Checkered Regression

Mastane Achab*

22 February 2022

This paper introduces the checkered regression model, a nonlinear generalization of logistic regression. More precisely, this new binary classifier relies on the multivariate function $\frac{1}{2}\left(1+\tanh(\frac{z_1}{2})\times\cdots\times\tanh(\frac{z_m}{2})\right)$, which coincides with the usual sigmoid function in the univariate case $m=1$. While the decision boundary of logistic regression consists of a single hyperplane, our method is shown to tessellate the feature space by any given number $m\geq 1$ of hyperplanes. In order to fit our model's parameters to some labeled data, we describe a classic empirical risk minimization framework based on the cross entropy loss that can be optimized through stochastic gradient descent. A multiclass version of our approach is also proposed.

1 Introduction

Logistic regression (LR) is one of the most standard approaches for binary classification: it simply learns a linear prediciton rule, through a convex minimization problem in the case of the cross entropy loss (see e.g. [1], [4]). Nevertheless, it suffers from a lack of representational power: indeed, it cannot handle nonlinear relations between features and labels. For that reason, artificial neural networks (a.k.a. deep learning models) are nowadays preferred over linear methods such as LR across a broad spectrum of machine learning applications, ranging from image or speech recognition to natural language processing ([3]). But in general, the optimization of a deep neural network is a highly non-convex problem for which we still have little theoretical understanding.

As an alternative to deep learning, this paper proposes a new binary classifier that strictly generalizes LR: we call it the *checkered regression* (CR) model. Loosely speaking, CR can be seen as a single hidden-layer neural network with tanh activations that are multiplied together, instead of being additively combined as is customary. This is somehow similar to the NALU module proposed in [6] that computes the elementwise product of tanh and sigmoid activations. Contrary to LR, the scope of CR expands beyond linear separability. As shall be seen in the next section, the predictions of a CR model can tile the feature space into a checkerboard-like pattern.

*mastane.achab@gmail.com

2 The checkered regression model

Let $m \geq 1$ and $\mathbf{1} = (1, \ldots, 1)$ be the all-ones vector of size m. We start with the formal definitions below.

Definition 1. (Checkoid function). *The checkoid function Ξ_m is defined for all $z = (z_1, \ldots, z_m) \in \mathbb{R}^m$ by*

$$\Xi_m(z) = \frac{1}{2}\left(1 + \prod_{k=1}^{m} \tanh\left(\frac{z_k}{2}\right)\right) = \frac{\sum_{\upsilon \in \{0,1\}^m \ s.t. \ \mathbf{1}^\intercal \upsilon \equiv 0[2]} e^{-\upsilon^\intercal z}}{(1 + e^{-z_1}) \times \cdots \times (1 + e^{-z_m})} \ .$$

Definition 2. (Checkered regression). *Let $\mathcal{X} \subseteq \mathbb{R}^d$ ($d \geq 1$). The checkered regression model with parameters $\omega = (\omega_1, \ldots, \omega_m) \in \mathbb{R}^{dm}$ is given by the posterior probabilities*

$$p_\omega\left(0|x\right) = 1 - p_\omega\left(1|x\right) = \Xi_m(\omega_1^\intercal x, \ldots, \omega_m^\intercal x) \quad \text{for all } x \in \mathcal{X} \ .$$

For $m = 1$, $\Xi_1 = \sigma$ is the sigmoid function and the checkered regression is simply a logistic regression. For general $m \geq 1$, we give next two properties of the checkoid function and CR.

Proposition 1. (Symmetry of Ξ_m). *Let $z = (z_1, \ldots, z_m) \in \mathbb{R}^m$, $k \in \{1, \ldots, m\}$ and $z' = (z'_1, \ldots, z'_m)$ with $z'_k = -z_k$ and $z'_j = z_j$ for $j \neq k$. Then,*

$$\Xi_m(z') = 1 - \Xi_m(z).$$

Proof. By the oddness of the hyperbolic tangent. □

Lemma 1. (Hamming distance parity). *Consider a checkered regression model with parameters $\omega = (\omega_1, \ldots, \omega_m)$. Let x and x' be two points in $\mathbb{R}^d$ both outside the m hyperplanes of the CR model, i.e. such that $\omega_k^\intercal x \neq 0$ and $\omega_k^\intercal x' \neq 0$ for all $1 \leq k \leq m$. Then, the two following conditions are equivalent.*

(i) The CR model predicts that x and x' share the same label:

$$sign\left(p_\omega\left(0|x\right) - \frac{1}{2}\right) = sign\left(p_\omega\left(0|x'\right) - \frac{1}{2}\right) \ .$$

(ii) The Hamming distance

$$\sum_{k=1}^{m} \mathbb{I}\left\{sign\left(\omega_k^\intercal x\right) \neq sign\left(\omega_k^\intercal x'\right)\right\}$$

is even.

