

On the Equivalence of Regression and Classification

Jayadeva^{*1}, Naman Dwivedi¹, Hari Krishnan C. K.¹, N. M. Anoop Krishnan²

¹ Department of Electrical Engineering, ²Department of Civil Engineering
Indian Institute of Technology Delhi, Hauz Khas, New Delhi 110016

Abstract

A formal link between regression and classification has been tenuous. Even though the margin maximization term $\|w\|$ is used in support vector regression, it has at best been justified as a regularizer. We show that a regression problem with M samples lying on a hyperplane has a one-to-one equivalence with a linearly separable classification task with $2M$ samples. We show that margin maximization on the equivalent classification task leads to a different regression formulation than traditionally used. Using the equivalence, we demonstrate a “regressability” measure, that can be used to estimate the difficulty of regressing a dataset, without needing to first learn a model for it. We use the equivalence to train neural networks to learn a linearizing map, that transforms input variables into a space where a linear regressor is adequate.

Keywords: Regression, Computation Theory, Optimal Map

I. INTRODUCTION

Both classification and regression formulations typically include L_1 or L_2 norms of the weight vector, i.e. $\|w\|_p$, $p = 1, 2$ in their optimization objectives. The primal Support Vector Classifier (SVC) classifier minimizes the \mathcal{L}_∞ norm of the weight vector, viz. $\frac{1}{2}\|w\|^2$, in order to maximize the margin viz. $\frac{2}{\|w\|}$. However, there is no equivalent notion of margin in the case of support vector regression (SVR), and the use of the same term in regression formulations is at best justified as a regularization term.

Torgo and Gama [1] split the regressed output range into a set of levels and assigned the regression task to a classification problem with multiple classes. A somewhat similar strategy was adopted by [2]. Equivalence in the context of SVC and ϵ -SVR was shown by [3]. They assume that “as in the classification case, the objective is to find a tradeoff between finding a hyperplane with the small norm and finding a hyperplane that performs regression well”. The margin of the regressing hyperplane is $\frac{\|w\|}{1-\epsilon}$, where w is the weight vector of the hyperplane.

Bi and Bennett [6] attempted to show that in the context of ϵ -SVR, every regression task with M samples (x^i, z_i) , $x^i \in \mathbb{R}^n$, $z_i \in \mathbb{R}$, $i = 1, 2, \dots, M$ can be mapped to an equivalent SVC with $2M$ samples. In the equivalent SVC task, M Class 1 samples are located at $(x^i, z_i + \epsilon)$, and another M Class (-1) samples are located at $(x^i, z_i - \epsilon)$.

A hyperplane $w^T x + b = 0$ that solves the ϵ -SVR must satisfy

$$\begin{aligned} w^T x + b &\geq z_i - \epsilon \\ w^T x + b &\leq z_i + \epsilon. \end{aligned}$$

Bi & Bennett show that in the equivalent SVC, a hyperplane boundary of the form

$$w^T x + b + \eta z = 0$$

must pass between Class 1 samples $\{x^i, z_i + \epsilon\}$, and Class -1 samples $\{x^i, z_i - \epsilon\}$. Figure 1 illustrates their approach for an example in which all samples to be regressed lie on a straight line.

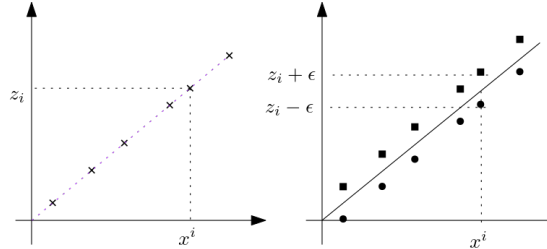


Fig. 1. Left: M regression samples on a line. Right: samples for the classification task obtained by duplicating and shifting samples by $\pm\epsilon$

Once such a hyperplane is learned by solving the classification problem, the regression estimate of an unknown sample x may be found from

$$z = \frac{-(w^T x + b)}{\eta} \quad (1)$$

However, note that with this equivalence, the margin is 2ϵ , independent of w , since the separating hyperplane must pass in between samples of the two classes.

The organization of the remainder of the paper, and key contributions, are summarized below.

Consider a regression dataset $\{x^i, z_i\}$, $i = 1, 2, \dots, M$, where $x^i \in \mathbb{R}^n$, $z_i \in \mathbb{R}$. We first consider the case when all samples lie exactly on a hyperplane.

- We show that the M class 1 samples located at $\frac{x^i}{z_i}$ and M class "-1" samples located at $-\frac{x^i}{z_i}$ are support vectors of an equivalent linearly separable binary classification problem.

This equivalence between regression and classification is discussed in Section II.

- Section IV discusses the proposed *regressability* measure, that estimates the difficulty of regressing a dataset without needing to first learn a regressor. The measure uses the equivalence shown in Section II.

- The equivalence allows us to train a neural network to learn a linearizing map $\phi(x)$. This is a map that transforms input variables x into a higher dimensional feature space $\phi(x)$, such that the regressed output z is a linear function of $\phi(x)$, i.e. $z = w^T \phi(x)$, where w is a weight or co-efficient vector. This is discussed in Section V.

- We learn a regressor in 2 steps. First, we learn a linearizing map as mentioned; we then learn a linear hyperplane in the feature space $\phi(x)$. This two step process obviates the need to tune hyperparameters that control the tradeoff between different terms of a loss function.
- Learning a linear regressor on such a feature map provides smoother interpolation, and in some cases, provides some extrapolation ability as well.

- Section VI discusses results and comparisons. Section VII contains concluding remarks.

II. REGRESSION IS CLASSIFICATION

III. APPROACH TO CONVERT REGRESSION TO EQUIVALENT CLASSIFICATION PROBLEM

We first consider a SVC problem with a linearly separable set of samples $\{\hat{x}^i, y_i\}$, $i = 1, 2, \dots, M$, where $\hat{x}^i \in \mathbb{R}^n$, $y_i \in \{-1, 1\}$. We also assume that the separating hyperplane $w^T x = 0$ passes through the origin; we revisit this assumption in the sequel. The primal hard margin support vector machine classifier solves

$$\text{Minimize}_w \frac{1}{2} \|w\|^2 \quad (2)$$

subject to the constraints

$$y_i (w^T x^i) \geq 1, \quad i = 1, 2, \dots, M \quad (3)$$

In this case, all support vectors satisfy

$$y_i (w^T x^i) = 1, \text{ i.e.} \quad (4)$$

$$(w^T x^i) = 1, \text{ for class 1 samples, and} \quad (5)$$

$$-(w^T x^i) = -1, \text{ for class -1 samples, and} \quad (6)$$

Now consider a regression dataset $\{x^i, z_i\}$, $i = 1, 2, \dots, M$, where $x^i \in \mathbb{R}^n$, $z_i \in \mathbb{R} \setminus 0$. Without loss of generality, we assume that the desired regressor passes through the origin. We first consider the case when all samples lie on a hyperplane. We have

$$(w^T x^i) = z_i. \quad (7)$$

This allows two possibilities

$$w^T \left(\frac{x^i}{z_i} \right) = 1 \quad (8)$$

$$w^T \left(-\frac{x^i}{z_i} \right) = -1 \quad (9)$$

$$(10)$$

Constraints (8) and (9) are identical to (5) and (6). The equivalence for the "hard margin" case is thus trivially established. In a nutshell, a regression dataset $\{x^i, z_i\}$, $i = 1, 2, \dots, M$, where $x^i \in \mathbb{R}^n$, $z_i \in \mathbb{R} \setminus 0$ is equivalent to a classification task with samples

$$\begin{pmatrix} x^i \\ z_i \end{pmatrix} \text{ in class 1, i.e. } y_i = 1, \text{ and} \quad (11)$$

$$\begin{pmatrix} -x^i \\ z_i \end{pmatrix} \text{ in class -1, i.e. } y_i = -1 \quad (12)$$

