Directional naive Bayes classifiers

Pedro L. López-Cruz · Concha Bielza · Pedro Larrañaga

Abstract

Directional data are ubiquitous in science. These data have some special properties that rule out the use of classical statistics. Therefore, different distributions and statistics, such as the univariate von Mises and the multivariate von Mises–Fisher distributions, should be used to deal with this kind of information. We extend the naive Bayes classifier to the case where the conditional probability distributions of the predictive variables follow either of these distributions. We consider the simple sce-nario, where only directional predictive variables are used, and the hybrid case, where discrete, Gaussian and direc-tional distributions are mixed. The classifier decision functions and their decision surfaces are studied at length. Artificial examples are used to illustrate the behavior of the classifiers. The proposed classifiers are then evaluated over eight datasets, showing competitive performances against other naive Bayes classifiers that use Gaussian distributions or discretization to manage directional data.

1 Introduction

Directional data can be found in almost every field of science [24, 25]. Information measured as angles is commonly used to capture the direction of some phenomenon of interest, e.g., biologists study the movement of animals, meteorologists measure the direction of air currents, geologists observe the orientation of magnetic fields in rocks, etc. Modern visualization techniques manifest valuable three-dimensional information in a number of domains, e.g., neuroscientists are interested in the direction of neuronal axons and dendrites, microbiologists analyze the angles formed by protein structures and astrologists study the position and movement of celestial bodies.

Directional information can be captured using either angles measured in radians (or compass degrees), or directional vectors in an n-dimensional Euclidean space. We will use the terms "angular" and "circular" to specifically refer to the first kind of representation. We should note that there is a correspondence between the two representations by transforming the Cartesian coordinates of a point to its spherical coordinates. We will use the term "linear", as opposed to "directional" or "angular", to refer to common continuous information, e.g., wind speed measured in kilometers per hour, mass measured in kilograms, etc.

Special techniques are necessary to work with directional information due to its distinctive properties [37, 50]. For instance, given the angles 1° and 359°, the classical linear mean would be 180°, which points in exactly the opposite direction. It is clear that the mean angle should be 0°. Also, different visualization tools are necessary to convey directional information, e.g., rose diagrams are used instead of regular histograms. The periodical behavior that comes from having a directional domain makes linear statistics unsuitable for this kind of data. Directional statistics provides the theoretical background and the techniques to successfully work with this information.

Supervised classification [18] studies the problem of assigning a class label to an object based on a set of features

P. L. López-Cruz (✉) · C. Bielza · P. Larrañaga
Computational Intelligence Group, Departamento de Inteligencia Artificial, Facultad de Informática, Universidad Politécnica de Madrid, Campus de Montegancedo sn, 28660 Boadilla del Monte, Madrid, Spain
e-mail: pedro.lcruz@upm.es

that characterize the object. A classifier is a model that uses a function to assign a class to a new object based on the values of its features, which are modeled as predictive variables. Supervised learning algorithms are used to find that function by analyzing a set of training objects with a known class label. A large number of classification paradigms have been proposed in the literature. Bayesian networks [42, 57] are a kind of probabilistic graphical model. They have been used to solve a wide range of problems because they can compactly represent the problem domain, and factorization enables efficient computations that would be intractable otherwise. Bayesian classifiers apply these techniques to supervised classification.

Although directional data can be found in a lot of different domains, supervised classification problems including directional information as predictive variables have not been systematically studied by the machine learning research community. In fact, only 5 out of the 135 datasets for supervised classification available in the UCI Machine Learning Repository [28] include some variable measured in angles (see Sect. 4). To the best of our knowledge, the directional variables in those problems have been treated as linear continuous variables without taking into account the characterizing properties of the data.

In this paper, we extend the naive Bayes (NB) classifier [51] for use with directional predictive variables. We study the decision functions of naive Bayes classifiers using von Mises distributions or von Mises–Fisher distributions to model directional data. We also consider hybrid scenarios with directional, linear and discrete predictive variables. The selective naive Bayes classifier [44] is adapted to work with these hybrid domains. We evaluate the proposed methods on a set of real problems and compare them with other Bayesian classifiers that use Gaussian or discrete probability distributions for modeling the angular variables. A thorough analysis of the results is performed.

Classification problems using directional probability distributions have mainly been studied in the field of discriminant analysis. Morris and Laycock [55] studied the discriminant analysis of von Mises and Fisher distributions. Eben [19] analyzed the discriminant analysis of two von Mises distributions with unknown means and equal concentrations. Recently, discriminant analysis for von Mises–Fisher distributions was studied in [22], and misclassification probabilities for the von Mises distribution were estimated in two scenarios, i.e., considering populations with equal or different concentrations. Discriminant analysis has been studied for other directional distributions as well, e.g., Watson's, Selby's and Arnold's distributions in the sphere [20, 23]. In a related paper, SenGupta and Roy [64] proposed a classification rule based on the mean chord-length between an observation and two different populations of angular data belonging to two different class labels. More recently, SenGupta and Ugwuowo [65] proposed a likelihood ratio test based on a bootstrapping approach to classify angular and linear data. These approaches show several differences to the one studied in this paper. First, discriminant analysis focuses on the computation of misclassification probabilities. Here, we derive the decision functions of the naive Bayes classifiers and study them from a geometric point of view by analyzing the shape of the decision surfaces they induce. Second, these works only consider one predictive variable for classification. We study the decision functions for naive Bayes classifiers with two angular variables modeled with conditional (to the class) von Mises distributions. We also study naive Bayes classifiers with conditional von Mises–Fisher distributions. Additionally, we consider hybrid naive Bayes classifiers including linear, angular and discrete predictive variables at the same time. We also address the feature subset selection problem by adapting the selective naive Bayes classifier [44]. Finally, previous works only show the application of the techniques to one problem or dataset. In this paper, we perform an extensive evaluation of the proposed models on a set of real problems. This provides insights on the behavior of the directional naive Bayes classifiers and more general conclusions can be drawn.

In this work, we only consider maximum likelihood estimators of the parameters for the (conditional) von Mises and von Mises–Fisher probability densities. However, Bayesian parameter estimation for directional variables has received much interest, see e.g., [10, 32, 35, 49].

This paper is organized as follows. Section 2 reviews the two most studied directional distributions: the von Mises and the von Mises–Fisher distributions. Several extensions of the NB classifier are introduced in Sect. 3, where their behavior is studied from a theoretical point of view. Section 4 includes the evaluation of these models using eight datasets and the statistical comparisons with other classifiers. Finally, conclusions and future research lines are discussed in Sect. 5. Detailed derivations of the formulas are included in the attached Appendices 1, 2, 3, 4 for completeness.

2 Directional distributions

The most straightforward way to model directional data is to adjust linear distributions by wrapping them around the circle or the sphere. Several probability distributions have been proposed using this approach, e.g., the wrapped normal distribution [11, 60] or the wrapped Cauchy distribution [45]. However, the interest in this kind of information has led statisticians to propose special probability distributions to model directional data [50]. In this section, we

review the von Mises distribution and the von Mises–Fisher distribution.

2.1 The Univariate von Mises distribution

The univariate von Mises distribution [52] is the best known angular distribution, as it is the circular analogue of the Gaussian distribution. A circular variable Φ which follows the von Mises distribution on the unit circle is denoted by $\Phi \sim vM(\mu_\Phi, \kappa_\Phi)$ and its probability density function for a given angle ϕ is

$$f(\phi; \mu_\Phi, \kappa_\Phi) = \frac{\exp(\kappa_\Phi \cos(\phi - \mu_\Phi))}{2\pi I_0(\kappa_\Phi)}, \quad (1)$$

where μ_Φ is the mean direction angle, $\kappa_\Phi \geq 0$ is the concentration of the values around μ_Φ, and $I_v(\kappa)$ is the modified Bessel function of the first kind of order $v \in \mathbb{R}$.

The distribution of the points in the circle becomes uniform when $\kappa_\Phi = 0$, whereas high values of κ_Φ yield points tightly clustered around μ_Φ. The von Mises distribution is unimodal and symmetrical around the mean direction. The mean direction is also the mode, and the antimode is at $\mu_\Phi \pm \pi$. Figure 1 shows a random sample of 100 points from a von Mises distribution. We used the functions provided in the Circular Statistics Toolbox for Matlab [4] to sample the set of angles from the von Mises distributions.

Given a set of m values $\{\phi^{(1)}, \ldots, \phi^{(m)}\}$ randomly sampled from $\Phi \sim vM(\mu_\Phi, \kappa_\Phi)$, the maximum likelihood estimators of the parameters of the distribution are the sample mean direction

$$\widehat{\mu}_\Phi = \arctan\frac{\overline{C}}{\overline{S}},$$

where

$$\overline{C} = \frac{1}{m}\sum_{i=1}^{m} \cos\phi^{(i)} \text{ and } \overline{S} = \frac{1}{m}\sum_{i=1}^{m} \sin\phi^{(i)},$$

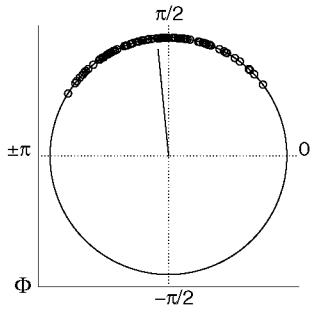


Fig. 1 Sample of 100 points from a von Mises distribution $vM(\pi/2, 5)$. The *black line* shows the sample mean direction $\widehat{\mu}_\Phi$ and its length is the mean resultant length $\overline{R}$

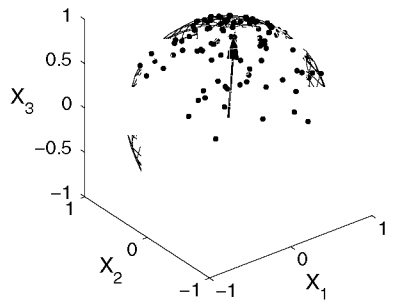


Fig. 2 Sample of 100 points from a von Mises–Fisher distribution $vMF((0,0,1)^{\mathrm{T}}, 5)$. The *black arrow* shows the sample mean direction $\widehat{\boldsymbol{\mu}}_{\mathbf{X}}$

and the concentration parameter $\widehat{\kappa}_\Phi = A^{-1}(\overline{R})$, where

$$A(\widehat{\kappa}_\Phi) = \frac{I_1(\widehat{\kappa}_\Phi)}{I_0(\widehat{\kappa}_\Phi)} = \overline{R} = \sqrt{\overline{C}^2 + \overline{S}^2}.$$

2.2 The multivariate von Mises–Fisher distribution

The unit hypersphere centered at the origin is defined by the set of n-dimensional points $\mathbb{S}^{n-1} = \{\mathbf{X} \in \mathbb{R}^n | \mathbf{X}^{\mathrm{T}}\mathbf{X} = 1\}$. A directional variable $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ which follows a multivariate von Mises–Fisher distribution on the unit hypersphere is denoted by $\mathbf{X} \sim vMF(\boldsymbol{\mu}_{\mathbf{X}}, \kappa_{\mathbf{X}})$, and its probability density function for a given unit n-dimensional vector $\mathbf{X}$ is

$$f(\mathbf{X}; \boldsymbol{\mu}_{\mathbf{X}}, \kappa_{\mathbf{X}}) = \frac{\kappa_{\mathbf{X}}^{\frac{n}{2}-1}}{\sqrt{(2\pi)^n} I_{\frac{n}{2}-1}(\kappa_{\mathbf{X}})} \exp(\kappa_{\mathbf{X}} \boldsymbol{\mu}_{\mathbf{X}}^{\mathrm{T}} \mathbf{X}), \quad (2)$$

where $\boldsymbol{\mu}_{\mathbf{X}}$ is the population mean direction vector satisfying $\boldsymbol{\mu}_{\mathbf{X}}^{\mathrm{T}}\boldsymbol{\mu}_{\mathbf{X}} = 1$ (i.e., $\|\boldsymbol{\mu}_{\mathbf{X}}\| = 1$), and $\kappa_{\mathbf{X}} \geq 0$ is the concentration parameter around $\boldsymbol{\mu}_{\mathbf{X}}$.

The von Mises–Fisher distribution reduces to the von Mises distribution when $n = 2$ and to the Fisher distribution [27] when $n = 3$. Like the von Mises distribution, the von Mises–Fisher distribution is also unimodal and symmetric around $\boldsymbol{\mu}_{\mathbf{X}}$, having the mode at $\boldsymbol{\mu}_{\mathbf{X}}$ and the antimode at $-\boldsymbol{\mu}_{\mathbf{X}}$. Figure 2 shows a set of 100 points from the distribution $vMF((0, 0, 1)^T, 5)$ defined in $\mathbb{S}^2$. To generate a sample from a von Mises–Fisher distribution, we use Jung's implementation[1] of the algorithm proposed in [69].

The maximum likelihood estimators for the parameters of the distribution $vMF(\boldsymbol{\mu}_{\mathbf{X}}, \kappa_{\mathbf{X}})$ given a sample of unit vectors $\{\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(m)}\}$ are the sample mean direction

$$\widehat{\boldsymbol{\mu}}_{\mathbf{X}} = \frac{\sum_{i=1}^{m} \mathbf{X}^{(i)}}{\left\|\sum_{i=1}^{m} \mathbf{X}^{(i)}\right\|},$$

and the concentration parameter $\widehat{\kappa}_{\mathbf{X}} = A_n^{-1}(\overline{R})$, where

[1] The source code is available at: http://www.unc.edu/sungkyu.

$$A_n(\widehat{\kappa}_{\mathbf{X}}) = \frac{I_{\frac{n}{2}}(\widehat{\kappa}_{\mathbf{X}})}{I_{\frac{n}{2}-1}(\widehat{\kappa}_{\mathbf{X}})} = \overline{R} = \frac{R}{m} = \frac{\left\|\sum_{i=1}^{m} \mathbf{X}^{(i)}\right\|}{m}.$$

Unfortunately, $\widehat{\kappa}_{\mathbf{X}}$ cannot be found analytically, and approximations have to be computed numerically [68].

3 Naive Bayes classifiers with directional predictive variables

In this section the von Mises naive Bayes (vMNB) classifier is introduced, which uses univariate von Mises distributions to model the conditional probability density functions of the angular variables. Next, the von Mises–Fisher naive Bayes (vMFNB) classifier is presented, where the conditional density functions of directional variables are modeled using multivariate von Mises–Fisher distributions. We derive the decision functions for each case and study the decision surfaces. Derivations of the decision functions and the surfaces that they induce are detailed in the Appendices 1 and 2. Hybrid scenarios with continuous and discrete predictive variables modeled using different probability distributions are a frequent occurrence in supervised classification. Therefore, we investigate the hybrid NB classifier in Sects. 3.4 and 3.5, where the predictive variables are modeled using directional distributions and discrete or Gaussian distributions. Finally, the selective naive Bayes (SelNB) classifier is adapted to work with directional distributions in Sect. 3.6.

