MATHEMATICAL MODELING AND COMPUTING, Vol. 8, No. 4, pp. 726–735 (2021)

Diabetes prediction using an improved machine learning approach

Lyaqini S., Nachaoui M.

LMA FST Beni-Mellal,
University Sultan Moulay Slimane, Morocco

(Received 23 May 2021; Accepted 7 June 2021)

This paper deals with a machine-learning model arising from the healthcare sector, namely diabetes progression. The model is reformulated into a regularized optimization problem. The term of the fidelity is the L^1 norm and the optimization space of the minimum is constructed by a reproducing kernel Hilbert space (RKSH). The numerical approximation of the model is realized by the Adam method, which shows its success in the numerical experiments (if compared to the stochastic gradient descent (SGD) algorithm).

Keywords: *supervised learning, smooth approximation, Adam algorithm, diabetes diagnosis, Tikhonov regularization, smooth optimization.*

2010 MSC: 68T05, 68T09 **DOI:** 10.23939/mmc2021.04.726

1. Introduction

According to the International Diabetes Federation, one of eleven people worldwide lives with diabetes. In recent years, the impact of diabetes has increased dramatically, making it a global threat. Today, diabetes is consistently the leading cause of death. Therefore, early detection of diabetes is very important so that measures can be taken in time and the progression of the disease can be prevented to avoid further complications [1,2]. Through efforts of artificial intelligence, which enables early detection and diagnosis of diabetes by an automated process, more beneficial than manual diagnosis [3–6].

The central contribution of this paper is the consideration of machine learning approaches in the disease progression of diabetes. The importance of the diabetes dataset the features, in order to determinate which covariates are important factors in disease progression is studied. Furthermore, two features being the most important in disease progression to build our predictive model are selected. The model is reformulated into a regularized optimization problem with the term of the fidelity is the L^1 norm and the optimization space of the minimum is constructed by a reproducing kernel Hilbert space [7,8]. Indeed, reproducing kernel Hilbert spaces are particularly important in the area of statistical learning theory because of the famous represented theorem which states that any function in an RKHS that minimizes an empirical risk function can be written as a linear combination of the kernel function evaluated at the training set. This is a useful result in practice, as it effectively simplifies the problem of empirical risk minimization from an infinite-dimensional to a finite dimensional optimization problem.

The use of L^1 loss function for supervised learning problem gives more consistent results [9–13]. This consolidates the idea of converting the supervised problem based on absolute loss function into a minimization one. However, the fidelity term of the resulting optimization problem is not differentiable which precludes the use of standard optimization methods. In order, to overcome the difficulty caused by the non-differentiability of the fidelity term, we introduce a smooth approximation technique [10, 14–16] to transform it into a differentiable and convex one. Furthermore, the fidelity term of the resulting optimization problem is twice differentiable and convex, which is solved directly using Tikhonov regularization and Adam algorithm [17, 18]. Finally, we present several numerical validations of the proposed computational approach and comparison with stochastic gradient descent (SGD) algorithm on the basis of the relative error. Obtained experimental result on the diabetes data sets indicates that the proposed approach is an efficient and helpful tool in machine learning.

The organization of this paper is as follows. In section 2, we present the setting problem and its reformulation as a minimization one, using absolute loss function. Section 3 is concerned to a smooth approximation of the absolute loss function. Furthermore, a numerical algorithm based on Adam algorithm is presented. In section 4, we evaluate the efficiency of the proposed approach using diabetes data sets and compare it with stochastic gradient descent (SGD) algorithm.

2. Formulating the supervised problem

The problem of supervised learning can be stated in this way: given a set of examples $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, where $x_i \in \mathcal{X} \subset \mathbb{R}^d$ called the input space and $y_i \in [-M, M]$ is the output space, $M > 0$, find the solution f^* of the minimization problem

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} V(f(x_i), y_i), \tag{1}$$

where V is a positive function, which measures the error between y and its prediction $f(x)$, called loss function and $\mathcal{H}$ is a reproducing kernel Hilbert space defined by the positive definite kernel K, which we discuss in the next subsection.

2.1. Reproducing kernel Hilbert space (RKHS)

In what follows we briefly summarize the properties of RKHS [8] needed in this work.

Let $\mathcal{X}$ be a Hilbert space of real valued functions defined on $\mathcal{X}$ and equipped with an inner product $\langle \ldots \rangle_{\mathcal{H}}$.

Definition 1. *A function $K \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a reproducing kernel of $\mathcal{H}$ if it verifies the following properties:*

- $K_x(\cdot) = K(\cdot, x) \in \mathcal{H}$, $\forall x \in \mathcal{X}$;
- $f(x) = \langle f, K_x \rangle_{\mathcal{H}}$, $\forall f \in \mathcal{H}$.

