Bayesian Classification (CDT-34)

Preprint · May 2020

DOI: 10.13140/RG.2.2.24960.25602/1

1 author:

Luciano da F. Costa
University of São Paulo
734 PUBLICATIONS **13,303** CITATIONS

SEE PROFILE

Bayesian Classification (CDT-34)

Luciano da Fontoura Costa
luciano@ifsc.usp.br

São Carlos Institute of Physics – DFCM/USP

30th May 2020

Abstract

Bayesian decision theory, or Bayesian classification, represents a particularly important approach to supervised pattern recognition as it can be shown to be optimal, provided the involved probability density functions can be perfectly estimated, in the sense of minimizing the chances of misclassifications. In this work, we outline the basic principles of Bayesian classification, including Bayes theorem as well as illustrations of the classification method with respect to the number of categories and feature space dimensions.

"Truth is ever to be found in simplicity, and not in the multiplicity and confusion of things."

Isaac Newton.

1 Introduction

Supervised classification corresponds to a class of pattern recognition methods in which information about the existing classes is incorporated into the method during a training stage, often involving the presentation of objects with known categories and/or respective prototypes (e.g. [1, 2, 3, 4]).

Among the several existing approaches to supervised classification, Bayesian decision theory, or Bayesian classification, represents an important choice because it can be shown to be optimal regarding the chances of misclassifications. This, however, requires perfect estimation of the involved probability density functions. By using Bayes theorem, it is possible to derive a classification criterion that takes into account the conditional density probability functions of the considered measurements (features) given each specific category.

The Bayesian classification criterion is particularly simple and generic to any number of categories or dimension of the feature space, which further contribute to making this method particularly interesting. One of the reasons it is not even more applied is that it relies on a good estimation of the involved conditional, which cannot be always guaranteed, especially in situations in which relatively few samples of the objects are available.

We start by briefly presenting the concept of conditional density probability functions and respective estimation, and proceeds by deriving Bayes theorem and presenting the Bayesian decision criterion. Several examples considering 2 or 3 categories and feature spaces with 1 or 2 dimensions are provided.

2 Conditional Density Probability Functions

A density probability function on M random variables X_i, $i = 1, 2, \ldots, M$, organized as a feature vector $\vec{X}$, is a non-negative scalar field $p(\vec{X})$ so that:

$$\int_{-\infty}^{\infty} p(\vec{X}) d\vec{X} = 1 \tag{1}$$

The set of features chosen to characterize the entities to be recognized provides a good example of random vector $\vec{X}$, which will have an associated density probability function $p(\vec{X})$ (e.g. [4]).

Given C categories, named c_i with $i = 1, 2, \ldots, C$, we can define the conditional density probability functions of the measurements in $\vec{X}$ with respect to each of the existing categories as:

$$p_1(\vec{X}|c_1), p_2(\vec{X}|c_2), \ldots, p_c(\vec{X}|c_C) \tag{2}$$

For instance, we can have two types of fruits, let's say grapes (c_1) and oranges (c_2), characterized in terms of

$M = 2$ measurements corresponding to their weight (X_1) and length (X_2), so that $\vec{X} = [X_1, X_2]^T$. In this particular case, we ill have the following conditional probability density functions:

$$p_1(\vec{X}|c_1), p_2(\vec{X}|c_2) \tag{3}$$

Bayesian classification require these conditional probabilities as input, so we need some means for estimating them. There are two main ways in which this can be accomplished: (i) *parametric*; and (ii) *non-parametric*.

In the case of parametric estimation, the nature of the density is known. For instance, we may know that the variables at hand follow a normal or uniform distribution. What remains to be estimated are their respective parameters, hence the name parametric estimation (e.g. [5, 6]).

Contrariwise, in non-parametric estimation we do not know about the type of density, so it needs to be estimated from the data. This can be done in several ways, such as by obtaining a multidimensional relative frequency histogram of the random vector variables in question, or by using some interpolation scheme such as those based on kernels. For instance, give a set of finite samples of the random vector, each represented by a respective Dirac delta function, a respective density probability function can be estimated by convolving these delta Dirac functions with a suitable kernel, such as a normal density (e.g. [6]).