3.1 The naive Bayes classifier

One of the simplest models for supervised classification is the *naive Bayes* [18, 51]. A NB classifier has two types of variables: the class variable C and a set of predictive variables $\mathbf{X} = \{X_1, \ldots, X_d, X_{d+1}, \ldots, X_n\}$. The class variable C is discrete and takes values in the set $\Omega(C)$. The predictive variables can be divided into two sets: the set of discrete variables $\{X_1, \ldots, X_d\}$ and the set of continuous variables $\{X_{d+1}, \ldots, X_n\}$. NB assumes that all the predictive variables are conditionally independent given the class variable:

$$p(C = c|\mathbf{X} = \mathbf{X}) \propto p(C = c) \prod_{i=1}^{d} p(X_i = x_i|C = c) \prod_{i=d+1}^{n} f_{X_i|C=c}(x_i).$$

Although conditional independence is a strong assumption, the NB classifier has shown competitive accuracies and surprisingly good results in a lot of real world problems [14]. NB uses the maximum a posteriori decision rule to assign a class value c^* to a new instance $\mathbf{X}$: $c^* = \arg\max_{c \in \Omega(C)} p(C = c|\mathbf{X} = \mathbf{X})$, and

$$\begin{aligned} &p(C=1|\mathbf{X}=\mathbf{X}) = p(C=2|\mathbf{X}=\mathbf{X}), \\ &r(\mathbf{X}) = p(C=1|\mathbf{X}=\mathbf{X}) - p(C=2|\mathbf{X}=\mathbf{X}) \\ &\quad = p(C=1)\prod_{i=1}^{d} p(X_i=x_i|C=1) \prod_{i=d+1}^{n} f_{X_i|C=1}(x_i) \\ &\quad - p(C=2)\prod_{i=1}^{d} p(X_i=x_i|C=2) \prod_{i=d+1}^{n} f_{X_i|C=2}(x_i) \end{aligned} \quad (3)$$

is the decision function.

If the class has more than two values, a decision surface is considered for each pair of values, and the subregions defined by all the surfaces are labeled accordingly. The decision surfaces of a NB classifier with binary predictive variables are hyperplanes [51]. Later on, the same result was shown for general discrete variables [58]. Duda and Hart [17] found polynomial decision surfaces when the NB has ordinal predictive variables.

Duda et al. [18] showed that the decision surface is also a hyperplane when the conditional joint probability distributions of the predictive variables is modeled with a multivariate Gaussian with class-independent covariance matrices, i.e., the covariance matrices are the same for each class value. On the other hand, the decision surfaces are hyperquadrics when the covariance matrices are different for each class value.

3.2 The von Mises naive Bayes

In this section, we derive the decision surfaces of the vMNB, where the conditional probability densities of the predictive variables are modeled using von Mises distributions. First, in Sect. 3.2.1 we study the simplest approach where one predictive variable is considered. Then, we extend our analysis to the scenario where two predictive variables are used (Sect. 3.2.2).

3.2.1 vMNB with one predictive angular variable

We start with the simplest scenario, where vMNB has a binary class and only one predictive angular variable Φ [46].

Theorem 1 *Let C be a binary class variable with values $\Omega(C) = \{1, 2\}$. Let Φ be one predictive angular variable defined in the domain $\Omega(\Phi) = (-\pi, \pi]$, with conditional probability density functions modeled as von Mises distributions $vM(\mu_{\Phi|c}, \kappa_{\Phi|c})$ for each class value $c \in \Omega(C)$. Then, vMNB finds the two following decision angles that separate the class subregions*

$$\begin{aligned} \phi' &= \alpha + \arccos(D/T), \\ \phi'' &= \alpha - \arccos(D/T), \end{aligned} \quad (4)$$

with the constants

$$D = -\ln \frac{p(C=1)I_0(\kappa_{\Phi|2})}{p(C=2)I_0(\kappa_{\Phi|1})},$$
$$\cos\alpha = a/T,$$
$$\sin\alpha = b/T,$$
$$T = \sqrt{a^2+b^2},$$
$$a = \kappa_{\Phi|1}\cos\mu_{\Phi|1} - \kappa_{\Phi|2}\cos\mu_{\Phi|2},$$
$$b = \kappa_{\Phi|2}\sin\mu_{\Phi|1} - \kappa_{\Phi|2}\sin\mu_{\Phi|2}.$$

Proof See Appendix 1, vMNB with one predictive variable.

Corollary 1 *The vMNB classifier with a binary class and one predictive angular variable Φ is a linear classifier using the decision line*

$$r(x,y) = (\kappa_{\Phi|1}\cos\mu_{\Phi|1} - \kappa_{\Phi|2}\cos\mu_{\Phi|2})x + (\kappa_{\Phi|2}\sin\mu_{\Phi|1} - \kappa_{\Phi|2}\sin\mu_{\Phi|2})y - D = 0,$$

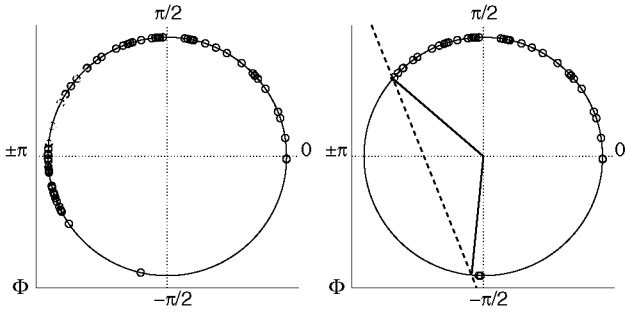
where $(x,y) = (\cos\phi, \sin\phi)$ are the Cartesian coordinates in $\mathbb{R}^2$ of the point defined by the angle ϕ on the unit circle.

Proof The proof is straightforward from Theorem 1 (see Appendix 1, vMNB with one predictive variable.).

The vMNB classifier with one predictive circular variable modeled using von Mises conditional distributions divides the circle into two regions using two angles. Also, we can see vMNB as a linear classifier finding the line that goes through the points on the circumference defined by ϕ' and ϕ''. Angle α in (4) can be interpreted as a weighted mean of the mean directions $\mu_{\Phi|c}$ for each class, using the values of the concentration parameters as weights. On the other hand, the length of the arc between the two angles (the distance that defines the size of the regions) is given by $\arccos(D/T)$, which depends on the concentrations, the mean directions and the prior probabilities of the class values. These prior probabilities are used in the logarithm in D. They influence the "size" of the class regions, moving the decision bounds so that more likely classes are given a larger subregion.

Figure 3a shows an example of a set of 100 points sampled from the conditional probability density distributions $\Phi|C=1 \sim \upsilon M(\pi/2, 2)$ and $\Phi|C=2 \sim \upsilon M(\pi, 5)$. The classes are considered equiprobable, i.e., $p(C=1) = p(C=2) = 0.5$. Figure 3b shows the class assigned to each angle by vMNB and the angles ($\phi' = 2.43$ and $\phi'' = -1.67$ rad) that define the class regions.

Particular cases To gain a thorough understanding of the classifier, we now study how these decision surfaces are defined for different values of parameters $\mu_{\Phi|c}$ and $\kappa_{\Phi|c}$. To study the decision bounds we consider that the classes are equiprobable, i.e., $p(C=1) = p(C=2) = 0.5$. This erases the influence of the prior probabilities of the class values.



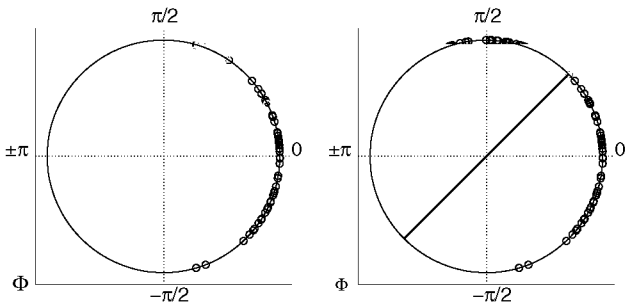
(a) True classification **(b) Predicted classification**

Fig. 3 True and predicted class for a sample of 100 angles. *Dark blue circles* represent points for class $C=1$ and *light blue circles* represent angles for class $C=2$. The *solid lines* in **b** show the angles defining the bounds of each class region. The *dashed line* is the decision line induced by vMNB (color figure online)

- Case 1: $\kappa_{\Phi|1} = \kappa_{\Phi|2}$ and $\mu_{\Phi|1} \neq \mu_{\Phi|2}$. When the two distributions share the same concentration value but have different mean directions, the decision angles are (see Appendix 1, vMNB with one predictive variable, Particular cases)

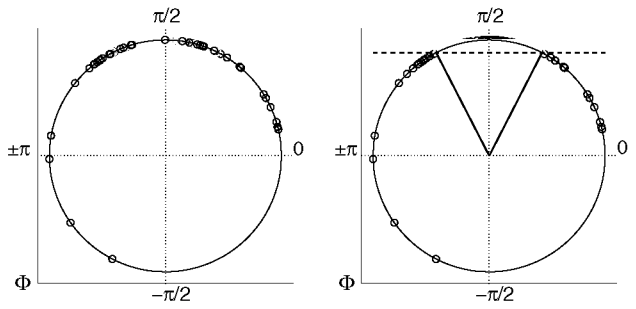
$$\phi' = \frac{1}{2}(\mu_{\Phi|1} + \mu_{\Phi|2}),$$
$$\phi'' = \frac{1}{2}(\mu_{\Phi|1} + \mu_{\Phi|2}) + \pi.$$

In this scenario, the decision surface is an axis that divides the circle into two semicircles (the angles are π rad apart). The axis goes through the center of the circle and is the bisector of the angle defined by the two mean directions. Figure 4a shows an example with a sample of 100 points drawn from the distributions $\Phi|C=1 \sim vM(0,5)$ and $\Phi|C=2 \sim vM(\pi/2,5)$. The classes are equiprobable a priori. vMNB finds an axis that forms an angle of $\pi/4$ with the horizontal axis and yields a semicircle for each class value (Fig. 4b).



(a) True classification **(b) Predicted classification**

Fig. 4 True and predicted class for a sample of 100 angles where the conditional densities share the same concentration (Case 1). *Dark blue circles* represent points for class $C=1$ and *light blue circles* represent angles for class $C=2$. The *green axis* separates each class region in **b** (color figure online)



(a) True classification **(b)** Predicted classification

Fig. 5 True and predicted class for a sample of 100 angles when the conditional densities share the same mean direction (Case 2). *Dark blue circles* represent points for class $C = 1$ and *light blue circles* represent angles for class $C = 2$. The *solid lines* in **b** show the angles defining each class region. The *dashed line* is the decision line induced by vMNB (color figure online)

- Case 2: $\kappa_{\Phi|1} \neq \kappa_{\Phi|2}$ and $\mu_{\Phi|1} = \mu_{\Phi|2} = \mu_\Phi$. vMNB finds the following angles when the mean directions are equal but the concentrations of the conditional distributions are different (see Appendix 1, vMNB with one predictive variable, Particular cases)

$$\phi' = \mu_\Phi + \arccos\frac{D}{\kappa_{\Phi|1} - \kappa_{\Phi|2}},$$
$$\phi'' = \mu_\Phi - \arccos\frac{D}{\kappa_{\Phi|1} - \kappa_{\Phi|2}}.$$

The two angles are defined according to the shared mean direction μ_Φ, and the "spread" of the arc that they form is determined by the difference in the concentration parameters. The region including the mean direction always corresponds to the class with a larger concentration. Figure 5a shows a set of 100 points sampled from the distributions $vM(\pi/2,2)$ and $vM(\pi/2,10)$. The two classes are equiprobable a priori. Figure 5b shows the classification provided by vMNB and the decision angles, which are both 0.47 rad away from the mean direction $\mu_\Phi = \pi/2$ (2.04 and 1.10 rad). The decision line is orthogonal to the mean direction μ_Φ and its position depends on the difference of the concentration values.

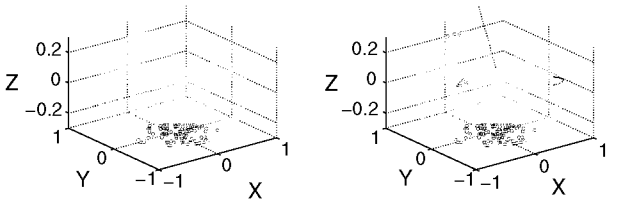
3.2.2 vMNB with two predictive angular variables

We now study the more complex scenario where two angular predictive variables Φ and Ψ are used in vMNB. The domain defined by the predictive variables is a torus $(-\pi, \pi] \times (-\pi, \pi]$.

Theorem 2 *Let C be a binary class with values in $\Omega(C) = \{1,2\}$. Let Φ and Ψ be two angular variables defined in the domain $(-\pi, \pi]$. Let the conditional probability density functions of the variables Φ and Ψ be von Mises distributions $vM(\mu_{\Phi|c}, \kappa_{\Phi|c})$ and $vM(\mu_{\Psi|c}, \kappa_{\Psi|c})$. Then, the decision surface induced by the von Mises naive Bayes classifier is given by the 2-degree multivariate polynomials*

$$\begin{aligned} &clx + dly - az^2 + bz\sqrt{l^2 - z^2} + bLz \\ &\quad + (aL + Dl)\sqrt{l^2 - z^2} + al^2 + DLl = 0, \\ &clx + dly - az^2 - bz\sqrt{l^2 - z^2} + bLz \\ &\quad - (aL + Dl)\sqrt{l^2 - z^2} + al^2 + DLl = 0, \end{aligned} \tag{5}$$

where (x, y, z) are the Cartesian coordinates in $\mathbb{R}^3$ of the points lying on the surface of the torus, and a, b, c, d, l, L and D are constants (see Appendix 1, vMNB with two predictive variables).



(a) True classification **(b)** Predicted classification

Fig. 6 True class and class predicted using vMNB for a sample of 1,000 points. Points with $C = 1$ are shown in *dark blue*, whereas points with $C = 2$ are shaded *light blue*. The decision boundaries are drawn in *green* (color figure online)

The decision surfaces in (5) are quadratic in z, so vMNB is not a linear classifier when two predictive angular variables are used. The complexity of the classifier increases from the base scenario with one predictive variable (Sect. 3.2.1). This behavior differs from the NB classifier with discrete variables, where the decision surfaces are always linear no matter the number of predictive variables.

We illustrate this scenario with an artificial example. Figure 6a shows a set of 1,000 points sampled from the distributions $\Phi|C = 1 \sim vM(\pi, 2)$ and $\Psi|C = 1 \sim vM(-2\pi/3, 6)$ (shown in dark blue) and $\Phi|C = 2 \sim vM(\pi/2, 5)$ and $\Psi|C = 2 \sim vM(\pi, 3)$ (shown in light blue), and mapped into a torus. The two classes are equiprobable a priori. Figure 6b shows the classification provided by vMNB and the complex decision bounds induced by it, where we can see the non-linear behavior of the classifier.