Such a Hilbert space $\mathcal{H}$ is called a reproducing kernel Hilbert space (RKHS). It is also known that a kernel $K(\cdot, \cdot)$ is a reproducing kernel if and only if it is symmetric and positive definite, that is:

$$\sum_{i,j=1}^{n} a_i a_j K(x_i, x_j) \geqslant 0,$$

for any $n \in \mathbb{N}$, $x_1, \ldots, x_n \in \mathcal{X}$ and $a_1, \ldots, a_n \in \mathbb{R}$.

Since $\mathcal{X}$ is a compact set of $\mathbb{R}^n$, then a kernel K defined on $\mathcal{X} \times \mathcal{X}$ is said to be a Mercer's kernel if it is continuous and positive semi-definite.

The space $\mathcal{H}$ can be read

$$\mathcal{H} = \overline{\operatorname{span}\{K_x(\cdot) | x \in X\}},$$

as the closure of the subspace of all linear combinations of $K_x(\cdot)$.

3. Smooth absolute loss function

In this section, we approximate the absolute loss function with a smooth one, which is twice differentiable and convex. Recently, the absolute loss function shows its effectiveness on machine learning task [10]. For this reason, we reformulate the supervised learning problem using the absolute loss function. Thus, the minimization problem (1) became

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} |f(x_i), y_i|. \tag{2}$$

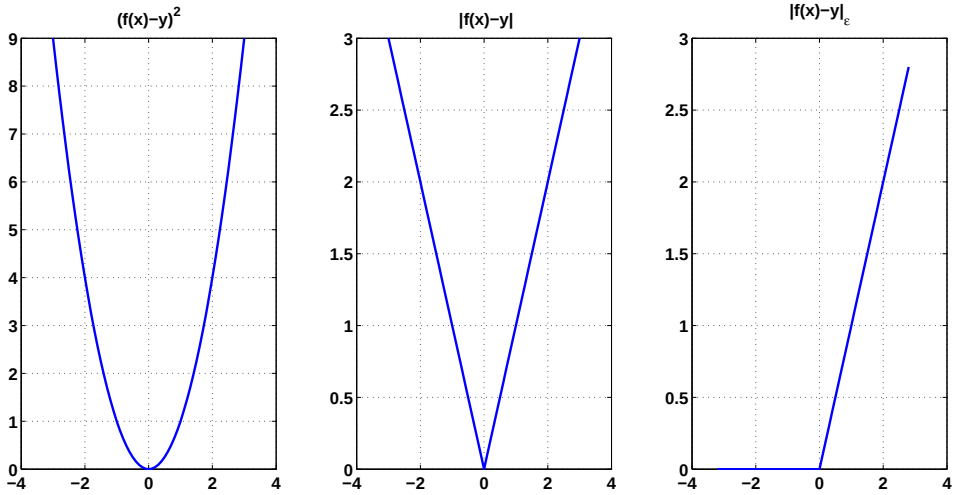

Fig. 1. Absolute loss function.

From the figure above, we see that the absolute loss functions are non-differentiable. This precludes the use of standard optimization tools, which required the assumption of differentiability of the objective function. For this reason, we approximate the absolute loss function with a differentiable one, using the so called smooth approximation [10, 19]. This in fact allowing us to use the gradient kinds method to solve the resulting optimization problems.

The absolute loss function can be accurately approximated by a smooth function which is twice differentiable and convex, which would be defined as follows.

In order to approximate the objective function of the minimization problem (2), the absolute function is written as

$$|u| = \max(u, 0) + \max(-u, 0).$$

For all $u \in \mathbb{R}$, the max function is approximated by a smooth function,

$$\max_{\alpha}(u, 0) = \left(u + \frac{1}{\alpha}\log\left[1 + \exp\left(-\alpha u\right)\right]\right), \qquad \forall \alpha > 0.$$

Then we get the following smooth approximation for the absolute function

$$\begin{aligned} |u|_\alpha &= \max_{\alpha}(u, 0) + \max_{\alpha}(-u, 0) \\ &= \frac{1}{\alpha}\Big[\log\big(1 + \exp(\alpha u)\big) + \log\big(1 + \exp(-\alpha u)\big)\Big] \\ &= \frac{1}{\alpha}\log\Big[\big(1 + \exp(-\alpha u)\big)\big(1 + \exp(\alpha u)\big)\Big] \\ &= \frac{1}{\alpha}\log\big(2 + \exp(-\alpha u) + \exp(\alpha u)\big), \qquad \forall \alpha > 0. \end{aligned}$$

Notice that the loss function is always a true function of only one variable u, with $u = \omega - y$.

