Neurocomputing 207 (2016) 141–149

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Discriminative directional classifiers

Kelwin Fernandes [a,b,*], Jaime S. Cardoso [a,b,*]

[a] INESC TEC, Porto, Portugal
[b] Universidade do Porto, Porto, Portugal

ARTICLE INFO

Article history:
Received 31 July 2015
Received in revised form
4 February 2016
Accepted 23 March 2016
Communicated by Ding Wang
Available online 13 May 2016

Keywords:
Directional statistics
Supervised classification
Discriminative models
Logistic regression

ABSTRACT

In different areas of knowledge, phenomena are represented by directional-angular or periodic-data; from wind direction and geographical coordinates to time references like days of the week or months of the calendar. These values are usually represented in a linear scale, and restricted to a given range (e.g. $[0, 2\pi)$), hiding the real nature of this information. Therefore, dealing with directional data requires special methods. So far, the design of classifiers for periodic variables adopts a generative approach based on the usage of the von Mises distribution or variants. Since for non-periodic variables state of the art approaches are based on non-generative methods, it is pertinent to investigate the suitability of other approaches for periodic variables. We propose a discriminative Directional Logistic Regression model able to deal with angular data, which does not make any assumption on the data distribution. Also, we study the expressiveness of this model for any number of features. Finally, we validate our model against the previously proposed directional naïve Bayes approach and against a Support Vector Machine with a directional Radial Basis Function kernel with synthetic and real data obtaining competitive results.

1. Introduction

Several phenomena and concepts in real life applications are represented by angular data or, as is referred in the literature, directional data. Some examples of directional information are the wind direction as analyzed by meteorologists, magnetic fields in rocks studied by geologists, geographic coordinates, among others [1]. Also, some entities are usually referenced in an angular manner; gynecologists denote the location to perform a biopsy, when performing a colposcopic screening, using the angle formed by the vertical axis of the cervix. Another example can be found in the area of computer vision, where color is often defined in cylindrical spaces like the Hue-Saturation-Value (HSV) color space. However, directional information is not constrained to scientific contexts; on a daily basis we naturally use angular variables. For example, time is usually represented by hours, days of the week, day of the month, season, etc. This reference system is cyclic by nature.

Directional variables are usually encoded as a periodic value in a given range (e.g. $[0, 2\pi)$, $[0°, 360°)$). This work focuses merely in this representation of directionality, where an angular variable is a real-value number with periodicity defined by a range. However, directional data can also be found in other representations, such as discrete categorical values ordered by a circular relation [2]. Also, some literature makes use of histograms which lie in a circular space instead of the linear one.

Working effectively with directional data requires dealing with techniques that are aware of the angular nature of the information [1]. For example, 0 and 2π are indeed the same angle and their average is not π but 0. In this sense, directional statistics concerns the problems derived from using traditional linear statistics with this type of data [1]. Even visualization of this type of data requires different representations to illustrate its periodic behavior (e.g. rose diagrams and circular histograms). In order to formalize the definition of a directional function, consider the predicate *dir* defined in Eq. (1), where $\mathbb{N}$ is the set of integers and $\mathbb{B} = \{\text{true}, \text{false}\}$:

$$\begin{aligned} &dir : \mathbb{N} \longrightarrow \mathbb{B} \\ &dir(i) = \text{true}, \quad \text{iff the } i\text{th feature is directional} \end{aligned} \tag{1}$$

We will say that the function f, with domain in $\mathbb{R}^n$, is directional with period $\vec{P}$ (i.e. the feature in the position i has period $\vec{P}_i$), if and only if Eq. (2) holds, where non-directional features are assumed to have infinite period (i.e. $\neg dir(i) \Rightarrow \vec{P}_i = \infty^+$):

$$f(\vec{\theta}) = f(\vec{\theta} + \vec{k} \circ \vec{P}), \quad \vec{k} \in \mathbb{Z}^n \tag{2}$$

Here on, we will restrict the periodicity of the directional values to $P_i = 1$, without loss of generality.

[*] Corresponding authors at: INESC TEC, Porto, Portugal.
E-mail addresses: kafc@inesctec.pt (K. Fernandes), jaime.cardoso@inesctec.pt (J.S. Cardoso).

http://dx.doi.org/10.1016/j.neucom.2016.03.076
0925-2312/© 2016 Elsevier B.V. All rights reserved.

Supervised learning can be understood as the process of learning a function f based on the so-called training data that comprises examples of the input vectors and their corresponding target values [3]. In this work, we are interested in the learning task known as classification, where the target can take a finite number of values. These values are usually denoted as classes or labels and the input vector defines a set of features that describe objects in the domain of the function. As the result of a supervised classification task, we obtain a classifier, which is used to assign a class to an object that has not been seen at the training stage. The ability to correctly label new instances is known as generalization [3]. Traditional models that do not take into account directionality may suffer drop of generalization in areas near to the period of the function. Furthermore, the function may return different decisions for different $\Delta + \overrightarrow{k} \circ \overrightarrow{P}$, $\overrightarrow{k} \in \mathbb{Z}^n$, and a fixed $\Delta \in \mathbb{R}^n$, despite all of them semantically represent the same angle.

In this work we propose a binary classifier aware of the directional constraint. The rest of this paper is organized as follows. Section 2 describes related work in the area of directional statistics and learning. Sections 3–5 detail the proposed model, its expressiveness and the optimization strategy, respectively. Section 6 summarizes the performed experiments to assess the relevance of the proposed model and, finally, Section 7 summarizes some conclusions and future work.

2. Related work

Most different types of problems and approaches in Machine Learning can be broadly defined as a classification, regression or clustering tasks. Classification and Regression are the most common supervised learning tasks. On the other hand, clustering is probably the best known unsupervised learning task, where the objective is to group data into non predefined categories based on some similarity criterion.

Previous attempts to address learning tasks with directional data have been carried out in each of the aforementioned areas. Most of them take advantage of circular distributions (such as von Mises and von Mises–Fisher). For instance, Banarjee et al. [4] proposed a generative mixture-model approach for clustering directional data using the von Mises–Fisher distribution. Moreover, they conclude that the spherical k-means is a special case of the mixture of von Mises–Fisher model. Fitting mixtures of angular distributions have been separately studied by Mooney et al. [5] and Mardia et al. [6].

Regression scenarios with directional data have been studied in several contexts [7–9]. Xu and Schoenberg [9] proposed a kernel regression method based on the von Mises distribution. Their method was used to discover the relationship between a single directional explanatory variable (wind direction) and a real-valued linear response variable (total area burned per day in wildfires). Fisher and Lee [7] studied the regression problem where the predictive variables are linear and the model outcome is directional. Their work also assumes that angular observations follow von Mises distributions and focuses on the estimation of the distribution parameters. Finally, Kato et al. [8] addressed the circular–circular problem, wherein both, predictive and target observations, have a circular nature.

