

Optimal distance separating halfspace ^{*}

Frank Plastria[†] Emilio Carrizosa[‡]

Sept 17 2002

Abstract

One recently proposed criterion to separate two datasets in discriminant analysis, is to use a hyperplane which minimises the sum of distances to it from all the misclassified data points. Here all distances are supposed to be measured by way of some fixed norm, while misclassification means lying on the wrong side of the hyperplane, or rather in the wrong halfspace. In this paper we study the problem of determining such an optimal halfspace.

In dimension d , we prove that there always exists an optimal separating halfspace passing through d affinely independent data points. This directly shows that the problem is polynomially solvable in fixed dimension by an algorithm of $O(n^{d+1})$.

If a different norm or gauge is used for each dataset in order to measure distances to the hyperplane, or if all distances are measured by a fixed (asymmetric) gauge, then one can still show that there always exists an optimal separating halfspace passing through $d - 1$ affinely independent data points.

The one-dimensional problem is extremely easy to solve: it suffices to find a balancing separating point, i.e. yielding an equal number (or weight) of misclassifieds for each dataset. It also follows that in any dimension any optimal separating halfspace always balances the misclassified points, where the balancing criterion now takes the shape of the used gauges into account.

Keywords. norm-distance to hyperplane, separating halfspace, discriminant analysis.

^{*}This paper was started in Oberwolfach, Germany in december 2001, and completed while the first author was visiting the Departamento de Matematica at the University of Bologna, Italy in september 2002. Both institutes are gratefully acknowledged for their support.

[†]BEIF - Department of Management Informatics, Vrije Universiteit Brussel, Pleinlaan 2, B 1050 Brussels, Belgium, e-mail: Frank.Plastria@vub.ac.be

[‡]Facultad de Matemáticas, Universidad de Sevilla, Tarfia s/n, 41012 Sevilla, Spain, e-mail: ecarrizosa@us.es

1 Gauge Distance to a Hyperplane

Let γ be a gauge on \mathbb{R}^d with unit ball B , i.e. B is a compact convex set containing the origin in its interior such that

$$\gamma(x) = \min\{ t \geq 0 \mid x \in tB \},$$

see e.g. [8, 3, 4]. Given a hyperplane H in \mathbb{R}^d the γ -distance of a point $a \in \mathbb{R}^d$ to H is defined as

$$d_\gamma(a, H) \stackrel{\text{def}}{=} \min\{ \gamma(x - a) \mid x \in H \}.$$

Let γ° be the dual (or polar) gauge of γ , given by

$$\gamma^\circ(v) \stackrel{\text{def}}{=} \max\{ \langle v ; y \rangle \mid \gamma(y) \leq 1 \},$$

which is well-defined and also a gauge on (the dual space of) \mathbb{R}^d , see e.g. [8]. This definition directly implies the following well-known generalized Cauchy-Schwartz inequality, see e.g. [8], p. 129.

$$\langle v ; y \rangle \leq \gamma^\circ(v)\gamma(y) \quad \forall v, y \in \mathbb{R}^d \quad (1)$$

in which for any fixed $v \neq 0$ equality holds iff $y = \lambda z$ for some $\lambda \geq 0$ and some $z \in \partial\gamma^\circ(v)$, where $\partial\gamma^\circ(v)$ denotes the (nonempty) subdifferential of the dual gauge at v , see e.g. [6]. Note that equality in (1) for $v, y \neq 0$ also implies $\langle v ; y \rangle > 0$.

We will denote the hyperplane of equation $\langle u ; x \rangle = \beta$ ($u \neq 0$) by $H^\circ(u, \beta)$. The following theorem, already obtained in [7] gives a simple expression for the gauge distance to a hyperplane. In order to be self contained we reproduce the proof here.