Let us denote by $V_\alpha(u)$ the smoothed loss function with parameter α of $V(u) = |u|$ given by

$$V_\alpha(u) = \frac{1}{\alpha}\log\big(2 + \exp(-\alpha u) + \exp(\alpha u))\big), \qquad \forall \alpha > 0.$$

As illustrated in Fig. 2, the smoothed absolute loss V_α approaches the absolute loss V as α goes to $+\infty$.

In the sequel, we consider the approximate optimization problem of (1) given by

$$\min_{f \in \mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} V_\alpha(f(x_i), y_i). \tag{3}$$

This problem is a strongly convex minimization problem. Moreover, the cost function in the problem (3) is differentiable, thus a Adam's method can be used to solve it.

The problem (3) is an ill-posed problem [20, 21]. A standard approach to imposing well-posed to a procedure is via the concept of regularization. The concept of regularization consists in searching for approximate solutions by setting regularity constraints on the reproducing Hilbert space $\mathcal{H}$. In particular, we use the Tikhonov regularization, replacing the minimization problem (3) with the following one,

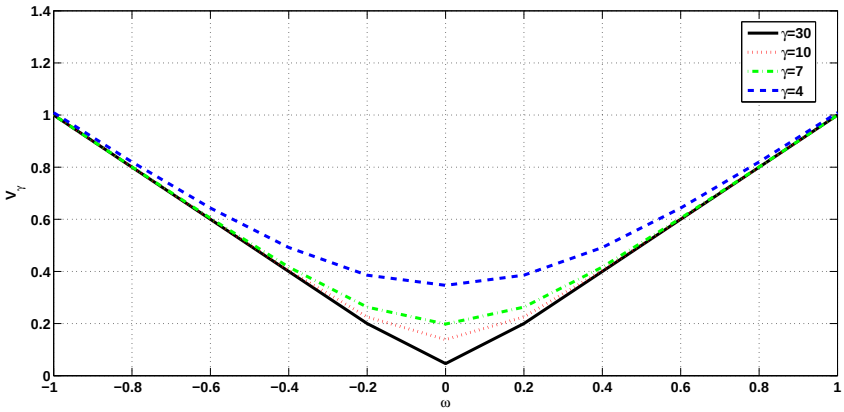

Fig. 2. Smoothed absolute loss function with different smoothing parameters.

$$\min_{f\in\mathcal{H}} \left(\frac{1}{n} \sum_{i=1}^{n} V_\alpha(f(x_i), y_i) + \lambda \|f\|_{\mathcal{H}}^2 \right), \tag{4}$$

where $\lambda > 0$ is the regularization parameter and $\|\cdot\|_{\mathcal{H}}$ is the norm in $\mathcal{H}$.

Due to the representer theorem [22], the solution f_λ^α of the problem (4), can be written as a finite linear combination of kernel evaluations in the data, namely

$$f_\lambda^\alpha(x) = \sum_{i=1}^{n} c_i K(x, x_i), \tag{5}$$

where $c_i \in \mathbb{R}$, $i = 1, \ldots, n$ and K the reproducing kernel of $\mathcal{H}$. Therefore, the solution to the possibly infinite dimensional optimization problem (4) can be found in the n-dimensional span of the functions $K_{x_i}(\cdot)$, $i = 1, \ldots, n$. To find the coefficients, it is sufficient to solve the following problem

$$\min_{c\in\mathbb{R}^n} \mathcal{J}_\lambda^\alpha(c), \tag{6}$$

where

$$\mathcal{J}_\lambda^\alpha(c) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{1}{\alpha} \log \left(2 + \exp\big(-\alpha(c^\top K_i - y_i) \big) + \exp\big(\alpha(c^\top K_i - y_i) \big) \right) \right) + \lambda\, c^\top K c.$$

Denote

$$K = \begin{pmatrix} K(x_1, x_1) & \cdots & K(x_1, x_n) \\ \vdots & \ddots & \vdots \\ K(x_n, x_1) & \cdots & K(x_n, x_n) \end{pmatrix}, \quad K_i = \begin{pmatrix} K_{x_i}(x_1) \\ \vdots \\ K_{x_i}(x_n) \end{pmatrix}, \quad c = \begin{pmatrix} c_1 \\ \vdots \\ c_n \end{pmatrix} \quad \text{and} \quad y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}.$$

There are many optimization tools for solving the problem (6) [23, 24]. In this paper we use Adam methods with line search methods for solving the resulting optimization problem (6).

Then we will treat the experimental results obtained after the simulation of diabetic data using the proposed approach.