Circular ordinal regression is an intermediate problem in this area, which lies between regression and classification. It considers a discrete number of labels which preserve a certain circular order. Devlaminck et al. [2] proposed two methods to solve this problem. The first one is an SVM variation, and the second method transforms the circular ordinal regression problem into multiclass classification. However, the directionality concerns in [2] are focused on the model outcome rather than on the feature space.

In the area of directional classification, different approaches have been considered: from Discriminant Analysis [10,11] to generative models [1,12,13]. SenGupta and Roy [14] proposed a distance-based classification rule using the chord-length between two points on the circle to classify unidimensional data. In more recent work, SenGupta and Ugwuowo [15] developed a multidimensional method for binary classification using directional data; they studied data on torus (two directional variables) and cylinder (one linear variable and one directional variable). Their approach has the limitation that it assumes as known the probabilities of misclassification [15].

Kirby and Miranda [16] proposed a variation on the classic feed-forward neural network by including the notion of a circular node, able to store and transmit angular information. In fact, their node is an abstraction for the combination of a pair of coupled nodes, whose combined values are constrained to lie on the unit circle. However, their solution is not invariant to the same inputs at different periods, namely, a pair of coupled nodes may return different responses to the same angular input. Furthermore, their model requires to manually define the hybrid architecture.

Finally, adaptions to generative models were studied in the past. First, Zemel et al. [13] extended the Boltzmann machine to consider cyclic units. On the other hand, López et al. proposed a directional naïve Bayes formulation [1,12]. Their contribution involves using the von Mises and von Mises–Fisher distributions for the directional variables instead of the classic Gaussian distribution. The effectiveness of this method relies on the independence assumption of the features and the adequacy of the von Mises distribution to model the behavior of the directional features.

In this work, we propose a Directional Logistic Regression, the discriminative counterpart to the Naïve Bayes model, which does not make assumptions on the distribution of the input data.

3. Directional logistic regression

Generative classifiers aim to model the joint probability $p(x, y)$, where x and y respectively denote the input and output variables. Traditional generative models would then make their predictions by choosing the label y that maximizes $p(x, y)$, computed using Bayes rules [17]. Instead, discriminative classifiers model the posterior probability $p(y|x)$. This computation is done in a direct manner or by learning a map from inputs x to the class labels [17].

As we have shown in Section 2, previous attempts to design classifiers for periodic data adopted a generative approach based on the von Mises distribution or variants [1]. Since state of the art approaches are based on non-generative methods for non-periodic variables [18], in this work we propose a discriminant approach to classify directional data. Our contribution stands as a directional-aware version of the Logistic Regression [19], which is the discriminant counterpart of the naïve Bayes classifier, previously used to address this problem. This relation is known as a Generative-Discriminative pair [17].

Eq. (3) defines the Directional Logistic Regression (dLR) model. This model can be understood as a Logistic Regression with a mapping from the original angular space to a linear one. As we show in Section 5, this mapping is learned simultaneously with the feature coefficients. Hereinafter, the two possible labels belong to $\{0, 1\}$, and n is the number of features:

$$f(\theta) = \frac{1}{1+e^{-k \cdot h(\theta)}}$$

$$h(\theta) = \omega_0 + \sum_{i=1}^{n} \omega_i g_i(\theta_i)$$

$$g_i(\theta_i) = \begin{cases} \sin(2\pi(\theta_i + \varphi_i)), & \text{if } dir(i) \\ \theta_i, & \text{otherwise.} \end{cases} \tag{3}$$

This model is a hybrid approach to Logistic Regression for modeling linear and directional data, whereby a mapping from angular variables to linear space is learned. The number of parameters involved in the proposed model is

$$n + 1 + (\#i \in \mathbb{N}^+ \mid i \le n : dir(i))$$

If all the variables are linear, the model is reduced to the traditional Logistic Regression with $n+1$ parameters. Also, we have included an extra k parameter that defines the slope of the sigmoid function, which does not change the predicted label but softens the decision boundary. Given the properties of the sine function, the model holds the directional condition.

4. Expressiveness of the model

In this section we analyze the model's expressiveness by studying the induced boundaries, as was done by López et al. [1] for the von Mises naïve Bayes model. We start with the scenario where the feature space is constrained to one directional feature (Section 4.1). Section 4.2 presents the most general scenario with an unconstrained number of directional and linear features. As previously mentioned, when all variables are linear, the model becomes a classical Logistic Regression and the subsequent decision surface is a hyperplane in the $\mathbb{R}^n$ space. Therefore, we are interested in settings where at least one variable is directional.

4.1. One-dimensional feature space with one angular variable

In this section we show the expressiveness of the Directional Logistic Regression (dLR) for the trivial case of unidimensional problems with a single angular variable. As we show below, it is easier to reason about the expressiveness of the model in the equivalent space where each variable is transformed into a pair of coordinates in a $(0,0)$-centered unit 2-sphere, where $x_i = \cos(2\pi\theta_i)$ and $y_i = \sin(2\pi\theta_i)$. This space will hereafter be referred to as the *transformed space* or *extended space*, while the original data representation will be denoted as the *original space*.

Without loss of generality, we assume that the model classifies an instance as positive if its outcome is larger than 0.5, thus leaving the final decision to the sign of the h function.

Theorem 1. *The dLR classifier with one predictive directional variable induces a separation boundary equivalent to a two dimensional line in the transformed space. Moreover, the set of induced decision lines is complete in the space of two dimensional lines.*

Proof.

$$\begin{aligned}
& h(\theta) = 0 \\
\equiv\ & \langle \text{Definition of } h \rangle \\
& \omega_0 + \omega_1 \sin\left(2\pi(\theta_1 + \varphi_1)\right) = 0 \\
\equiv\ & \langle \text{Sum of two angles} \rangle \\
& \omega_0 + \omega_1 \left(\sin(2\pi\theta_1)\cos(2\pi\varphi_1) + \cos(2\pi\theta_1)\sin(2\pi\varphi_1) \right) = 0 \\
\equiv\ & \langle x_1 = \cos(2\pi\theta_1), \quad y_1 = \sin(2\pi\theta_1) \rangle \\
& \omega_0 + \omega_1 \left(y_1 \cos(2\pi\varphi_1) + x_1 \sin(2\pi\varphi_1) \right) = 0 \\
\equiv\ & \langle \text{Arithmetic} \rangle \\
& \omega_1 \cos(2\pi\varphi_1) y_1 = -\omega_1 \sin(2\pi\varphi_1) x_1 - \omega_0 \\
\equiv\ & \langle \text{Arithmetic} \rangle \\
& y_1 = -\tan(2\pi\varphi_1) x_1 - \frac{\omega_0}{\omega_1 \cos(2\pi\varphi_1)}
\end{aligned}$$

