sample into a training sample of size n (90% of the original sample size) and a test sample with the remaining observations.

To quantify the performance of each method, we compute the mean squared error (MSE) in the prediction of Y in the test sample.

Bike sharing demand

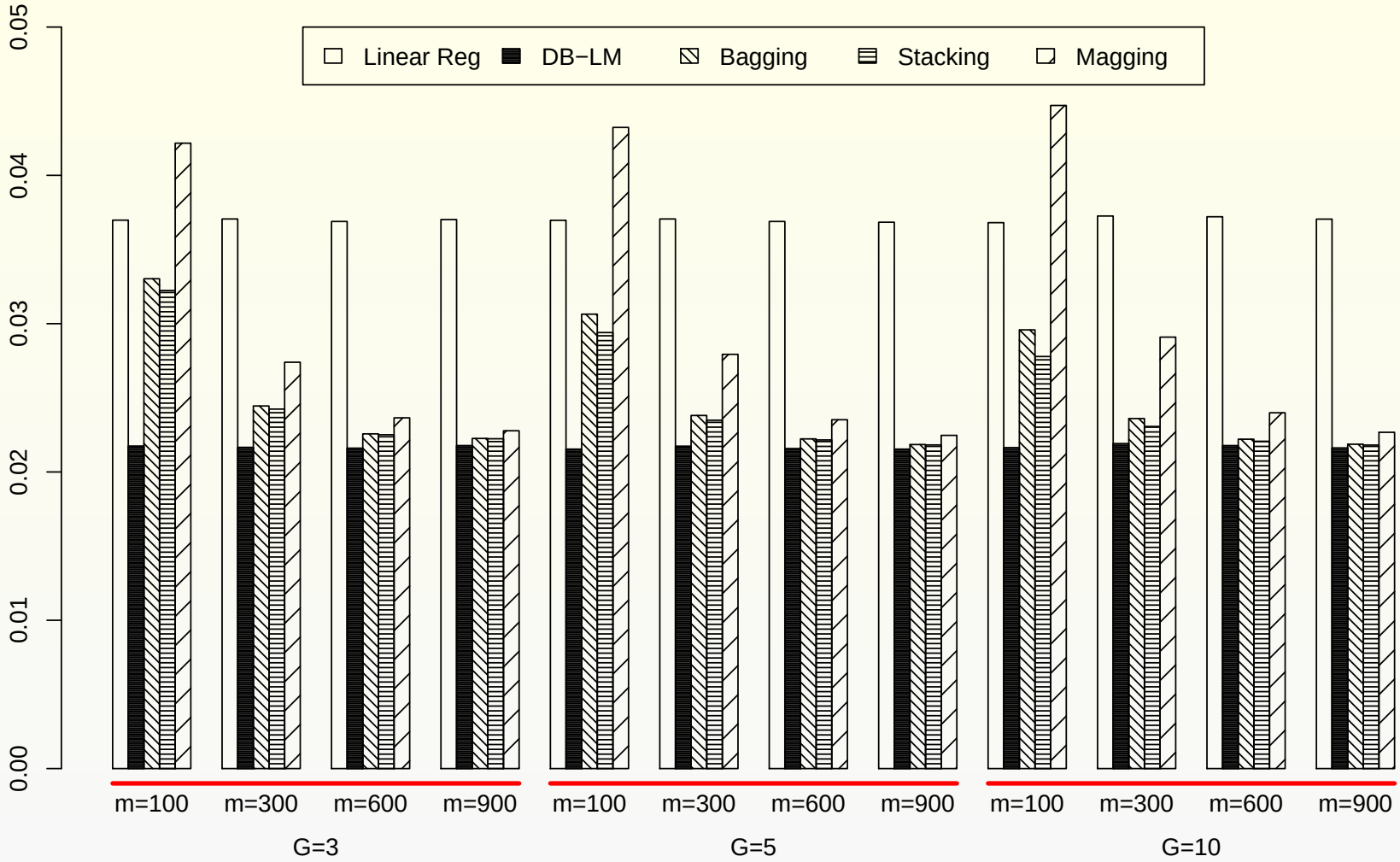
This dataset is provided by the Capital Bikeshare program, which operates in the District of Columbia. The program records several details, such as travel duration, departure and arrival locations and time elapsed between departure and arrival, for each rental in the bike sharing system.

We analyse the data on a daily basis for the period between 1 January 2013 and 31 December 2018.

The response variable Y is the daily count of users (casual and registered) scaled by the mean daily number of users (in the year corresponding to the day). The total number of days in the data set is 2903.