3.3 The von Mises–Fisher naive Bayes

The same approach can be used when the data in our problem are directional unit vectors in $\mathbb{R}^n$. These directional vectors can also be represented as points in the unit hypersphere $\mathbb{S}^{n-1} = \{\mathbf{X} \in \mathbb{R}^n | \|\mathbf{X}\| = 1\}$ and modeled using the von Mises–Fisher distribution. Then, the classifier has only one n-dimensional predictive variable $\mathbf{X}$.

Theorem 3 *Let C be a binary class variable with values $\Omega(C) = \{1, 2\}$. Let $\mathbf{X}$ be a n-dimensional variable defined in the unit hypersphere $\mathbb{S}^{n-1} = \{\mathbf{X} \in \mathbb{R}^n | \|\mathbf{X}\| = 1\}$. Let the conditional probability densities $\mathbf{X}|C = c$ follow a von Mises–Fisher distribution $vMF(\boldsymbol{\mu}_{\mathbf{X}|c}, \kappa_{\mathbf{X}|c})$. Then, the von Mises–Fisher naive Bayes is a linear classifier yielding the decision hyperplane*

$$(\kappa_{\mathbf{X}|1}\boldsymbol{\mu}_{\mathbf{X}|1} - \kappa_{\mathbf{X}|2}\boldsymbol{\mu}_{\mathbf{X}|2})^{\mathrm{T}}\mathbf{X} + \ln\frac{p(C=1)(\kappa_{\mathbf{X}|1})^{\frac{n}{2}-1} I_{\frac{n}{2}-1}(\kappa_{\mathbf{X}|2})}{p(C=2)(\kappa_{\mathbf{X}|2})^{\frac{n}{2}-1} I_{\frac{n}{2}-1}(\kappa_{\mathbf{X}|1})} = 0. \quad (6)$$

Proof See Appendix 2.

Therefore, the decision surface in (6) is a hyperplane in $\mathbb{R}^n$ that divides the space into the two regions for classification. The intersection of the hyperplane and the hypersphere is a circumference with the points that have the same posterior probability of being assigned to either class. The hyperplane can also be characterized by a non-zero normal vector and a point $\mathbf{X}_0$ belonging to the hyperplane. That characterization is easier to interpret. The hyperplane found by vMFNB is given by

$$(\kappa_{\mathbf{X}|1}\boldsymbol{\mu}_{\mathbf{X}|1} - \kappa_{\mathbf{X}|2}\boldsymbol{\mu}_{\mathbf{X}|2})^{\mathrm{T}}(\mathbf{X} - \mathbf{X}_0) = 0. \quad (7)$$

Figure 7a shows an example in $\mathbb{S}^2$ of a set with 1,000 points from $\mathbf{X}|C = 1 \sim vMF((-1, 0, -0.2)^{\mathrm{T}}, 10)$ (dark blue) and $\mathbf{X}|C = 2 \sim vMF((-0.5, -0.5, 1)^{\mathrm{T}}, 20)$ (light blue). The classes are considered equiprobable a priori. If we replace the values of the parameters in the hyperplane expression (6), we get the plane $10x_2 - 22x_3 = -9.3069$. Alternatively, if we use the equation with the normal vector and the point (7), the plane that we get has the normal vector $(0, 10, -22)^{\mathrm{T}}$ and contains the point $\mathbf{X}_0 = (0, 0, 0.4230)^{\mathrm{T}}$. Figure 7b shows the classification given by vMFNB, the decision hyperplane and the circumference that bounds the class regions.

Figueiredo [22] also derived the decision function in (6). However, as far as we know, it is the first time that the geometric interpretation of the induced decision surface is studied at length, and the following special scenarios are analyzed.

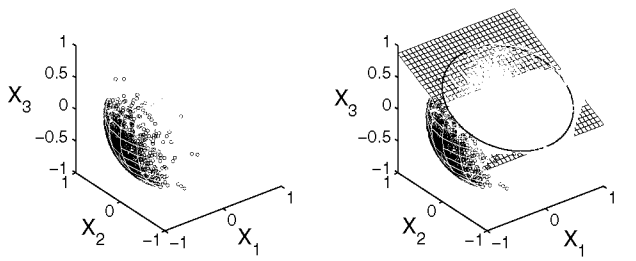
3.3.1 Particular cases

We study the shape of the decision hyperplanes for some special cases when the conditional probability distributions share the value of one parameter. Like the analysis for the vMNB (Sect. 3.2.1), the classes are assumed to be equiprobable a priori.

- Case 1: $\kappa_{\mathbf{X}|1} = \kappa_{\mathbf{X}|2}$ and $\boldsymbol{\mu}_{\mathbf{X}|1} \neq \boldsymbol{\mu}_{\mathbf{X}|2}$. When the concentration parameter values are the same but the mean directions are different, the hyperplane equation simplifies to (see Appendix 2, Particular cases)

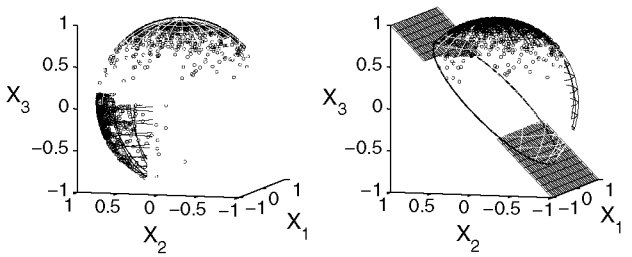
$$(\boldsymbol{\mu}_{\mathbf{X}|1} - \boldsymbol{\mu}_{\mathbf{X}|2})^{\mathrm{T}}\mathbf{X} = 0. \quad (8)$$

Equation (8) defines a hyperplane that goes through the origin (center of the sphere), dividing it into two hemispheres. The plane goes through the "middle point" of the segment that contains the points in the hypersphere corresponding to the mean directions, like the bisector in vMNB. In fact, we can write the hyperplane equation as $(\boldsymbol{\mu}_{\mathbf{X}|1} - \boldsymbol{\mu}_{\mathbf{X}|2})^{\mathrm{T}}(\mathbf{X} - 0) = 0$. Accordingly, vMFNB finds a hyperplane with the normal vector $(\boldsymbol{\mu}_{\mathbf{X}|1} - \boldsymbol{\mu}_{\mathbf{X}|2})^{\mathrm{T}}$, which is the vector connecting the points in the hypersphere defined by the two mean directions. Additionally, the hyperplane contains the origin point ($\mathbf{0}$). In this case, since the plane goes through the center of the sphere, the intersection is a great circle (a.k.a. Riemannian circle), that is, one of the circles with the same radius as the sphere. The great circle and the hypersphere share the same center. Figure 8a shows a set of 1,000 points from the distributions $\mathbf{X}|C = 1 \sim vMF((0, 0, 1)^{\mathrm{T}}, 7)$ (dark blue) and $\mathbf{X}|C = 2 \sim vMF((0, 1, 0)^{\mathrm{T}}, 7)$ (light blue). The classes have the same probability a priori. The classification provided by



(a) True classification (b) Predicted classification

Fig. 7 True class and class predicted using the von Mises–Fisher NB for a sample of 1,000 points. Class $C = 1$ points are shown in *dark blue*, whereas class $C = 2$ data are drawn in *light blue* (color figure online)



(a) True classification (b) Predicted classification

Fig. 8 True class and class predicted when the conditional densities share the same concentration (Case 1). Class $C = 1$ points are shown in *dark blue*, whereas class $C = 2$ data are represented in *light blue* (color figure online)

vMFNB can be seen in Fig. 8b, where the decision hyperplane is given by the equation $-x_2 + x_3 = 0$.

- Case 2: $\kappa_{\mathbf{X}|1} \neq \kappa_{\mathbf{X}|2}, \boldsymbol{\mu}_{\mathbf{X}|1} = \boldsymbol{\mu}_{\mathbf{X}|2} = \boldsymbol{\mu}_{\mathbf{X}}$. In the scenario where the concentration parameters have different values but the mean directions are the same, vMFNB finds the hyperplane (see Appendix 2, Particular cases)

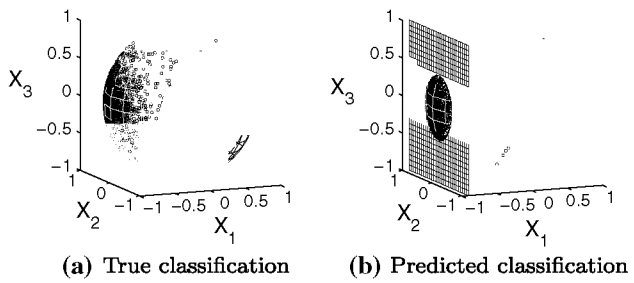
$$\boldsymbol{\mu}_{\mathbf{X}}^{\mathrm{T}}(\mathbf{X} - \mathbf{X}_0) = 0$$

$$\text{with } \mathbf{X}_0 = \boldsymbol{\mu}_{\mathbf{X}} \frac{1}{\kappa_{\mathbf{X}|1} - \kappa_{\mathbf{X}|2}} \ln \frac{(\kappa_{\mathbf{X}|1})^{\frac{n}{2}-1} I_{\frac{n}{2}-1}(\kappa_{\mathbf{X}|2})}{(\kappa_{\mathbf{X}|2})^{\frac{n}{2}-1} I_{\frac{n}{2}-1}(\kappa_{\mathbf{X}|1})}.$$

Therefore, vMFNB finds a hyperplane perpendicular to the mean direction and containing point $\mathbf{X}_0$. Point $\mathbf{X}_0$ is also located in the direction of the mean and its exact position depends on the values of the concentration parameters $\kappa_{\mathbf{X}|1}$ and $\kappa_{\mathbf{X}|2}$. Figure 9a shows 1,000 points sampled from the distributions $\mathbf{X}|C = 1 \sim vMF((-1,0,0)^{\mathrm{T}}, 20)$ (dark blue) and $\mathbf{X}|C = 2 \sim vMF((-1,0,0)^{\mathrm{T}}, 5)$ (light blue). The two class values are equiprobable a priori. If we replace the values of the parameters in the hyperplane expression, we get the plane $x_1 = -0.9076$. Alternatively, if we use the equation with the normal vector and the point, the plane that we get has the normal vector $(-1, 0, 0)^{\mathrm{T}}$ and contains the point $\mathbf{X}_0 = (-0.9076, 0, 0)^{\mathrm{T}}$. Figure 9b shows the classification given by the vMFNB classifier, the decision hyperplane and the circumference that bounds the class regions.

3.4 Hybrid Gaussian–von Mises–Fisher naive Bayes

A very interesting scenario arises when combining directional and non-directional data. This is a frequent situation when we can measure both the magnitude and the direction of a given phenomenon, e.g., the direction and the velocity of wind currents or the strength and orientation of a magnetic field. We study the hybrid NB classifier where the directional variable $\mathbf{X}$ is modeled using von Mises–Fisher distributions and the linear variable $\mathbf{Y}$ is modeled using multivariate Gaussian distributions. The conditional probability distributions of the predictive variables given the class value c are $\mathbf{X}|C = c \sim vMF(\boldsymbol{\mu}_{\mathbf{X}|c}, \kappa_{\mathbf{X}|c})$ and $\mathbf{Y}|C = c \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{Y}|c}, \boldsymbol{\Sigma}_{\mathbf{Y}|c})$.



(a) True classification (b) Predicted classification

Fig. 9 True class and class predicted when the conditional densities share the same mean direction (Case 2). *Dark blue circles* refer to class $C = 1$ points and class $C = 2$ data are drawn in *light blue* (color figure online)

The multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is defined by its two parameters: the mean $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$. The decision function $r(\mathbf{Y})$ of a Gaussian NB [59] found by substituting this probability density function in (3) is:

$$\begin{aligned} r(\mathbf{Y}) = &-\frac{1}{2}(\mathbf{Y} - \boldsymbol{\mu}_{\mathbf{Y}|1})^{\mathrm{T}}(\boldsymbol{\Sigma}_{\mathbf{Y}|1})^{-1}(\mathbf{Y} - \boldsymbol{\mu}_{\mathbf{Y}|1}) \\ &+\frac{1}{2}(\mathbf{Y} - \boldsymbol{\mu}_{\mathbf{Y}|2})^{\mathrm{T}}(\boldsymbol{\Sigma}_{\mathbf{Y}|2})^{-1}(\mathbf{Y} - \boldsymbol{\mu}_{\mathbf{Y}|2}) \\ &+\ln \frac{p(C=1)\left|\boldsymbol{\Sigma}_{\mathbf{Y}|2}\right|^{1/2}}{p(C=2)\left|\boldsymbol{\Sigma}_{\mathbf{Y}|1}\right|^{1/2}}. \end{aligned} \tag{9}$$

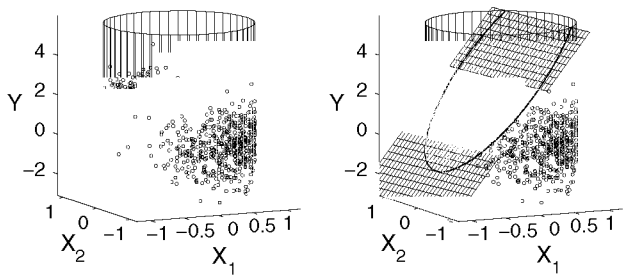
Duda et al. [19] show that the surfaces induced by that function are hyperplanes when $\boldsymbol{\Sigma}_{\mathbf{Y}|1} = \boldsymbol{\Sigma}_{\mathbf{Y}|2}$ and general hyperquadrics when $\boldsymbol{\Sigma}_{\mathbf{Y}|1} \neq \boldsymbol{\Sigma}_{\mathbf{Y}|2}$.

To compute the decision function for the hybrid Gaussian–von Mises–Fisher NB, we have to substitute the von Mises–Fisher (2) and the Gaussian probability density functions in the decision function expression (3). Assuming conditional independence between $\mathbf{X}$ and $\mathbf{Y}$ given the class C, the decision function obtained after operating is the sum of two decision functions $r(\mathbf{X}, \mathbf{Y}) = r(\mathbf{X}) + r(\mathbf{Y})$, obtained in (6) and (9), but considering the prior probabilities $p(C = c)$ in only one of the components. The shape of the surface induced by the function $r(\mathbf{X}, \mathbf{Y})$ is determined by the most complex of the two components in the sum. We have shown that the decision surfaces defined by $r(\mathbf{X})$ are hyperplanes. Therefore, if the conditional probability distributions of the linear variable $\mathbf{Y}$ have the same covariance matrices, we have that the hybrid Gaussian–von Mises–Fisher NB finds a hyperplane to bound the class regions. On the other hand, if the covariance matrices are different, the decision surface is a general hyperquadric, ranging from simple hyperplanes to complex hyperhyperboloids [18]. We use an artificial example to illustrate this behavior.

The simplest model of this hybrid NB includes one circular variable $\mathbf{X} = (X_1, X_2)$ defined in the unit circumference $\mathbb{S}^1 = \{(x_1, x_2) \in \mathbb{R}^2 | x_1^2 + x_2^2 = 1\}$ and one linear variable Y defined in $\mathbb{R}$. The domain of the problem is the Cartesian product $\mathbb{S}^1 \times \mathbb{R}$, which defines a cylinder with unit radius. In this example the variable Y is 1-dimensional, so the covariance matrix is just the variance $\boldsymbol{\Sigma}_{\mathbf{Y}|c} = \sigma^2_{\mathbf{Y}|c}, c \in \{1, 2\}$.

