Discrete Applied Mathematics 160 (2012) 1441–1452

Contents lists available at SciVerse ScienceDirect

Discrete Applied Mathematics

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

Minimizing the error of linear separators on linearly inseparable data

Boris Aronov [a], Delia Garijo [b], Yurai Núñez-Rodríguez [c], David Rappaport [d], Carlos Seara [e,*], Jorge Urrutia [f]

[a] Department of Computer Science and Engineering, Polytechnic Institute of NYU, USA
[b] Dept. de Matemática Aplicada I, Universidad de Sevilla, Spain
[c] Lakes Environmental Software, 60 Bathurst Dr. Unit 6, Waterloo, ON, N2V2A9, Canada
[d] School of Computing, Queen's University, Kingston, Canada
[e] Dept. de Matemàtica Aplicada II, Universitat Politècnica de Catalunya, Spain
[f] Instituto de Matemáticas, Universidad Nacional Autónoma de México, Mexico

ARTICLE INFO

Article history:
Received 27 August 2011
Received in revised form 28 February 2012
Accepted 7 March 2012
Available online 30 March 2012

Keywords:
Linearly inseparable
Classifiers
Error minimizers

ABSTRACT

Given linearly inseparable sets R of red points and B of blue points, we consider several measures of how far they are from being separable. Intuitively, given a potential separator ("classifier"), we measure its quality ("error") according to how much work it would take to move the misclassified points across the classifier to yield separated sets. We consider several measures of work and provide algorithms to find linear classifiers that minimize the error under these different measures.

1. Introduction

Current massive data collection methods have provided researchers with a wealth of data, together with the challenge of making sense of it. Partitioning or clustering data as a method of data analysis is an important tool in providing meaning to large amounts of data. Performing this type of analysis is multifaceted, and can range from applications in geography and land use, pattern recognition, medical health studies, economics, detecting similarity between genres of music, and data mining to assist in targeted marketing strategies, to name but a few.

Partitioning data using separators or classifiers to perform cluster analysis on training sets is a standard technique, for example it is used in pattern recognition applications [22]. Thus the problem of determining if two disjoint point sets are separable has been widely studied in the literature. See for instance Megiddo [35] for the seminal fixed dimension linear programming method for linear separability. Elizondo [24] surveys a variety of techniques and discusses the application of linear separability to machine learning. O'Rourke et al. [36] and Boissonnat et al. [7] consider algorithms for circular separability, and Hurtado et al. [30] and Arkin et al. [3] examine a variety of separability criteria.

In some applications the training data may contain some points that have been misclassified resulting in the situation where no natural[1] partition scheme classifies the data. In this case classification is attempted where some amount of error is tolerated. Within that context, Aronov and Har-Peled [4] studied the following problem: given a bicolored point set, find a ball that contains the maximum number of red points without containing any blue points. Cortés et al. [18] address the

* Corresponding author.
E-mail addresses: aronov@poly.edu (B. Aronov), dgarijo@us.es (D. Garijo), yurai@cs.queensu.ca (Y. Núñez-Rodríguez), daver@cs.queensu.ca (D. Rappaport), carlos.seara@upc.edu (C. Seara), urrutia@matem.unam.mx (J. Urrutia).

[1] Here by a *natural partition* we mean a partition given by a simple classifier like disk, square, or a halfplane.

0166-218X/$ – see front matter
doi:10.1016/j.dam.2012.03.009

problem of finding two boxes S_R and S_B such that the number of red and blue points in S_R and S_B respectively is maximized, while ignoring the points in $S_R \cap S_B$. Mathematical programming techniques have been used in the operations research community to solve similar problems [5,19,31,40].

In this paper we present algorithms that minimize the error when using a linear separator. Given two linearly inseparable point sets we attempt to find a hyperplane which splits the union of the sets into disjoint subsets in such a way that some *error functions* are minimized. We call such hyperplanes *optimal classifiers*. The notion of *optimality* is left intentionally informal as the precise properties that should be optimized are application dependent. We will examine several different criteria for choosing an optimal classifier. We will proceed on the assumption that the dimension d of the problem is a small constant and be mostly concerned about the asymptotic dependence of the speed of our algorithms on the size n of the point sets.

Let R be a set of r red points and B a set of b blue points in $\mathbb{R}^d$. Let $n := r + b$ be the total number of points and assume that the point sets are *disjoint* and *in general position*, that is, no $d + 1$ of the points lie in the same hyperplane in $\mathbb{R}^d$. We say that R and B are *(linearly) separable* if there exists a *(linear) separator*, which is an oriented hyperplane so that the red points lie to its left and the blue points lie to its right. (Formally, each side of the hyperplane is a closed halfspace delimited by it, so points on the hyperplane are considered to lie on both sides simultaneously.) If there is no separator for R and B, then we say that the sets are *inseparable*.