4. Experimental results

To evaluate this algorithm it is often desirable to have some standardized benchmark data sets. In our case we choose to evaluate the proposed method through real-life data from the UCI machine learning repository. To evaluate the efficiency of the proposed algorithm, we compared its cost function and its relative error with the Stochastic gradient descent algorithm. The goal is to show that the proposed approach is faithfully and faster to predict the considered models.

Let use the Gaussian kernel, given by

$$K(x,x') = \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - x_i')^2\right),$$

as reproducing kernel of the space $\mathcal{H}$, where σ is a free parameter and n is the size of training set.

Thereafter, we will study the effectiveness of the proposed method on a real world example, called Diabetes Data Set [25]. These data comprise observations on 442 patients, with the response of interest being a quantitative measure of disease progression one year after baseline.

Table 1. Diabetes data sets.

Data sets	Number of Instances	Number of Attributes
Diabetes	442	10

There are ten baseline variables age, sex, body-mass index, average blood pressure, and six blood serum measurements, were obtained for each one of $n = 442$ diabetes patients, as well as the response of interest, a quantitative measure of disease progression one year after baseline. The aim of this work is to build a model that predicts the y response from the ten baseline variables, that produces accurate baseline predictions of response for future patients, and that the shape of the model suggests which covariates are important factors in disease progression [25].

Table 2. Diabetes data description.

Attributes	Description	Type of data
Age	age in years	Digital
Sex		Boolean
bmi	body-mass index	Digital
bp	average blood pressure	Digital
s1	tc, T-Cells (a type of white blood cells)	Digital
s2	ldl, low-density lipoproteins	Digital
s3	hdl, high-density lipoproteins	Digital
s4	tch, thyroid stimulating hormone	Digital
s5	ltg, lamotrigine	Digital
s6	glu, blood sugar level	Digital

The first 10 columns have been normalized to have mean 0 and Euclidean norm 1 and the output column y has been centered. For the implementation, we construct our training set by taking 360 random observations form all data sets. Then by using the proposed approach, we generate the turbulence model. Let's first find the correlation of each of these feature pairs and visualize the correlations using a heatmap. As shown in the heatmap, age, sex, bmi and bp all correlate significantly with the outcome variable.

As it can be seen, from figures below that the predicted model faithfully follows the behavior of unseen observations with high accuracy.

Table 3. The relative error obtained by three different methods on diabetes data sets.

	Adam	SGD
Relative Error	0.0024	0.05

In Table 3, we present the the relative error given by $\frac{\|f-y\|_2}{\|y\|_2}$. The quantitative results presented in this table show that the proposed approach is not only the best one on the training set but also has a very good testing set accuracy. As a conclusion, we can say that the proposed method becames more efficient in reaching accurate optimal solution.