Theorem 1 *For any gauge γ and any hyperplane $H^\circ(u, \beta)$ we have*

$$d_\gamma(a, H^\circ(u, \beta)) = \begin{cases} \frac{\beta - \langle u ; a \rangle}{\gamma^\circ(u)} & \text{when } \langle u ; a \rangle \leq \beta \\ \frac{\langle u ; a \rangle - \beta}{\gamma^\circ(-u)} & \text{when } \langle u ; a \rangle > \beta \end{cases}$$

Any γ -closest point of $H^\circ(u, \beta)$ to a is found as the unique intersection point of $H^\circ(u, \beta)$ with the line through a having as direction any subgradient of γ° at u when $\langle u ; a \rangle \leq \beta$, and at $-u$ when $\langle u ; a \rangle > \beta$.

Proof. Assume first that $\langle u ; a \rangle \leq \beta$. For any $x \in H^\circ(u, \beta)$ we have, after substituting v by u ($\neq 0!$) and y by $x - a$ in the generalized Cauchy-Schwartz inequality (1), that

$$\gamma(x - a) \geq \frac{\langle u ; x - a \rangle}{\gamma^\circ(u)} = \frac{\beta - \langle u ; a \rangle}{\gamma^\circ(u)}.$$

where equality happens at $x \in H^\circ(u, \beta)$ iff

$$x - a \text{ is of the form } \lambda z, \text{ for some } z \in \partial\gamma^\circ(u). \quad (2)$$

Moreover, such an x exists. Indeed, since $u \neq 0$, for any given $z \in \partial\gamma^\circ(u)$ we have $\gamma^\circ(z) = 1$, so by (1) that

$$\langle u ; z \rangle = \gamma(z)\gamma^\circ(u) = \gamma^\circ(u) > 0,$$

thus the function $\lambda \geq 0 \mapsto \langle u ; a + \lambda z \rangle - \beta = \langle u ; a \rangle - \beta + \lambda \langle u ; z \rangle$ has a unique root in $[0, +\infty[$. In other words, there exist some $\lambda \geq 0$ and $x \in H^\circ(u, \beta)$ satisfying (2).

In case $\langle u ; a \rangle > \beta$ we apply the result above to $(-u, -\beta)$. \square

When γ is symmetric ($\gamma(-x) = \gamma(x)$ for all $x \in \mathbb{R}^d$), i.e. when it is a norm, then its dual enjoys the same property, and the following simplified formula directly arises (compare with [5], which uses a proof based on the Kuhn-Tucker conditions)

Corollary 2 *For any norm ν we have*

$$d_\nu(a, H^\circ(u, \beta)) = \frac{|\beta - \langle u ; a \rangle|}{\nu^\circ(u)}$$

2 Optimal separating halfspace

Let there be given two finite datasets $A, B \subset \mathbb{R}^d$ and a gauge γ on \mathbb{R}^d , used for measuring distances.

Any pair $\sigma \stackrel{\text{def}}{=} (u, \beta) \in \mathbb{R}^d \setminus \{0\} \times \mathbb{R}$ defines the halfspaces and hyperplanes

$$H^\#(\sigma) = H^\#(u, \beta) \stackrel{\text{def}}{=} \{ x \in \mathbb{R}^d \mid \langle u ; x \rangle \# b \}$$

where $\# \in \{\leq, <, =, \geq, >\}$. Note that these sets all remain the same when σ is multiplied by any strictly positive constant, but not when the sign is inverted. In fact we will use the halfspace $H^<(\sigma)$ to discriminate elements from A , as opposed to those of B . Therefore we say we separate by a halfspace (or oriented hyperplane), and the sign of σ is thus part of the information. We will also speak of the ‘halfspace’ σ by an abuse of terminology.

Given σ , we define the following sets.

- $A_\sigma^c \stackrel{\text{def}}{=} A \cap H^<(\sigma)$, the set of correctly classified points from A
- $A_\sigma^m \stackrel{\text{def}}{=} A \cap H^>(\sigma)$, the set of misclassified points from A
- $B_\sigma^c \stackrel{\text{def}}{=} B \cap H^>(\sigma)$, the set of correctly classified points from B
- $B_\sigma^m \stackrel{\text{def}}{=} B \cap H^<(\sigma)$, the set of misclassified points from B
- $A_\sigma^n \stackrel{\text{def}}{=} A \cap H^=(\sigma)$, the set of non-classified points from A
- $B_\sigma^n \stackrel{\text{def}}{=} B \cap H^=(\sigma)$, the set of non-classified points from A