Then, given that the range of the tangent function is $\mathbb{R}$, the decision boundary can be rewritten as the two dimensional line equation $y = mx + b$, with any possible slope $m = -\tan(2\pi\varphi_1)$ and y-intercept $b = \frac{\omega_0}{\omega_1 \cos(2\pi\varphi_1)}$. □

Theorem 1 shows that the expressiveness of the dLR for unidimensional problems with one predictive directional variable in the transformed space is defined by the entire set of two dimensional lines. However, the decision boundary in the original space is not linear; it is translated as two decision angular-thresholds, ϕ and ϕ', such that, if the angular distance between them is Δ, one of the possible induced models in the original space is represented by the parameter configuration:

$$\varphi_1 = \pm \frac{1}{2\pi} \arcsin\left(\sqrt{\frac{1 + \cos(2\pi\Delta)}{2}} \right) - \phi$$

$$\omega_0 = \mp \sqrt{\frac{(1 + \cos(2\pi\Delta))}{2}}$$

$$\omega_1 = 1$$

where ω_0 takes the positive version of the equation if the distance between both thresholds is greater than half of the period (ϕ_1 the negative side) and vice versa. Notice that there is an infinite number of models with the same decision boundary, since we can scale ω by any non-zero factor and obtain the same predictions.

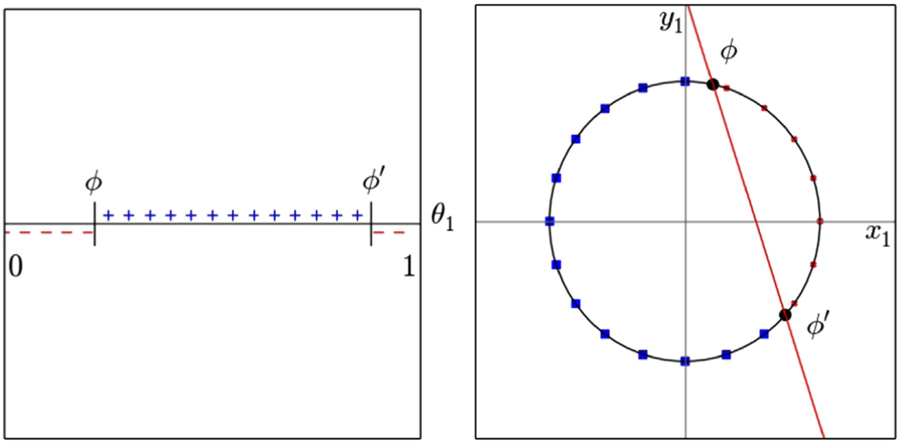

Fig. 1. Decision boundary for a problem with one directional variable. *Left:* decision boundary in the original space represented by two decision thresholds. *Right:* decision boundary in the extended space represented by the 2-dimensional line.

This property is also true for the standard logistic regression. An example of the model expressiveness for this trivial case is illustrated in Fig. 1.

4.2. N-dimensional feature space with K angular variables

We now analyze the general scenario where the feature space has an unrestricted number of directional and non-directional variables. For the sake of simplicity, we assume that the first K features are directional and the remaining linear (referred to as hypothesis H_0 in the proof of the Theorem 2). This assumption does not suppose a loss of generality given that the model is invariant to the arrangement of the features. As we did before, we analyze the expressiveness of the model in the transformed space.

Theorem 2. *The dLR classifier with N predictive variables, being K $\leq N$ of them directional, induces a separation boundary equivalent to a (N+K)-dimensional hyperplane in the transformed space.*

Proof.

$$\omega_0 + \sum_{i=1}^{N} \omega_i g_i(\theta_i) = 0$$

$$\equiv \langle \text{Range Split} \rangle$$

$$\omega_0 + \sum_{i=1}^{K} \omega_i g_i(\theta_i) + \sum_{i=K+1}^{N} \omega_i g_i(\theta_i) = 0$$

$$\equiv \langle H_0,\ \text{Definition of } g \rangle$$

$$\omega_0 + \sum_{i=1}^{K} \omega_i \sin(2\pi(\theta_i + \varphi_i)) + \sum_{i=K+1}^{N} \omega_i \theta_i = 0$$

$$\equiv \langle \text{Sum of two angles},\ \text{Arithmetic} \rangle$$

$$\omega_0 + \sum_{i=K+1}^{N} \omega_i \theta_i + \sum_{i=1}^{K} \omega_i \left(\sin(2\pi\theta_i)\cos(2\pi\varphi_i) + \cos(2\pi\theta_i)\sin(2\pi\varphi_i) \right) = 0$$

$$\equiv \langle x_i = \cos(2\pi\theta_i), y_i = \sin(2\pi\theta_i),\ \text{Arithmetic} \rangle$$

$$\omega_0 + \sum_{i=K+1}^{N} \omega_i \theta_i + \sum_{i=1}^{K} \left(\omega_i \sin(2\pi\varphi_i) x_i + \omega_i \cos(2\pi\varphi_i) y_i \right) = 0 \qquad \square$$

An interesting and usual two dimensional scenario arises when angular measurements are accompanied by a scale factor or magnitude (e.g. wind direction and speed, forces, etc.), thereby inducing a cylinder as the geometric space where input vectors lie. Fig. 2 shows an example of the decision region in both, the original $\mathbb{R}^2$ space and the transformed $\mathbb{R}^3$ space, where one variable is directional.

5. Optimization strategy

For the purpose of this work, the traditional gradient descent learning strategy from the Logistic Regression was adapted to the proposed directional version. Let us assume we have a set of labelled input data S, where each instance $\langle \theta, y \rangle \in S \subseteq \mathbb{R}^n \times \{0, 1\}$, is a pair of an input vector θ and its corresponding label y.

From this scenario, we consider the traditional regularized Logistic loss function (Log loss) used in (multinomial) Logistic Regression (cf. Eq. (4)):

$$J(\omega, \varphi) = -\frac{1}{|S|} \sum_{\langle \theta, y \rangle \in S} cost(y, \theta) + \frac{\lambda}{2n} \sum_{i=1}^{n} \omega_i^2 \tag{4}$$

$$cost(y, \theta) = y \log(f(\theta)) + (1 - y) \log(1 - f(\theta)) \tag{5}$$

This function can be enhanced in order to include different misclassification costs by considering the weighted sum of the errors. In order to fit the model, the goal of our optimization task is to find the best parameter configuration ω, φ such that:

$$\arg\min_{\omega, \varphi} J(\omega, \varphi)$$

Using a gradient descent strategy requires the computation of the partial derivatives of the goal function J with respect to each model parameter. The corresponding derivatives are shown below in Eqs. (6a)–(6d). A more detailed explanation about the deduction steps involved in the computation of these derivatives is presented in Appendix B:

$$\frac{\partial}{\partial \omega_0} J(\omega, \varphi) = \frac{k}{|S|} \sum_{\langle \theta, y \rangle \in S} (f(\theta) - y) \tag{6a}$$

$$\frac{\partial}{\partial \omega_{i>0}} J(\omega, \varphi) = \frac{k}{|S|} \sum_{\langle \theta, y \rangle \in S} (f(\theta) - y) \cdot g_i(\theta_i) + \frac{\lambda}{n} \omega_i \tag{6b}$$

$$\frac{\partial}{\partial \varphi_i} J(\omega, \varphi) = \frac{k \cdot \omega_i}{|S|} \sum_{\langle \theta, y \rangle \in S} (f(\theta) - y) \frac{\partial}{\partial \varphi_i} g_i(\theta_i) \tag{6c}$$

$$\frac{\partial}{\partial \varphi_i} g_i(\theta_i) = \begin{cases} 2\pi \cos(2\pi(\theta_i + \varphi_i)), & \text{if } dir(i) \\ 0, & \text{otherwise} \end{cases} \tag{6d}$$

