4OR-Q J Oper Res (2014) 12:77–85
DOI 10.1007/s10288-013-0245-0

RESEARCH PAPER

Linear separation and approximation by minimizing the sum of concave functions of distances

Frank Plastria · Emilio Carrizosa

Received: 13 May 2013 / Published online: 20 September 2013
© Springer-Verlag Berlin Heidelberg 2013

Abstract One recently proposed criterion to separate two data sets in Classification is to use a hyperplane that minimizes the sum of distances to it from all the misclassified data points, where misclassification means lying on the wrong side of the hyperplane, or rather in the wrong halfspace. In this paper we study an extension of this problem: we seek the hyperplane minimizing the sum of concave nondecreasing functions of the distances of misclassified points to it. It is shown that an optimal hyperplane exists containing at least d affinely independent points. This extends the result known for the minimization of the sum of distances, and enables to use combinatorial local-search heuristics for this problem. As a corollary, the same result is obtained for the approximation problem in which a hyperplane minimizing the sum of concave nondecreasing functions of the distances from a set of data points is sought.

Keywords Linear separation · Linear approximation · Distance minimization

Mathematics Subject Classification 90B85 · 62-07 · 68U05

1 Nonlinear distance functions

Two nonempty finite data sets A, B in $\mathbb{R}^d$ are said to be linearly separable if there exist $u \in \mathbb{R}^d$, $u \neq 0$ and $\beta \in \mathbb{R}$, such that the hyperplane,

The research of the authors was partially supported by the projects FWOAL453 and VUB-GOA62 (Belgium) and MTM2009-14039 and FQM-329 (Junta de Andalucía, Spain).

F. Plastria (✉)
MOSI, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium
e-mail: Frank.Plastria@vub.ac.be

E. Carrizosa
Fac. de Matemáticas, Universidad de Sevilla, Tarfia s/n, 41012 Sevilla, Spain
e-mail: ecarrizosa@us.es

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

satisfies

$$\begin{array}{ll} \langle u\ ;\ a\rangle \geq \beta & \forall a \in A \\ \langle u\ ;\ b\rangle \leq \beta & \forall b \in B \end{array} \tag{1}$$

In this paper we consider the case where the sets A, B are not linearly separable, and thus a hyperplane $H(u,\beta)$ satisfying some relaxation of (1) is sought.

A first strategy, proposed by Mangasarian (1994), amounts to finding a hyperplane maximizing the number of constraints in (1) that are satisfied,

$$\begin{array}{ll} \max & |A^*| + |B^*| \\ \text{s.t.} & \langle u\ ;\ a\rangle \geq \beta \quad \forall a \in A^* \\ & \langle u\ ;\ b\rangle \leq \beta \quad \forall b \in B^* \\ & A^* \subset A,\ B^* \subset B \\ & u \neq 0,\ \beta \in \mathbb{R} \end{array} \tag{2}$$

A different strategy, proposed in Mangasarian (1999) and later analyzed in Carrizosa and Plastria (2008), Karam et al. (2007), is based on the minimization of distances to correct classification, as described below. See also Plastria and Carrizosa (2012) for a minmax approach.

Given $u \in \mathbb{R}^d$, $u \neq 0$ and $\beta \in \mathbb{R}$, let $H(u,\beta)^{\geq}$ and $H(u,\beta)^{\leq}$ denote the two closed halfspaces with common boundary $H(u,\beta)$,

$$H(u,\beta)^{\geq} \stackrel{\text{def}}{=} \left\{x \in \mathbb{R}^d\ :\ \langle u\ ;\ x\rangle \geq \beta\right\}$$
$$H(u,\beta)^{\leq} \stackrel{\text{def}}{=} \left\{x \in \mathbb{R}^d\ :\ \langle u\ ;\ x\rangle \leq \beta\right\}$$

According to (1) any point $a \in A$ (resp. $b \in B$) is misclassified when $a \notin H(u,\beta)^{\geq}$ (resp. $b \notin H(u,\beta)^{\leq}$).