Any halfspace σ^* minimizing the sum of γ -distances from all misclassified points $A_{\sigma^*}^m \cup B_{\sigma^*}^m$ to its boundary hyperplane $H^=(\sigma^*)$ is called a γ -optimal separating halfspace, i.e.

$$\sigma^* \in \arg \min \{ f(\sigma) \mid \sigma \in \mathbb{R}^d \setminus \{0\} \times \mathbb{R} \}$$

where

$$f(\sigma) \stackrel{\text{def}}{=} \sum_{a \in A_\sigma^m} d_\gamma(a, H^=(\sigma)) + \sum_{b \in B_\sigma^m} d_\gamma(b, H^=(\sigma))$$

The set of all halfspaces in \mathbb{R}^d is denoted by \mathcal{H} . As mentioned above, for any $\lambda > 0$ we have $H^\#(\lambda\sigma) = H^\#(\sigma)$, so the function

$$\mathbf{H}^< : \mathbb{R}^d \setminus \{0\} \times \mathbb{R} \rightarrow \mathcal{H} : \sigma = (u, \beta) \mapsto H^<(\sigma)$$

is surjective, with inverse image of each halfspace a ray $\mathbb{R}_0^+ u$ for some $u \neq 0$ in \mathbb{R}^d . In particular each such ray contains a single u with $\|u\| = 1$, where $\|\cdot\|$ denotes the standard euclidean norm, or any other norm or gauge, e.g. γ .

Therefore \mathcal{H} is homeomorphic to $S_{d-1} \times \mathbb{R}$, where

$$S_{d-1} \stackrel{\text{def}}{=} \{ u \in \mathbb{R}^d \mid \|u\| = 1 \}$$

Therefore, when γ is a norm ν , corollary 2 shows that determining a ν -optimal separating halfspace amounts to solving the following mathematical program, introduced by Mangasarian [5]

$$\min \left\{ \sum_{a \in A_{(u,\beta)}^m} (\langle u ; a \rangle - \beta) + \sum_{b \in B_{(u,\beta)}^m} (-\langle u ; b \rangle + \beta) \mid \nu^\circ(u) = 1 \right\}$$

which looks rather awful: it is the minimization of a piecewise linear function over a unit ball.

This formulation is not valid for a gauge, for which, according to theorem 1, one should rather write the problem as follows.

$$\min \{ f(u, \beta) \mid \|u\| = 1 \} \tag{3}$$

where

$$f(u, \beta) = \sum_{a \in A} \max\left(\frac{\langle u ; a \rangle - \beta}{\gamma^\circ(-u)}, 0\right) + \sum_{b \in B} \max\left(\frac{-\langle u ; b \rangle + \beta}{\gamma^\circ(u)}, 0\right)$$

and $\|\bullet\|$ is any gauge, e.g. the euclidean norm.

This formulation allows us to prove

Theorem 3 *A γ -optimal separating halfspace always exists.*

Proof. Defining $M_A = \max\{\gamma(a) \mid a \in A\}$, we have for any u

$$|\langle u ; a \rangle| = | \langle -u ; a \rangle | \leq \gamma^\circ(-u)\gamma(a) \leq \gamma^\circ(-u)M_A \quad \forall a \in A$$

and, with M_B similarly defined, we find that for any u

$$\begin{aligned} f(u, 0) &= \sum_{a \in A} \max\left(\frac{\langle u ; a \rangle}{\gamma^\circ(-u)}, 0\right) + \sum_{b \in B} \max\left(\frac{-\langle u ; b \rangle}{\gamma^\circ(u)}, 0\right) \\ &\leq \sum_{a \in A} M_A + \sum_{b \in B} M_B \\ &= |A|M_A + |B|M_B \stackrel{\text{def}}{=} M \end{aligned}$$