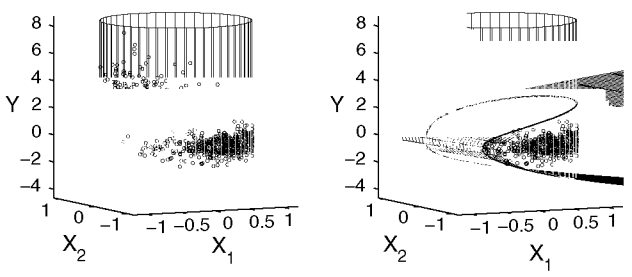
3.4.1 Particular cases

We analyzed the two cases described above for this model.



(a) True classification (b) Predicted classification

Fig. 10 True class and class predicted using the hybrid Gaussian–von Mises–Fisher NB classifier for a sample of 1,000 points when the conditional Gaussian distributions share the same variance. *Dark blue circles* refer to class $C = 1$ points, whereas class $C = 2$ data are drawn in *light blue* (color figure online)



(a) True classification (b) Predicted classification

Fig. 11 True class and class predicted using the hybrid Gaussian–von Mises–Fisher NB classifier for a sample of 1,000 points when the conditional Gaussian distributions have different variances. *Dark blue circles* refer to class $C = 1$ points, whereas class $C = 2$ data are drawn in *light blue* (color figure online)

- Case 1: $\sigma^2_{Y|1} = \sigma^2_{Y|2} = \sigma^2$. Substituting the probability density functions of the von Mises–Fisher and Gaussian distributions in the decision function (3) and arranging all the terms, we get the following expression defining a hyperplane:

$$\begin{aligned} r(x_1, x_2, y) = {} & (\kappa_{\mathbf{X}|1}\mu_{X_1|1} - \kappa_{\mathbf{X}|2}\mu_{X_1|2})x_1 \\ & + (\kappa_{\mathbf{X}|1}\mu_{X_2|1} - \kappa_{\mathbf{X}|2}\mu_{X_2|2})x_2 \\ & + \frac{\mu_{Y|1} - \mu_{Y|2}}{\sigma^2}y + \frac{\mu^2_{Y|2} - \mu^2_{Y|1}}{2\sigma^2} \\ & + \ln\frac{p(C=1)I_0(\kappa_{\mathbf{X}|2})}{p(C=2)I_0(\kappa_{\mathbf{X}|1})}. \end{aligned}$$

  Figure 10a shows the true classification for 1,000 points sampled using the distributions $\mathbf{X}|C = 1 \sim vMF((0.2, -0.8)^{\mathrm{T}}, 5)$ and $Y|C = 1 \sim \mathcal{N}(0, 1)$ for points in class 1, and the distributions $\mathbf{X}|C = 2 \sim vMF((-0.8, -0.5)^{\mathrm{T}}, 10)$ and $Y|C = 2 \sim \mathcal{N}(2, 1)$ for points with $C = 2$. Figure 10b shows the classes predicted by the hybrid Gaussian–von Mises–Fisher NB classifier and the hyperplane that separates the two class regions.
- Case 2: $\sigma^2_{Y|1} \neq \sigma^2_{Y|2}$. In this scenario, the decision function obtained by the hybrid Gaussian–von Mises–Fisher NB is given by the following expression, which is quadratic for y:

$$\begin{aligned} r(x_1, x_2, y) = {} & (\kappa_{\mathbf{X}|1}\mu_{X_1|1} - \kappa_{\mathbf{X}|2}\mu_{X_1|2})x_1 \\ & + (\kappa_{\mathbf{X}|1}\mu_{X_2|1} - \kappa_{\mathbf{X}|2}\mu_{X_2|2})x_2 \\ & + \frac{\sigma^2_{Y|1} - \sigma^2_{Y|2}}{2\sigma^2_{Y|1}\sigma^2_{Y|2}}y^2 + \frac{\sigma^2_{Y|2}\mu_{Y|1} - \sigma^2_{Y|1}\mu_{Y|2}}{\sigma^2_{Y|1}\sigma^2_{Y|2}}y \\ & - \frac{\mu^2_{Y|1}}{2\sigma^2_{Y|1}} + \frac{\mu^2_{Y|2}}{2\sigma^2_{Y|2}} \\ & + \ln\frac{p(C=1)I_0(\kappa_{\mathbf{X}|2})\sigma_{Y|2}}{p(C=2)I_0(\kappa_{\mathbf{X}|1})\sigma_{Y|1}}. \end{aligned}$$

  Figure 11a shows a sample of 1,000 points using the distributions $\mathbf{X}|C = 1 \sim vMF((0.2, -0.8)^{\mathrm{T}}, 5)$ and $Y|C = 1 \sim \mathcal{N}(0, 0.25)$ for points in class 1, and the distributions $\mathbf{X}|C = 2 \sim vMF((-0.8, -0.5)^{\mathrm{T}}, 10)$ and $Y|C = 2 \sim \mathcal{N}(2, 4)$ for the points in the class 2. The classes are considered equiprobable a priori. The classification provided by the hybrid NB and the hyperquadratic decision surface that bounds the class regions are shown in Fig. 11b.

3.5 Hybrid discrete Gaussian–von Mises–Fisher naive Bayes

Categorical data is also commonly found in different fields of science [1]. For example, binary variables can be used to indicate the presence or absence of a given trait in the phenomenon whose direction we are measuring. Discrete variables coding some qualitative aspect of the phenomenon can also be interesting for classification. Additionally, continuous variables with arbitrary distributions are usually discretized to make their analysis easier.

The NB classifiers presented above can be directly extended to the case including categorical predictive variables. Assuming that there are d discrete predictive variables $\{X_1, ..., X_d\}$, the classifier induces a set of decision surfaces, one for each possible combination of the values of the discrete variables. When analyzing a new instance $\mathbf{z}$, we first have to check the values of the discrete values to select the corresponding decision surface and use the values of the continuous variables for classification purposes.

Adding discrete predictive variables would modify the independent term of the equation that specifies the decision surface, i.e., the probabilities of the discrete conditional distributions change the position of the decision surface but not its shape. Therefore, in the case of linear classifiers (vMNB, vMFNB and hybrid NB with equal covariance matrices), the decision hyperplanes found for every

combination of values for the discrete predictors are all parallel to each other.

We use a simple artificial example of a NB classifier with two predictive variables to illustrate this point. We have a circular variable $\mathbf{X} = (X_1, X_2)$ defined in the unit circumference $\mathbb{S}^1 = \{(x_1, x_2) \in \mathbb{R}^2 | x_1^2 + x_2^2 = 1\}$ and one categorical variable Y that takes two values, e.g., $Y \in \{1,2\}$. The class variable C is binary and its values are considered equiprobable. The conditional probability distributions of the predictive variables for class $C = 1$ are: $\mathbf{X}|C = 1 \sim vMF((0.2, -0.8)^{\mathrm{T}}, 5)$ and $p(Y = 1 \mid C = 1) = 0.15$. The conditional probability distributions for points with $C = 2$ are $\mathbf{X}|C = 2 \sim vMF((-0.8, -0.5)^{\mathrm{T}}, 10)$ and $p(Y = 1 \mid C = 2) = 0.6$.

A set of 50 points are drawn from those distributions and the true and predicted classifications are shown in Fig. 12a, b, respectively. Figure 12c shows the points where $Y = 1$ and the decision line that bounds the class regions, whereas Fig. 12d shows the same information for points $Y = 2$. The decision lines are clearly parallel. Note that the above analysis is also valid including linear multivariate Gaussian distributions, although they have not been included in the artificial example for simplicity's sake.

3.6 Selective von Mises naive Bayes

Naive Bayes classifiers are affected by redundant variables [44]. Finding good predictive variables can significantly increase the accuracy of NB. Langley and Sage [44] proposed the selective naive Bayes (SelNB) algorithm. SelNB finds the variables inducing the most accurate NB structure in a wrapper fashion. Pérez et al. [59] proposed a filter-wrapper approach to induce SelNB classifiers. First, the filter algorithm ranks the predictive variables using the mutual information (MI) between each variable and the class. Then each step of the wrapper algorithm induces a new classifier including the next predictive variable in the ranking. The algorithm uses classification accuracy (computed with an inner tenfold cross-validation procedure) to evaluate the models and selects the best classifier.

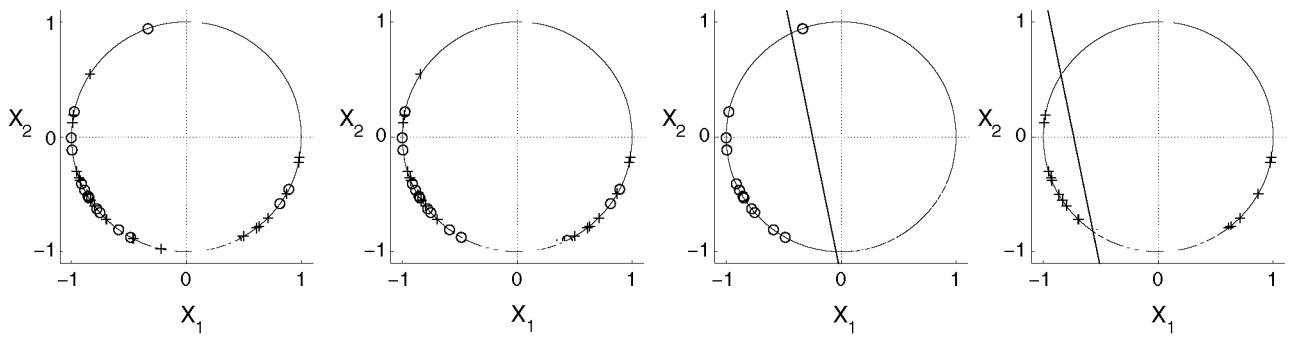
SelNB computes $\mathrm{MI}(X_i, C)$ between each predictive variable X_i and the class variable C. $\mathrm{MI}(X_i, C)$ is the reduction of the entropy of the class given that we know the value of X_i. This measure represents the information that variable X_i gives about C. Therefore, higher values of MI relate to more informative variables. Appendix 3 details the computation of $\mathrm{MI}(X_i, C)$.

The classifier learned by SelNB is a NB classifier which does not include all the predictive variables. This algorithm can discard irrelevant variables but still suffers from redundant variables. On the other hand, the wrapper algorithm proposed in [44] can discard both irrelevant and redundant variables. On the downside, however, it is less computationally efficient, since n^2 combinations of n predictive variables have to be tested in the worst-case scenario. The filter-wrapper algorithm uses a greedy heuristic to rank the variables according to the information they provide about the class. Accordingly, it has to test at most n classifiers. If the number of variables n is very large, we can limit the number of variables by setting $n_{\max} < n$ in the wrapper step, and only $n_{\max}$ subsets of variables are tested.

The complexity of the decision surfaces induced by SelNB depends on the number and the type of the variables selected in the final NB structure, as discussed in the previous sections.

4 Experimental results

This section reports the results of the experimental evaluation of the classifiers presented in this paper. Eight



(a) True classification of the points.

(b) Predicted classification using the hybrid NB classifier.

(c) Predicted classification of the points and decision line for $Y = 1$.

(d) Predicted classification of the points and decision line for $Y = 2$.

Fig. 12 True class and class predicted by the hybrid discrete von Mises–Fisher NB classifier. Class $C = 1$ data are highlighted in *dark blue*, whereas class $C = 2$ data are highlighted in *light blue*. *Circles* are used to represent data with $Y = 1$ and crosses refer to data with $Y = 2$ (color figure online)

datasets were considered for evaluation (see Appendix 4 for a detailed description). The performance of the different algorithms and the statistical comparison of the results are included in Sect. 4.1. Section 4.2 illustrates the differences between using Gaussian and von Mises distributions to model angular data.

4.1 Comparison of classifiers

In this section we evaluate the performance of vMNB against other NB classifiers which ignore the angular nature of the data. We compared the following algorithms:

- vMNB: NB classifier using Gaussian distributions for linear continuous variables and von Mises distributions for angular variables.
- SelvMNB: Selective NB classifier, where the linear variables are modeled using Gaussian distributions and the angular variables are modeled using von Mises distributions.
- GNB: Gaussian NB classifier where the probability density functions of all the continuous variables given the class values are modeled using Gaussian distributions.
- SelGNB: Selective Gaussian NB classifier that uses Gaussian distributions for all the continuous predictive variables.
- dNB: Discrete NB classifier where all the continuous variables are discretized using Fayyad and Irani's algorithm [21]. This classifier was run in Weka [34].