Given a norm γ in $\mathbb{R}^d$, let d_γ be the distance in $\mathbb{R}^d$ induced by γ, and consider the problem of finding the hyperplane minimizing the sum of distances of the points to their respective halfspace of correct classification,

$$\begin{array}{ll} \min & \displaystyle\sum_{a\in A} d_\gamma(a, H^{\geq}(u,\beta)) + \sum_{b\in B} d_\gamma(b, H^{\leq}(u,\beta)) \\ \text{s.t.} & u \neq 0,\ \beta \in \mathbb{R} \end{array} \tag{3}$$

Observe that

$$d_\gamma(a, H^{\geq}(u,\beta)) = \frac{(-\langle u\ ;\ a\rangle + \beta)^+}{\gamma^\circ(u)} \tag{4}$$

$$d_\gamma(b, H^{\leq}(u,\beta)) = \frac{(\langle u\ ;\ b\rangle - \beta)^+}{\gamma^\circ(u)}, \tag{5}$$

Carrizosa and Plastria (2008), where $(t)^+ = \max\{t, 0\}$, and γ° is the norm dual to γ.

In this paper we address an extension of models (2) and (3). A different way of combining (2)–(3) can be found in Chen and Mangasarian (1996).

Here, instead of summing the misclassification *distances* as in (3), we take the sum of misclassification *costs,* where cost is a function of distance with certain properties. More precisely, consider the set of functions $\mathcal{F}$,

$$\mathcal{F} \stackrel{\text{def}}{=} \{f : \mathbb{R}^+ \longrightarrow \mathbb{R}^+ \mid f \text{ is concave, with } f(0) = 0\}. \tag{6}$$

Any $f \in \mathcal{F}$ satisfies $f(s) \geq 0$ for all $s \geq 0$, and when applied to distances yields a cost (which is zero for zero distances). On the other hand, concavity implies a diminishing effect of larger distances, which is a necessary condition for robustness. See Carrizosa and Rodríguez-Chía (1997) for further motivation for this model of cost functions.

The definition implies that any $f \in \mathcal{F}$ is nondecreasing in $\mathbb{R}^+$, Plastria (2009).

For each $a \in A$, (resp. $b \in B$,) let f_a (resp. f_b) be a function in $\mathcal{F}$, and consider the problem

$$\begin{array}{l} \min \displaystyle\sum_{a\in A} f_a(d(a, H^{\geq}(u, \beta))) + \sum_{b\in B} f_b(d(b, H^{\leq}(u, \beta))) \\ \text{s.t.} \quad u \neq 0, \ \beta \in \mathbb{R}, \end{array}$$

which, by (4)–(5), can be written as

$$\begin{array}{l} \min \ F(u, \beta) \stackrel{\text{def}}{=} \displaystyle\sum_{a\in A} f_a\left(\frac{(-\langle u\ ;\ a\rangle + \beta)^+}{\gamma^\circ(u)}\right) + \sum_{b\in B} f_b\left(\frac{(\langle u\ ;\ b\rangle - \beta)^+}{\gamma^\circ(u)}\right) \\ \text{s.t.} \quad u \neq 0, \ \beta \in \mathbb{R}. \end{array} \tag{7}$$

A possible choice for functions f in (7) is $f(t) = t^p$ $(0 < p \leq 1)$. In particular, for $p = 1$, one obtains (3), whereas values of p, $0 < p < 1$ have been advocated, in a related context, by Stam (1997); the step function

$$f(t) = \begin{cases} 1, & \text{if } t > 0 \\ 0, & \text{if } t = 0 \end{cases} \tag{8}$$

which counts the number of misclassified points, yields (2); other functions of possible interest such as $f(t) = \min\{t/C, 1\}$ or $f(t) = 1 - e^{-tC}$ $(C > 0)$ remain, as far as we know, unexplored.

The purpose of this note is to show that, under these concavity conditions, there always exists an optimal solution (u, β) to (7) such that $H(u, \beta)$ contains sufficiently many data points. It follows that complete enumeration can be used as a polynomial resolution technique.

2 The localization property

Before we show the localization property, the following technical lemmas are needed.

Lemma 1 *Let K be a convex set in $\mathbb{R}^{d+1}$. Then, the set $\mathbb{R}_{++} \cdot K$ defined as*

$$\mathbb{R}_{++} \cdot K \stackrel{def}{=} \{z \in \mathbb{R}^{d+1} : z = ty \text{ for some } t > 0, y \in K\}$$

is convex.

This is in fact simply the (pointed) conical hull of K (i.e. without the origin in case $0 \notin K$)

Lemma 2 *Let ν be a norm in $\mathbb{R}^{d+1}$. Let S be a pointed cone in $\mathbb{R}^{d+1}$ such that $\nu(x) > 0 \, \forall x \in S$. Let $v_1, \ldots, v_k \in \mathbb{R}^{d+1}$ be such that $\langle v_i \, ; \, x \rangle \geq 0$ for all $x \in S$, and let $f_1, \ldots, f_r$ be concave nondecreasing functions in $\mathbb{R}_+$. Then, the function φ,*

$$\varphi(x) = \sum_{i=1}^{r} f_i \left(\frac{\langle v_i \, ; \, x \rangle}{\nu(x)} \right)$$

is quasiconcave on S.