Then, we can use a gradient-based optimization strategy to fit the model. In our case, we have used a Gradient Descent variation

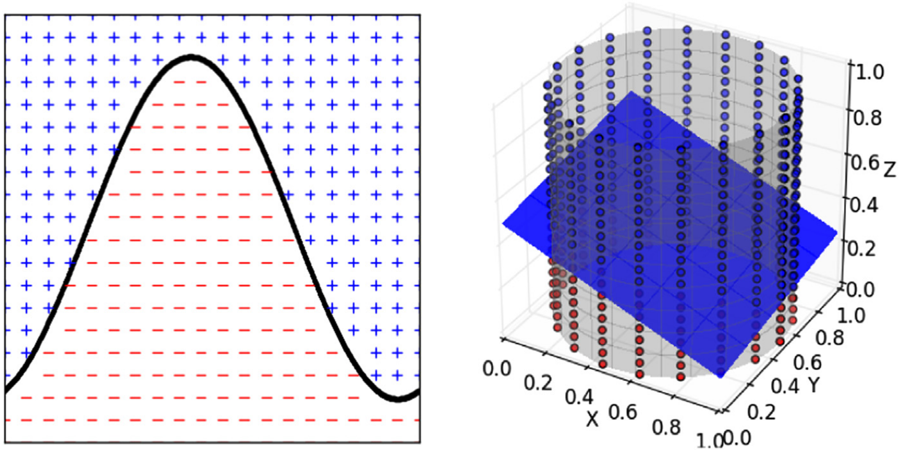

Fig. 2. Decision boundary for a mixed problem in $\mathbb{R}^2$. *Left:* non-linear decision boundary in the original space. *Right:* decision boundary in the extended space represented by a three dimensional plane.

with decaying learning rate and increasing slope of the sigmoid function to boost the algorithm's convergence (see Algorithm 1). In order to avoid having to change the learning rate as the slope of the sigmoid function changes, we removed the constant k from the derivatives, which preserves the direction of the gradient but simplifies parameter tuning. In gradient descent optimization techniques, monotonously decreasing the learning rate towards zero guarantees the convergence of the iterative process. In our setting, given that the search space is not convex, the process may converge to a local minimum. However, as will be shown in the experimental evaluation, the proposed algorithm is able to reach competitive results.

Algorithm 1. Gradient descent with variable sigmoid's slope.

```
function GRADIENT DESCENT (samples, labels)
    $\omega, \varphi \leftarrow$ initialize_model()
    $\omega^*, \varphi^* \leftarrow \omega, \varphi$
    $J^* \leftarrow J(\omega, \varphi)$
    $k, \varepsilon \leftarrow 1, \varepsilon_{init}$

    for $i \leftarrow 1$ to max_iterations do
        $\omega, \varphi \leftarrow \omega - \alpha \cdot \frac{\partial}{\partial \omega} J(\omega, \varphi),\ \varphi - \alpha \cdot \frac{\partial}{\partial \varphi} J(\omega, \varphi)$
        $J_{next} \leftarrow J(\omega, \varphi)$
        if $k < k_{max} \wedge |J_{next} - J^*| < \varepsilon$ then
            $k, \varepsilon \leftarrow k+1, \varepsilon \cdot \varepsilon_{\Delta}$
        end if
        if $J_{next} < J^*$ then
            $\omega^*, \varphi^* \leftarrow \omega, \varphi$
            $J^* \leftarrow J_{next}$
        else
            $\alpha \leftarrow \alpha \cdot$ decaying_rate
        end if
    end for

    return $\omega^*, \varphi^*$
end function
```

6. Experiments

In this section we detail the experimental evaluation of the proposed directional Logistic Regression (dLR) classifier and its non-directional version Logistic Regression (LR) against their generative counterparts von Mises naïve Bayes and Gaussian naïve Bayes classifiers [1]. These methods can be summarized as follows:

1. *GNB*: Gaussian NB classifier that models continuous variables using Gaussian distributions.
2. *vMNB*: NB classifier that models linear variables using Gaussian distributions and directional variables using von Mises distributions.

Furthermore, López et al. [1] validated a feature selection strategy proposed by Langley and Sage [20] as a wrapper of their NB approach. Also, they evaluated the performance of the NB classifier by discretizing all the continuous variables. However, given that the goal of this section is to validate the performance of the proposed discriminative method against its generative counterpart, we considered the plain GNB and vMNB methods. The study of feature selection and discretization strategies are out of the scope of this work and might improve the results shown below. In the following experiments, the κ parameter of the von Mises distribution was approximated by 100 iterations (a much larger number of iterations than required to have good convergence values) of Newton's method proposed by Sra [21].

Also, we compare our model with a Support Vector Machine (SVM) [22] using a directional squared exponential (i.e. Gaussian Radial Basis Function – RBF) kernel [23]. This kernel considers the distance between a given pair of points, wherein the distance between two directional variables is considered in an angular manner instead of the traditional Euclidean distance. The regularization parameter (C) and the γ parameter of the squared exponential kernel were chosen by cross-validation among seven different values in the logarithmic scale between 10^{-2} and 10^2.

On the other hand, both Logistic Regression variants had an initial learning rate value (α) of 0.1 and a maximum number of 20,000 iterations for the synthetic data and 10,000 iterations for real data, but most datasets required much less iterations to converge. The model was initialized using small random values ($\omega_i \in [-0.05, 0.05]$ and $\varphi_i \in [-0.05, 0.05]$). The regularization constant $C = \lambda^{-1}$ was chosen following the same strategy used in the training of the SVM.

6.1. Experiments with synthetic data

We evaluated the performance of the classifiers using one directional predictive variable and two possible responses (binary classification), under three different statistical distributions (e.g. uniform, triangular and von Mises). Then, for each possible distribution we randomly generated 75 synthetic binary datasets with 100 samples (50 samples per class). Afterwards, we validated the accuracy of each model using a training and test validation assessment using the classic 70–30 partition. We compared the two aforementioned naïve Bayes versions with the two versions of the Logistic Regression. Also, we assessed the proposed strategy by comparing the results with a brute force search that compares each possible pair of thresholds (by maximizing the margin between two observations belonging to different classes) and minimizes the training error (g-dLR), which represents the best value that could be achieved by optimizing the model according to its training classification error. It should be clear that the brute force optimization is not an option in practice when several features are used.

