Measuring the error of linear separators on linearly inseparable data

Boris Aronov* Delia Garijo† Yurai Núñez-Rodríguez‡ David Rappaport§
Carlos Seara¶ Jorge Urrutia‖

Given two sets of red and blue points in $\mathbb{R}^d$, we say that they are *linearly separable* if there exists a hyperplane, or *linear separator*, that partitions $\mathbb{R}^d$ such that each part contains only red or only blue points. If there is no such separator, then we say that the two sets are *linearly inseparable*. In this work, given inseparable sets, we seek *approximate separators* with the "best" approximation. The notion of "best" is left intentionally informal as the precise properties that should be optimized are application dependent.

Let R be a set of p red points and B a set of q blue points in $\mathbb{R}^d$. Let $n = p + q$ and assume that the points are in general position, that is, no $d + 1$ of the points lie in the same hyperplane in $\mathbb{R}^d$. Let H denote a hyperplane which misclassifies $R \cup B$ and partitions it into two non-empty subsets: the *left* subset in which the red points are *well classified* and the blue points are *misclassified*, and the *right* subset which plays a complementary role. Given $R \cup B$ and H we may be left with a set $\Omega \subset R \cup B$ of points misclassified by H. We use $s(H)$ to represent the quality of H as an approximate separator, the *cost* of H. Our main goal is to study different criteria for finding approximate separators that minimize the cost under one of the following assumptions:

1. $s(H)$ is the maximum distance from H to a point in Ω: MinMax criterion.
2. $s(H)$ is the sum of distances from H to every point in Ω: MinSum criterion.
3. $s(H)$ is the sum of squares of distances from H to every point in Ω: MinSum^2 criterion.
4. $s(H)$ is the cardinality of Ω: MinMis criterion.

The complexities obtained in this work for computing optimal separators assuming each of the above criteria in different dimensions are shown in Table 1. For the sake of brevity we only describe the main ideas and results obtained for the MinSum criterion.

Table 1: Summary of results.

Dimension	MinMax	MinSum	MinSum^2	MinMis
$d = 1$	$\Theta(n)$	$\Theta(n)$	$\Theta(n)$	$\Theta(n \log n)$
$d = 2$	$\Theta(n \log n)$	$O(n \log n + n^{4/3} \log^{1+\epsilon} n)$ $O(n \log n + n^{4/3})$ $(*)$	$O(n^2)$	$O(n^2)$
$d = 3$	$O(n^2)$	$O(n^{5/2} \log^6 n)$ $(*)$	$O(n^3)$	$O(n^3)$
$d \geq 4$	$O(n^{\lceil d/2 \rceil})$	$O(n^d)$	$O(n^d)$	$O(n^d)$

$(*)$ expected time

A one-dimensional algorithm for finding an optimal approximate separator offers intuitive insight to solutions in higher dimensions. This phenomenon is easily explained by the following observation.

*Dept. of Computer and Information Science, Polytechnic University, USA, aronov@poly.edu.
†Depto. de Matemtica Aplicada I, Universidad de Sevilla, Spain, dgarijo@us.es.
‡School of Computing, Queen's University, Kingston, Canada, yurai@cs.queensu.ca.
§School of Computing, Queen's University, Kingston, Canada, daver@cs.queensu.ca.
¶Dept. de Matemàtica Aplicada II, Universitat Politècnica de Catalunya, Spain, carlos.seara@upc.edu.
‖Instituto de Matemáticas, Universidad Nacional Autónoma de México, Mexico, urrutia@matem.unam.mx.

Remark 1. *Consider an optimal approximate separator H for P in $\mathbb{R}^d$. Let P^- (H^-) denote the projection of P (resp. H) to a line perpendicular to H. Then H^- is an optimal approximate separator of P^- and furthermore $s(H) = s(H^-)$.*

Therefore every optimal solution in higher dimensions has an equivalent one-dimensional solution, but the number of candidate solutions that have to be evaluated to determine an optimal one usually increases with the dimension. Thus, we first consider the problem in $\mathbb{R}$. Let R and B be inseparable sets of red and blue points on a line. The following lemma characterizes the solution of the MinSum problem for dimension $d = 1$.

Lemma 1. *An optimal separator point a lies between two consecutive points of $R \cup B$ such that the number of misclassified blue points is equal to the number of misclassified red points.*

Lemma 2. *Let $|R| = p$, $|B| = q$. The* MinSum *problem in $\mathbb{R}$ has infinitely many optimal solutions. An optimal separator is any point a of the closed interval defined by the two consecutive points of $R \cup B$ in the positions p and $p + 1$ counting from left to right.*

Theorem 1. *The one-dimensional* MinSum *problem can be solved in $\Theta(n)$ time.*

We now consider the problem in $\mathbb{R}^2$. Let ℓ be an optimal separator line for R and B according to the MinSum criterion. Let ℓ^+ (ℓ^-) be the open half-plane above (below) ℓ. The following lemma is a straightforward consequence of Remark 1 and Lemmas 1 and 2.

Lemma 3. *The number of misclassified blue points on ℓ^+ is equal to the number of misclassified red points on ℓ^-. The two-dimensional* MinSum *problem has an infinite number of optimal separators.*

Theorem 2. *The two-dimensional* MinSum *problem can be solved deterministically in time $O(n \log n + n^{4/3} \log^{1+\epsilon} n)$ for an arbitrarily small constant $\epsilon > 0$ or in $O(n \log n + n^{4/3})$ expected time.*

We can extend the two-dimensional discussion to three dimensions obtaining the $O(n^{5/2} \log^6 n)$ expected time specified in Table 1.

Let $\mathcal{A}(P)$ be the dual arrangement of lines obtained by dualizing the n points of the set $R \cup B$. By Lemmas 2 and 3 an optimal separator always exists and can be found between the p and $(p+1)$-levels of $\mathcal{A}(P)$. This fact holds for any dimension. Thus, for dimension $d \geq 4$, the upper bound for the size of the p-level is only slightly better than $O\left(n^{\lfloor d/2 \rfloor} p^{\lceil d/2 \rceil}\right)$ [2]. More concretely, an upper bound is $O(n^{d-\alpha_d})$ for a very small $\alpha_d = 1/(4d-3)^d$. As Agarwal *et al.* [1] observed, the bound can be made sensitive to p, namely $O(n^{\lfloor d/2 \rfloor} p^{\lceil d/2 \rceil - \alpha_d})$. Matoušek *et al.* [3] give an $O(n^{4-2/45})$ upper bound for $d = 4$.

Theorem 3. *The d-dimensional* MinSum *problem can be solved deterministically in time*

$$O\left(n^{\lfloor d/2 \rfloor} p^{\lceil d/2 \rceil} \left(\frac{\log n}{\log p}\right)^{O(1)}\right) = O(n^d).$$

References

[1] P. K. Agarwal, B. Aronov, T. M. Chan, and M. Sharir. On levels in arrangements of lines, segments, planes, and triangles. *Discrete Comput. Geom.* 19, 1998, 315–331.

[2] K. L. Clarkson and P. W. Shor. Applications of random sampling in computational geometry, II. *Discrete Comput. Geom.* 4, 1989, 387–421.

[3] J. Matoušek, M. Sharir, S. Smorodinsky, and U. Wagner. On k-sets in four dimension. *Discrete Comput. Geom.* 35, 2006, 177–191.