3 Bayes Theorem

The conditional probability of observing an event B given that another event A has already occurred is expressed as $P(B|A)$. These two events are subsets of a sample space Ω. The preliminary occurrence of even A can be understood as a *redefinition* of the sampling space from Ω to A, implying the subsequent event B to become effectively restricted to the intersection between B and the new sampling space A, i.e.:

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \tag{4}$$

By symmetry, we also have:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{5}$$

Combining the two previous equations, we can derive that:

$$\frac{P(A|B)}{P(B|A)} = \frac{\frac{P(A \cap B)}{P(B)}}{\frac{P(A \cap B)}{P(A)}} = \frac{P(A)}{P(B)} \tag{6}$$

The above equation corresponds to Bayes theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{7}$$

The main importance of this theorem is that it allows us to change the time dependence between the two subsequent events.

4 Bayesian Decision Criterion

Consider we have two classes c_1 and c_2, both described by a feature vector $\vec{X}$, as well as the probabilities

$p(c_1|\vec{X}) \iff$ probability of c_1 given the features $\vec{X}$
$p(c_2|\vec{X}) \iff$ probability of c_2 given the features $\vec{X}$

We can define the following likelihood ratio:

$$r = \frac{p(c_1|\vec{X})}{p(c_2|\vec{X})}$$

It is reasonable to adopt as classification criterion taking class c_1 whenever $r > 1$.

However, in typical application problems, the probabilities $p(c_1|\vec{X})$ and $p(c_2|\vec{X})$ are not available. However, we typically can estimate $p(\vec{X}|c_1)$ and $p(\vec{X}|c_2)$ by using a parametric or non-parametric approach. Bayes theorem provides the critical link between these two types of probabilities, more specifically, we can write:

$$p(c_1|\vec{X}) = \frac{p(\vec{X}|c_1)P(c_1)}{p(\vec{x})}$$
$$p(c_2|\vec{X}) = \frac{p(\vec{X}|c_2)P(c_2)}{p(\vec{x})}$$

therefore establishing the link between the *a priori* and *a posteriori* probabilistic information. Now, we can rewrite the likelihood ratio as:

$$r = \frac{p(c_1|\vec{X})}{p(c_2|\vec{X})} = \frac{\frac{p(\vec{X}|c_1)P(c_1)}{p(\vec{x})}}{\frac{p(\vec{X}|c_2)P(c_2)}{p(\vec{x})}} = \frac{p(\vec{X}|c_1)P(c_1)}{p(\vec{X}|c_2)P(c_2)}$$

Thus, if $r > 1$ we take class c_1, otherwise we take c_2. Bayes decision theory consists in the systematic application of the above criterion to as many categories as necessary and considering feature vectors in any dimension. It can be formally shown (e.g. [1, 3]) that the above criterion minimizes the probability of misclassifications, implying Bayesian classification to be optimal regarding this important aspect. It is important to keep in mind that this optimality property depends on having access, or being able to estimate, the exact conditional probability distributions implied by the Bayesian approach.

Summarizing the above results, we have that given C classes c_i, $i = 1, 2, \ldots, C$, and an M-dimensional feature vector $\vec{X}$, the respective conditional probability density functions $p(\vec{X}|c_i$ and mass probabilities (c_i), as well as a new object with respective feature vector $\vec{Y}$ to have its

category determined, the Bayesian criterion for choosing the category of Y can be expressed as:

$$\text{take } c_i \text{ so that } p(\vec{X}|c_i)P(c_i) \geq p(\vec{X}|c_j)P(c_j)$$

for any $j = 1, 2, \ldots, C$.

Several examples of application of this criterion, considering one and two-dimensional feature spaces, as well as several categories, are provided in the following sections, but before that we have a brief discussion on the errors potentially involved in Bayesian classification.

5 Misclassification Errors

An important point to bear in mind is that the optimality of the Bayesian classification approach by no means guarantee that no errors will be made, but only that the chance of such misclassification errors will be minimized.

Therefore, it is interesting to briefly discuss what errors are involved in Bayesian classification (as well as any other supervised classification approach).

Consider the situation shown in Figure 1.