Table 1 summarizes the accuracy results for these experiments. In general, the Grid Search strategy obtained the best results for each possible distribution. As expected, the dLR classifier trained with the gradient descent algorithm outperforms both generative models for all the distribution but the von Mises distribution. Furthermore, the difference between the gradient-based and the grid strategy suggests that there is still room for improving the optimization stage, although the optimization is doing a good job. The worst results were obtained by the Logistic Regression as it

Table 1
Average classification error per model with unidimensional synthetic datasets.

Distr.	GNB	vMNB	LR	dLR	g-dLR
Uniform	91.25 ± 4.85	92.56 ± 6.15	82.99 ± 7.63	$\mathbf{93.44} \pm 3.02$	*96.07 ± 2.00*
Triangular	93.56 ± 4.14	94.82 ± 2.62	86.99 ± 8.53	$\mathbf{95.34} \pm 2.43$	*96.78 ± 1.89*
von Mises	95.25 ± 2.56	$\mathbf{96.25} \pm 2.42$	87.34 ± 9.46	95.56 ± 2.52	*96.47 ± 2.25*

does not have enough expressiveness to discriminate these directional datasets.

6.2. *Experiments with real data*

Then, we validated the advantages of the proposed approach using thirteen real datasets. For this purpose, we compared the two naïve Bayes variations and the SVM with directional RBF kernel against the classic Logistic Regression and the directional version proposed in this work. For computational reasons we only validated the gradient-based optimization strategy, given that the Grid-Search approach, used in the previous experiments, would be computationally intractable. Table 2 summarizes the dimensionality of the evaluated datasets (e.g. number of variables, class values and instances). Also, Appendix A details some aspects regarding the data acquisition and preprocessing that was carried in order to turn these datasets feasible for these models.

Multiclass instances were handled using a one-versus-one approach for both versions of the Logistic Regression. All the experiments detailed below were executed with a stratified 5-fold cross-validation technique (by preserving the percentage of samples for each class) and results of 40 different runs were averaged. Results of these experiments are summarized in Table 3, exhibiting average accuracy and standard deviation for 40 independent runs. The best model for each dataset is represented bold.

When comparing generative models, we obtained similar results to those obtained by López et al., namely vMNB achieves similar or better results than the GNB in most datasets [1]. The directional version of the Logistic Regression classifier reports a broad and significant advantage when compared with the non-directional approach, achieving up to 22% more percentage points in the eBay dataset than the traditional LR.

Table 2
Summary of the main characteristics of the datasets used in this work. Including number of features per type (i.e. Directional – Dir, Linear – Lin, Discrete – Disc) and number of samples per dataset (#).

Dataset	Number of variables			Class	#
	Dir	Lin	Disc	values	
Colposcopy	3	6	0	3	150
Behavior	140	426	20	4	261
Arrhythmia	4	191	66	2	430
eBay	1	2	0	11	528
Megaspores	1	0	0	2	960
Characters	5	31	0	10	1000
OnlineNews	1	12	0	2	1000
Continents	2	0	0	5	3481
Wall	6	6	0	4	5456
Temperature0	1	1	1	3	8764
Temperature1	2	1	0	3	8764
Temperature2	5	1	0	3	8764
MAGIC	1	10	0	2	19,020

In general, the best results in the entire set of problems were achieved either by the dLR (6 datasets) or by the SVM model (7 datasets). As can be seen in Table 3, dLR obtained better results than the SVM model mainly in the datasets with fewer instances. Given that the RBF kernel can be understood as a projection on a feature space with infinite number of dimensions, the SVM model can generate highly nonlinear decision regions in contrast with the $N+K$-hyperplanes generated by dLR. Thereby, dLR offers a much more succinct representation to reason about directional data without compromising accuracy. Also, in some contexts it is preferred to use simpler (linear) models, specially when computational resources are limited or when there are interpretability requirements.

Moreover, dLR achieved better results than its non-directional and generative counterparts in almost all datasets, being only surpassed in the Temperature0 dataset. Furthermore, when combining the best descriptors for the basic Temperature dataset, considering the season as a nominal value encoded as an integer for the vMNB and, as a directional variable for the dLR, the dLR classifier achieves the best performance. On average, the proposed model achieved accuracy values 8.15% higher than the vMNB.

The main disadvantage of the proposed model, when compared with its generative counterpart, is the computational time required in the training stage. While naïve Bayes approaches require basic fitting of statistical distributions, dLR is learned by means of an iterative procedure, with asymptotic complexity $\mathcal{O}(I \times |S| \times N)$, where I is the maximum number of iterations, $|S|$ is the number of samples in the training set and N is the number of features. However, once trained, dLR is computationally competitive as it has linear complexity on the number of features–$\mathcal{O}(N)$.

7. Conclusions

Different concepts in real life applications are represented by directional variables. These concepts are not restricted to the scientific domain, but can be easily found in daily routines, such as representing of time in a periodic repetitive calendar (e.g. hour, day of the week, month, etc.). Traditional classifiers, which are unaware of the angular nature of these variables, might not properly model the data. Thereby, some directional classifiers have been proposed in the past, most of them using generative approaches [1,13] and the directional von Mises distribution [10,1].

Table 3
Average accuracy per model using 5-fold cross-validation.

Dataset	GNB	vMNB	LR	dLR	SVM
Colposcopy	74.71 ± 7.08	70.93 ± 7.83	73.66 ± 6.97	**80.61** ± 6.49	80.39 ± 7.51
Behavior	47.21 ± 9.43	49.26 ± 9.20	82.46 ± 3.48	**82.68** ± 3.56	82.63 ± 3.71
Arrhythmia	67.06 ± 4.03	67.05 ± 4.07	78.31 ± 3.99	78.38 ± 4.04	**78.66** ± 3.87
eBay	77.45 ± 3.37	83.88 ± 3.75	62.33 ± 4.42	**84.86** ± 3.21	84.11 ± 3.21
Megaspores	76.72 ± 2.54	76.61 ± 2.71	62.50 ± 0.00	**76.78** ± 2.58	76.32 ± 2.72
Characters	70.94 ± 2.62	73.40 ± 2.99	94.99 ± 1.59	**95.77** ± 1.35	95.75 ± 1.27
OnlineNews	55.37 ± 2.12	55.29 ± 2.03	56.25 ± 2.94	**56.26** ± 2.95	52.80 ± 0.10
Continents	94.66 ± 0.72	94.90 ± 1.08	94.79 ± 0.74	95.87 ± 0.72	**97.72** ± 0.48
Wall	45.69 ± 2.01	51.07 ± 2.79	58.06 ± 1.39	66.53 ± 1.29	**86.41** ± 0.94
Temperature0	68.56 ± 0.83	69.99 ± 1.80	59.15 ± 0.78	56.14 ± 0.92	**72.76** ± 0.82
Temperature1	64.44 ± 0.83	65.04 ± 1.49	59.15 ± 0.90	70.28 ± 0.89	**71.21** ± 0.91
Temperature2	12.84 ± 0.09	67.70 ± 1.66	59.65 ± 0.70	79.21 ± 0.87	**82.28** ± 0.79
MAGIC	72.68 ± 0.53	73.01 ± 0.52	79.08 ± 0.50	80.77 ± 0.49	**87.35** ± 0.46