Proof. By successive applications of Proposition 1. □

Lemma 1 shows that the decision regions of a CR model form a hyperplane tessellation of the feature space $\mathcal{X}$, where the tiles are binary labeled in a checkered fashion.

Cross entropy loss Given a random pair (X, Y) valued in $\mathcal{X} \times \{0, 1\}$ and a collection $((X_1, Y_1), \ldots, (X_n, Y_n))$ of i.i.d. copies of (X, Y), we follow the empirical risk minimization paradigm ([2]). In other words, we adjust the parameters ω of our CR model to match the true posterior probabilities

$$\mathbb{P}(Y = 0|X = x) = 1 - \mathbb{P}(Y = 1|X = x),$$

by minimizing some empirical risk. By analogy with LR, we take the cross entropy loss:

$$\min_{\omega \in \mathbb{R}^{dm}} \widehat{L}(\omega) := \frac{1}{n} \sum_{i=1}^{n} -Y_i \log p_\omega(1|X_i) - (1 - Y_i) \log p_\omega(0|X_i) \ , \tag{1}$$

which also corresponds to maximum likelihood estimation.

Gradient A natural way of solving the optimization problem (1) is through stochastic gradient descent (SGD). At each iteration of SGD, we need to compute the gradient (with respect to ω) of the loss induced by a single training pair (X_i, Y_i). The next result allows to do this differentiation.

Proposition 2. (Partial derivatives). *Consider the cross entropy loss of CR: for all $z = (z_1, \ldots, z_m) \in \mathbb{R}^m$,*

$$\ell(z) = -\log(\Xi_m(z)) \ .$$

Then for any $1 \leq k \leq m$, the k-th partial derivative of ℓ is

$$\frac{\partial \ell}{\partial z_k}(z) = \sigma(z_k) \cdot \left(1 - \frac{\Xi_{m-1}(z_{-k})}{\Xi_m(z)}\right) \ ,$$

where $z_{-k} = (z_j)_{j \neq k}$.

In particular, Proposition 2 implies that the partial derivatives of ℓ are bounded in the interval $(-1, 1)$.

Multiclass CR We now define a muticlass version of our model that can be used in classification problems with more than two classes. Let $\mathcal{Y} = \{0, \ldots, c-1\}$ be the set of classes with $c \geq 3$.

Definition 3. (Multiclass checkered regression). *The multiclass checkered regression model with parameters $\Omega = (\Omega_1, \ldots, \Omega_m) \in \mathbb{R}^{dm(c-1)}$ is given by the posterior probabilities*

$$\forall y \in \mathcal{Y}, \quad p_\Omega(y|x) = \Xi_{m,y}(\Omega_1 x, \ldots, \Omega_m x) \ ,$$

where the multiclass checkoid function $\Xi_{m,y}$ is defined for all $Z \in \mathbb{R}^{(c-1)\times m}$ as

$$\Xi_{m,y}(Z) = \frac{\sum_{\upsilon \in \mathcal{Y}^m \ s.t. \ \mathbf{1}^\intercal \upsilon \equiv y[c]} e^{-tr(E_\upsilon Z)}}{\left(1 + \sum_{j=1}^{c-1} e^{-Z_{j,1}}\right) \times \cdots \times \left(1 + \sum_{j=1}^{c-1} e^{-Z_{j,m}}\right)} \ ,$$

where E_υ is the $m \times (c-1)$ matrix given by $[E_\upsilon]_{k,j} = \mathbb{I}\{j = \upsilon_k\}$ for $1 \leq k \leq m, 1 \leq j \leq c-1$.

For $m = 1$, multiclass CR coincides with the well-known multiclass (or 'softmax') LR model.

References

[1] C. M. Bishop. *Pattern recognition and machine learning*. springer, 2006.

[2] L. Devroye, L. Györfi, and G. Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

[3] I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. MIT press, 2016.

[4] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[5] Y. Plan and R. Vershynin. Dimension reduction by random hyperplane tessellations. *Discrete & Computational Geometry*, 51(2):438–461, 2014.

[6] A. Trask, F. Hill, S. E. Reed, J. Rae, C. Dyer, and P. Blunsom. Neural arithmetic logic units. *Advances in neural information processing systems*, 31, 2018.