Capital Bikeshare
Total count of daily users (both registered and not)
Season: winter (1), spring (2), summer (3), autumn (4)
Year, codified to 0 (=2011), 1 (=2012), 2 (=2013),. . . , 7 (=2018)
Month, codified to 1,2,. . . ,12
National holiday (1) or not (0)
Weekday, codified to 0 (=Sunday), 1(=Monday), . . . , 6 (=Saturday)
Working day (1) or weekend day (0)

NOAA at DCA
Average daily wind speed (miles per hour)
Precipitation (inches to hundredths)
Maximum temperature (in Fahrenheit)
Minimum temperature (in Fahrenheit)
Ceiling height dimension (in meters)
Mean daily temperature (in Celsius)
Sea level pressure (in hPa)
Relative humidity (in %)



Mean squared prediction errors for the Capital Bikeshare data.

Linear regression	DB-LM	Random forests
0.037	0.022	0.017

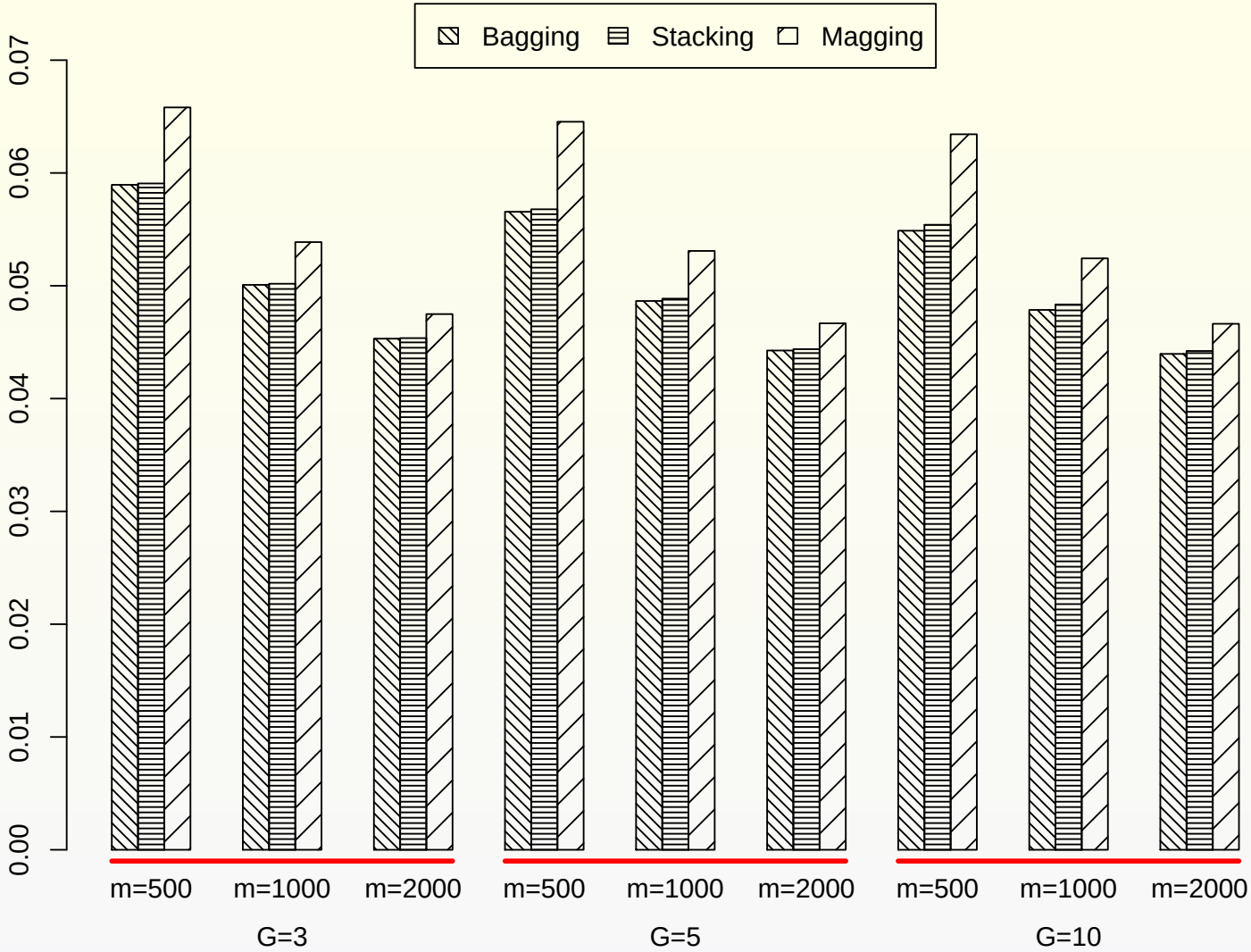
King County house sales (Kaggle)

Contains 21597 house prices from King County (Washington State, USA), sold between May 2014 to May 2015.