In this work, we proposed a discriminative binary classifier that is able to receive mixed data (directional and linear). This classifier adds to the classic Logistic Regression (LR) awareness about the angular nature of the data. As we demonstrated in the experimental assessment of the proposed model with both synthetic and real data, it can achieve competitive results when compared against traditional non-directional LR, against previous generative approaches and against Support Vector Machines using a directional Radial Basis Function kernel. Other advantages of the dLR model accruing from retaining the access to the posterior probabilities include risk minimization, reject option, compensating for class priors, combining models, etc. Non-probabilistic methods, like the SVM, need to involve an intermediate step where a map from the decision regions to the actual probability is estimated [3].

Therefore, the directional Logistic Regression (dLR) classifier offers promising results when dealing with directional data, and there is room for future improvement. For instance, in the near future we plan to introduce the concept of functional margin from Support Vector Machines into this model. This analysis is relevant since the decision region obtained in the extended space does not correspond to the right margin that maximizes the distance between the support vectors and the decision boundary in the original space. Fig. 3 shows the different regions obtained by maximizing the margin in the extended space (dashed line) and in the original space (solid line). For visualization purposes, the decision boundary computed in the original space is transformed into its equivalent line in the extended space. This artifact is a result of comparing the distance between the support vector and the decision boundary in the two dimensional Euclidean space instead of using the angular distance between the support vector and the boundary-sphere intersection points.

Also, we plan to adapt our dLR model to be intended as a directional perceptron within an Artificial Neural Network. The opportunity of studying the effect of dynamic frequency regimes in this classifier is an open problem, as the dLR classifier was defined in a way that it is able to encode only a single period of the directional variables. Finally, there is room for exploring more advanced optimization techniques that may improve the performance of this model.

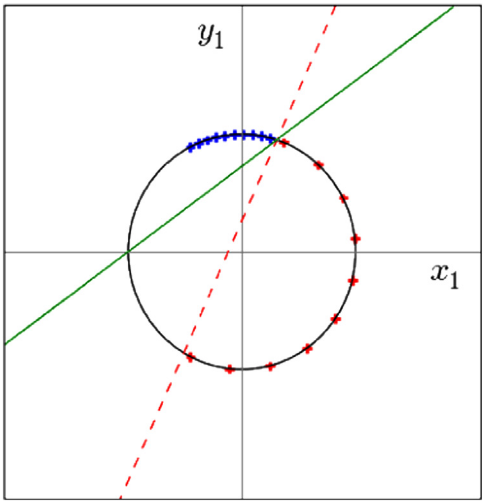

Fig. 3. Decision boundary for a linear SVM in the extended space (dashed line) and in the original space (solid line).

Acknowledgments

This work is co-financed by: NORTE-01-0145-FEDER-000016 of the North Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, and by National Funds through the FCT - Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) within project SFRH/BD/93012/2013.

Appendix A. Datasets

This section describes the datasets used in this work. Although all these datasets are publicly available, some of them required preprocessing. Datasets *Arrhythmia*, *Characters*, *OnlineNews*, *MAGIC* and *Wall* can be found in the Machine Learning repository UCI (https://archive.ics.uci.edu/ml/) [24]. The following list is ordered by the number of samples in the data.

A.1. Colposcopy

Digital colposcopy is a widely used technology to detect cervical intraepithelial neoplasia. This dataset explores the classification of colposcopic images according to their acquisition modality (i.e. Hinselmann, Green and Schiller) [25]. From each image, the average value of each channel from the HSV color space was extracted (μ_H, μ_S, μ_V). Also, we included the points at one standard deviation from the mean $(\mu_H \pm \sigma_H, \mu_S \pm \sigma_S, \mu_V \pm \sigma_V)$. The HSV space is represented by a cylinder, wherein the Hue channel has a directional nature.

A.1.1. Behavior

Human Behavior Analysis is explored in this dataset from a group perspective. Pereira et al. [26,27] proposed a set of features and an encoding to describe group trajectories. These features include directional values (e.g. orientation, gaze, etc.), which represent individual information and relational information between the individuals and their group to describe four behaviors (e.g. equally interested, unbalance interest, balance interest and chatting). As directional information is masked by using a bag-of-words method over the trajectories and encoding the final features as a frequency histogram, we preprocessed this dataset by encoding each trajectory by its most representative 5 words and their frequency, thereby including the original directional information. Also, we included the number of words with non-zero frequency as a feature.

A.2. Arrhythmia

This dataset focuses on the presence (and its type) or absence of cardiac arrhythmia from electrocardiograms [28]. As done by López et al. [1], we transformed the problem into a binary classification task (e.g. presence, absence), removed unclassified samples and removed the variable 14 ($>83\%$ missing values). The remaining missing values were filled using the median. López et al. also removed some variables that they define as non-informative (87 variables). Given that the criteria used to determine these variables was unspecified, we decided to keep the dataset unchanged.

A.3. eBay

This dataset was collected by van de Weijer et al. [29]. The main purpose of this dataset is to learn the color name of a given real-world object. Objects are represented by an image from the eBay auction site (www.ebay.com) and are labelled according to their main color (11 colors considered). We extracted the average Hue, Saturation and Value (HSV color space) from the hand-segmented image as the image descriptor.

A.4. Megaspores

Classification of megaspores according to their group in the biological taxonomy (binary classification task) using, as predictive variable, the angle of their wall elements [30].

A.5. Characters

This dataset results from a modification of the "*Artificial Characters*" dataset from UCI [24]. Originally, each sample in the dataset described an artificially generated capital letter of the English alphabet (i.e. A, C, D, E, F, G, H, L, P, R) by a set of segments which resemble an automatic image segmentation algorithm. Each segment is encoded by the two dimensional starting and ending points, its length and its angle. Given that the number of segments is not fixed, we clustered the segments of each sample using the k-means [31] clustering algorithm (with $k=5$). Then, each character is represented by the sequence of cluster centroids together with the frequency of the original segments assigned to each cluster. The sequence is ordered by the frequency values. Also, we included the number of original segments in the sample.