Let $S \stackrel{\text{def}}{=} \max\{1, \gamma^\circ(u), \gamma^\circ(-u) \mid \|u\| = 1\}$, then, as soon as $\beta > 2SM$, for any u with $\|u\| = 1$, we have $\beta/\gamma^\circ(-u) \geq \beta/S > 2M$, and therefore $\frac{\langle u ; a \rangle - \beta}{\gamma^\circ(-u)} \leq M - 2M < 0$ for all $a \in A$, leading to

$$\max\left(\frac{\langle u ; a \rangle - \beta}{\gamma^\circ(-u)}, 0\right) = 0$$

while for all $b \in B$

$$\max\left(\frac{-\langle u ; b \rangle + \beta}{\gamma^\circ(u)}, 0\right) > -M + 2M = M$$

hence

$$\begin{aligned} f(u, \beta) &> 0 + \sum_{b \in B} M \\ &\geq M \end{aligned}$$

which always surpasses $f(u, 0)$, so cannot be optimal. When $\beta \leq -2SM$, the same conclusion is obtained similarly.

This shows that we may add the constraint $|\beta| \leq 2SM$ to the problem without loss of optimality, which then, together with $\|u\| = 1$, consists of minimising a continuous function over a compact set, and thus admits an optimal solution. \square

Rephrasing this problem slightly differently allows us to derive some nice structural properties of the optimal solution set.

3 Fixed norm distance

We start by looking at the norm case. We assume that the discrimination problem is non-trivial, i.e. the datasets A and B cannot be fully (not necessarily strictly) separated by a hyperplane, in other words their convex hulls have an intersection with nonempty interior

Theorem 4 *Suppose $\dim(\text{conv}(A) \cap \text{conv}(B)) = d$. For distances measured by a fixed norm ν some ν -optimal separating halfspace exists the boundary hyperplane of which passes through d affinely independent points of $A \cup B$.*

Proof. Let $\sigma^0 = (u^0, \beta^0)$ define a ν -optimal separating halfspace, existence of which has been guaranteed in theorem 3. Without loss of generality we may assume $\nu^\circ(u^0) = 1$.

Define $\mathcal{T} \subset \mathbb{R}^d \times \mathbb{R}$ as the set of halfspaces leading to a similar classification as σ of all data points, more precisely by

$$(u, \beta) \in \mathcal{T} \quad \text{iff} \quad \begin{cases} \langle u ; a \rangle \leq \beta & \forall a \in A_{\sigma^0}^c \cup A_{\sigma^0}^n \\ \langle u ; a \rangle \geq \beta & \forall a \in A_{\sigma^0}^m \\ \langle u ; b \rangle \geq \beta & \forall b \in B_{\sigma^0}^c \cup B_{\sigma^0}^n \\ \langle u ; b \rangle \leq \beta & \forall b \in B_{\sigma^0}^m \end{cases}$$

and the following affine function

$$\begin{aligned} g &: \mathbb{R}^d \times \mathbb{R} \rightarrow \mathbb{R} \\ &: (u, \beta) \mapsto \sum_{a \in A_{\sigma^0}^m} (\langle u ; a \rangle - \beta) + \sum_{b \in B_{\sigma^0}^m} (-\langle u ; b \rangle + \beta) \end{aligned}$$

Note that $g(u^0, \beta^0) > 0$ due to the assumption that A and B cannot be fully separated.

Consider the set P in $\mathbb{R}^d \times \mathbb{R}$ defined as

$$P = \{ (u, \beta) \in \mathcal{T} \mid g(u, \beta) = g(u^0, \beta^0) \}.$$

P is a closed polyhedral set in $\mathbb{R}^d \times \mathbb{R}$: it is defined by linear inequalities and one linear equality in $d+1$ variables. P is also bounded. Indeed, if unbounded, there would exist some $(u, \beta) \neq (0, 0)$ such that $(\lambda u + u^0, \lambda \beta + \beta^0) \in P$, for all $\lambda \geq 0$. This means we would have

$$\begin{aligned} \langle u ; a \rangle &\leq \beta & \forall a \in A_{\sigma^0}^c \cup A_{\sigma^0}^n \\ \langle u ; a \rangle &\geq \beta & \forall a \in A_{\sigma^0}^m \\ \langle u ; b \rangle &\geq \beta & \forall b \in B_{\sigma^0}^c \cup B_{\sigma^0}^n \\ \langle u ; b \rangle &\leq \beta & \forall b \in B_{\sigma^0}^m \\ g(u, \beta) &= 0 \end{aligned}$$

which, by the definition of g , also implies that $\langle u ; a \rangle - \beta = 0$ for all $a \in A_{\sigma^0}^m$, and $\langle u ; b \rangle - \beta = 0$ for all $b \in B_{\sigma^0}^m$, which would mean that the hyperplane $H^=(u, \beta)$ fully separates A and B , contrary to our assumption.