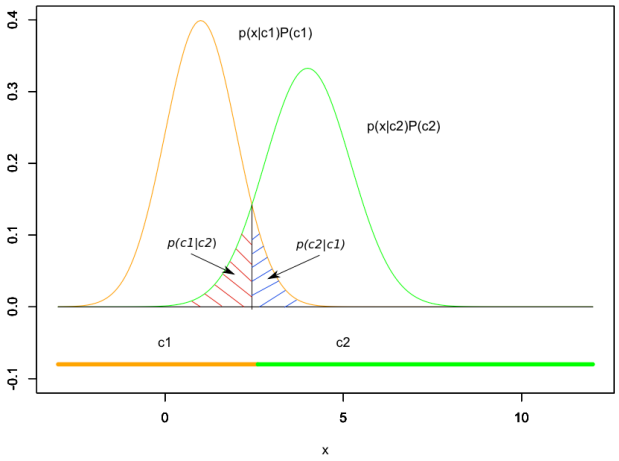


Figure 1: The conditional probability functions, respectively weighted by the mass class probabilities, of two hypothetical categories of entities to be recognized. The blue hashed area corresponds to the probability of deciding on class c_2 given that the original entity was of class c_1, which corresponds to a classification error. The red hashed area corresponds to the probability of incorrectly classifying an object from class c_2 as being of class c_1.

The blue hashed area corresponds to the probability of taking an entity from class c_1 as being of class c_2, while the red hashed area is the probability of taking an object from class c_2 as belonging to class c_1. The Bayesian classification approach minimizes the sum of these two errors.

Observe that the misclassification errors are related to the overlap between the involved conditional probability functions.

6 One-Dimensional Feature Spaces

Figure 2 depicts the conditional one-dimensional probability density functions, weighted by respective mass probabilities, of two hypothetical types of objects to be recognized while taking into account the respective values of a feature X.

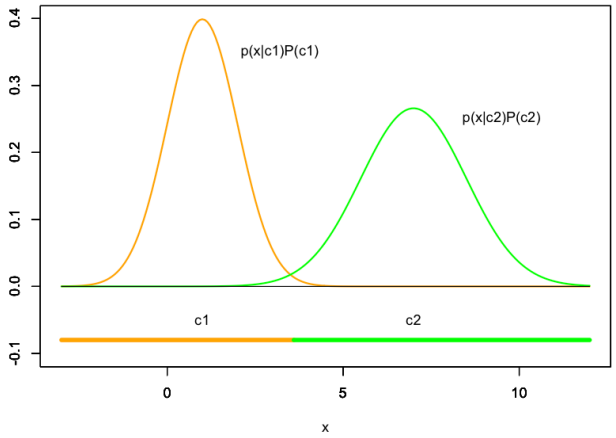


Figure 2: Bayesian decision regions defined by two one-dimensional conditional probability density functions, weighted by respective mass probabilities.

The decision regions implied by the application of the Bayesian decision theory are shown underneath the respective densities. Whenever a new object has a measurement falling within each of these intervals, it is taken to have the respectively associated category. For instance, if a new object has $X = 7$, it will be taken to belong to class c_2.

Figure 3 illustrates another interesting situation involving two categories characterized by the one-dimensional random variable X.

We observe that Bayesian classification does not imply the obtained classification regions to be continuous. For instance, in the case of the previous figure, we have that the decision region associated to category c_1 has been split into two separated regions.

Figure 4 shows a classification problem involving three categories with objects characterized in terms of the one-dimensional feature X.

7 Multi-Dimensional Feature Spaces

Bayesian classification can be straightforwardly extended to feature vectors of any dimension. For instance, Figure 5 illustrates the application of the Bayesian approach to

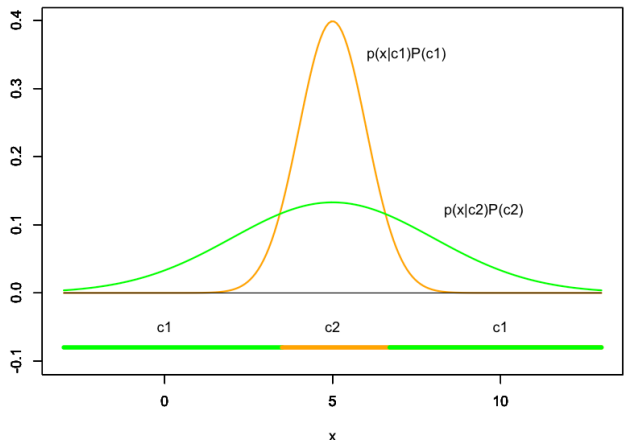


Figure 3: The classification regions obtained by the application of the Bayesian decision criterion do not need to be continuous for each category.