We use a stratified tenfold cross-validation technique to estimate the accuracy of the classifiers. The cross-validation procedure was run ten times independently. Therefore, 100 accuracy values are obtained. Table 1 shows the mean accuracy and the standard deviation for each dataset and each method. Table 2 shows the complexity of the final Bayesian classifiers induced by the methods in the complete datasets averaged over ten independent runs, i.e., the number of parameters in the models, the number of predictive variables, the percentage of angular variables in the final classifier, and the elapsed time needed to learn the Bayesian classifiers. We find that the performance of classifiers using von Mises distributions for the angular predictive variables (vMNB and SelvMNB) is similar to or better than when Gaussian conditional probability distributions are used for those variables (GNB and SelGNB). dNB using supervised discretization achieves competitive results against SelvMNB and SelGNB and yields the best results in four datasets (`Protein10`, `MAGIC`, `Arrhythmia` and `Covertype`). Note that the discretization algorithm can inherently perform some sort of feature selection by discretizing a variable in only one value. This could explain why dNB achieves such good results. Note also that dNB needs to estimate more parameters than SelvMNB and SelGNB in all the datasets but two (`Megaspores` and `Temperature`). For `Covertype`, SelvMNB and SelGNB achieved the same accuracy in all the folds. Neither algorithm selected either of the two angular variables (see Table 2), so SelvMNB and SelGNB induce exactly the same classifier for this problem and no significant differences can be found between them. The number of parameters in vMNB and GNB are the same because both Gaussian and von Mises distributions have two parameters and no feature subset selection is performed. However, GNB is slightly faster than vMNB because estimating the concentration of a von Mises density involves more operations than variance estimation for Gaussian densities. SelvMNB is also slower than SelGNB even when the number of selected variables is the same. Apart from having slower parameter estimation equations, the method used for sampling a von Mises density is computationally less efficient than the sampling algorithms for Gaussian densities. These sampling methods are used when computing the mutual information between each predictive variable and the class (see Appendix 3). vMNB frequently outperforms GNB in those datasets with a higher percentage of angular variables, e.g., `Protein1`, `Protein10` or `Auslan`. This highlights the importance of using von Mises distributions for modeling angular data. SelvMNB and SelGNB included a similar percentage of angular variables in the final Bayesian classifiers. In most

Table 1 Mean accuracy and standard deviation of the classifiers evaluated on different datasets using ten runs of a stratified tenfold cross-validation

Algorithm	vMNB	SelvMNB	GNB	SelGNB	dNB
`Megaspores`	76.50 ± 3.56	76.50 ± 3.56	**76.60** ± 3.58	**76.60** ± 3.58	75.22 ± 3.37
`Protein1`	98.04 ± 0.16	**98.39** ± 0.15	97.63 ± 0.18	97.96 ± 0.17	97.78 ± 0.21
`Protein10`	83.98 ± 0.55	86.14 ± 0.54	80.77 ± 0.60	82.11 ± 0.48	**86.91** ± 0.50
`Temperature`	**74.08** ± 1.40	74.07 ± 1.38	72.47 ± 1.26	72.47 ± 1.26	72.80 ± 1.33
`Auslan`	64.47 ± 3.01	81.72 ± 2.80	64.39 ± 3.03	**82.24** ± 2.44	78.24 ± 2.81
`MAGIC`	72.75 ± 0.92	75.26 ± 0.82	72.68 ± 0.92	74.92 ± 0.88	**77.73** ± 0.81
`Arrhythmia`	76.52 ± 6.63	78.19 ± 6.09	76.47 ± 6.56	78.17 ± 6.16	**78.93** ± 6.30
`Covertype`	65.43 ± 0.46	67.07 ± 0.41	65.56 ± 0.45	67.07 ± 0.41	**68.49** ± 0.44

Table 2 Complexity analysis of the Bayesian classifiers

	Megaspores	Protein1	Protein10	Temperature	Auslan	MAGIC	Arrhythmia	Covertype
vMNB								
# params	5	9	243	23	22,894	41	685	454
# vars	1	2	30	3	120	10	174	54
% ang vars	100	100	100	33.33	50	10	2.30	3.70
Time	0.0006	0.0103	0.1494	0.0074	0.7274	0.0089	0.0496	0.8598
SelvMNB								
# params	5	5	71	23	6,687	13	363.40	20
# vars	1	1	8.50	3	34.70	3	90.60	1
% ang vars	100	100	100	33.33	58.98	33.33	1.60	0
Time	0.1009	0.2375	9.4965	0.1113	159.7019	0.2860	8.8707	61.9942
GNB								
# params	5	9	243	23	22,894	41	685	454
# vars	1	2	30	3	120	10	174	54
% ang vars	100	100	100	33.33	50	10	2.30	3.70
Time	0.0004	0.0052	0.0618	0.0029	0.4881	0.0075	0.0493	0.8449
SelGNB								
# params	5	5	51.80	23	5,832	13.40	469.20	20
# vars	1	1	6.10	3	30.20	3.10	117.10	1
% ang vars	100	100	100	33.33	60.26	32.50	2.57	0
Time	0.0031	0.0300	4.3668	0.0206	127.4423	0.1692	8.3954	61.6814
dNB								
# params	5	9	243	23	22,894	41	685	454
# vars	1	2	31	3	120	10	174	54
% ang vars	100	100	96.77	33.33	50	10	2.30	3.70

For each dataset and each Bayesian classifier, the table shows the number of parameters of the classifier (# params), the number of predictive variables (# vars), the percentage of angular variables out of the total number (# vars) of variables (% ang vars) and the elapsed time in seconds (time) used to learn the Bayesian classifier. The results are averaged over ten runs. The complete datasets were used to learn the Bayesian classifiers. We used Weka sofware to learn dNB, so the learning times are not comparable and have not been included

Table 3 Average ranking of the algorithms computed over all the datasets

Algorithm	Average ranking
SelvMNB	2.125
dNB	2.375
SelGNB	2.8125
vMNB	3.3125
GNB	4.375

scenarios, the same variables were selected by both SelvMNB and SelGNB. Therefore, when SelvMNB yields better results than SelGNB, it means that von Mises densities model the data in a better way than Gaussian densities (see Sect. 4.2).

Table 3 shows how each algorithm ranked on average across all datasets. SelvMNB is the highest-ranking algorithm, and we find that both vMNB and SelvMNB rank higher than their linear counterparts, GNB and SelGNB, respectively.

Table 4 Adjusted p values of post hoc tests when performing all pairwise comparisons between classifiers

H_1	p_{Neme}	p_{Holm}	p_{Shaf}	p_{Berg}
SelvMNB $\neq$ GNB	**0.0443**	**0.0443**	**0.0443**	**0.0443**
GNB $\neq$ dNB	0.1141	0.1027	0.0685	0.0685
GNB $\neq$ SelGNB	0.4811	0.3849	0.2886	0.1924
vMNB $\neq$ SelvMNB	1.0000	0.9315	0.7985	0.7985
vMNB $\neq$ GNB	1.0000	1.0000	1.0000	0.7985
vMNB $\neq$ dNB	1.0000	1.0000	1.0000	0.7985
SelvMNB $\neq$ SelGNB	1.0000	1.0000	1.0000	1.0000
vMNB $\neq$ SelGNB	1.0000	1.0000	1.0000	1.0000
SelGNB $\neq$ dNB	1.0000	1.0000	1.0000	1.0000
SelvMNB $\neq$ dNB	1.0000	1.0000	1.0000	1.0000

Statistically significant results at $\alpha = 0.05$ are highlighted in bold

Statistical methods for comparing algorithms over a set of problems were proposed in [12, 31] to find statistical differences in the performance of all pairs of algorithms. Table 4 shows the adjusted p values reported by these

Table 5 Results of a Wilcoxon sign-rank test using the sorted difference in a tenfold cross-validation averaged over 10 runs

	vMNB vs. GNB		vMNB vs. dNB		GNB vs. dNB	
	H_1	p value	H_1	p value	H_1	p value
Megaspores	$<$	0.2734	$>^*$	0.0420	$>^*$	0.0244
Protein1	$>^*$	0.0010	$>^*$	0.0010	$<^*$	0.0068
Protein10	$>^*$	0.0010	$<^*$	0.0010	$<^*$	0.0010
Temperature	$>^*$	0.0020	$>^*$	0.0098	$<$	0.1875
Auslan	$>$	0.3125	$<^*$	0.0010	$<^*$	0.0010
MAGIC	$>^*$	0.0195	$<^*$	0.0010	$<^*$	0.0010
Arrhythmia	$>$	0.4375	$<$	0.1611	$<$	0.1377
Covertype	$<^*$	0.0186	$<^*$	0.0010	$<^*$	0.0010
	SelvMNB vs. SelGNB		SelvMNB vs. dNB		SelGNB vs. dNB	
	H_1	p value	H_1	p value	H_1	p value
Megaspores	$<$	0.2734	$>^*$	0.0420	$>^*$	0.0244
Protein1	$>^*$	0.0010	$>^*$	0.0010	$>^*$	0.0098
Protein10	$>^*$	0.0010	$<^*$	0.0020	$<^*$	0.0010
Temperature	$>^*$	0.0020	$>^*$	0.0098	$<$	0.1875
Auslan	$<$	0.2461	$>^*$	0.0020	$>^*$	0.0020
MAGIC	$>^*$	0.0098	$<^*$	0.0010	$<^*$	0.0010
Arrhythmia	$>$	0.5098	$<$	0.4229	$<$	0.3477
Covertype	$\neq$	1.0000	$<^*$	0.0010	$<^*$	0.0010

methods. Considering all the datasets, only the differences between the performance of GNB and SelvMNB are statistically significant (significance level $\alpha = 0.05$).

Some datasets (see Table 6 in Appendix 4) include few angular variables (Covertype, Arrhythmia, MAGIC). Modeling these variables with a von Mises distribution or a Gaussian distribution is likely to have little impact on classifier accuracy. Therefore, it is worthwhile comparing algorithm performance on each dataset individually. Bouckaert [9] recommends using a t test with a sorted runs sampling scheme to evaluate replicability of classifier learning experiments. He also states that this procedure yields an acceptable type I error and good power. We used a non-parametric alternative and applied a Wilcoxon sign-rank test with the sorted runs sampling scheme. Table 5 shows the p values of the Wilcoxon sign-rank test over the sorted difference of accuracies for a tenfold cross-validation averaged over 10 runs. The null hypothesis is that the median of the averaged differences is zero, i.e., both algorithms perform similarly. The alternative hypotheses (H_1) were selected according to the results reported in Table 1. Statistically significant results at a significance level $\alpha = 0.05$ are highlighted with an asterisk (*). vMNB significantly outperformed GNB in four datasets (Protein1, Protein10, Temperature and MAGIC), whereas GNB only outperformed vMNB in the Covertype problem. We found no statistical differences between the two classifiers for the Megaspores, Auslan and Arrhythmia datasets. Similar results were found when comparing SelvMNB and SelGNB. However, SelGNB did not significantly outperform SelvMNB in any dataset, whereas SelvMNB outperformed SelGNB in four datasets. These two algorithms induce the same classifier for the Covertype dataset, so there were no statistical differences between the two methods for that dataset (p value = 1.0 in Table 5). The dNB classifier with discretized predictive variables yields very good results. dNB significantly outperforms vMNB in four datasets, whereas vMNB significantly outperforms dNB in three datasets. On the other hand, SelvMNB significantly outperforms dNB in

Table 6 Datasets used in this study

Dataset	# angular vars	# linear vars	# discrete vars	# class values	# instances
Megaspores	1	0	0	2	960
Protein1	2	0	0	2	49,676
Protein10	30	0	0	4	49,314
Temperature	1	1	1	3	8,753
Auslan	60	60	0	95	2,565
MAGIC	1	10	0	2	19,020
Arrhythmia	5	175	73	2	430
Covertype	2	8	44	7	100,000

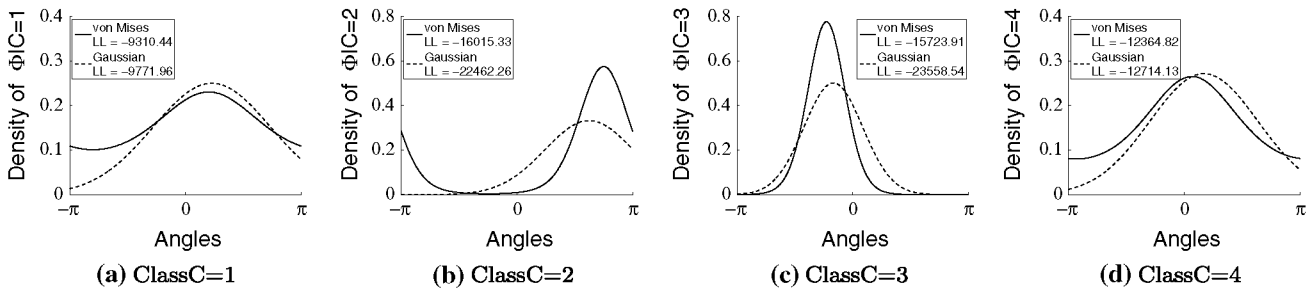


Fig. 13 von Mises (*solid*) and Gaussian (*dashed*) conditional distributions fitted for variable 11 in `Protein10` dataset

four datasets, whereas dNB significantly outperforms SelvMNB in three datasets. vMNB and SelvMNB perform better against dNB than their linear counterparts, GNB and SelGNB, respectively.

4.2 Goodness-of-fit analysis

To understand why vMNB performs better, we illustrate the differences between using linear and angular distributions to model directional data. We took variable 11 in the `Protein10` dataset, which was selected by both SelvMNB and SelGNB as an important predictive variable for classification. `Protein10` has four class values. NB fits one conditional probability density for each class value c. Figure 13 plots the Gaussian (dashed lines) and the von Mises (solid lines) conditional distributions fitted from the data. We can see important mismatches for class values $C = 2$ and $C = 3$. Figure 13b shows how the Gaussian distribution ignores the periodicity of the data and yields a density of $3.97{\cdot}10^{-5}$ for an angle of $-180°$, and a density of 0.2049 for 180°. Therefore, Gaussian distributions yield two different densities for the same angle. On the other hand, the von Mises distribution in Fig. 13c is more peaked and yields higher densities than the Gaussian distribution for values close to the mean.

The legends in Fig. 13 include the log-likelihood of the models given the data: $LL = \sum_i \log f_{\Phi|c}(\phi^{(i)})$. von Mises distributions always yield higher LL than Gaussian distributions. In fact, the highest differences in the log-likelihood between von Mises and Gaussian distributions can be found for class values 2 and 3.

Gaussian and von Mises distributions are very similar when the concentration of the values is high. However, using Gaussian distributions to model angles can negatively affect NB's behavior. We use an artificial example to illustrate this point. We generate a dataset with one angular predictive variable and a binary class with values $\Omega(C) = \{1, 2\}$. The classes are equiprobable a priori. Instances from class $C = 1$ follow the distribution $\Phi|C = 1 \sim vM(\pi, 2.5)$, whereas instances from class $C = 2$ follow the distribution $\Phi|C = 2 \sim vM(\pi/2, 2.5)$. Figure 14 shows the conditional

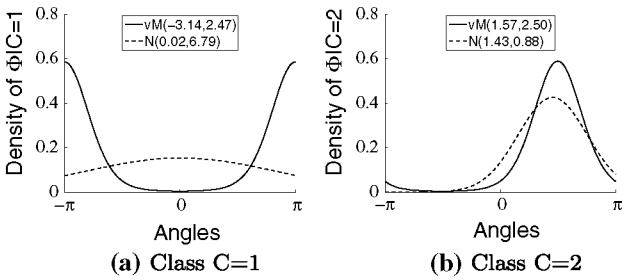


Fig. 14 von Mises (*solid*) and Gaussian (*dashed*) conditional distributions fitted for the artificial dataset

density functions of the von Mises and the Gaussian distributions fitted to a sample of 2,000 instances. Figure 14a shows that the Gaussian distribution ignores the periodicity of the data, overestimates the variance and incorrectly estimates the mean direction. This yields errors in NB's classification. For example, GNB classifies angle $\phi = \pi$ with class $C = 2$, whereas it should apparently belong to class $C = 1$ because the mean direction of the distribution that generates class $C = 1$ is $\mu_{\Phi|1} = \pi$. On the other hand, GNB labels the angle $\phi = 0$ with the class $C = 1$. The angle $\phi = 0$ is closer to the mean direction of the distribution with class $C = 2$ ($\mu_{\Phi|2} = \pi/2$), so it should be classified with $C = 2$.

5 Conclusion

Directional data can be found everywhere in science. Directional information has a number of properties that make it necessary to develop and use different techniques than the ones used with linear information.

In this paper, we extended one of the simplest and best known models for classification, the naive Bayes classifier, to the case where directional data are used as predictive variables. First, we reviewed the most common distributions in directional statistics: the von Mises distribution and the von Mises–Fisher distribution. Understanding the implications of the naive Bayes assumption and the theoretical properties of the classifier is the key to interpreting

its behavior and establishing its problem-solving potential [58]. Therefore, we analyzed the decision functions of the NB classifiers using directional predictive variables and studied the surfaces induced by those decision functions at length for different values of the parameters. We also studied the more general scenarios where a hybrid NB classifier accounts for discrete, linear (Gaussian) and directional predictive variables.