Therefore P is a polytope of (at most) dimension d . It contains the optimal solution (u^0, β^0) , so any solution optimising f on P is also an optimal solution. Moreover,

$$\begin{aligned} \forall (u, \beta) \in P : f(u, \beta) &= \frac{g(u, \beta)}{\nu^\circ(u)} \\ &= \frac{g(u^0, \beta^0)}{\nu^\circ(u)} \end{aligned}$$

So minimizing f on P turns out to be equivalent to maximizing the function $(u, \beta) \mapsto \nu^\circ(u)$ on P . Since this function is convex, it attains its maximum on the polytope P at some extreme point (u^1, β^1) of P . Therefore (u^1, β^1) satisfies as equality $d+1$ linearly independent constraints defining P , and, since one constraint is already an equality, it satisfies (at least) d linearly independent inequality constraints as equality. This means it is a point common to d independent linear subspaces of $\mathbb{R}^d \times \mathbb{R}$, each one of which has as normal a vector of type $(c, -1)$, where $c \in A \cup B$, showing these c are affinely independent. It follows that the hyperplane $H^=(u^1, \beta^1)$ passes through d affinely independent points of $A \cup B$. \square

As a consequence, determining an optimal separating hyperplane, may be done through a finite procedure, by enumeration and evaluation of all $O(n^d)$ choices of d points out of the $n = |A \cup B|$ datapoints. Evaluating one particular choice basically consists in inverting a $d \times d$ matrix for determining the equation of the hyperplane, and calculating the position of all n points, including the distance when misclassified, a task of complexity $O(n)$ in fixed dimension. Thus for d fixed, we obtain an all-over complexity of $O(n^{d+1})$.

4 Mixed norm or gauge distance

Statistically it makes sense to use a different distance measure for measuring membership to A and B , reflecting the possibly quite different shapes of these point sets. To allow for this situation, and generalising it to possibly asymmetric distance measures, we redefine the objective to be minimized by way of the two gauges γ_A and γ_B as

$$f(\sigma) \stackrel{\text{def}}{=} \sum_{a \in A_\sigma^m} d_{\gamma_A}(a, H^=(\sigma)) + \sum_{b \in B_\sigma^m} d_{\gamma_B}(b, H^=(\sigma))$$

or

$$f(\sigma) = \sum_{a \in A_{(u, \beta)}^m} \frac{\langle u ; a \rangle - \beta}{\gamma_A^\circ(-u)} + \sum_{b \in B_{(u, \beta)}^m} \frac{-\langle u ; b \rangle + \beta}{\gamma_B^\circ(u)}$$

Obviously, this also includes the case with only one gauge distance measure used for all data points.

The proof of theorem 3 is easily adapted to this slightly extended situation, thus guaranteeing existence of an optimal separating hyperplane. Theorem 4 does, however, not extend and we obtain the following only slightly weaker result.

Theorem 5 *Suppose $\dim(\text{conv}(A) \cap \text{conv}(B)) = d$. For mixed distances measured by the gauges γ_A and γ_B some optimal separating halfspace exists the boundary hyperplane of which passes through $d - 1$ affinely independent points of $A \cup B$.*

Proof. Let $\sigma^0 = (u^0, \beta^0)$ be an optimal separating halfspace and define the following functions f^0 and g on $\mathbb{R}^d \setminus \{0\} \times \mathbb{R}$:

$$\begin{aligned} f^0(u, \beta) &\stackrel{\text{def}}{=} f_A(u, \beta) + f_B(u, \beta) \\ &\stackrel{\text{def}}{=} \frac{1}{\gamma_A^\circ(-u)} \sum_{a \in A_{(u, \beta)}^m} (\langle u ; a \rangle - \beta) + \frac{1}{\gamma_B^\circ(u)} \sum_{b \in B_{(u, \beta)}^m} (-\langle u ; a \rangle + \beta) \end{aligned}$$

and

$$g(u, \beta) = \frac{1}{\gamma_A^\circ(-u^0)} \sum_{a \in A_{(u^0, \beta^0)}^m} (\langle u ; a \rangle - \beta) + \frac{1}{\gamma_B^\circ(u^0)} \sum_{b \in B_{(u^0, \beta^0)}^m} (-\langle u ; a \rangle + \beta)$$