It is also evident that if the regression samples lie on a hyperplane, samples of the equivalent classification task are linearly separable. Let the solution to the classification problem be a hyperplane w . Clearly, for all support vectors, $w^T x = \pm 1$. For the solution to the regression task, the same w suffices, since $w^T x = z$.

We now elucidate the approach with an intuitive example. Consider samples drawn randomly from the line $z = mx$, $m \in \mathbb{R}$; any set of samples will have $z_i = mx_i$. The equivalent classification task will have class -1 support vectors at $\frac{-1}{m}$, and class 1 support vectors at $\frac{1}{m}$, i.e. there will be only 2 support vectors at $\frac{-1}{m}$ and $\frac{1}{m}$. The separating hyperplane in the equivalent classification problem is the line

$$mx = 0 \quad (13)$$

. The regression estimate z^* for a test sample x^* is given by $z^* = mx^*$. In the equivalent classification problem, the samples corresponding to x^* are given by

$$\begin{pmatrix} x^* \\ z^* \end{pmatrix} \text{ in class 1, and} \quad (14)$$

$$\begin{pmatrix} -x^* \\ z^* \end{pmatrix} \text{ in class -1} \quad (15)$$

The regression problem and the classification task are shown in Fig. 2. Note that the margin in this case is $\frac{2}{m}$.

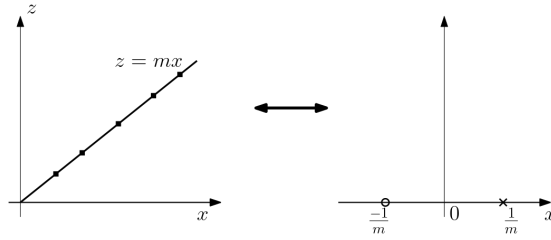


Fig. 2. Left: Regression samples taken from the line $z = mx$. Right: Samples for the equivalent classification task lie at $\pm \frac{1}{m}$. Note that the margin for the classification task is $\frac{2}{m}$, that tends to zero as the line becomes more vertical.

Observe that as the slope of the line m increases, the margin reduces. Any small perturbation of the sample location x to $x + \epsilon$ will lead to an error in the regression estimate by $m(x + \epsilon)$, i.e. the error will increase as the line becomes more vertical. When $m \rightarrow \infty$, z cannot be estimated from x , and the margin of the equivalent classification task $\rightarrow 0$.

In real world tasks, the permissible regression error is relative to the regressed value, e.g. when plotting the current-voltage (I-V) relationship curve for a diode, the currents may range from microamperes 10^{-6} for small values of the voltage V across the diode, to amperes for large values of V . An error of a μA is negligible when $I = 1 A$, but clearly unacceptable for $I = 1 \mu A$. Our approach to regression shows that the equivalent classification task normalizes samples with respect to their location on the regressing hyperplane.

Consider a regression task with samples $\{x^i, z_i\}$, $i = 1, 2, \dots, M$, where $x^i \in \mathbb{R}^n$, $z_i \in \mathbb{R} \setminus 0$.

$$x_+^i = \begin{pmatrix} x^i \\ z_i \end{pmatrix} \text{ in class 1, i.e. } y_i = 1, \text{ and} \quad (16)$$

$$x_-^i = \begin{pmatrix} -x^i \\ z_i \end{pmatrix} \text{ in class -1, i.e. } y_i = -1 \quad (17)$$

We now show the SVC formulation and a formal equivalence. The SVC formulation is in the spirit of a least squares SVM [4] or proximal SVM [5], but employs a L1 error measure. Consider the SVC formulation where we assume that the

separating hyperplane passes through the origin. The classification dataset is given by $\{x^i, y_i\}$, $i = 1, 2, \dots, M$, where $x^i \in \mathbb{R}^n$, $y_i \in \{-1, 1\}$

$$\text{Minimize } \frac{1}{2} \|w\|^2 + C \sum_{i=1}^M (q_i^+ + q_i^-) \quad (18)$$

‘subject to the constraints

$$y_i(w^T x^i) + (q_i^+ - q_i^-) = 1 \quad (19)$$

$$q_i^+, q_i^- \geq 0, \quad i = 1, 2, \dots, M \quad (20)$$

$$(21)$$

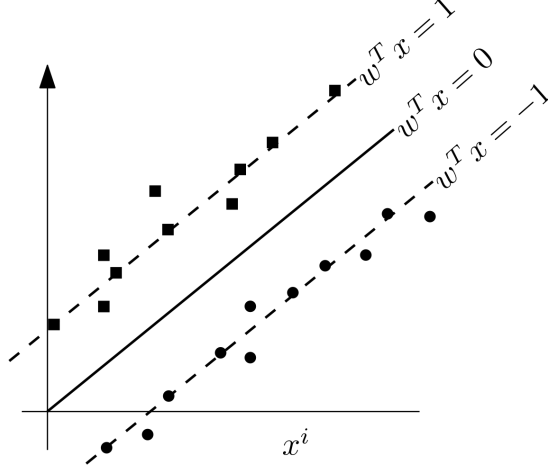


Fig. 3. The separating hyperplane is given by $w^T x = 0$. Hyperplanes $w^T x = 1$ and $w^T x = -1$ are proximal to samples of the two classes. A sample x^i not lying on either of these hyperplanes will be at a non-zero distance from its proximal plane, given by $w^T x^i = z_i - (q_i^+ - q_i^-)$

The separating hyperplane is given by $w^T x = 0$; hyperplanes $w^T x = 1$ and $w^T x = -1$ are proximal to the two classes, as in the case of the least squares SVM [4] or proximal SVM [5]. The separating and proximal planes are illustrated in Fig. 3.