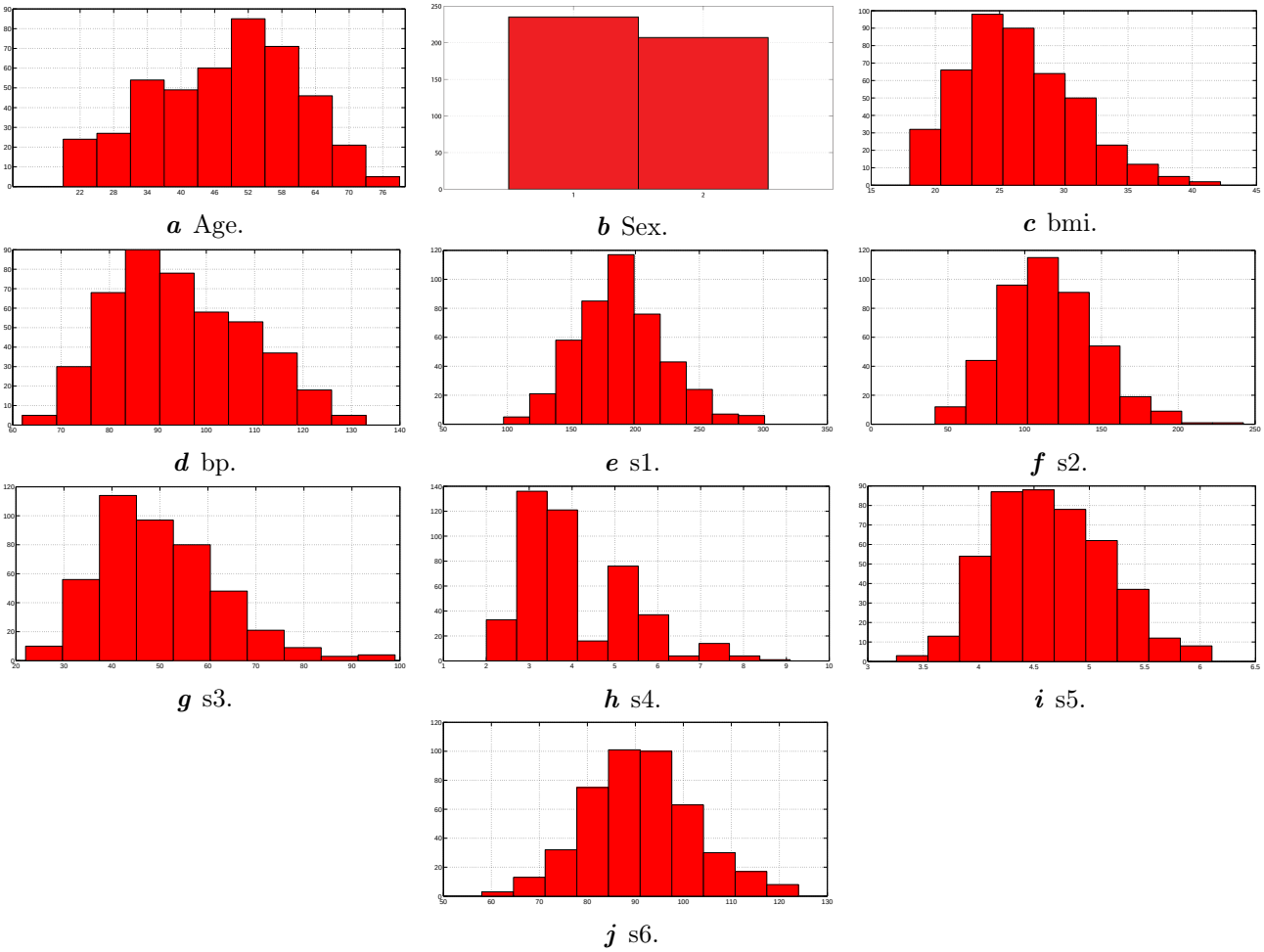

Fig. 3. Diabetes data set information.

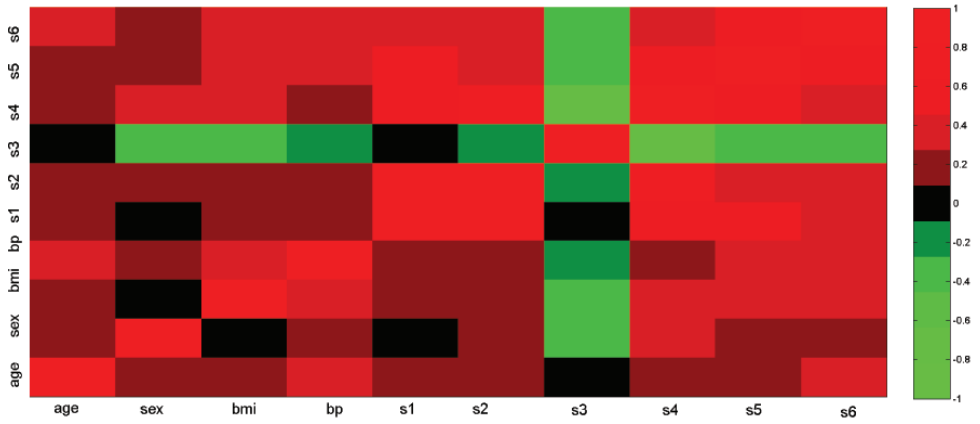

Fig. 4. Real and predicted model using Adam and Stochastic gradient descent algorithms, for $\sigma = 0.5$, $\alpha = 10$ and $\lambda = 10^{-5}$.

4.1. Feature importance

In this section, we study the importance of the features, in order to determinate which covariates are important factors in disease progression. Furthermore, we select the two features which are the most important in disease progression to build our predictive model.

One can see from Fig. 9 that the two important factors in disease progression are **bmi** (body-mass index) and **S3** (high-density lipoproteins). Consequently, we use Adam's algorithm to build our predictive model using only these two features.

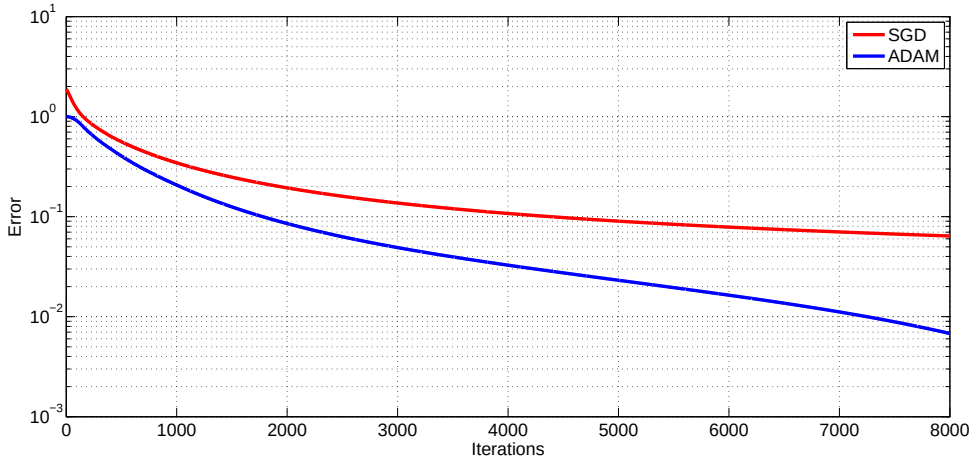

Fig. 5. Relative Error.