We showed that the NB classifier with one directional predictive variable, using either the univariate von Mises or the multivariate von Mises–Fisher distribution, is a linear classifier. The decision surface induced by the classifier is a hyperplane (or a set of hyperplanes if more than two class values are considered) that separates the class regions. Therefore, it should be especially well suited for solving problems with linearly separable classes. When two angular predictive variables are considered, the vMNB classifier induces more complex quadratic decision surfaces. In the hybrid setting where von Mises–Fisher and Gaussian distributions are used to model the predictive variables, we showed that the complexity of the decision surfaces depends on the parameters of the Gaussian distribution. Thus, the decision surfaces are hyperplanes when the covariance matrices of the conditional predictive distributions are equal and hyperquadrics when they are not [18]. Artificial examples were used to illustrate the behavior of the different classifiers and to show the decision surfaces they induce. NB performance is reduced when irrelevant or redundant predictive variables are used [44]. Therefore, we adapted the selective NB algorithm to the use of directional distributions.

We evaluated the vMNB classifier over 8 datasets and compared it against the corresponding NB classifiers that use Gaussian distributions or discretization for modeling angular variables. SelvMNB was the best ranking algorithm. Statistical tests were performed to find significant differences in the performance of the classifiers. vMNB and SelvMNB performed similarly or better than the classifiers using linear distributions in all but one dataset.

The naive Bayes classifier's conditional independence assumption is quite restrictive and clearly limits the kind of problems that these models can solve. Several Bayesian classifiers that relax the conditional independence assumption have been proposed in the literature, e.g., the tree-augmented naive Bayes [30], the seminaive Bayes [56], the k-dependence Bayesian classifier [63] or the general Bayesian network classifier. Extending these models to the use of directional variables is by no means trivial, since it has been shown that both marginal and conditional distributions cannot be von Mises distributions [47, 49]. Therefore, this is an open and interesting research field.

Directional data can be found in other machine learning scenarios, e.g., clustering and regression problems. Clustering with directional data has been extensively studied in recent papers, see e.g., [2, 3, 53]. Also, many works are available on regression models where the target variable to predict is angular and the predictive variables are either angular [41] or linear [16, 26]. Regression models with spherical target and predictive variables have also been studied in [15, 61]. Recently, circular ordinal regression, where the target variable is discrete, but defined in a circular ordered domain has been approached in [13] using support vector machines. Directional information has also been used in neural networks, where Zemel et al. [70] proposed an extension of the Boltzmann machine with angular units.

On the other hand, Bayesian networks have also been applied to classical regression problems [29, 54]. Hybrid models that include different types of probability distributions have attracted much interest, and different approaches have been proposed [8, 36, 62, 66]. Directional distributions add yet another possibility to the range of distributions that can be considered in hybrid Bayesian networks. Hybrid probability distributions for modeling the joint density of angular and linear variables [38] could be used in hybrid Bayesian networks for regression. When several angular variables are included in the Bayesian network, the fact that we cannot model both marginal and conditional distributions as von Mises distributions is again a crucial problem in these models.

Directional statistics opens a number of interesting challenges and opportunities within machine learning research, particularly for probabilistic graphical models. We hope that further research in this area and the implementation of more complex models will provide an excellent tool for solving difficult problems in a wide range of fields.

Appendix 1: von Mises NB classifier decision function

vMNB with one predictive variable

We start by equaling the posterior probability of each class value using the probability density function of the von Mises distribution (1):

$$p(C=1)\frac{1}{2\pi I_0(\kappa_{\Phi|1})}\exp(\kappa_{\Phi|1}\cos(\phi-\mu_{\Phi|1})) = p(C=2)\frac{1}{2\pi I_0(\kappa_{\Phi|2})}\exp(\kappa_{\Phi|2}\cos(\phi-\mu_{\Phi|2})).$$

Simplify the constant 2π, take logarithms and arrange all terms on the same side of the equation:

$$\kappa_{\Phi|1}\cos(\phi-\mu_{\Phi|1})-\kappa_{\Phi|2}\cos(\phi-\mu_{\Phi|2}) + \ln\frac{p(C=1)}{I_0(\kappa_{\Phi|1})}-\ln\frac{p(C=2)}{I_0(\kappa_{\Phi|2})}=0.$$

Substitute $\cos(\beta-\gamma)=\cos(\beta)\cos(\gamma)+\sin(\beta)\sin(\gamma)$ and operate the logarithms:

$$\kappa_{\Phi|1}\Big[\cos\phi\cos\mu_{\Phi|1}+\sin\phi\sin\mu_{\Phi|1}\Big] - \kappa_{\Phi|2}\Big[\cos\phi\cos\mu_{\Phi|2}+\sin\phi\sin\mu_{\Phi|2}\Big] + \ln\frac{p(C=1)I_0(\kappa_{\Phi|2})}{p(C=2)I_0(\kappa_{\Phi|1})}=0.$$

Arrange using $\cos\phi$ and $\sin\phi$ as common terms:

$$(\kappa_{\Phi|1}\cos\mu_{\Phi|1}-\kappa_{\Phi|2}\cos\mu_{\Phi|2})\cos\phi + (\kappa_{\Phi|1}\sin\mu_{\Phi|1}-\kappa_{\Phi|2}\sin\mu_{\Phi|2})\sin\phi + \ln\frac{p(C=1)I_0(\kappa_{\Phi|2})}{p(C=2)I_0(\kappa_{\Phi|1})}=0.$$

Substitute

$$a=\kappa_{\Phi|1}\cos\mu_{\Phi|1}-\kappa_{\Phi|2}\cos\mu_{\Phi|2},$$
$$b=\kappa_{\Phi|1}\sin\mu_{\Phi|1}-\kappa_{\Phi|2}\sin\mu_{\Phi|2},$$
$$D=-\ln\frac{p(C=1)I_0(\kappa_{\Phi|2})}{p(C=2)I_0(\kappa_{\Phi|1})},$$

and get:

$$a\cos\phi+b\sin\phi=D.$$

Trigonometrically, this is equivalent to:

$$T\cos(\phi-\alpha)=D,$$

where $T=\sqrt{a^2+b^2}, \cos\alpha=a/T, \sin\alpha=b/T, \tan\alpha=b/a$. Isolating ϕ from the equation, we get:

$$\phi'=\alpha+\arccos(D/T),$$
$$\phi''=\alpha-\arccos(D/T).$$

The NB classifier finds two angles that bound the class regions.

Particular cases

We have also derived these angles when the conditional probability distributions share one of the parameters. We consider that the classes are equiprobable. If they are not equiprobable, the prior probabilities of the class values influence the value of D, modifying the class subregions so that more likely classes have larger subregions.

- Case 1: $\kappa_{\Phi|1}=\kappa_{\Phi|2}=\kappa_\Phi$ and $\mu_{\Phi|1}\neq\mu_{\Phi|2}$. When the concentration parameter is the same in the two distributions, we have the following values for the constants:

  $$a=\kappa_\Phi(\cos\mu_{\Phi|1}-\cos\mu_{\Phi|2}),$$
  $$b=\kappa_\Phi(\sin\mu_{\Phi|1}-\sin\mu_{\Phi|1}),$$
  $$D=-\ln\frac{p(C=1)I_0(\kappa_{\Phi|2})}{p(C=2)I_0(\kappa_{\Phi|1})}=-\ln 1=0.$$

  Substituting in the expression of the arccosine, we get:

  $$\arccos(D/T)=\arccos 0=\pi/2.$$

  To compute α, we take the trigonometric identities:

  $$\cos\beta-\cos\gamma=-2\sin\left(\frac{1}{2}(\beta+\gamma)\right)\sin\left(\frac{1}{2}(\beta-\gamma)\right),$$
  $$\sin\beta-\sin\gamma=2\sin\left(\frac{1}{2}(\beta-\gamma)\right)\cos\left(\frac{1}{2}(\beta+\gamma)\right),$$

  which we substitute in the following expression:

  $$\begin{aligned}\tan\alpha=\frac{b}{a}&=\frac{\kappa_\Phi(\sin\mu_{\Phi|1}-\sin\mu_{\Phi|2})}{\kappa_\Phi(\cos\mu_{\Phi|1}-\cos\mu_{\Phi|2})}\\&=\frac{2\sin(\frac{1}{2}(\mu_{\Phi|1}-\mu_{\Phi|2}))\cos(\frac{1}{2}(\mu_{\Phi|1}+\mu_{\Phi|2}))}{-2\sin(\frac{1}{2}(\mu_{\Phi|1}+\mu_{\Phi|2}))\sin(\frac{1}{2}(\mu_{\Phi|1}-\mu_{\Phi|2}))}\\&=-\frac{\cos(\frac{1}{2}(\mu_{\Phi|1}+\mu_{\Phi|2}))}{\sin(\frac{1}{2}(\mu_{\Phi|1}+\mu_{\Phi|2}))}\\&=-\cot\left(\frac{1}{2}(\mu_{\Phi|1}+\mu_{\Phi|2})\right)\\&=\tan\left(\frac{1}{2}(\mu_{\Phi|1}+\mu_{\Phi|2})+\frac{\pi}{2}\right),\end{aligned}$$

  $$\alpha=\frac{1}{2}(\mu_{\Phi|1}+\mu_{\Phi|2})+\frac{\pi}{2}.$$

Now we can compute the decision angles found by the classifier:

$$\phi=\alpha\pm\arccos(D/T)=\frac{1}{2}(\mu_{\Phi|1}+\mu_{\Phi|2})+\frac{\pi}{2}\pm\frac{\pi}{2}.$$

The two decision angles are:

$$\phi'=\frac{1}{2}(\mu_{\Phi|1}+\mu_{\Phi|2}),$$
$$\phi''=\frac{1}{2}(\mu_{\Phi|1}+\mu_{\Phi|2})+\pi.$$

These two angles correspond to the bisector angle of the two mean directions.

- Case 2: $\kappa_{\Phi|1} \neq \kappa_{\Phi|2}$ and $\mu_{\Phi|1} = \mu_{\Phi|2} = \mu_\Phi$. In this scenario the mean directions are equal, so the constants reduce to:

$$
\begin{aligned}
a &= (\kappa_{\Phi|1} - \kappa_{\Phi|2}) \cos \mu_\Phi, \\
b &= (\kappa_{\Phi|1} - \kappa_{\Phi|2}) \sin \mu_\Phi, \\
D &= -\ln \frac{p(C=1) I_0(\kappa_{\Phi|2})}{p(C=2) I_0(\kappa_{\Phi|1})} = -\ln \frac{I_0(\kappa_{\Phi|2})}{I_0(\kappa_{\Phi|1})}, \\
T &= \sqrt{a^2 + b^2} \\
&= \sqrt{(\kappa_{\Phi|1} - \kappa_{\Phi|2})^2 \cos^2 \mu_\Phi + (\kappa_{\Phi|1} - \kappa_{\Phi|2})^2 \sin^2 \mu_\Phi} \\
&= \sqrt{(\kappa_{\Phi|1} - \kappa_{\Phi|2})^2 (\cos^2 \mu_\Phi + \sin^2 \mu_\Phi)} \\
&= \kappa_{\Phi|1} - \kappa_{\Phi|2}.
\end{aligned}
$$

  We compute α by substituting in the expression:

$$
\begin{aligned}
\tan \alpha &= \frac{b}{a} = \frac{(\kappa_{\Phi|1} - \kappa_{\Phi|2}) \sin \mu_\Phi}{(\kappa_{\Phi|1} - \kappa_{\Phi|2}) \cos \mu_\Phi} = \tan \mu_\Phi, \\
\alpha &= \mu_\Phi.
\end{aligned}
$$

  Therefore, the resulting decision angles are given by:

$$
\begin{aligned}
\phi &= \alpha \pm \arccos(D/T), \\
\phi' &= \mu_\Phi + \arccos \frac{D}{\kappa_{\Phi|1} - \kappa_{\Phi|2}}, \\
\phi'' &= \mu_\Phi - \arccos \frac{D}{\kappa_{\Phi|1} - \kappa_{\Phi|2}}.
\end{aligned}
$$

Clearly, the two angles are defined with respect to the common mean direction, and their distance to that mean direction depends on the concentration parameter values.

vMNB with two predictive variables

In this scenario, we have two circular predictive variables Φ and Ψ. The domain defined by these variables is a torus $(-\pi, \pi] \times (-\pi, \pi]$. As in the simpler case above, we compute the decision surfaces induced by the classifier by equaling the posterior probability of the two class values

$$p(C=1 | \Phi = \phi, \Psi = \psi) = p(C=2 | \Phi = \phi, \Psi = \psi).$$

Using Bayes' rule and the conditional independence assumption, we get

$$
\begin{aligned}
&p(C=1) f_{\Phi|C=1}(\phi; \mu_{\Phi|1}, \kappa_{\Phi|1}) f_{\Psi|C=1}(\psi; \mu_{\Psi|1}, \kappa_{\Psi|1}) \\
&\quad = p(C=2) f_{\Phi|C=2}(\phi; \mu_{\Phi|2}, \kappa_{\Phi|2}) f_{\Psi|C=2}(\psi; \mu_{\Psi|2}, \kappa_{\Psi|2}).
\end{aligned}
$$

We substitute the von Mises density (1) and get:

$$
\begin{aligned}
&p(C=1) \frac{\exp(\kappa_{\Phi|1} \cos(\phi - \mu_{\Phi|1}))}{2\pi I_0(\kappa_{\Phi|1})} \frac{\exp(\kappa_{\Psi|1} \cos(\psi - \mu_{\Psi|1}))}{2\pi I_0(\kappa_{\Psi|1})} \\
&\quad = p(C=2) \frac{\exp(\kappa_{\Phi|2} \cos(\phi - \mu_{\Phi|2}))}{2\pi I_0(\kappa_{\Phi|2})} \frac{\exp(\kappa_{\Psi|2} \cos(\psi - \mu_{\Psi|2}))}{2\pi I_0(\kappa_{\Psi|2})}.
\end{aligned}
$$

We simplify the constant 2π, take logarithms and arrange all the terms on the same side of the equation:

$$
\begin{aligned}
&\kappa_{\Phi|1} \cos(\phi - \mu_{\Phi|1}) + \kappa_{\Psi|1} \cos(\psi - \mu_{\Psi|1}) \\
&\quad - \kappa_{\Phi|2} \cos(\phi - \mu_{\Phi|2}) - \kappa_{\Psi|2} \cos(\psi - \mu_{\Psi|2}) \\
&\quad + \ln \frac{p(C=1) I_0(\kappa_{\Phi|2}) I_0(\kappa_{\Psi|2})}{p(C=2) I_0(\kappa_{\Phi|1}) I_0(\kappa_{\Psi|1})} = 0.
\end{aligned}
$$