Let $P = \{p_1, \dots, p_n\} := R \cup B$. Let h be a hyperplane $x_1a_1 + \cdots + x_da_d = a_0$, and let h^- be the halfplane containing the points $(x_1, \dots, x_d)$ such that $x_1a_1 + \cdots + x_da_d \le a_0$, and h^+ be the halfspace that contains the points satisfying $x_1a_1 + \cdots + x_da_d \ge a_0$. We will say that h^- lies to the left of h, while h^+ lies to the right of h. If h were a separator of P, we would have $R \subset h^-$ and $B \subset h^+$. As it is not, it *misclassifies* the red points $R(h) := R \setminus h^-$ and the blue points $B(h) := B \setminus h^+$. We use $\Xi = \Xi(h) := R(h) \cup B(h)$ to denote the set of points misclassified by h. We use $s(h)$ to represent the quality of h as a classifier; it depends on h and $\Xi = \Xi(h)$. Our goal is to find a hyperplane that minimizes the cost under one of the following four measures, where $d(\cdot, \cdot)$ denotes the Euclidean distance between points in $\mathbb{R}^d$ and $d(p, X)$ denotes the Euclidean distance from a point p to a set X:

MinMax: Maximum Euclidean distance from h to a point in Ξ, i.e.,

$$s_\infty(h) := \max_{p\in\Xi(h)} d(p, h) = \max\{\max_{p\in R} d(p, h^-), \max_{p\in B} d(p, h^+)\}.$$

MinSum: Sum of the Euclidean distances from h to points in Ξ, i.e.,

$$s_1(h) := \sum_{p\in\Xi(h)} d(p, h) = \sum_{p\in R} d(p, h^-) + \sum_{p\in B} d(p, h^+).$$

MinSum^2: Sum of squares of the Euclidean distances from h to points in Ξ, i.e.,

$$s_2(h) := \sum_{p\in\Xi(h)} d^2(p, h) = \sum_{p\in R} d^2(p, h^-) + \sum_{p\in B} d^2(p, h^+).$$

MinMis: Just the cardinality of Ξ, i.e., the number of misclassified points,

$$s_0(h) = |R \setminus h^-| + |B \setminus h^+|.$$

We are interested in finding an optimal classifier, which we define to be a halfspace h_{OPT} minimizing the quantity $s(h)$; it is not always unique. Notice that since $d(p, h^\pm)$ is a continuous function of h, so are $s_\infty(h)$, $s_1(h)$, and $s_2(h)$.

We will use a standard duality transform. It maps a point $p \in \mathbb{R}^d$ to a non-vertical hyperplane $p^* \subset \mathbb{R}^d$, and vice versa, that is, it maps a non-vertical hyperplane h to the point h^* such with $(h^*)^* = h$ and $(p^*)^* = p$; moreover p is above h if and only if h^* is above p^*.

Outline of the paper. We present algorithms to find optimal classifiers using the four measures described above. In Section 2 we present solutions for the one-dimensional problem, as this will provide some illuminating intuition for proceeding to problems of higher dimension. We devote Section 3 to describing some crucial observations that relate the separability problems in one dimension to those in higher dimensions. In Sections 4–7 we study each of the measures in arbitrary dimension. The algorithms are based on existing techniques from the computational geometry literature. We show that finding an optimal classifier using the MinMax measure is equivalent to determining the penetration depth between two convex polyhedra, and can therefore be solved using existing methods. For optimizing classifiers using the MinSum, MinSum^2, and MinMis measures we use duality and levels in arrangements to systematically enumerate candidate solutions. The computational complexities of the algorithms are summarized in the following Table 1.

2. One dimension

We first consider the one-dimensional case of our set of problems. The input sets R and B lie on the real line. Then a classifier is a point h. We will assume that h^+ is the half-line $[h, +\infty)$ and h^- is the half-line $(-\infty, h]$; the reverse case is handled by a symmetric argument. For simplicity, we will omit the symmetric cases in the statement of our lemmas. We seek the point (or points) h_{OPT} minimizing $s(h)$.

Notice that $d(p, h^+)$ is convex as a function of h, as is its square ($(p - h)^2$ for $h \ge p$, and 0 for $h < p$); the same holds for $d(p, h^-)$. Since the first three error measures are defined as the maximum, sum, and the sum of squares of these functions