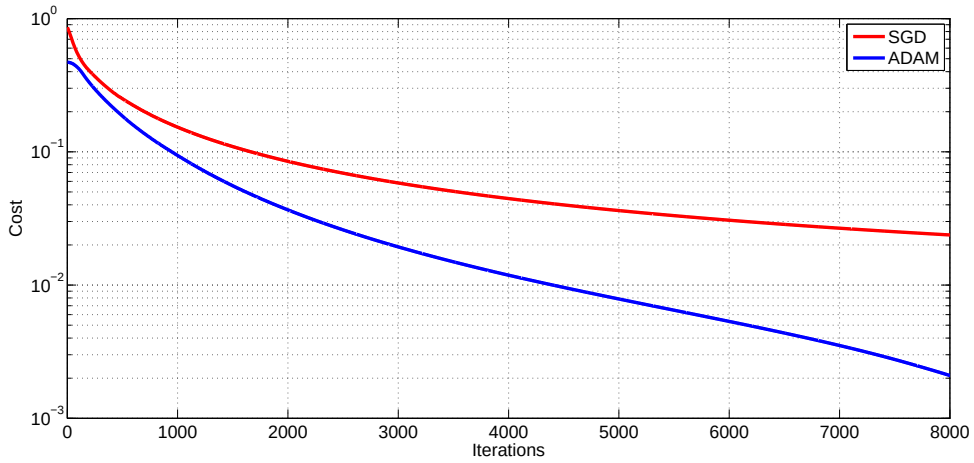

Fig. 6. Cost function.

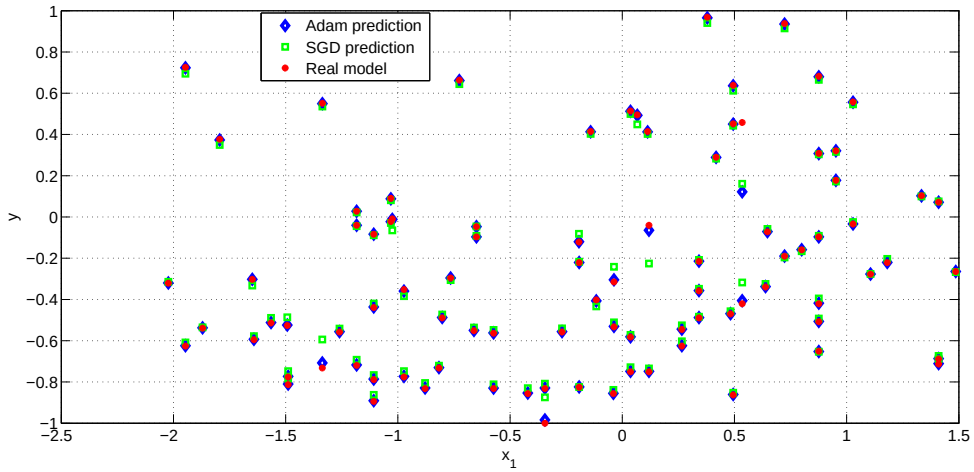

Fig. 7. Real and predicted model using Adam and Stochastic gradient descent algorithms, for $\sigma = 0.5$, $\alpha = 10$ and $\lambda = 10^{-5}$.

Fig. 10, shows that the model obtained using features **bmi** and **s3** of the diabetes dataset is also better at predicting disease progression.