The dual is derived in A and reproduced here for convenience.

$$\text{Minimize } \frac{1}{2} \sum_{i=1}^M \sum_{j=1}^M \lambda_i \lambda_j y_i y_j (x^i)^T (x^j) - \sum_{i=1}^M \lambda_i \quad (22)$$

subject to the constraints

$$-C \leq \lambda_i \leq C \quad (23)$$

$$(24)$$

From the K.K.T. conditions (see Appendix refAppendix1), we note that

$$-C < \lambda_i < C \implies q_i^+ = q_i^- = 0 \implies y_i(w^T x^i) = 1 \quad (25)$$

$$\lambda_i = -C \implies q_i^- = 0 \implies y_i(w^T x^i) < 1 \quad (26)$$

$$\lambda_i = C \implies q_i^+ = 0 \implies y_i(w^T x^i) > 1 \quad (27)$$

$$(28)$$

The equivalent classification samples

$$x_+^i = \left(\frac{x^i}{z_i} \right) \text{ in class 1, i.e. } y_i = 1, \text{ and} \quad (29)$$

$$x_-^i = \left(\frac{-x^i}{z_i} \right) \text{ in class -1, i.e. } y_i = -1 \quad (30)$$

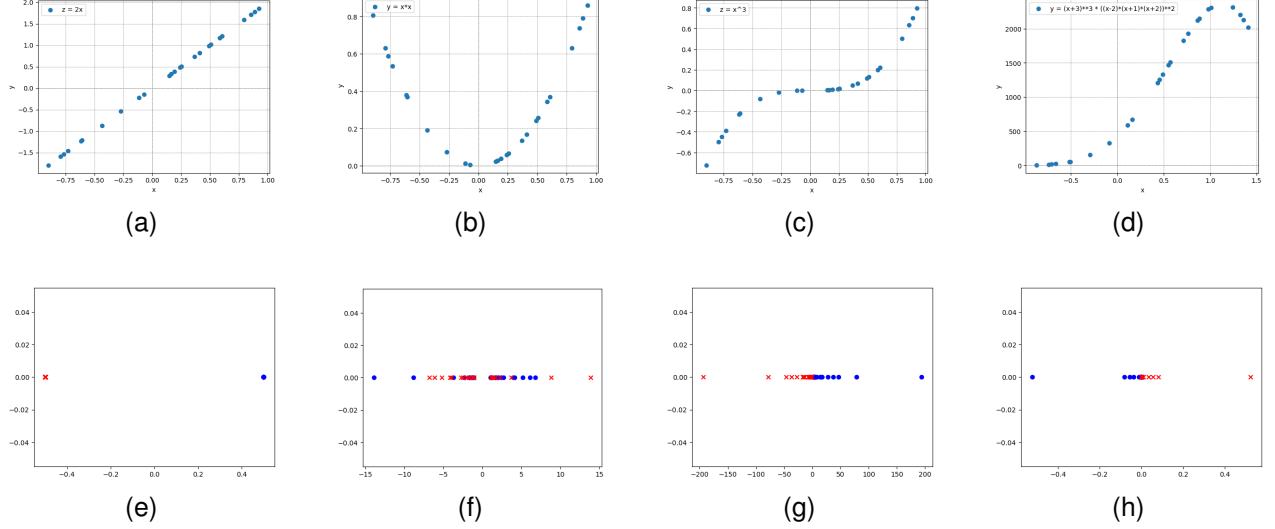


Fig. 4. (a) $z = 2x$, (b) $z = x^2$, (c) $z = x^3$, (d) $z = ((x+3)^3)((x-2)(x+1)(x+2))^2$; (e)–(h) equivalent binary classification problems corresponding to (a)–(d), with samples of the two classes distinguished by colour.

are then used to solve the SVC (22)–(23), which now has $2M$ samples. For convenience, we re-index the samples from $-M$ to M (excluding 0), so that $x^i = x^i_+$, $i = 1, 2, \dots, M$ and $x^i = x^i_-$, $i = -1, 2, \dots, -M$. The Lagrange multipliers λ_i , $i = \pm 1, \pm 2, \dots, \pm M$, are also correspondingly re-indexed. Note that

$$-C < \lambda_i < C \implies w^T \left(\frac{x^i}{z_i} \right) = 1 \implies w^T x^i = z_i, \quad i = 1, 2, \dots, M \quad (31)$$

$$\implies w^T \left(\frac{-x^i}{z_i} \right) = -1, \implies -C < \lambda_i < C, \quad i = -1, -2, \dots, -M \quad (32)$$

$$\lambda_i = -C \implies q_i^- = 0 \implies w^T \left(\frac{x^i}{z_i} \right) < 1 \implies w^T x^i < z_i, \quad i = 1, 2, \dots, M \quad (33)$$

$$\implies w^T \left(\frac{-x^i}{z_i} \right) > -1 \implies \lambda_i = -C, q_i^+ = 0 \quad i = -1, -2, \dots, -M \quad (34)$$

$$\lambda_i = C \implies q_i^+ = 0 \implies w^T \left(\frac{x^i}{z_i} \right) > 1 \implies w^T x^i > z_i, \quad i = 1, 2, \dots, M \quad (35)$$

$$\implies w^T \left(\frac{-x^i}{z_i} \right) < -1 \implies \lambda_i = C, q_i^- = 0 \quad i = -1, -2, \dots, -M \quad (36)$$

$$(37)$$

In short, samples (x^i, z_i) that satisfy on $w^T x^i = z_i$ correspond to Lagrange multipliers that satisfy $-C \leq \lambda_i \leq C$; those lying above the regressor, i.e. $w^T x^i < z_i$ correspond to $\lambda_i = -C$, and those lying below the regressor, i.e. $w^T x^i > z_i$ correspond to $\lambda_i = C$.

Figure 4 shows regression samples drawn randomly from four known functions of one variable, and equivalent samples for corresponding classification tasks, obtained by using (III). Figure 4(a) shows samples from the line $z = 2x$. In this case, the equivalent classification task has only 2 samples, one at $x = -\frac{1}{2}$, and the other at $x = \frac{1}{2}$, as shown in Fig. 4(e). The pairs (b), (f), (c), (g), and (d), (h) illustrate other functions.

It is evident that other than the case when samples are drawn from a straight line, other functions lead to non-linearly separable classification problems. Dong and Kothari [6] proposed a novel approach to estimating the relative difficulty of classifying a labelled set of samples, that they termed as "classifiability". It is not a wrapper approach, and does not require a classifier to be learned, in order to compute the measure. In Section IV, we briefly discuss classifiability, and using our equivalence, propose a "regressability" measure that estimates the difficulty of regressing a set of samples. Illustrative examples highlighted in the section provide insight into why some datasets may be much harder than others.