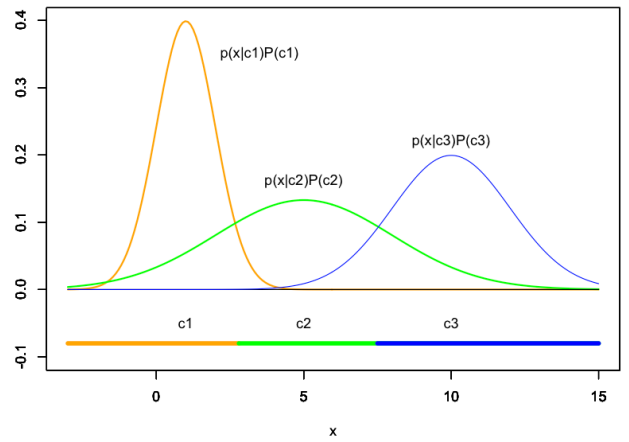


Figure 4: Bayesian classification extends naturally to any number of involved categories, three in the case of this example.

two categories characterized by a two-dimensional feature vector $\vec{X} = [X_1, X_2]^T$.

In (a) we have the two conditional probability density functions $p(\vec{X}|c_1)P(c_1)$ in orange) and $p(\vec{X}|c_2)P(c_2)$ (in green). The respectively obtained decision regions are shown in (b).

Figure 6 shows a Bayesian classification problem involving three categories characterized by respective two-dimensional feature vector $\vec{X} = [X_1, X_2]^T$. As with the one-dimensional case, non-contiguous classification regions can be obtained in the feature space (X_1, X_2) domain.

8 Concluding Remarks

Supervised classification is an important task in pattern recognition, with ample applications in the most diverse areas. Among the several methods for supervised classification, Bayesian classification stands out as a reference because of its optimality regarding the minimization of misclassifications, provided the involved conditional probability densities can be exactly estimated.

In the present work, we developed a concise introduction to Bayesian classification, starting with a brief review of density probability functions and the Bayes theorem, and proceeding with the presentation of the Bayesian decision criterion. Several examples concerning 2 or 3 categories in one and two-dimensional feature spaces have also been presented as a means to help with the familiarization of the concepts and presented methodology.

It should be taken into account that, despite its optimality, Bayesian decision theory is in practice constrained by the accuracy possibly achieved while estimating of the respectively involved multivariate probability density and mass functions. In such situations, it is interesting to consider alternative supervised classification approaches such as neuronal networks (e.g. [7]).

Acknowledgments.

Luciano da F. Costa thanks CNPq (grant no. 307085/2018-0) for sponsorship. This work has benefited from FAPESP grant 15/22308-2. The author thanks S. H. Hawley for commenting on this text.

References

[1] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley Interscience, 2000.

[2] K. Koutrombas and S. Theodoridis. *Pattern Recognition*. Academic Press, 2008.

[3] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[4] L. da F. Costa. Pattern cognition, pattern recognition. Researchgate, Dec 2019. `https://www.researchgate.net/publication/338168835_Pattern_Cognition_Pattern_Recognition_CDT-19`. [Online; accessed 29-Feb-2020].

[5] L. da F. Costa. Statistical modeling. `https://www.researchgate.net/publication/334726352_Statistical_Modeling_CDT-13`, 2019. [Online; accessed 10-Apr-2020.].

[6] L. da F. Costa. Multivariate statistical modeling. `https://www.researchgate.net/publication/`

340442989_Multivariate_Statistical_Modeling_CDT-26, 2019. [Online; accessed 10-Apr-2020.].

[7] S. Haykin. *Neural Networks And Learning Machines*. Pearson, 3rd edition, 2018.

Costa's Didactic Texts – CDTs

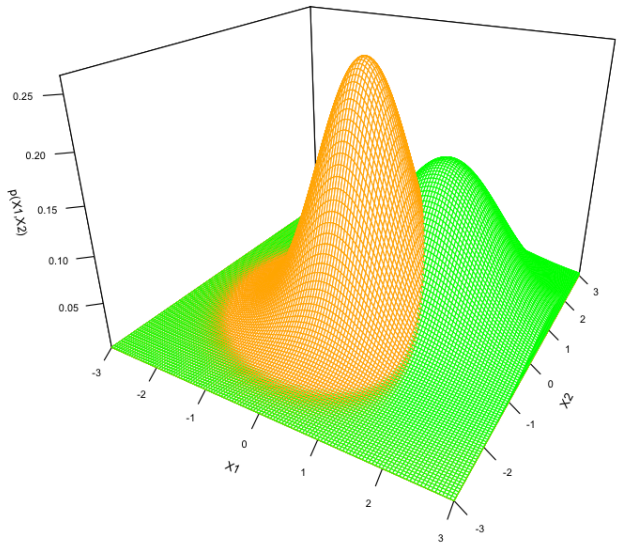
CDTs intend to be a halfway point between a formal scientific article and a dissemination text in the sense that they: (i) explain and illustrate concepts in a more informal, graphical and accessible way than the typical scientific article; and (ii) provide more in-depth mathematical developments than a more traditional dissemination work.