Observe that $g(u^0, \beta^0) = f^0(u^0, \beta^0) = f(u^0, \beta^0)$, while g is linear and $f^0 = f_A + f_B$ is non-linear.

Let the subset $P \subset \mathbb{R}^d \setminus \{0\} \times \mathbb{R}$ be defined by the constraints

$$\begin{aligned} \langle u ; a \rangle &\leq \beta & \forall a \in A_{\sigma^0}^c \cup A_{\sigma^0}^n \\ \langle u ; a \rangle &\geq \beta & \forall a \in A_{\sigma^0}^m \\ \langle u ; b \rangle &\geq \beta & \forall b \in B_{\sigma^0}^c \cup B_{\sigma^0}^n \\ \langle u ; b \rangle &\leq \beta & \forall b \in B_{\sigma^0}^m \\ g(u, \beta) &= f(u^0, \beta^0) \end{aligned}$$

It is easy to see by similar arguments as used in Theorem 4 that P is a polytope of (at most) dimension d . By definition of P we have $f^0 = f$ on P . But P contains the minimal solution (u^0, β^0) of f on $\mathbb{R}^d \setminus \{0\} \times \mathbb{R}$, so any minimum of f^0 on P will also be a global optimum of f and yield an optimal separating halfspace.

Since $f^0 = f_A + f_B$, any solution minimizing f^0 on P is an efficient solution to the biobjective problem of minimizing both f_A and f_B on P . Each of these functions is quasiconcave on P (see e.g. [1]), since their upper level sets at level α are respectively given by inequalities of the form

$$\alpha \gamma_A^\circ(-u) - \sum_{a \in A_{(u, \beta^0)}^m} (\langle u ; a \rangle - \beta) \leq 0$$

and

$$\alpha \gamma_B^\circ(u) - \sum_{b \in B_{(u^0, \beta^0)}^m} (-\langle u ; a \rangle + \beta) \leq 0$$

which, by convexity of γ_A° and γ_B° , define convex sets in $\mathbb{R}^d \setminus \{0\} \times \mathbb{R}$.

It was shown in [2] that in this case the set of *edges* (one-dimensional faces) of P constitutes a *dominator* of P , i.e. for any $(u, \beta) \in P$ there exists some (u', β') on some edge of P with $f_A(u', \beta') \leq f_A(u, \beta)$ and $f_B(u', \beta') \leq f_B(u, \beta)$, which implies $f^0(u', \beta') \leq f^0(u, \beta)$. And any edge of P is the intersection of $d - 1$ hyperplanes bounding linearly independent halfspaces defining P .

Therefore there exists some minimum of f^0 on P (and hence a global minimum of f) satisfying as equality $d - 1$ linearly independent inequalities among those defining P . But this corresponds to a halfspace with boundary hyperplane passing through $d - 1$ affinely independent points of $A \cup B$. \square

5 The one-dimensional case

The simplest case happens of course when the data points just consist of single values, i.e. $d = 1$. We study a slightly extended model in which each data points $c \in A \cup B$ also has a weight w_c . Note that all halfspaces are halflines, either of the form $\sigma = (1, \beta), H_\sigma^\leq =]-\infty, \beta]$ or $\sigma = (-1, \beta), H_\sigma^\leq = [-\beta, +\infty[$. We consider these two cases separately, and note that both are totally similar, simply inverting the roles of A and B (or equivalently, by sign inversion).