IV. HOW DIFFICULT IS LEARNING A REGRESSION DATASET : REGRESSABILITY

Figure 5(a) shows a classification problem in two variables; Fig. 5(b) shows the same data with the third dimension being the class label. Figure 6(a) and (b) show the same for another problem, where the data is better separated by a smoother decision

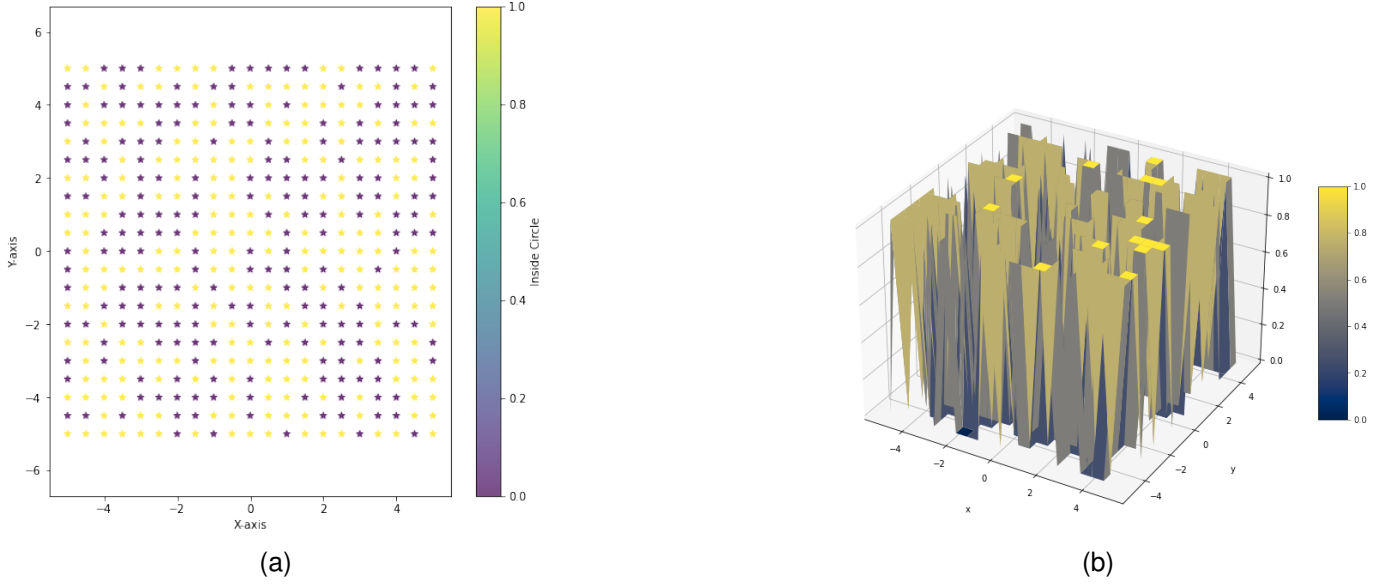


Fig. 5. (a) Samples of a classification problem and (b) the same samples with the class label indicated by the surface height.

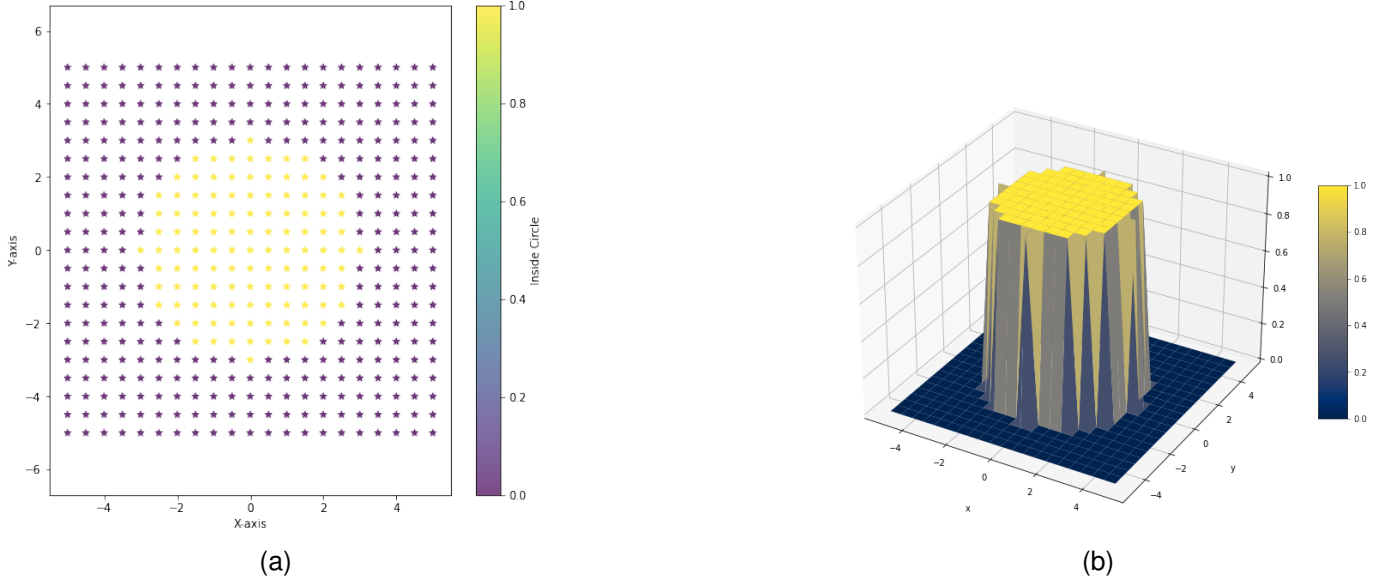


Fig. 6. (a) Samples of a simple classification problem and (b) the same samples with the surface height equal to the class label.

boundary.

Dong and Kothari [6] argued that the roughness of the class label surface is a measure of the difficulty of classifying the data, since the surface will be smooth when neighbours with similar labels are adjacent to each other.

Dong and Kothari [6] examined the second order joint conditional density function $f(\omega_1, \omega_2 | d)$, i.e. the probability of going from class ω_1 to class ω_2 within a distance d . Formally, in a binary classification problem, consider a training sample x^i , and another sample x^j lying in its neighbourhood, say within a distance d . Assuming that x^i and x^j are independent, the joint probability matrix

$$J^i = \begin{bmatrix} P(\omega_1 | x^j, \omega_1 | x^i) & P(\omega_2 | x^j, \omega_1 | x^i) \\ P(\omega_1 | x^j, \omega_2 | x^i) & P(\omega_2 | x^j, \omega_2 | x^i) \end{bmatrix} \quad (38)$$

will be strongly diagonal when patterns in the neighborhood of x^i belong to x^i 's class. As the class label surface becomes more rough, the off-diagonal entries become larger. Classifiability in the neighbourhood of x^i is computed as

$$\begin{aligned} C(x^i) = & P(\omega_1 | x^j) P(\omega_1 | x^i) \\ & + P(\omega_2 | x^j) P(\omega_2 | x^i) - P(\omega_2 | x^j) P(\omega_1 | x^i) \\ & - P(\omega_1 | x^j) P(\omega_2 | x^i) \end{aligned} \quad (39)$$