A.6. OnlineNews

This dataset comprises several features that describe online news published by the website Mashable (www.mashable.com). These features are used as predictive variables of the article popularity [32], which has been modelled as a binary variable according to its number of shares in social networks. We reduced the dataset to maintain only the 1000 first articles. Also, we consider only the features related to the date of the week when the article was published, the number of shares of previous articles with the same keywords and the number of shares of articles referenced in the content.

A.7. Continents

This dataset proposes the problem of predicting the continent where a given geographic coordinate (latitude and longitude) belongs. The dataset contains 3481 points from 178 countries using the `LatLong` service (http://www.latlong.net/).

A.8. Wall

This dataset faces the problem of Robot Navigation using the classic wall-following strategy in a clockwise fashion [33]. The robot senses the space by using 24 ultrasound sensors arranged equispaced around its "waist" [33]. In order to model this problem using directional data, we preprocessed the input vectors and extracted the three angles with minimum and maximum sensed value along their observations. The four possible decisions are: `Slight-Right-Turn`, `Sharp-Right-Turn`, `Slight-Left-Turn` and `Move-Forward`.

A.9. Temperature (0, 1 and 2)

Data was obtained by the Texas Commission on Environmental Quality website (www.tceq.state.tx.us) from a weather station located in the city of Houston in an hourly basis during 2012. The prediction task is defined as forecasting the outdoor temperature as low ($T \le 50°F$), medium ($50° < T < 70°$ and high ($T \ge 70°$). A previous version of this dataset (2010) was used by López et al. [1] in their work by considering as predictive variables the season (nominal), wind direction (angle) and wind speed. This version of the dataset is denoted in the results as *Temperature0*. Given that weather seasons are periodic, we modified this dataset (*Temperature1*) by considering the season as a directional variable. Finally, we studied the extended dataset (*Temperature2*) that also considers the month, day and hour of the acquisition.

A.10. MAGIC

Binary prediction class to distinguish images of hadronic showers initiated by primary gammas from those caused by cosmic rays in the upper atmosphere [34]. These images are modeled using ellipses. The parameters that describe these ellipses are considered as predictive variables, wherein one of them has directional nature, representing the angle of the major axis in the ellipse with the vector that connects its center with the camera center.

Appendix B. Partial derivatives

This section illustrates the calculations of the partial derivatives of the cost function involved in the gradient descent optimization method. The following deduction is similar to the one used in the calculus of the non-directional logistic regression derivatives.

In order to simplify the analysis, we compute below the derivative of the sigmoid function f with slope k,

$$\begin{aligned}
&f'(z)\\
=&\langle \text{Definition of } f,\ \text{Eq. ((3))}\rangle\\
&\left(\frac{1}{1+e^{-kh(z)}}\right)'\\
=&\langle \text{Power, exponential and chain rule}\rangle\\
&\frac{e^{-kh(z)}}{(1+e^{-kh(z)})^2}kh'(z)\\
=&\langle \text{Zero property of } +,\ \text{Additive inverse, Arithmetic}\rangle\\
&\left(\frac{1+e^{-kh(z)}}{(1+e^{-kh(z)})^2}-\frac{1}{(1+e^{-kh(z)})^2}\right)kh'(z)\\
=&\langle \text{Factorization, Arithmetic}\rangle\\
&\frac{1}{1+e^{-kh(z)}}\left(1-\frac{1}{1+e^{-kh(z)}}\right)kh'(z)\\
=&\langle \text{Definition of } f\rangle\\
&f(z)(1-f(z))kh'(z)
\end{aligned}$$

In the following analysis a dummy variable z will be introduced to denote any model parameter (i.e. ω_i, φ_i). Now, we can proceed to compute the derivative of the *cost* function (defined in Eq. (5)) as follows:

$$\begin{aligned}
&\frac{\partial}{\partial z}cost(y,\theta)\\
=&\langle \text{Definition of } cost\rangle\\
&\frac{\partial}{\partial z}(y\log(f(\theta))+(1-y)\log(1-f(\theta)))\\
=&\langle \text{Derivative of log and } f,\ \text{chain rule}\rangle\\
&y\frac{1}{f(z)}f(z)(1-f(z))k\frac{\partial}{\partial z}h(\theta)-\\
&(1-y)\frac{1}{1-f(z)}f(z)(1-f(z))k\frac{\partial}{\partial z}h(\theta)\\
=&\langle \text{Arithmetic, Factorization}\rangle\\
&(y(1-f(z))-(1-y)f(z))k\frac{\partial}{\partial z}h(\theta)
\end{aligned}$$

$$= \langle \text{Arithmetic} \rangle$$
$$(y - f(z))k\frac{\partial}{\partial z}h(\theta)$$

The partial derivatives of J (cf. Eq. (4)) with respect to each model parameter can be easily computed from this point:

$$\frac{\partial}{\partial z}J(\omega, \varphi)$$
$$= \langle \text{Definition of } J \rangle$$
$$\frac{\partial}{\partial z}\left(-\frac{1}{|S|}\sum_{\langle \theta, y\rangle \in S} cost(y, \theta) + \frac{\lambda}{2n}\sum_{i=1}^{n}\omega_i^2 \right)$$
$$= \langle \text{Derivative of } cost \rangle$$
$$\frac{k}{|S|}\sum_{\langle \theta, y\rangle \in S}(f(\theta) - y)\frac{\partial}{\partial z}h(\theta) + \frac{\partial}{\partial z}\left(\frac{\lambda}{2n}\sum_{i=1}^{n}\omega_i^2 \right)$$

Then, given that the partial derivatives of h with respect to each model parameter are defined as follows:

$$\frac{\partial}{\partial \omega_0}h(\theta) = 1 \tag{B.1}$$

$$\frac{\partial}{\partial \omega_{i>0}}h(\theta) = g_i(\theta) \tag{B.2}$$

$$\frac{\partial}{\partial \varphi_i}h(\theta) = \omega_i \frac{\partial}{\partial \varphi_i}g_i(\theta) \tag{B.3}$$

$$\frac{\partial}{\partial \varphi_i}g_i(\theta) = \begin{cases} 2\pi\cos(2\pi(\theta_i + \varphi_i)), & \text{if } dir(i) \\ 0, & \text{otherwise,} \end{cases} \tag{B.4}$$

the final derivatives are the expressions aforementioned in Eqs. (6a)–(6d).

References

[1] P.L. López-Cruz, C. Bielza, P. Larrañaga, Directional naive Bayes classifiers, Pattern Anal. Appl. (2013) 1–22.

[2] D. Devlaminck, W. Waegeman, B. Bauwens, B. Wyns, P. Santens, G. Otte, From circular ordinal regression to multilabel classification, in: Proceedings of the 2010 Workshop on Preference Learning, European Conference on Machine Learning, 2010.

[3] C.M. Bishop, et al., Pattern Recognition and Machine Learning, vol. 4, Springer, New York, 2006.