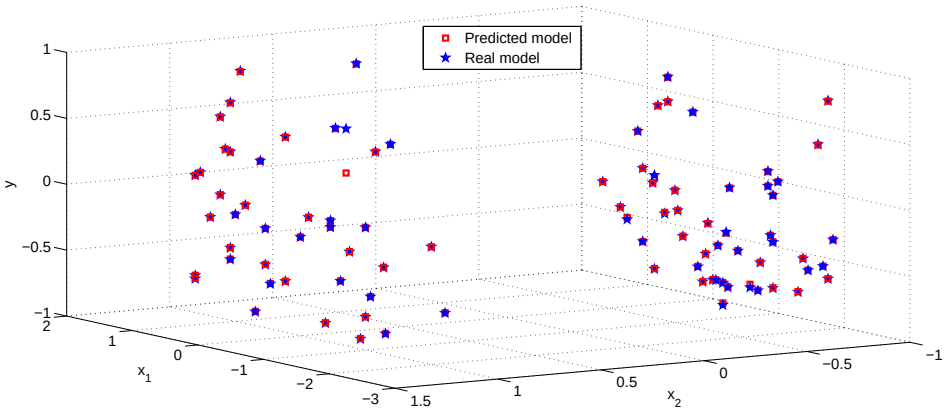

Fig. 8. Real and predicted model using features 1 and 2 using Adam algorithm, for $\sigma = 0.5$, $\alpha = 10$ and $\lambda = 10^{-5}$.

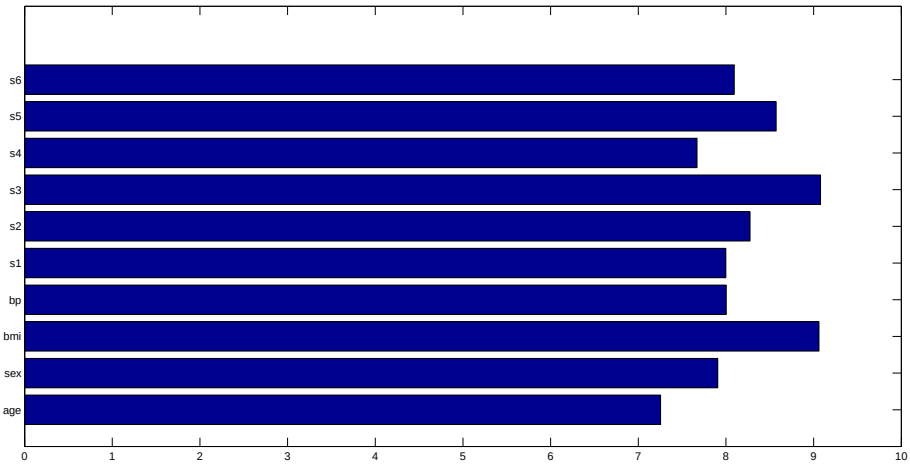

Fig. 9. Feature importance.

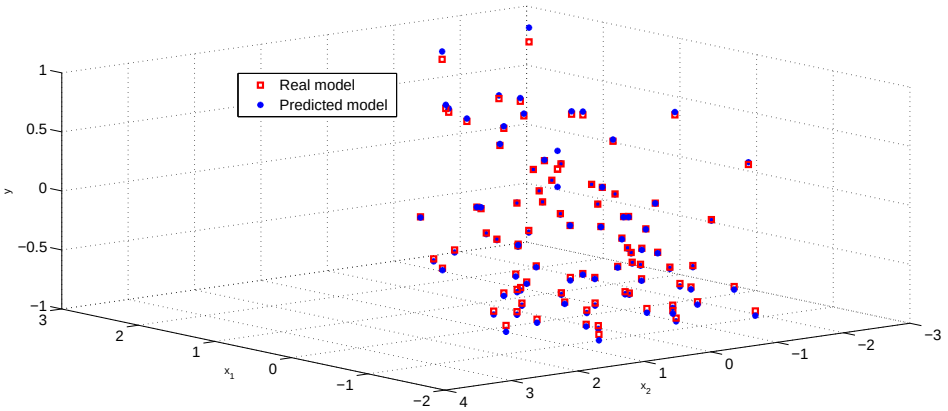

Fig. 10. Sparsity Example: Fitting only features **bmi** and **s3**.

5. Conclusion

In this paper, Tikhonov regularization and Adam's algorithm is proposed to predict the diabetes disease progression. The comparison with the stochastic gradient descent (SGD) algorithm on the basis of the relative error shows the performance of the proposed algorithm in the prediction of the disease progression. As a perspective we will focus on the integration of other methods for solving the non-smooth optimization problem directly without using the smoothing approximation.

[1] Ricci P., Blotière P. O., Weill A., Simon D., Tuppin P., Ricordeau P., Allemand H. Diabète traité: quelles évolutions entre 2000 et 2009 en France. Bull. Epidemiol. Hebd. **42** (42–43), 425–431 (2010).

[2] Isnard R., Legrand L., Pousset F. Insuffisance cardiaque et diabète: données épidémiologiques, phénotype et impact sur le pronostic. Médecine des Maladies Métaboliques. **15** (3), 246–251 (2021).

[3] Kavakiotis I., Tsave O., Salifoglou A., Maglaveras N., Vlahavas I., Chouvarda I. Machine learning and data mining methods in diabetes research. Computational and Structural Biotechnology Journal. **15**, 104–116 (2017).