Proof First observe that φ is well defined in S since $\nu(x) > 0 \, \forall x \in S$.

Define, for $\alpha \geq 0$, the set L_α as the upper level set of φ, namely:

$$L_\alpha = \{x \in S : \varphi(x) \geq \alpha\}$$

It suffices to show that, for any α, the set L_α is convex.

Convexity of L_α will be shown by showing that

$$L_\alpha = \mathbb{R}_{++} \cdot \left(\left\{ y \in S : \sum_{i=1}^{r} f_i \left(\langle v_i \, ; \, y \rangle \right) \geq \alpha \right\} \cap \{y : \nu(y) \leq 1\} \right). \tag{9}$$

By the concavity of the functions f_i, the functions $f_i(\langle v_i \, ; \, y \rangle)$ are concave, so their sum is also concave and its upper level sets are convex. Hence, by Lemma 1 the righthandside set is convex, and then the result follows.

Let us show (9). We have by definition of L_α that

$$L_\alpha = \bigcup_{\varepsilon > 0} \left\{ x \in S : \sum_{i=1}^{r} f_i \left(\frac{\langle v_i \, ; \, x \rangle}{\nu(x)} \right) \geq \alpha, \ \nu(x) = \varepsilon \right\} \tag{10}$$

$$= \bigcup_{\varepsilon > 0} \left\{ x \in S : \sum_{i=1}^{r} f_i \left(\frac{\langle v_i \, ; \, x \rangle}{\varepsilon} \right) \geq \alpha, \ \nu(x) = \varepsilon \right\} \tag{11}$$

$$= \bigcup_{\varepsilon > 0} \left\{ x \in S : \sum_{i=1}^{r} f_i \left(\frac{\langle v_i \, ; \, x \rangle}{\varepsilon} \right) \geq \alpha, \ \nu(x) \leq \varepsilon \right\}, \tag{12}$$

where the last equality follows from the following: it is evident that the set in (11) is included in the one in (12); any x in this latter set satisfies $x \in S$, and there exists

some $\varepsilon > 0$ such that $\nu(x) \leq \varepsilon$, and

$$\sum_{i=1}^{r} f_i\left(\frac{\langle v_i\ ;\ x\rangle}{\varepsilon}\right) \geq \alpha$$

By assumption all $\langle v_i\ ;\ x\rangle \geq 0$, so $\frac{\langle v_i\ ;\ x\rangle}{\nu(x)} \geq \frac{\langle v_i\ ;\ x\rangle}{\varepsilon}$ and, since all f_i are nondecreasing,

$$\sum_{i=1}^{r} f_i\left(\frac{\langle v_i\ ;\ x\rangle}{\nu(x)}\right) \geq \alpha.$$

This shows that x belongs to the set in (11) for the choice $\varepsilon = \nu(x)$.

Hence,

$$\begin{aligned} L_\alpha &= \bigcup_{\varepsilon>0} \left\{ x \in S :\ x = \varepsilon y,\ \sum_{i=1}^{r} f_i(\langle v_i\ ;\ y\rangle) \geq \alpha,\ \nu(y) \leq 1 \right\} \\ &= \bigcup_{\varepsilon>0} \left\{ x \in \mathbb{R}^{d+1} :\ x = \varepsilon y,\ y \in S,\ \sum_{i=1}^{r} f_i(\langle v_i\ ;\ y\rangle) \geq \alpha,\ \nu(y) \leq 1 \right\}, \end{aligned}$$

which is exactly statement (9), and thus the result follows.

Theorem 3 *Problem (7) has an optimal solution (u, β) such that (u, β) contains at least d affinely independent points of $A \cup B$.*

Proof The proof goes parallel to the ones given in Carrizosa and Plastria (2008), Plastria and Carrizosa (2001) for related problems, now using Lemma 2 as supporting result.

Take an arbitrary solution (u_0, β_0), $u_0 \neq 0$, and we will find a solution (u, β) with $H(u, \beta)$ containing at least d affinely independent points of $A \cup B$ and $F(u, \beta) \leq F(u_0, \beta_0)$.