[4] A. Banerjee, I.S. Dhillon, J. Ghosh, S. Sra, Clustering on the unit hypersphere using von Mises–Fisher distributions, J. Mach. Learn. Res. (2005) 1345–1382.

[5] J.A. Mooney, P.J. Helms, I.T. Jolliffe, Fitting mixtures of von Mises distributions: a case study involving sudden infant death syndrome, Comput. Stat. Data Anal. 41 (3) (2003) 505–513.

[6] K.V. Mardia, J.T. Kent, Z. Zhang, C.C. Taylor, T. Hamelryck, Mixtures of concentrated multivariate sine distributions with applications to bioinformatics, J. Appl. Stat. 39 (11) (2012) 2475–2492.

[7] N. Fisher, A. Lee, Regression models for an angular response, Biometrics (1992) 665–677.

[8] S. Kato, K. Shimizu, G.S. Shieh, A circular–circular regression model, Stat. Sin. 18 (2) (2008) 633.

[9] H. Xu, K. Nichols, F.P. Schoenberg, Kernel regression of directional data with application to wind and wildfire data in Los Angeles county, California, Forest Sci. 57 (4) (2011) 343–352.

[10] A. Figueiredo, Discriminant analysis for the von Mises–Fisher distribution, Commun. Stat.—Simul. Comput. 38 (9) (2009) 1991–2003.

[11] A. Figueiredo, P. Gomes, Discriminant analysis based on the watson distribution defined on the hypersphere, Statistics 40 (5) (2006) 435–445.

[12] P.L. López-Cruz, C. Bielza, P. Larrañaga, The von mises naive bayes classifier for angular data, in: Advances in Artificial Intelligence, Springer, Berlin Heidelberg, 2011, pp. 145–154.

[13] R.S. Zemel, C.K. Williams, M.C. Mozer, Lending direction to neural networks, Neural Netw. 8 (4) (1995) 503–512.

[14] A. SenGupta, S. Roy, A simple classification rule for directional data, in: Advances in Ranking and Selection, Multiple Comparisons, and Reliability, Springer, 2005, pp. 81–90.

[15] A. Sengupta, F.I. Ugwuowo, A classification method for directional data with application to the human skull, Commun. Stat.—Theory Methods 40 (3) (2011) 457–466.

[16] M. Kirby, R. Miranda, Circular nodes in neural networks, Neural Comput. 8 (2) (1996) 390–402.

[17] A. Ng, M. Jordan, On discriminative vs. generative classifiers: a comparison of logistic regression and naive Bayes, Adv. Neural Inf. Process. Systems 14 (2002) 841.

[18] P.M. Long, R.A. Servedio, Discriminative learning can succeed where generative learning fails, in: Learning Theory, Springer, Berlin Heidelberg, 2006, pp. 319–334.

[19] P. McCullagh, J.A. Nelder, P. McCullagh, Generalized Linear Models, vol. 2, Chapman and Hall, London, 1989.

[20] P. Langley, S. Sage, Induction of selective Bayesian classifiers, in: Proceedings of the Tenth International Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann Publishers Inc., Burlington, Massachusetts, 1994, pp. 399–406.

[21] S. Sra, A short note on parameter approximation for von mises-fisher distributions: and a fast implementation of i s (x), Comput. Stat. 27 (1) (2012) 177–190.

[22] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (3) (1995) 273–297.

[23] B. Schölkopf, K.-K. Sung, C.J. Burges, F. Girosi, P. Niyogi, T. Poggio, V. Vapnik, Comparing support vector machines with Gaussian kernels to radial basis function classifiers, IEEE Trans. Signal Process. 45 (11) (1997) 2758–2765.

[24] M. Lichman, UCI Machine Learning Repository, 2013, ⟨http://archive.ics.uci.edu/ml⟩.

[25] K. Fernandes, J.S. Cardoso, J. Fernandes, Temporal segmentation of digital colposcopies, in: Pattern Recognition and Image Analysis, Springer, Cham, Switzerland, 2015, pp. 262–271.

[26] E.M. Pereira, L. Ciobanu, J.S. Cardoso, Context-based trajectory descriptor for human activity profiling, in: 2014 IEEE International Conference on Systems, Man and Cybernetics (SMC), IEEE, 2014, pp. 2385–2390.

[27] E.M. Pereira, L. Ciobanu, J.S. Cardoso, Social signaling descriptor for group behaviour analysis, in: Pattern Recognition and Image Analysis, Springer, Cham, Switzerland, 2015, pp. 13–22.

[28] H.A. Guvenir, S. Acar, G. Demiroz, A. Cekin, A supervised machine learning algorithm for arrhythmia analysis, in: Computers in Cardiology 1997, IEEE, 1997, pp. 433–436.

[29] J. Van De Weijer, C. Schmid, Applying color names to image description, in: IEEE International Conference on Image Processing, 2007, ICIP 2007, vol. 3, IEEE, 2007, pp. III–493.

[30] W.L. Kovach, Quantitative methods for the study of lycopod megaspore ultrastructure, Rev. Palaeobot. Palynol. 57 (3) (1989) 233–246.

[31] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the em algorithm, J. R. Stat. Soc. Ser. B (Methodol.) (1977) 1–38.

[32] K. Fernandes, P. Vinagre, P. Cortez, A proactive intelligent decision support system for predicting the popularity of online news, in: Progress in Artificial Intelligence, Springer, 2015, pp. 535–546.

[33] A.L. Freire, G.A. Barreto, M. Veloso, A.T. Varela, Short-term memory mechanisms in neural network learning of robot navigation tasks: a case study, in: Robotics Symposium (LARS), 2009 6th Latin American, IEEE, 2009, pp. 1–6.

[34] R. Bock, A. Chilingarian, M. Gaug, F. Hakl, T. Hengstebeck, M. Jiřina, J. Klaschka, E. Kotrč, P. Savicky`, S. Towers, et al., Methods for multidimensional event classification: a case study using images from a cherenkov gamma-ray telescope, Nucl. Instrum. Methods Phys. Res. Sect. A: Acceler. Spectrom. Detect. Assoc. Equip. 516 (2) (2004) 511–528.

Kelwin Fernandes received his BSc in Computer Engineering in 2012 from Universidad Simón Bolívar in Caracas, Venezuela. Currently, he is a PhD student at the University of Porto and a researcher at INESC TEC in Porto, Portugal. His main research interests include machine learning, computer vision and artificial intelligence.

Jaime S. Cardoso holds a Licenciatura (5-year degree) in Electrical and PhD in Computer Vision in 2006, all from the University of Porto. Cardoso is an associate professor with Habilitation at the Faculty of Engineering of the University of Porto (FEUP) and the leader of the ‘Information Processing and Pattern Recognition’ Area in the Centre for Telecommunications and Multimedia of INESC TEC. His research can be summed up in three major topics: computer vision, machine learning and decision support systems. Cardoso has co-authored 150+ papers, 40+ of which in international journals.