[4] Perveen S., Shahbaz M., Keshavjee K., Guergachi A. Metabolic syndrome and development of diabetes mellitus: Predictive modeling based on machine learning techniques. IEEE Access. **7**, 1365–1375 (2018).

[5] Luo G. Automatically explaining machine learning prediction results: a demonstration on type 2 diabetes risk prediction. Health Information Science and Systems. **4** (1), Article number: 2 (2016).

[6] Benhamou P. Y., Lablanche S. Diabète de type 1: perspectives technologiques. Mise Au Point. 11–16 (2018).

[7] Hofmann T., Schölkopf B., Smola A. J. Kernel methods in machine learning. The Annals of Statistics. **36** (3), 1171–1220 (2008).

[8] Aronszajn N. Theory of Reproducing Kernels. Transactions of the American Mathematical Society. **68**, 337–404 (1950).

[9] Rosasco L., De Vito E., Caponnetto A., Piana M., Verri A. Are Loss Functions All the Same? Neural Computation. **16** (5), 1063–1076 (2004).

[10] Lyaqini S., Quafafou M., Nachaoui M., Chakib A. Supervised learning as an inverse problem based on non-smooth loss function. Knowledge and Information Systems. **62**, 3039–3058 (2020).

[11] Lyaqini S., Nachaoui M., Quafafou M. Non-smooth classification model based on new smoothing technique. Journal of Physics: Conference Series. **1743**, 012025 (2021).

[12] Nachaoui M. Parameter learning for combined first and second order total variation for image reconstruction. Advanced Mathematical Models & Applications. **5** (1), 53–69 (2020).

[13] El Mourabit I., El Rhabi M., Hakim A., Laghrib A., Moreau E. A new denoising model for multi-frame super-resolution image reconstruction. Signal Processing. **132**, 51–65 (2017).

[14] Chen C., Mangasarian O. L. A class of smoothing functions for nonlinear and mixed complementarity problems. Computational Optimization and Applications. **5** (2), 97–138 (1996).

[15] Lee Y. J., Hsieh W. F., Huang C. M. "/spl epsi/-SSVR: a smooth support vector machine for epsilon-insensitive regression. IEEE Transactions on Knowledge & Data Engineering. **17** (5), 678–685 (2005).

[16] Hajewski J., Oliveira S., Stewart D. Smoothed Hinge Loss and ℓ^1 Support Vector Machines. 2018 IEEE International Conference on Data Mining Workshops (ICDMW). 1217–1223 (2018).

[17] Défossez A., Bottou L., Bach F., Usunier N. On the convergence of Adam and Adagrad. arXiv preprint arXiv:2003.02395 (2020).

[18] Fei Z., Wu Z., Xiao Y., Ma J., He W. A new short-arc fitting method with high precision using Adam optimization algorithm. Optik. **212**, 164788 (2020).

[19] Rosales R., Schmidt M., Fung G. Fast Optimization Methods for L1 Regularization: A Comparative Study and Two New Approaches (2007).

[20] Hadamard J. Lectures on Cauchy's problem in linear partial differential equations. New Haven, Yale University Press (1923).

[21] Girosi F., Jones M., Poggio T. Regularization theory and neural networks architectures. Neural computation. **7** (2), 219–269 (1995).

[22] Schölkopf B., Herbrich R., Smola A. J. A generalized representer theorem. International conference on computational learning theory. 416–426 (2001).

[23] Boyd S., Vandenberghe L. Convex Optimization. Cambridge University Press, New York, USA (2004).

[24] Ruder S. An overview of gradient descent optimization algorithms. arXiv preprint arXiv:1609.04747 (2016).

[25] Efron B., Hastie T., Johnstone I., Tibshirani R. Least angle regression. Annals of statistics. **32** (2), 407–499 (2004).

Прогнозування діабету за допомогою вдосконаленого машинного навчання

Лякіні С., Нахауі М.

Математична лабораторія та застосунки Бені-Меллаль, Університет Султан Мулай Слиман, Марокко

У статті розглядається модель машинного навчання, що походить з області охорони здоров'я, а саме: прогресування діабету. Модель переформульовується в регуляризовану задачу оптимізації. Член правдоподібності — це норма L^1, а оптимізаційний простір мінімуму побудований за допомогою відтворюючого ядра гільбертового простору (ВЯГП). Чисельне наближення моделі реалізується методом Адама, який є успішним у чисельних експериментах (порівняно з алгоритмом стохастичного градієнтного спуску (СГС)).

Ключові слова: *контрольоване навчання, гладке наближення, алгоритм Адама, діагностика діабету, регуляризація Тихонова, гладка оптимізація.*