Define the sets

$$\begin{aligned} A^{\geq}_{(u_0,\beta_0)} &= \{a \in A : \langle u_0\ ;\ a\rangle - \beta_0 \geq 0\} \\ A^{\leq}_{(u_0,\beta_0)} &= \{a \in A : \langle u_0\ ;\ a\rangle - \beta_0 \leq 0\} \\ B^{\geq}_{(u_0,\beta_0)} &= \{b \in B : \langle u_0\ ;\ b\rangle - \beta_0 \geq 0\} \\ B^{\leq}_{(u_0,\beta_0)} &= \{b \in B : \langle u_0\ ;\ b\rangle - \beta_0 \leq 0\}. \end{aligned}$$

By construction, since the sets A, B are assumed not to be linearly separable, one has that

$$\begin{aligned} &\text{If } A^{\geq}_{(u_0,\beta_0)} = \emptyset, \text{ then } B^{\geq}_{(u_0,\beta_0)} \neq \emptyset \\ &\text{If } A^{\leq}_{(u_0,\beta_0)} = \emptyset, \text{ then } B^{\leq}_{(u_0,\beta_0)} \neq \emptyset. \end{aligned} \tag{13}$$

Moreover, let $C(u_0, \beta_0)$ be the polyhedron in $\mathbb{R}^{d+1}$ defined by the inequalities and equality

$$\begin{array}{c} \langle u\ ;\ a\rangle - \beta \geq 0 \quad \forall a \in A^{\geq}_{(u_0,\beta_0)} \\ \langle u\ ;\ a\rangle - \beta \leq 0 \quad \forall a \in A^{\leq}_{(u_0,\beta_0)} \\ \langle u\ ;\ b\rangle - \beta \geq 0 \quad \forall b \in B^{\geq}_{(u_0,\beta_0)} \\ \langle u\ ;\ b\rangle - \beta \leq 0 \quad \forall b \in B^{\leq}_{(u_0,\beta_0)} \\ \displaystyle\sum_{a\in A^{\geq}_{(u_0,\beta_0)}} (\langle u\ ;\ a\rangle - \beta) - \sum_{a\in A^{\leq}_{(u_0,\beta_0)}} (\langle u\ ;\ a\rangle - \beta) \\ + \displaystyle\sum_{b\in B^{\geq}_{(u_0,\beta_0)}} (\langle u\ ;\ b\rangle - \beta) - \sum_{b\in B^{\leq}_{(u_0,\beta_0)}} (\langle u\ ;\ b\rangle - \beta) = \delta_0, \end{array} \tag{14}$$

where δ_0 is the constant

$$\begin{aligned} \delta_0 \stackrel{\text{def}}{=} & \sum_{a\in A^{\geq}_{(u_0,\beta_0)}} (\langle u_0\ ;\ a\rangle - \beta_0) - \sum_{a\in A^{\leq}_{(u_0,\beta_0)}} (\langle u_0\ ;\ a\rangle - \beta_0) \\ & + \sum_{b\in B^{\geq}_{(u_0,\beta_0)}} (\langle u_0\ ;\ b\rangle - \beta_0) - \sum_{b\in B^{\leq}_{(u_0,\beta_0)}} (\langle u_0\ ;\ b\rangle - \beta_0)\,. \end{aligned}$$

By construction, $\delta_0 \geq 0$. Moreover, $\delta_0 > 0$, since, otherwise, $H(u_0, \beta_0)$ would contain $A \cup B$, contradicting the assumption that A, B are not linearly separable.

On the other hand, $C(u_0, \beta_0) \neq \emptyset$, since $(u_0, \beta_0) \in C(u_0, \beta_0)$.

Moreover, the polyhedron $C(u_0, \beta_0)$ is bounded. Indeed, otherwise, it would contain a direction $(u, \beta) \neq (0, 0)$, which should satisfy

$$\begin{array}{c} \langle u\ ;\ a\rangle - \beta \geq 0 \quad \forall a \in A^{\geq}_{(u_0,\beta_0)} \\ \langle u\ ;\ a\rangle - \beta \leq 0 \quad \forall a \in A^{\leq}_{(u_0,\beta_0)} \\ \langle u\ ;\ b\rangle - \beta \geq 0 \quad \forall b \in B^{\geq}_{(u_0,\beta_0)} \\ \langle u\ ;\ b\rangle - \beta \leq 0 \quad \forall b \in B^{\leq}_{(u_0,\beta_0)} \\ \displaystyle\sum_{a\in A^{\geq}_{(u_0,\beta_0)}} (\langle u\ ;\ a\rangle - \beta) - \sum_{a\in A^{\leq}_{(u_0,\beta_0)}} (\langle u\ ;\ a\rangle - \beta) \\ + \displaystyle\sum_{b\in B^{\geq}_{(u_0,\beta_0)}} (\langle u\ ;\ b\rangle - \beta) - \sum_{b\in B^{\leq}_{(u_0,\beta_0)}} (\langle u\ ;\ b\rangle - \beta) = 0, \end{array} \tag{15}$$

implying all inequalities above are equalities, and then $A \cup B$ would be contained in a hyperplane, which contradicts our non separability hypothesis.