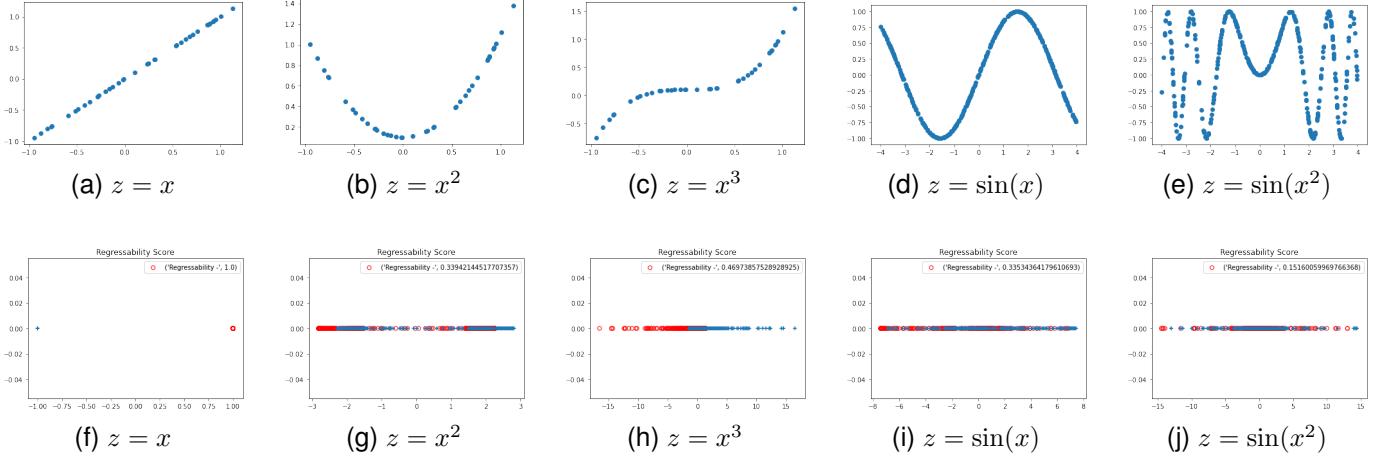


Fig. 7. (a)–(e) Plots of selected 1-D functions, and (f)–(j) equivalent classification problems with computed regressability scores.

The classifiability of the dataset is given by

$$L = \sum_i P(x^i) C(x^i), \quad (40)$$

where $P(x^i)$ is the number of patterns in the neighbourhood of x^i divided by the number of samples N . Given a regression dataset \mathbf{R} , we first determine the equivalent cClassification dataset \mathbf{C} . Regressability of \mathbf{R} is determined as the classifiability of the equivalent classification dataset \mathbf{C} . Figure 7 illustrates the concept on a few examples.

Figure 7 provides some insight into regressability. The linear function $z = x$ has a regressability score of 1.0; the equivalent classification samples are linearly separable. The regressability of $z = x^2$ is lower than that of $z = x^3$, because the monotonic slope of $z = x^3$ makes it easier to approximate with a line. This observation is with the caveat that trends may differ in the case of noisy data. Finally, $z = \sin(x)$ has higher regressability than $z = \sin(x^2)$. Table IV includes a few more illustrative examples. Data samples in the table are taken from standard functions; the noise added is from a Gaussian distribution with mean 0 and variance 1.

TABLE I
REGRESSABILITY EXAMPLES

Function	Regressability
$Y = X + \text{Noise}$	0.99
$Y = X^2 + \text{Noise}$	0.34
$Y = X^3 + \text{Noise}$	0.31
$Y = X^5 + \text{Noise}$	0.18
$Y = (((X + 1)^2) * (X - 3)^2) + \text{Noise}$	0.26
$Y = (((X + 3)^3) * ((X - 2) * (X + 1) * (X + 2))^2) + \text{Noise}$	0.28
$Y = (((X + 3)^2) * ((X + 1) * (X + 2))^2) + \text{Noise}$	0.42
$Y = \sin(X) + \text{Noise}$	0.43
$Y = \sin(X^2) + \text{Noise}$	0.15
$Y = X^2 + \sin(X^2) + \text{Noise}$	0.34

The availability of an equivalent classification problem enables a large repertoire of theoretical results and techniques traditionally used for classification, to be adapted for regression. For example, the notion of margin can be understood clearly in the context of equivalent samples. In the next section, we use the preceding discussion, to develop a different approach to regression using neural networks. We train a neural network to learn a linearizing map. Given a set of binary classification samples at the input of a neural network, image vectors at the output layer are such that samples of the same class are mapped close to each other in the output space, and those of different classes are as distant from each other as possible. The loss function tries to minimize the within-class scatter, and maximize the between-class scatter.

V. LEARNING A LINEARIZING MAP

Figure 8 illustrates the proposed approach. Consider a neural network that maps input samples x into a space $\phi(x)$ at the output layer. An input regression sample (x^i, z_i) is mapped to its image $\phi(x^i)$ at the output layer. The equivalent binary classification samples at the output are $\frac{\phi(x^i)}{z_i}$ and $\frac{-\phi(x^i)}{z_i}$. The neural network loss function is designed to minimize the scatter within samples of the same class and maximize the distance between samples of different classes. Figure 9(a) shows the plot

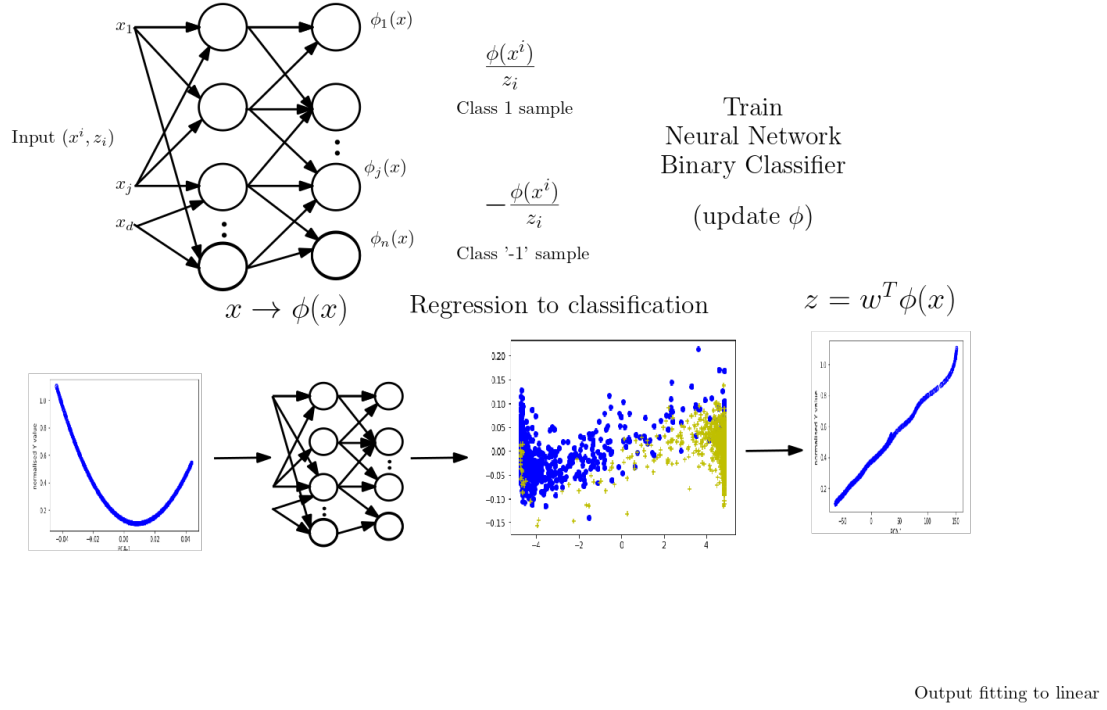


Fig. 8. Learning a linearizing map $\phi(x)$ given samples (x^i, z_i)