It is hoped that CDTs can also incorporate new insights and analogies concerning the reported concepts and methods. We hope these characteristics will contribute to making CDTs interesting both to beginners as well as to more senior researchers.

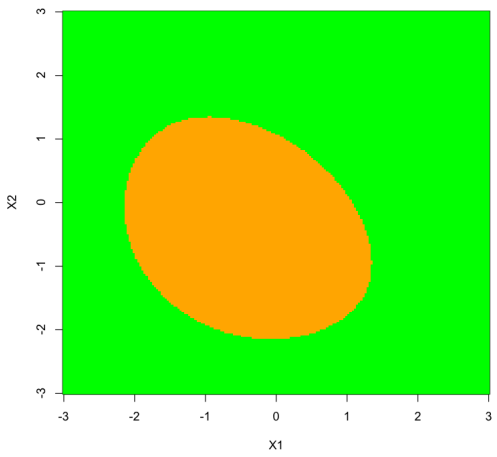
Each CDT focuses on a limited set of interrelated concepts. Though attempting to be relatively self-contained, CDTs also aim at being relatively short. Links to related material are provided in order to provide some complementation of the covered subjects.

Observe that CDTs, which come with absolutely no warranty, are non distributable and for non-commercial use only.

Please check for new versions of CDTs, as they can be revised. Also, CDTs can be cited, e.g. by including the respective DOI. The complete set of CDTs can be found at: `https://www.researchgate.net/project/Costas-Didactic-Texts-CDTs`.



(a)



(b)

Figure 5: (a) : The conditional density functions $p(\vec{X}|c_1)P(c_1)$ (in orange) and $p(\vec{X}|c_2)P(c_2)$ (in green), characterized by respective feature vector $\vec{X} = [X_1, X_2]^T$. (b) The respectively obtained classification regions in the space (X_1, X_2).

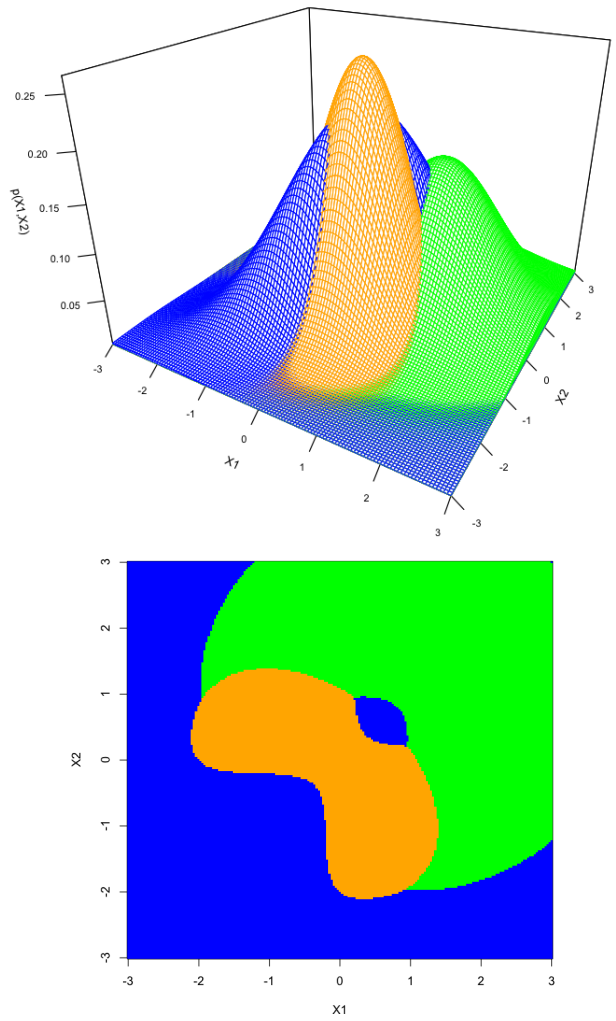


Figure 6: A situation involving Bayesian classification of three categories characterized by respective two-dimensional feature vector $\vec{X} = [X_1, X_2]^T$.