We substitute the trigonometric identity $\cos(\beta - \gamma) = \cos(\beta)\cos(\gamma) + \sin(\beta)\sin(\gamma)$ and arrange the terms:

$$
\begin{aligned}
&(\kappa_{\Phi|1} \cos \mu_{\Phi|1} - \kappa_{\Phi|2} \cos \mu_{\Phi|2}) \cos \phi \\
&\quad + (\kappa_{\Phi|1} \sin \mu_{\Phi|1} - \kappa_{\Phi|2} \sin \mu_{\Phi|2}) \sin \phi \\
&\quad + (\kappa_{\Psi|1} \cos \mu_{\Psi|1} - \kappa_{\Psi|2} \cos \mu_{\Psi|2}) \cos \psi \\
&\quad + (\kappa_{\Psi|1} \sin \mu_{\Psi|1} - \kappa_{\Psi|2} \sin \mu_{\Psi|2}) \sin \psi \\
&\quad + \ln \frac{p(C=1) I_0(\kappa_{\Phi|2}) I_0(\kappa_{\Psi|2})}{p(C=2) I_0(\kappa_{\Phi|1}) I_0(\kappa_{\Psi|1})} = 0.
\end{aligned}
$$

We define the following constants:

$$
\begin{aligned}
a &= \kappa_{\Phi|1} \cos \mu_{\Phi|1} - \kappa_{\Phi|2} \cos \mu_{\Phi|2}, \\
b &= \kappa_{\Phi|1} \sin \mu_{\Phi|1} - \kappa_{\Phi|2} \sin \mu_{\Phi|2}, \\
c &= \kappa_{\Psi|1} \cos \mu_{\Psi|1} - \kappa_{\Psi|2} \cos \mu_{\Psi|2}, \\
d &= \kappa_{\Psi|1} \sin \mu_{\Psi|1} - \kappa_{\Psi|2} \sin \mu_{\Psi|2}, \\
D &= -\ln \frac{p(C=1) I_0(\kappa_{\Phi|2}) I_0(\kappa_{\Psi|2})}{p(C=2) I_0(\kappa_{\Phi|1}) I_0(\kappa_{\Psi|1})},
\end{aligned}
$$

and substitute them to get

$$a \cos \phi + b \sin \phi + c \cos \psi + d \sin \psi = D.$$

The Cartesian coordinates of the points defined by the angles ϕ and ψ on the surface of a torus are

$$
\begin{aligned}
x &= (L + l \cos \phi) \cos \psi, \\
y &= (L + l \cos \phi) \sin \psi, \\
z &= l \sin \phi,
\end{aligned}
$$

where L is the distance from the center of the torus to the center of the revolving circumference that generates the torus, and l is the radius of the revolving circumference. We isolate the trigonometric functions and get

$$\sin\phi = z/l,$$
$$\cos\phi = \pm\sqrt{1-\sin^2\phi} = \pm\sqrt{1-\left(\frac{z}{l}\right)^2} = \pm\frac{1}{l}\sqrt{l^2-z^2},$$
$$\sin\psi = \frac{y}{L+l\cos\phi},$$
$$\cos\psi = \frac{x}{L+l\cos\phi}.$$

Substituting these expressions, we get the two following equations corresponding to the two signs of $\cos\phi$:

$$\frac{a}{l}\sqrt{l^2-z^2}+\frac{b}{l}z+\frac{c}{L+\sqrt{l^2-z^2}}x + \frac{d}{L+\sqrt{l^2-z^2}}y+D=0.$$
$$-\frac{a}{l}\sqrt{l^2-z^2}+\frac{b}{l}z+\frac{c}{L-\sqrt{l^2-z^2}}x + \frac{d}{L-\sqrt{l^2-z^2}}y+D=0.$$

Operating and arranging the terms, we get

$$clx+dly-az^2+bz\sqrt{l^2-z^2}+bLz + (aL+Dl)\sqrt{l^2-z^2}+al^2+DLl=0,$$
$$clx+dly-az^2-bz\sqrt{l^2-z^2}+bLz - (aL+Dl)\sqrt{l^2-z^2}+al^2+DLl=0.$$

These expressions are quadratic in z. Therefore, we conclude that von Mises NB with two predictive variables is a much more complex and flexible classifier than von Mises NB with one predictive variable.

Appendix 2: von Mises–Fisher NB classifier decision function

To study the decision function for the von Mises–Fisher NB classifier we proceed as in Appendix 1. We equal the posterior probabilities of the class values using the probability density function in Eq. (1):

$$r(\mathbf{X})=0 \Leftrightarrow p(C=1)\frac{(\kappa_{\mathbf{X}|1})^{\frac{n}{2}-1}}{\sqrt{(2\pi)^n}I_{\frac{n}{2}-1}(\kappa_{\mathbf{X}|1})}\exp(\kappa_{\mathbf{X}|1}\boldsymbol{\mu}_{\mathbf{X}|1}^{\mathrm{T}}\mathbf{X}) = p(C=2)\frac{(\kappa_{\mathbf{X}|2})^{\frac{n}{2}-1}}{\sqrt{(2\pi)^n}I_{\frac{n}{2}-1}(\kappa_{\mathbf{X}|2})}\exp(\kappa_{\mathbf{X}|2}\boldsymbol{\mu}_{\mathbf{X}|2}^{\mathrm{T}}\mathbf{X}).$$

Simplify the constants and take logarithms:

$$\ln\frac{p(C=1)(\kappa_{\mathbf{X}|1})^{\frac{n}{2}-1}}{I_{\frac{n}{2}-1}(\kappa_{\mathbf{X}|1})}+\kappa_{\mathbf{X}|1}\boldsymbol{\mu}_{\mathbf{X}|1}^{\mathrm{T}}\mathbf{X} = \ln\frac{p(C=2)(\kappa_{\mathbf{X}|2})^{\frac{n}{2}-1}}{I_{\frac{n}{2}-1}(\kappa_{\mathbf{X}|2})}+\kappa_{\mathbf{X}|2}\boldsymbol{\mu}_{\mathbf{X}|2}^{\mathrm{T}}\mathbf{X}.$$

Arrange all the terms on the same side of the equation and operate the logarithms to get the following hyperplane equation:

$$(\kappa_{\mathbf{X}|1}\boldsymbol{\mu}_{\mathbf{X}|1}-\kappa_{\mathbf{X}|2}\boldsymbol{\mu}_{\mathbf{X}|2})^{\mathrm{T}}\mathbf{X} + \ln\frac{p(C=1)(\kappa_{\mathbf{X}|1})^{\frac{n}{2}-1}I_{\frac{n}{2}-1}(\kappa_{\mathbf{X}|2})}{p(C=2)(\kappa_{\mathbf{X}|2})^{\frac{n}{2}-1}I_{\frac{n}{2}-1}(\kappa_{\mathbf{X}|1})}=0.$$

Particular cases

Considering that both class values have the same prior probability and that one of the parameters has the same value in both distributions, Case 1 and Case 2 can be simplified as follows. When the prior probabilities are different, the hyperplanes move away from the mean direction of the most likely class value, making their subregions larger.

- Case 1: $\kappa_{\mathbf{X}|1}=\kappa_{\mathbf{X}|2}=\kappa_{\mathbf{X}}$ and $\boldsymbol{\mu}_{\mathbf{X}|1}\neq\boldsymbol{\mu}_{\mathbf{X}|2}$. When the distributions share the concentration parameter, we get the expression:

  $$(\kappa_{\mathbf{X}}\boldsymbol{\mu}_{\mathbf{X}|1}-\kappa_{\mathbf{X}}\boldsymbol{\mu}_{\mathbf{X}|2})^{\mathrm{T}}\mathbf{X}+\ln\frac{p(C=1)\kappa_{\mathbf{X}}^{\frac{n}{2}-1}I_{\frac{n}{2}-1}(\kappa_{\mathbf{X}})}{p(C=2)\kappa_{\mathbf{X}}^{\frac{n}{2}-1}I_{\frac{n}{2}-1}(\kappa_{\mathbf{X}})}=0.$$

  The logarithm reduces to 0 and we can take $\kappa_{\mathbf{X}}$ as common term:

  $$\kappa_{\mathbf{X}}(\boldsymbol{\mu}_{\mathbf{X}|1}-\boldsymbol{\mu}_{\mathbf{X}|2})^{\mathrm{T}}\mathbf{X}=0.$$

  Therefore, given that $\kappa>0$ (otherwise the distributions are uniform), the hyperplane equation reduces to:

  $$(\boldsymbol{\mu}_{\mathbf{X}|1}-\boldsymbol{\mu}_{\mathbf{X}|2})^{\mathrm{T}}\mathbf{X}=0.$$

  That equation specifies a hyperplane that contains the origin point ($\mathbf{0}$) and goes through the middle point of the sector that connects the points of the hypersphere defined by the mean directions $\boldsymbol{\mu}_{\mathbf{X}|1}$ and $\boldsymbol{\mu}_{\mathbf{X}|2}$.
- Case 2: $\kappa_{\mathbf{X}|1}\neq\kappa_{\mathbf{X}|2}$ and $\boldsymbol{\mu}_{\mathbf{X}|1}=\boldsymbol{\mu}_{\mathbf{X}|2}=\boldsymbol{\mu}_{\mathbf{X}}$. In the case where the mean directions have the same value, we can derive the following equation:

$$(\kappa_{\mathbf{X}|1}\boldsymbol{\mu}_{\mathbf{X}}^{\mathrm{T}} - \kappa_{\mathbf{X}|2}\boldsymbol{\mu}_{\mathbf{X}}^{\mathrm{T}})\mathbf{X} + \ln\frac{p(C=1)(\kappa_{\mathbf{X}|1})^{\frac{n}{2}-1}I_{\frac{n}{2}-1}(\kappa_{\mathbf{X}|2})}{p(C=2)(\kappa_{\mathbf{X}|2})^{\frac{n}{2}-1}I_{\frac{n}{2}-1}(\kappa_{\mathbf{X}|1})} = 0.$$

We can take $\boldsymbol{\mu}_{\mathbf{X}}^{\mathrm{T}}$ as a common term:

$$(\kappa_{\mathbf{X}|1} - \kappa_{\mathbf{X}|2})\boldsymbol{\mu}_{\mathbf{X}}^{\mathrm{T}}\mathbf{X} + \ln\frac{(\kappa_{\mathbf{X}|1})^{\frac{n}{2}-1}I_{\frac{n}{2}-1}(\kappa_{\mathbf{X}|2})}{(\kappa_{\mathbf{X}|2})^{\frac{n}{2}-1}I_{\frac{n}{2}-1}(\kappa_{\mathbf{X}|1})} = 0.$$

Dividing by $(\kappa_{\mathbf{X}|1} - \kappa_{\mathbf{X}|2})$, we get:

$$\boldsymbol{\mu}_{\mathbf{X}}^{\mathrm{T}}\mathbf{X} + \frac{1}{\kappa_{\mathbf{X}|1} - \kappa_{\mathbf{X}|2}}\ln\frac{(\kappa_{\mathbf{X}|1})^{\frac{n}{2}-1}I_{\frac{n}{2}-1}(\kappa_{\mathbf{X}|2})}{(\kappa_{\mathbf{X}|2})^{\frac{n}{2}-1}I_{\frac{n}{2}-1}(\kappa_{\mathbf{X}|1})} = 0.$$

The hyperplane defined by that equation is perpendicular to the shared mean direction vector $\boldsymbol{\mu}_{\mathbf{X}}$, and its position is given by the relationships between the concentration parameters.

Appendix 3: Mutual information computation

The mutual information between two variables X and Y is defined as

$$\begin{aligned}\mathrm{MI}(X,Y) &= \int_X\int_Y \rho(x,y)\log\frac{\rho(x,y)}{\rho(x)\rho(y)}\mathrm{d}x\mathrm{d}y \\ &= \mathbb{E}_{(X,Y)}\left[\log\frac{\rho(x,y)}{\rho(x)\rho(y)}\right],\end{aligned} \tag{10}$$

where ρ is a generalized probability function.

In supervised classification problems, we have to estimate $\mathrm{MI}(X_i, C)$ from a set of data pairs $\left(x_i^{(j)}, c^{(j)}\right), j = 1,\ldots,m$. When X_i is a discrete variable, an estimator of the mutual information in (10) is given by

$$\mathrm{MI}(X_i,C) = \frac{1}{m}\sum_{j=1}^{m}\log\frac{\widehat{p}\left(x_i^{(j)}, c^{(j)}\right)}{\widehat{p}\left(x_i^{(j)}\right)\widehat{p}\left(c^{(j)}\right)}, \tag{11}$$

where $\widehat{p}$ are the probabilities estimated from the counts in the dataset.

When the predictive variable X_i is continuous, we take an approach consistent with conditional independence assumptions and we model the conditional probability densities of $X_i|C = c$ as Gaussian or von Mises distributions, depending on the nature of the variable, i.e., linear or angular. Therefore, the marginal density of X_i is a mixture of Gaussian or von Mises distributions, respectively. Algorithm 1 shows the process for computing $\mathrm{MI}(X_i, C)$.