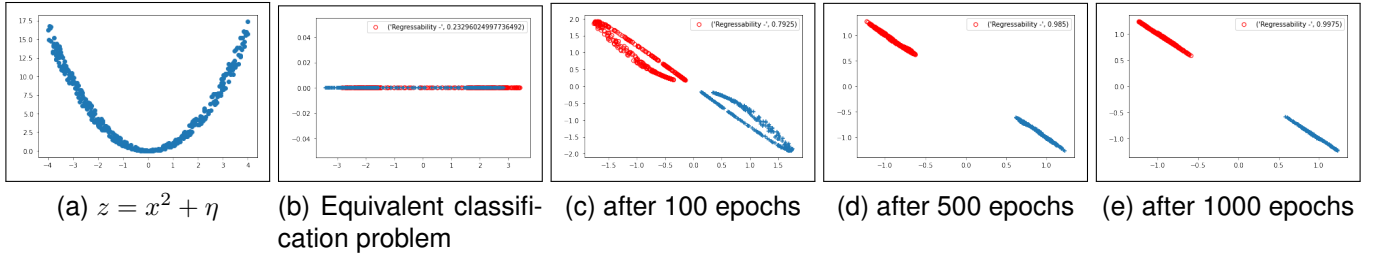


Fig. 9. Neural network training: (a) samples from $z = x^2 + \text{noise}$; (b) samples of the equivalent classification problem; (c)–(e) projection of output-layer samples $\phi(x^i)$ onto the first principal component, with the vertical axis indicating the regression target (z), taken after 100, 500, and 1000 training epochs, respectively.

of the function $z = x^2$. Gaussian noise with zero mean and unit variance has been added to the target z . Figure 9(b) shows how the equivalent binary classification samples are distributed; colours indicate class labels. Figure 9(c)–(e) show the output layer image samples i.e. $\phi(x^i)$, projected onto the first two principal components of the space of image samples. Note that the regressability of samples in the $\phi()$ space improves as the neural network gets trained over epochs.

Figures 10 and 11 illustrate the same with functions $z = x^3$ and $z = \sin(x^2)$. One advantage of learning a linearizing map is better interpretability. The output layer features provide insight into a regressed function since it can be expressed as a linear combination of them.

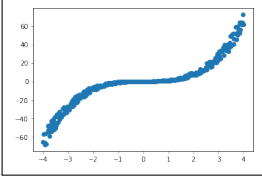
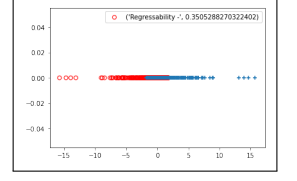
Define

$$S_b = (\mu_1 - \mu_2) (\mu_1 - \mu_2)^T \quad (41)$$

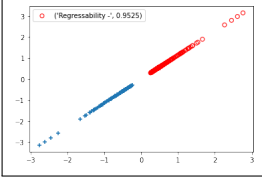
$$S_w = \frac{1}{2M} \sum_{i=1}^2 \sum_{x^j \in \text{Class } i}^M (\phi(x_j) - \mu_i) (\phi(x_j) - \mu_i)^T \quad (42)$$

where μ_1 and μ_2 are the means of class 1 and class 2, respectively. The neural network uses the J_4 loss function proposed by Fukunaga [7], that is given by

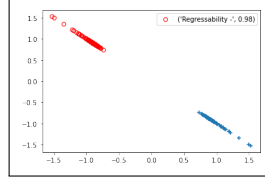
$$J_4 = \frac{S_w}{S_b} \quad (43)$$

(a) $z = x^3 + \eta$ 

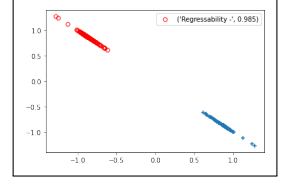
(b) Equivalent classification problem



(c) after 100 epochs

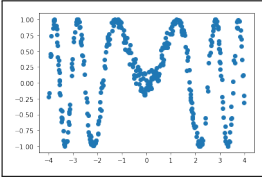
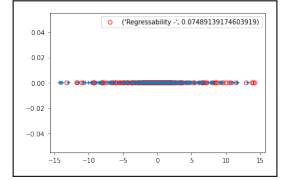


(d) after 500 epochs

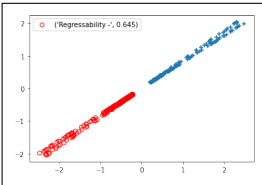


(e) after 1000 epochs

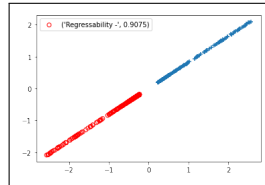
Fig. 10. Neural network training: (a) samples taken from $z = x^3 + \text{noise}$; (b) samples of the equivalent classification problem; (c)–(e) projection of output-layer samples $\phi(x^i)$ onto the first principal component, with the vertical axis indicating the regression target (z), taken after 100, 500, and 1000 training epochs, respectively.

(a) $z = \sin(x^2) + \eta$ 

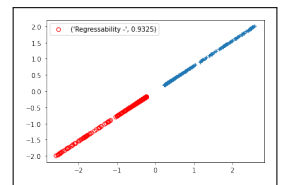
(b) Equivalent classification problem



(c) after 100 epochs



(d) after 500 epochs



(e) after 1000 epochs

Fig. 11. Neural network training: (a) samples taken from $z = \sin(x^2) + \text{noise}$; (b) samples of the equivalent classification problem; (c)–(e) projection of output-layer samples $\phi(x^i)$ onto the first principal component, with the vertical axis indicating the regression target (z), taken after 100, 500, and 1000 training epochs, respectively.

Minimizing J_4 minimizes the within-class scatter and maximizes the between-class scatter. Algorithm 1 summarizes the training procedure for a neural network using the J_4 loss function.

Figures 12 and 13 show how the map evolves when a neural network is trained with samples of $z = x^3$ and $z = x^2$, respectively. The network has 3 layers, with one input neuron, 5 hidden layer neurons, and 10 neurons in the output (ϕ) layer. A tanh activation was used, and the learning rate was 0.01. The samples consisted of 1200 points randomly chosen in the domain $(-2.0, 3.0)$.