In the first case the problem we face is to minimise

$$f(\beta) \stackrel{\text{def}}{=} \sum_{a \in A} w_a \max\{a - \beta, 0\} + \sum_{b \in B} w_b \max\{\beta - b, 0\}$$

which is convex, and piecewise linear with breakpoints at all points of $A \cup B$. Therefore the minimum is reached at any point β^* where left derivative of f is nonpositive and right derivative is nonnegative. These directional derivatives at the point β are respectively

$$\begin{aligned} \sum_{a \in A, a > \beta} w_a - \sum_{b \in B, b \leq \beta} w_b \\ \sum_{a \in A, a \geq \beta} w_a - \sum_{b \in B, b < \beta} w_b \end{aligned}$$

which immediately leads to following lemma.

Theorem 6 *Let β^+ be such that*

$$\begin{aligned} \sum_{a \in A, a > \beta^+} w_a &\leq \sum_{b \in B, b \leq \beta^+} w_b \\ \sum_{a \in A, a \geq \beta^+} w_a &\geq \sum_{b \in B, b < \beta^+} w_b \end{aligned}$$

and β^- such that

$$\begin{aligned} \sum_{a \in A, a < \beta^-} w_a &\leq \sum_{b \in B, b \geq \beta^-} w_b \\ \sum_{a \in A, a \leq \beta^-} w_a &\geq \sum_{b \in B, b > \beta^-} w_b \end{aligned}$$

then either $(1, \beta^+)$ or $(-1, \beta^-)$ defines an optimal separating halfline according to which of the values $f(1, \beta^+)$ or $f(-1, \beta^-)$ is the lower. \square

The conditions of lemma 6 express that the total weight of the misclassified points for each dataset should be in balance. This is a very simple criterion, and makes it very easy to find the optimal separating halfline.

The following simple algorithm finds both some β^+ and some β^- in one pass. It suffices to sort all elements of $A \cup B$. Four counters are used to store the misclassified weight for each set and each direction. For the initial sweep position at $-\infty$ these are initialized to 0 for B and to $\sum_{a \in A} w_a$ for A in direction $u = 1$, and to 0 for A and $\sum_{b \in B} w_b$ for B in direction $u = -1$. Each time the increasing sweep crosses a new point of $A \cup B$ one updates these counters accordingly. As soon as the difference between the two counters corresponding to a same direction changes sign, one has arrived at a balancing point for this direction. These two solutions may then be evaluated to choose the best one.

Clearly this is an $O(n \log n)$ algorithm in general when including the sorting, and $O(n)$ when already sorted. It is also possible to devise a linear time method, using a technique similar to the well-known linear time median finding algorithm.

Example 7 Consider the two 1-dimensional (unweighted) datasets

$$A = \{1, 2, 3, 4, 5, 7, 10, 13, 16\}$$

$$B = \{6, 8, 9, 11, 12, 14, 15, 17, 18, 19, 20\}$$

One easily finds that β^+ can be chosen anywhere within the interval $[9, 10]$ with 3 misclassifieds from each set, whereas $\beta^- = 11$ is the unique way to have 7 misclassifieds when inverting the classification.

Evidently the solution with lowest objective value is $(1, \beta^+)$.

It should, however, not always be the case that the optimal separating halfline misclassifies less than half of the datapoints!

Example 8 Consider the two 1-dimensional unweighted datasets

$$A = \{0, 100, 101\} \text{ and } B = \{98, 99, 1000\}$$

Evidently, any $\beta \in [99, 100]$ balances both datasets in both directions. But $(1, \beta)$ gives a lower objective value than $(-1, \beta)$, so should be preferred.

However, the number of misclassifieds is then 2 for both datasets, while the inverse solution has only 1 misclassified in each.

This example shows that Mangasarian's sum of distances is perhaps not always a satisfactory objective for discrimination analysis purposes. One might therefore rather consider the following rule:

the misclassifieds of both datasets should be in balance, and consist of no more than half of the total weight of the smaller weight dataset, ensuring this is also the case for the other dataset.

For one dimensional data this much simpler criterion depends only on the ordering of all data points. This means it is also useful for (possibly weighted) data on an ordinal scale, where only order queries between pairs of values are available, and no numerical calculations of differences or distances are allowed. This idea is currently under further investigation.