Algorithm 1 *Estimation of $MI(X_i, C)$ with continuous X_i.*

Inputs: A set $\left\{\left(x_i^{(j)}, c^{(j)}\right),\ j = 1,\ldots,m\right\}$ of training data pairs

Steps:

1. *Estimate the prior probability for each class value $\widehat{p}(C = c) = n_c/m$, where n_c is the number of instances in the data belonging to class c.*
2. *Estimate the maximum likelihood parameters of the conditional density functions of X_i given each $c \in \Omega(C)$:*
   - *If X_i is a linear variable, fit a Gaussian probability density $\widehat{f}_{X_i|c}(x; \widehat{\mu}_{X_i|c}, \widehat{\sigma}^2_{X_i|c})$.*
   - *If X_i is an angular variable, fit a von Mises probability density $\widehat{f}_{X_i|c}(x; \widehat{\mu}_{X_i|c}, \widehat{\kappa}_{X_i|c})$.*
3. *Sample M pairs of points $\left(x_i^{*(j)}, c^{*(j)}\right)$ from the distribution $\widehat{\rho}(x_i, c) = \widehat{f}_{X_i|c}(x_i)\widehat{p}(C = c)$. For each class value c, sample $M \cdot \widehat{p}(C = c)$ instances from the density $\widehat{f}_{X_i|c}(x_i)$ and build the pairs (x_i, c).*
4. *Compute the mutual information as*

$$\begin{aligned}MI(X_i,C) &= \frac{1}{M}\sum_{j=1}^{M}\log\frac{\widehat{f}_{X_i|c^{*(j)}}\left(x_i^{*(j)}\right)\widehat{p}\left(C = c^{*(j)}\right)}{\widehat{f}_{X_i}\left(x_i^{*(j)}\right)\widehat{p}\left(C = c^{*(j)}\right)} \\ &= \frac{1}{M}\sum_{j=1}^{M}\log\frac{\widehat{f}_{X_i|c^{*(j)}}\left(x_i^{*(j)}\right)}{\sum_{k\in\Omega(C)}\widehat{p}(C=k)\widehat{f}_{X_i|k}\left(x_i^{*(j)}\right)}.\end{aligned} \tag{12}$$

Appendix 4: Dataset analysis and preprocessing

A thorough inspection of the datasets for supervised classification available in the UCI Machine Learning Repository [28] reported only 5 out of 135 datasets containing some variable measured in angles (bottom half of Table 6). We found no reference to these directional data having be given special treatment. For this reason, we assume that they have been studied as linear continuous variables without taking into account their special properties. We omitted the Breast Tissue dataset [39, 67] from the study because it was not clear whether the "PhaseAngle" variable really represents an angle and how it was measured. Additionally, another four datasets not included in the UCI repository were considered for evaluation (top half of

Table 6). A description of the datasets used in this study follows:

UCI datasets

- **Australian Sign Language** (`Auslan`): Identification of 95 Australian Sign Language signs using position (x, y, z) and orientation angles (roll, pitch, yaw) of both hands [40]. Therefore, 12 measurements are studied. According to [40], the bending measurements are not very reliable, and they were omitted as predictive variables. This is a time series classification problem. The position and orientation of the hands are measured at different times, yielding approximately 54 data frames for each sign. We resampled a set of 10 evenly distributed frames and used them as predictive variables. According to the description, there are 95 different signs (class values), and each sign is repeated 27 times. However, the `her` sign only appears three times, whereas the `his-hers` sign appears 24 times. Therefore, we have assumed that they are the same sign and have considered them all as `his-hers` signs.
- **MAGIC Gamma Telescope** (`MAGIC`): Discrimination of the images of hadronic showers initiated by primary gammas from those caused by cosmic rays in the upper atmosphere [7]. The images of the hadronic showers captured by the telescope are preprocessed and modeled as ellipses. The predictive variables describe the shape of the ellipses. The dataset includes one angular variable that captures the angle of the major axis in the ellipse with the vector that connects the center of the ellipse with the center of the camera.
- `Arrhythmia`: Identification of the presence and absence of cardiac arrhythmia from electrocardiograms (ECG). The original dataset has 16 class values: one for healthy items, 14 types of cardiac arrhythmias and one class value for unclassified items [33]. We erased the unclassified items and built a binary class (normal vs. arrhythmia). The predictive variables describe clinical measurements, patient data and ECG recordings. The angular variables describe the vector angles from the front plane of four ECG waves. We removed variable 14, which had more than 83% missing values, and used Weka's `ReplaceMissingValues` filter [22] to fill in the missing values of variables 11–13 and 15 with the mode. We also removed some non-informative discrete and continuous variables.
- `Covertype`: Prediction of the kind of trees that grow in a specific area given some attributes describing the geography of the land [6]. The two angular variables describe the aspect (orientation) of the land from the true north and the slope of the ground. The original dataset has 581,012 samples and we used a Weka-supervised resampling method (without replacement) to reduce the dimensionality of the dataset to 100,000 samples.

Other datasets

- `Megaspores`: Classification of megaspores into two classes (their group in the biological taxonomy) according to the angle of their wall elements [43]. The dataset is an example included in Oriana software.[2]
- `Protein1`: Prediction of secondary structure including one aminoacid, using the dihedral angles (ϕ, ψ) of the residue as predictive information. We only considered α-helix and β-sheet structures, making the class binary. The data were retrieved from the protein geometry database [5].
- `Protein10`: Prediction of secondary structure including one aminoacid, using the dihedral angles (ϕ, ψ) and the planarity angle (ω). We considered the three angles in ten consecutive residues. We classified the four most common structures: α-helices, β-sheets, bends and turns. The data were retrieved from the protein geometry database [5].
- `Temperature`: Prediction of the outdoor temperature from the season, wind speed and wind direction. We used hourly measurements from a weather station located in the city of Houston. Data for the year 2010 were retrieved, and we removed the hours with missing values for any of the four variables. The information was collected from the Texas Commission on Environmental Quality website.[3] The class variable (outdoor temperature) was measured in degrees Fahrenheit and discretized into the following three values: low ($T \leq 50$), medium ($50 < T < 70$) and high ($T \geq 70$).

References

1. Agresti A (2007) An introduction to categorical data analysis, 2nd edn. Wiley, New York
2. Amayri O, Bouguila N. (2013) Beyond hybrid generative discriminative learning: spherical data classification. Pattern Anal Appl, in press
3. Banerjee A, Dhillon IS, Ghosh J, Sra S (2005) Clustering on the unit hypersphere using von Mises–Fisher distributions. J Mach Learn Res 6:1345–1382
4. Berens P (2009) CircStat: a MATLAB toolbox for circular statistics. J Stat Softw 31(10):1–21

[2] The Oriana software is available at: http://www.kovcomp.co.uk/oriana.

[3] The Texas Commission on Environmental Quality website is available at: http://www.tceq.state.tx.us.

5. Berkholz DS, Krenesky PB, Davidson JR, Karplus PA (2010) Protein geometry database: a flexible engine to explore backbone conformations and their relationships to covalent geometry. Nucleic Acids Res 38(Suppl 1):D320–D325
6. Blackard JA, Dean DJ (1999) Comparative accuracies of artificial neural networks and discriminant analysis in predicting forest cover types from cartographic variables. Comput Electron Agric 24(3):131–151
7. Bock RK, Chilingarian A, Gaug M, Hakl F, Hengstebeck T, Jiřina M, Klaschka J, Kotrč E, Savický P, Towers S, Vaiciulis A, Wittek W (2004) Methods for multidimensional event classification: a case study using images from a Cherenkov gamma-ray telescope. Nucl Instrum Methods in Phys Res Sect A-Accel Spectrom Detect Assoc Equip 516(2–3):511–528
8. Bøttcher SG (2004) Learning Bayesian networks with mixed variables. PhD thesis, Aalborg University
9. Bouckaert RR (2004) Estimating replicability of classifier learning experiments. In: Brodley CE (ed) Proceedings of the 21st international conference on machine learning, ACM
10. Damien P, Walker S (1999) A full Bayesian analysis of circular data using the von Mises distribution. Can J Stat Rev Can Stat 27(2):291–298
11. deHaas–Lorentz GL (1913) Die Brownsche Bewegung und einige verwandte Erscheinungen. Friedr. Vieweg und Sohn
12. Demšar J (2006) Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res 7:1–30
13. Devlaminck D, Waegeman W, Bauwens B, Wyns B, Santens P, Otte G (2010) From circular ordinal regression to multilabel classification. In: Proceedings of the 2010 workshop on preference learning, European conference on machine learning
14. Domingos P, Pazzani M (1997) On the optimality of the simple Bayesian classifier under zero–one loss. Mach Learn 29:103–130
15. Downs TD (2003) Spherical regression. Biometrika 90(3):655–668
16. Downs TD, Mardia KV (2002) Circular regression. Biometrika 89(3):683–697
17. Duda RO, Hart PE (1973) Pattern classification and scene analysis. Wiley, New York
18. Duda RO, Hart PE, Stork DG (2001) Pattern classification, 2nd edn. Wiley, New York
19. Eben K (1983) Classification into two von Mises distributions with unknown mean directions. Aplikace Matematiky 28(3):230–237
20. El Khattabi S, Streit F (1996) Identification analysis in directional statistics. Comput Stat Data Anal 23:45–63
21. Fayyad UM, Irani KB (1993) Multi-interval discretization of continuous-valued attributes for classification learning. In: Bajcsy R (ed) Proceedings of the 13th international joint conference on artificial intelligence, Morgan Kaufmann, San Mateo, pp 1022–1027
22. Figueiredo A (2009) Discriminant analysis for the von Mises–Fisher distribution. Commun Stat Simul Comput 38(9): 1991–2003
23. Figueiredo A, Gomes P (2006) Discriminant analysis based on the Watson distribution defined on the hypersphere. Stat: J Theor Appl Stat 40(5):435–445
24. Fisher NI (1987) Statistical analysis of spherical data. Cambridge University Press, Cambridge
25. Fisher NI (1993) Statistical analysis of circular data. Cambridge University Press, Cambridge
26. Fisher NI, Lee AJ (1992) Regression models for an angular response. Biometrics 48:665–677
27. Fisher RA (1953) Dispersion on a sphere. Proc R Soc Lond Ser A Math Phys Sci 217:295–305
28. Frank A, Asuncion A (2010) UCI machine learning repository. http://archive.ics.uci.edu/ml
29. Frank E, Trigg L, Holmes G, Witten IH (2000) Technical note: naive Bayes for regression. Mach Learn 41(1):5–25
30. Friedman N, Geiger D, Goldszmidt M (1997) Bayesian network classifiers. Mach Learn 29:131–163
31. García S, Herrera F (2008) An extension on "Statistical comparisons of classifiers over multiple data sets" for all pairwise comparisons. J Mach Learn Res 9:2677–2694
32. Guttorp P, Lockhart RA (1988) Finding the location of a signal: a Bayesian analysis. J Am Stat Soc 83:322–330
33. Güvenir HA, Acar B, Demiröz G, Çekin A (1997) A supervised machine learning algorithm for arrhythmia analysis. In: Murray A, Swiryn S (eds) Computers in cardiology 1997, pp 433–436
34. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009) The WEKA data mining software: an update. SIGKDD Explorations 11(1)
35. Hornik K, Grün B (2013) On conjugate families and Jeffreys priors for von Mises–Fisher distributions. J Stat Plan Infer 143(5): 992–999
36. Jaakola TS (1997) Variational methods for inference and estimation in graphical models. PhD thesis, Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology
37. Jammalamadaka SR, SenGupta A (2001) Topics in circular statistics. World Scientific, Singapore
38. Johnson RA, Wehrly TE (1978) Some angular-linear distributions and related regression models. J Am Stat Assoc 73(363):602–606
39. Jossinet J (1996) Variability of impedivity in normal and pathological breast tissue. Med Biol Eng Comput 34(5):346–350
40. Kadous MW (2002) Temporal classification: extending the classification paradigm to multivariate time series. PhD thesis, School of Computer Science and Engineering, University of New South Wales
41. Kato S, Shimizu K, Shieh G (2008) A circular-circular regression model. Stat Sin 18(2):633–645
42. Koller D, Friedman N (2009) Probabilistic graphical models. Principles and techniques. The MIT Press, Boston
43. Kovach WL (1989) Quantitative methods for the study of lycopod megaspore ultrastructure. Rev Palaeobot Palynol 57(3–4):233–246
44. Langley P, Sage S (1994) Induction of selective Bayesian classifiers. In: López de Mántaras R, Poole D (eds) Proceedings of the 10th conference on uncertainty in artificial intelligence, Morgan Kaufmann, San Mateo, pp 399–406
45. Lévy MP (1939) L'addition des variables aléatoires définies sur une circonférence. Bull Soc Math Fr 67:1–41
46. López-Cruz PL, Bielza C, Larranaga P (2011) The von Mises naive Bayes classifier for angular data. In: Proceedings of the 14th conference of the Spanish Association for Artificial Intelligence, CAEPIA 2011, LNCS 7023, pp 145–154
47. Mardia KV (1975) Statistics of directional data. J R Stat Soc Ser B Stat Methodol 37(3):349–393
48. Mardia KV (2006) On some recent advancements in applied shape analysis and directional statistics. In: Barber S, Baxter PD, Mardia KV (eds) Systems biology and statistical bioinformatics, Leeds University Press, Leeds, pp 9–17
49. Mardia KV (2010) Bayesian analysis for bivariate von Mises distributions. J Appl Stat 37(3):515–528
50. Mardia KV, Jupp PE (2000) Circular statistics. Wiley, New York
51. Minsky M (1961) Steps toward artificial intelligence. Proc Inst Radio Eng 49:8–30
52. von Mises R (1918) Uber die "Ganzzahligkeit" der Atomgewichte und verwandte Fragen. Phys Z 19:490–500
53. Mooney JA, Helms PJ, Jolliffe IT (2003) Fitting mixtures of von Mises distributions: a case study involving sudden infant death syndrome. Comput Stat Data Anal 41(3–4):505–513
54. Morales M, Rodríguez C, Salmerón A (2007) Selective naive Bayes for regression based on mixtures of truncated exponentials. Int J Uncertainty Fuzziness Knowl Based Syst 15(6):697–716
55. Morris JE, Laycock PJ (1974) Discriminant analysis of directional data. Biometrika 61(2):335–341

56. Pazzani MJ (1995) Searching for dependencies in Bayesian classifiers. In: Fisher D, Lenz HJ (eds) Learning from Data: Artificial Intelligence and Statistics V. In: Proceedings of the 5th International Workshop on Artificial Intelligence and Statistics, Springer, pp 239–248
57. Pearl J (1988) Probabilistic reasoning in intelligent systems. Morgan Kaufmann, San Mateo
58. Peot MA (1996) Geometric implications of the naive Bayes assumption. In: Horvitz E, Jensen FV (eds) Proceedings of the 12th conference on uncertainty in artificial intelligence, Morgan Kaufmann, San Mateo, pp 414–419
59. Pérez A, Larrañaga P, Inza I (2006) Supervised classification with conditional Gaussian networks: increasing the structure complexity from naive Bayes. Int J Approx Reason 43:1–25
60. Perrin F (1928) Étude mathématique du mouvement Brownien de rotation. Ann Sci Ec Norm Super 45:1–51
61. Rivest LP, Chang T (2006) Regression and correlation for 3×3 rotation matrices. Can J Stat Rev Can Stat 34(2):187–202
62. Romero V, Rumí R, Salmerón A (2006) Learning hybrid Bayesian networks using mixtures of truncated exponentials. Int J Approx Reason 42:54–68
63. Sahami M (1996) Learning limited dependence Bayesian classifiers. In: Simoudis E, Han J, Fayyad UM (eds) Proceedings of the 2nd international conference on knowledge discovery and data mining, AAAI Press, pp 335–338
64. SenGupta A, Roy S (2005) A simple classification rule for directional data. In: Balakrishnan N, Nagaraja HN, Kannan N (eds) Advances in ranking and selection, multiple comparisons, and reliability, statistics for industry and technology, Birkhäuser, Boston, pp 81–90
65. SenGupta A, Ugwuowo FI (2011) A classification method for directional data with application to the human skull. Commun Stat Theory Methods 40:457–466
66. Shenoy PP, West JC (2011) Inference in hybrid Bayesian networks using mixtures of polynomials. Int J Approx Reason 52(5):641–657
67. da Silva JE, Marques de Sá J, Jossinet J (2000) Classification of breast tissue by electrical impedance spectroscopy. Med Biol Eng Comput 38(1):26–30
68. Sra S (2012) A short note on parameter approximation for von Mises–Fisher distributions: and a fast implementation of $I_s(x)$. Comput Stat 27(1):177–190
69. Wood AT (1994) Simulation of the von Mises–Fisher distribution. Commun Stat Simul Comput 23(1):157–164
70. Zemel RS, Williams CKI, Mozer MC (1995) Lending direction to neural networks. Neural Netw 8(4):503–512