Each subfigure shows z_i plotted against the projection of $\phi(x^i)$ on the first principal component of output layer samples. The evolution indicates that as training progresses, the neural network learns a map so that the regressed output z is a linear combination of neuron outputs in the final layer, i.e. $z = w^T \phi(x)$, where w is a set of real co-efficients.

VI. EXPERIMENTAL RESULTS

In order to provide a comprehensive comparison on the performance of the proposed approach, we refer to the very extensive experimental survey of regression methods by Delgado et al. [8]. They consider 77 popular regression methods categorized into 19 families. These datasets were meticulously categorized into four distinct groups: "Small Easy," "Small Difficult," "Large Easy," and "Large Difficult." We chose datasets from the "Large Difficult" group, to facilitate comparisons with the most

Algorithm 1 J_4 Regression Algorithm

Input: Samples $(x^i, z_i), i = 1, 2, \dots, M$
Parameters: ; Neural Network Architecture (no. of layers, neurons per layer), learning rate η . **Output:** Feature map at the output layer, denoted by $\phi(x)$, where x is the input to the first layer.

- 1: **while** not converged in epoch i **do**
 - 2: Feed forward the sample from input space to output space $x \rightarrow \phi(x)$
 - 3: Convert $(\phi(x^i), z_i), i = 1, 2, 3..M$ into an equivalent classification problem
 - 4: Class +1: $\frac{\phi(x^i)}{z_i}$
 - 5: Class -1: $\frac{\phi(x^i)}{-z_i}$
 - 6: **for** each class c_i **do**
 - 7: calculate μ_i : The sample mean for the i -th class
 - 8: Calculate the J_4 loss using (V)
 - 9: Update weights of all layers by back-propagating the gradient of the loss function.
 - 10: **end for**
 - 11: **end while**
 - 12: Denote the map learnt by the neural network as $\phi(x)$, where x is the vector of inputs to the first layer of the neural network. Determine the co-efficients of a linear regressor $z = w^T \phi(x)$ to minimize the Mean Squared Error loss over all samples.
-

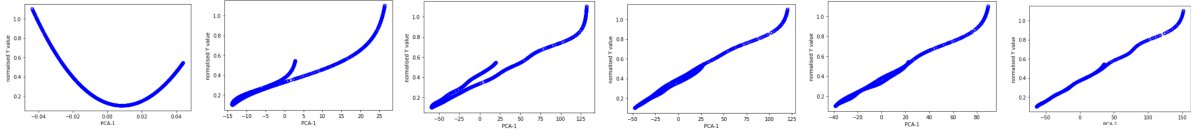


Fig. 12. Samples of $z_i = (x^i)^2$ plotted vs. the projection of $\phi(x^i)$ onto the first principal component of output layer samples. The evolution indicates that as training progresses, the neural network learns a map so that the regressed output z is a linear combination of neuron outputs in the final layer, i.e. $z = w^T \phi(x)$, where w is a set of real co-efficients. Snapshots are at epoch 0, 2000, 4000, 6000, 8000, and 10000.

challenging and widely distributed examples. The training and testing methodology, including data folds, follows that used by [8]. Table II summarizes relevant statistics about the selected datasets.

Dataset Name	No of Samples	Dimensions
1 Airfoil	1,503	3
2 3DRoad	4,34,874	4
3 Beijing pm25	41,758	12
4 Buzz Twitter	583,250	77
5 Cuff less	61,000	3
6 Facebook comment	40,949	54
7 KEGG Relation	54,413	22
8 Online news	39,644	59
9 Greenhouse net	955,167	15
10 Physico protein	45,730	9
11 pm25 beijing dongsi	24,237	13
12 pm25 beijing dongsi huan	20,166	13
13 pm25 beijing nongzhanguan	24,137	13
14 pm25 beijing us post	49,579	13
15 pm25 chengdu shahepu	23,142	13
16 pm25 chengdu caotangsi	22,997	13

TABLE II
DATASET INFORMATION

Table III summarizes comparisons of the proposed J_4 regression approach with results of the Average Neural Network (AvNNet) taken from [8]. The AvNNet has been used as the baseline since it is the neural network based approach (see [8]) in their survey.

The tabulated results indicate that the imposition of a linearity constraint during the training of a Deep Neural Network led to an enhancement in model performance. Specifically, the R2 score achieved through the application of the J_4 regression method surpassed that of the Averaged Neural Network Model utilizing Mean Squared Error loss.

VII. CONCLUSION

In this work, we proposed a novel approach to regression tasks by incorporating linearity constraints inspired by the mathematical principles governing straight lines. Our methodology focused on transforming regression problems into equivalent

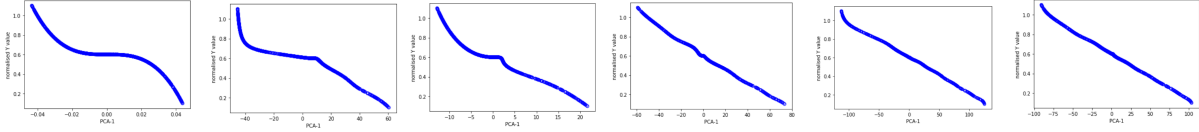


Fig. 13. Samples of $z_i = (x^i)^3$ plotted vs. the projection of $\phi(x^i)$ onto the first principal component of output layer samples. The evolution indicates that as training progresses, the neural network learns a map so that the regressed output z is a linear combination of neuron outputs in the final layer, i.e. $z = w^T \phi(x)$, where w is a set of real co-efficients. Snapshots are at epoch 0, 2000, 4000, 6000, 8000, and 10000.

TABLE III
TEST DATA R2 SCORE FOR AVNN AND J4 REGRESSOR ON DIFFERENT DATASETS

Sr No	Data Set	AvNN(Avg Neural Network)	J4 Regressor
1	Airfoil	0.892	0.902
2	3DRoad	0.4681	0.9055
3	Beijing_pm25	0.6386	0.6531
4	Buzz_Twitter	0.3929	0.9491
5	Cuff_less	0.7242	0.6465
6	Facebook_comment	0.8205	0.7167
7	KEGG_Relation	0.9089	0.9483
8	Online_news	0.1445	-0.00012
9	Greenhouse_net	0.1259	0.483
10	Physico_protein	0.4728	0.5342
11	pm25_beijing_dongsi	0.6174	0.6436
12	pm25_beijing_dongsihuan	0.6495	0.6767
13	pm25_beijing_nongzhanguan	0.6312	0.6704
14	pm25_beijing_us_post	0.6229	0.6445
15	pm25_chengdu_shahepu	0.4903	0.6042
16	pm25_chengdu_caotangsi	0.5121	0.5967

classification problems and enforcing linearity in the feature space. We utilized the J_4 loss function to achieve convergence of class points and maximize the separation between them.

Our approach demonstrated successful linearity in the feature space, allowing for the prediction of data points in both interpolated and extrapolated regions. We further extended our methodology to handle interpolation and extrapolation using Radial Basis Function Neural Networks, showcasing its effectiveness in capturing input data distribution and ensuring smooth predictions.