The response variable is the logarithm of the house sale price.

19 predictors:

Date of the home sale
Number of bedrooms
Number of bathrooms
Square footage (sqft) of the interior living space
Sqft of the land space
Number of floors
1 if the house overlooks the waterfront; 0 if not
Index from 0 to 4 grading the view from the property
Index from 1 to 5 on the condition of the house
Index from 1 to 13, grading quality of construction and design
Sqft of the interior housing space above ground level
Sqft of the interior housing space below ground level
The year the house was initially built
The year of the house's last renovation
Zipcode area
Latitude
Longitude
Sqft of interior living space for the nearest 15 neighbors
Sqft of the land lots of the nearest 15 neighbors



Mean squared prediction errors for the King County house sales

Linear regression	Random forests
0.101	0.047

Bagging and DB-GLM: a real data set

To illustrate the performance of the ensemble DB classifier, first we analyzed the *Online Shoppers Purchasing Intention* dataset from the UCI ML Repository. It records 12330 e-shopping sessions, of which 10422 did not end up with a purchase and the other 1908 ended up with shopping. There are 10 numerical and 8 qualitative attributes of each session (month, operating system, browser, region, ...).

Experiment: $n = 2000$ observations were randomly selected (without replacement) from the dataset. In each of 100 iterations we separated this sample into a training sample of 1800 observations and a test sample of size 200. The training sample was used to fit (i) the classical GLM model with only quantitative features, (ii) the DB-GLM using all the features and (iii) the DB-GLM on $G = 5$ subsamples of size $m = 400$, later aggregated using bagging. Both logit and probit links were considered.

Link function	Classical GLM (quantitative features)	DB-GLM (whole sample)	DB-GLM and bagging
logit	0.88940	0.88380	0.88395
probit	0.88930	0.88465	0.88350

Correct classification rates of the test sample.

Bagging and DB-GLM: a simulation study

We simulated mixed-type features from two populations using the R package `Umpire` (Coombes *et al.* 2021). In all cases we sampled n "vectors" $\mathbf{Z}$ of 20 mixed-type features, with prior probabilities of the populations sampled from a Dirichlet distribution. We consider two models:

> Model 1: The same proportion (1/3) of continuous, binary and nominal features.
>
> Model 2: The proportion of continuous and binary features is the same (25% approximately) and the remaining (50%) of the features are qualitative ones.

In each experiment the sample of size n is simulated only once, and, from then on, these observations are treated as if they were a real data set.

Preliminary study: We drew a sample of 2000 observations from Model 2. In each of 100 iterations we drew (without replacement) a training sample of size 1800 and a test sample of 200. We chose $G = 4$ and $m = 500$. We fitted the same three classifiers (with the logit link).

The sample proportions of the two populations were 56% and 44% and the number of continuous, binary and nominal variables were 5, 5 and 10, respectively.

Correct classification rates for Model 2 with $n = 2000$.

Classical logistic (quantitative features)	DB logistic (whole sample)	DB logistic and bagging
0.7401	0.9734	0.9780

We have sampled $n = 10000$ observations from Models 1 and 2.

In each of the 100 runs, the sample is split into a training sample of size 8000 and a test sample of size 2000.

We fit two models: the logistic regression with only quantitative features and the DB logistic regression model on $G = 20$ subsamples of size $m = 500$ plus bagging.

Correct classification rates for Models 1 and 2 with $n = 10000$.

	Classical logistic (quantitative features)	DB logistic and bagging
Model 1	0.8220	0.9244
Model 2	0.9208	0.9601

Conclusion

Distance-based regression is able to deal with very general types of regressors or features if a suitable dissimilarity is defined between them. The DB-LM (resp., DB-GLM) improves the performance of linear (resp., generalized linear) regression when qualitative predictors are available and informative about the response.

Fitting the DB regression is unfeasible when sample sizes are too large. For medium sample sizes subsampling and aggregation techniques (bagging, stacking and magging) applied to DB linear regression attain the same error rates as DB-LM applied to the whole sample, but reduce drastically the computing time. This conclusion also holds for bagging and DB logistic regression.

Ensemble techniques allow to use DB prediction models with large sample sizes, where the fitting to the global sample is out of reach but the use of qualitative predictors contributes to improve the prediction or classification performance.