6 Optimal balancing halfspaces

The idea of balancing the misclassifieds of both datasets may be extended to higher dimensions, and turns out to be quite important, as will readily be seen. For the sake of generality we also consider here the weighted version of the separation problem.

Given the two datasets A and B with corresponding weights w_a ($a \in A$) and w_b ($b \in B$) and gauges γ_A and γ_B we define a balancing halfspace as follows.

Definition 9 The halfspace $\sigma = (u, \beta)$ balances A and B if

$$\frac{1}{\gamma_A^\circ(-u)} \sum_{a \in A^{>(u,\beta)}} w_a \leq \frac{1}{\gamma_B^\circ(u)} \sum_{b \in B^{\leq(u,\beta)}} w_b \quad (4)$$

$$\frac{1}{\gamma_A^\circ(-u)} \sum_{a \in A^{\geq(u,\beta)}} w_a \geq \frac{1}{\gamma_B^\circ(u)} \sum_{b \in B^{<(u,\beta)}} w_b \quad (5)$$

$$(6)$$

Theorem 10 All optimal separating halfspaces balance A and B .

Proof. Consider any fixed u , and look for an optimal choice for a corresponding β . The objective for fixed u then looks like

$$f^u(\beta) = \sum_{a \in A_{(u,\beta)}^m} w_a \frac{\langle u ; a \rangle - \beta}{\gamma_A^\circ(-u)} + \sum_{b \in B_{(u,\beta)}^m} w_b \frac{-\langle u ; b \rangle + \beta}{\gamma_B^\circ(u)} \quad (7)$$

$$= \sum_{a \in A, \langle u ; a \rangle > \beta} w_a \frac{\langle u ; a \rangle - \beta}{\gamma_A^\circ(-u)} - \sum_{b \in B, \langle u ; b \rangle < \beta} w_b \frac{-\langle u ; b \rangle + \beta}{\gamma_B^\circ(u)} \quad (8)$$

$$= \sum_{a' \in A', a' > \beta} w_{a'}(a' - \beta) - \sum_{b \in B, b' \leq \beta} w_{b'}(-b' + \beta) \quad (9)$$

where $a' = \langle u ; a \rangle$, $w_{a'} = \frac{w_a}{\gamma_A^\circ(-u)}$ and $w_{b'} = \frac{w_b}{\gamma_B^\circ(u)}$. This is a one-dimensional separation problem on A' , B' with weights $w_{a'}$ and $w_{b'}$, on which lemma 6 applies, i.e. any optimal solution β^* balances A' and B' , in other words, the halfspace (u, β^*) balances A and B .

Applying this for any optimal u the result follows. \square

Combined with the theorems in sections 3 and 4, one sees that only balancing halfspaces must be retained, further strongly reducing the number of candidate halfspaces to be considered.

References

- [1] AVRIEL, M., DIEWERT, W.E., SCHAIBLE, S., and ZHANG, I., *Generalized Concavity*, Plenum Press, New York, New York, 1988.
- [2] CARRIZOSA, E., and PLASTRIA, F., Dominators for Multiple-objective Quasiconvex Maximization Problems, *Journal of Global Optimization*, Vol. 18, No. 1, pp. 35–58, 2000.
- [3] DURIER, R., and MICHELOT, C., Geometrical Properties of the Fermat-Weber Problem, *European Journal of Operational Research*, Vol. 20, pp. 332–343, 1985.
- [4] HIRIART-URRUTY, J.-B., and LEMARÉCHAL, C., *Convex analysis and minimization algorithms*, Springer, 1993.
- [5] MANGASARIAN, O.L., Arbitrary-Norm Separating Plane, *Operations Research Letters*, Vol. 24, pp. 15–23, 1999.
- [6] MICHELOT, C., The Mathematics of Continuous Location, *Studies in Locational Analysis*, Vol. 5, pp. 59–83, 1993.
- [7] PLASTRIA, F. and CARRIZOSA, E., Gauge-distances and median hyperplanes, *Journal of Optimization Theory and Applications*, Vol. 110, pp.173-182, 2001.
- [8] ROCKAFELLAR, T., *Convex Analysis*, Princeton University Press, Princeton, New Jersey, 1970.