The experimental results on various datasets, including challenging "Large Difficult" datasets, revealed that our J_4 regression model outperformed the Averaged Neural Network Model with Mean Squared Error loss in terms of R2 scores. This suggests that enforcing linearity constraints during training can improve the performance of regression models, especially in scenarios with complex and widely distributed data.

In conclusion, our proposed regression methodology, with its focus on linearity and the J_4 loss function, presents a promising approach for addressing regression tasks. Further research and application of this methodology in diverse domains could provide valuable insights into its generalizability and effectiveness in UCI datasets.

APPENDIX

APPENDIX I

Consider a classification dataset given by $\{x^i, y_i\}$, $i = 1, 2, \dots, M$, where $x^i \in \mathbb{R}^n$, $y_i \in \{-1, 1\}$. We rewrite the SVC formulation (22)-(23), where we assume that the separating hyperplane passes through the origin.

$$\text{Minimize } \frac{1}{2} \|w\|^2 + C \sum_{i=1}^M (q_i^+ + q_i^-) \quad (44)$$

subject to the constraints

$$y_i(w^T x^i) + (q_i^+ - q_i^-) = 1 \quad (45)$$

$$q_i^+, q_i^- \geq 0, \quad i = 1, 2, \dots, M \quad (46)$$

$$(47)$$

The Lagrangian for the above primal is given by

$$\begin{aligned} \mathcal{L}(w, q^+, q^-, \lambda, \eta^+, \eta^-) = & \frac{1}{2} \|w\|^2 + C \sum_{i=1}^M (q_i^+ + q_i^-) - \sum_{i=1}^M \eta_i^+ q_i^+ - \sum_{i=1}^M \eta_i^- q_i^- \\ & + \sum_{i=1}^M \lambda_i [1 - y_i(w^T x^i) - q_i^+ + q_i^-] \end{aligned} \quad (48)$$

$$-\infty \leq \lambda_i \leq \infty; \eta_i^+ \geq 0; \eta_i^- \geq 0 \quad (49)$$

The Karush-Kuhn-Tucker (K.K.T.) conditions are given by [9]

$$\nabla_w \mathcal{L} = 0 \implies w - \sum_{i=1}^M \lambda_i y_i x^i = 0 \implies w = \sum_{i=1}^M \lambda_i y_i x^i \quad (50)$$

$$\frac{\partial \mathcal{L}}{\partial q_i^+} = 0 \implies C - \eta_i^+ - \lambda_i = 0 \implies \eta_i^+ + \lambda_i = C \quad (51)$$

$$\frac{\partial \mathcal{L}}{\partial q_i^-} = 0 \implies C - \eta_i^- + \lambda_i = 0 \implies \lambda_i - \eta_i^- = C \quad (52)$$

$$-\infty \leq \lambda_i \leq \infty \quad (53)$$

$$\eta_i^+, \eta_i^- \geq 0 \quad (54)$$

From (51), (52), (53) and (54), we have

$$\eta_i^+ = C - \lambda_i \geq 0 \implies \lambda_i \leq C \quad (55)$$

$$\eta_i^- = C + \lambda_i \geq 0 \implies \lambda_i \geq -C \quad (56)$$

The complementarity K.K.T. conditions are given by

$$\lambda_i [1 - y_i(w^T x^i) - q_i^+ + q_i^-] = 0 \quad (57)$$

$$\eta_i^+ q_i^+ = 0 \quad (58)$$

$$\eta_i^- q_i^- = 0, \quad i = 1, 2, \dots, M \quad (59)$$

Substituting from (50)-(52) into (48), we obtain the dual as

$$\mathbf{Max}_\lambda - \frac{1}{2} \sum_{i=1}^M \sum_{j=1}^M \lambda_i \lambda_j y_i y_j (x^i)^T x^j + \sum_{i=1}^M \lambda_i \quad (60)$$

subject to the constraints

$$-C \leq \lambda_i \leq C \quad (61)$$

which may be rewritten as

$$\mathbf{Min}_\lambda \frac{1}{2} \sum_{i=1}^M \sum_{j=1}^M \lambda_i \lambda_j y_i y_j (x^i)^T x^j - \sum_{i=1}^M \lambda_i \quad (62)$$

subject to the constraints

$$-C \leq \lambda_i \leq C \quad (63)$$

Finally, we re-visit our assumption that the regressor is of the form $z = w^T x$, i.e. the regressor passes through the origin and does not have an offset. Consider a regressor of the form $z = w^T x + b$, where b is the offset. We assume that at least one sample, say, x^0 , lies on the line, i.e. $z_0 = w^T x^0 + b$. Note that at this point we do not know w or b , since these are determined after solving the optimization problem, and the equivalent classification samples need to be determined before solving for the classifier.

Sample x^0 is treated as a reference point. All samples (x^i, z_i) , $i = 1, 2, \dots, M$ are replaced by $(x^i - x^0, z_i - z_0)$. Note that $w^T(x - x^0)$ passes through the origin, and b is not needed.

In principle, we could take any sample and treat it as the reference x^0 . Of course, x^0 is excluded from training data. However, in practice, some samples may be noisy or outliers; choosing such samples is imprudent. We therefore choose

$$x^0 = \frac{1}{M} \sum_{i=1}^M x^i \quad (64)$$

$$z_0 = \frac{1}{M} \sum_{i=1}^M z_i \quad (65)$$

Although this choice is not always optimal, it averages any noise present in sample locations x^i or regression targets z_i .

REFERENCES

- [1] L. Torgo and J. Gama, "Regression by classification," in Advances in Artificial Intelligence: 13th Brazilian Symposium on Artificial Intelligence, SBIA'96 Curitiba, Brazil, October 23–25, 1996 Proceedings 13, pp. 51–60, Springer, 1996.
- [2] R. Salinan and V. Kecman, "Regression as classification," in 2012 Proceedings of IEEE Southeastcon, pp. 1–6, IEEE, 2012.
- [3] M. Pontil, R. Rifkin, and T. Evgeniou, "From regression to classification in support vector machines," 1998.
- [4] J. A. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," Neural processing letters, vol. 9, pp. 293–300, 1999.
- [5] G. Fung and O. L. Mangasarian, "Proximal support vector machine classifiers," in Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 77–86, 2001.
- [6] M. Dong and R. Kothari, "Feature subset selection using a new definition of classifiability," Pattern Recognition Letters, vol. 24, no. 9-10, pp. 1215–1225, 2003.
- [7] K. Fukunaga, Introduction to statistical pattern recognition. Elsevier, 2013.
- [8] M. Fernández-Delgado, M. S. Sirsat, E. Cernadas, S. Alawadi, S. Barro, and M. Febrero-Bande, "An extensive experimental survey of regression methods," Neural Networks, vol. 111, pp. 11–34, 2019.
- [9] S. Chandra, Jayadeva, and A. Mehra, Numerical optimization with applications. Alpha Science International, 2009.