One also has that no $(0, \beta) \in C(u_0, \beta_0)$. Indeed, $(0, 0) \notin C(u_0, \beta_0)$ because $\delta_0 > 0$, and thus $(0, 0)$ does not satisfy the last condition defining $C(u_0, \beta_0)$. Suppose further that, on the contrary, some $\beta \neq 0$ exists such that $(0, \beta) \in C(u_0, \beta_0)$. If it were $\beta > 0$, $(0, \beta)$ could not satisfy the first and third block of constraints in (14), implying $A^{\geq}(u_0, \beta_0) = B^{\geq}(u_0, \beta_0) = \emptyset$, whereas for $\beta < 0$, the second and fourth block of

inequalities should be void, yielding $A^{\leq}(u_0, \beta_0) = B^{\leq}(u_0, \beta_0) = \emptyset$. Both cases are in contradiction with (13), thus we conclude that no β exists with $(0, \beta) \in C(u_0, \beta_0)$.

Let ν be the convex positively homogeneous function in $\mathbb{R}^{d+1}$ defined as

$$\nu(u, \beta) = \gamma^{\circ}(u).$$

We just showed in other words that $\nu(u, \beta) > 0$ for all $(u, \beta) \in C(u_0, \beta_0)$.

Now, consider Problem (7) restricted to the nonempty bounded polyhedron $C(u_0, \beta_0)$.

One has that for all $(u, \beta) \in C(u_0, \beta_0)$

$$\begin{aligned} F(u, \beta) &= \sum_{a \in A^{\leq}_{(u_0, \beta_0)}} f_a\left(\frac{\beta - \langle a\ ;\ u\rangle}{\gamma^{\circ}(u)}\right) + \sum_{b \in B^{\geq}_{(u_0, \beta_0)}} f_b\left(\frac{\langle b\ ;\ u\rangle - \beta}{\gamma^{\circ}(u)}\right) \\ &= \sum_{a \in A^{\leq}_{(u_0, \beta_0)}} f_a\left(\frac{\beta - \langle a\ ;\ u\rangle}{\nu(u, \beta)}\right) + \sum_{b \in B^{\geq}_{(u_0, \beta_0)}} f_b\left(\frac{\langle b\ ;\ u\rangle - \beta}{\nu(u, \beta)}\right) \end{aligned} \tag{16}$$

By Lemma 2, F is quasiconcave on the bounded polyhedron $C(u_0, \beta_0)$, thus there exists some extreme point (u^*, β^*) of $C(u_0, \beta_0)$ such that

$$F(u^*, \beta^*) \leq F(u, \beta) \qquad \forall (u, \beta) \in C(u_0, \beta_0),$$

and, in particular, $F(u^*, \beta^*) \leq F(u_0, \beta_0)$.

Since (u^*, β^*) is a vertex of $C(u_0, \beta_0)$, there must exist a subset of $d + 1$ linearly independent constraints from (14) satisfied with equality at (u^*, β^*). In other words $H(u^*, \beta^*)$ must contain a subset of affinely independent points from $A \cup B$ with cardinality at least d. □

As a consequence of Theorem 3, finding an optimal solution to Problem (7) can be reduced to inspecting a finite set of candidate solutions, since any d affinely independent points univoquely define a hyperplane. This is illustrated in Fig. 1. Setting in (7) f as $f(t) = t^p$ for $p = 0.1, 0.2, \ldots, 1$, we obtain, by complete enumeration of candidate lines, the depicted optimal separating lines. Each line is optimal for the respective p-values 0.1–0.2, 0.3, 0.4–0.7 and 0.8–1. It may be observed in the example that, the lower the value of p, the lower the number of missclassified points, viz., respectively 10, 10, 11 and 12, as was to be expected since f converges to the step function (8) for $p \downarrow 0$.

Such a complete enumeration strategy is only feasible for low-dimensional problems. For more general settings, the localization property may be advantageously used in local-search heuristic approaches, as was already done in Plastria and De Bruyne (2010) for the case of linear distance functions.

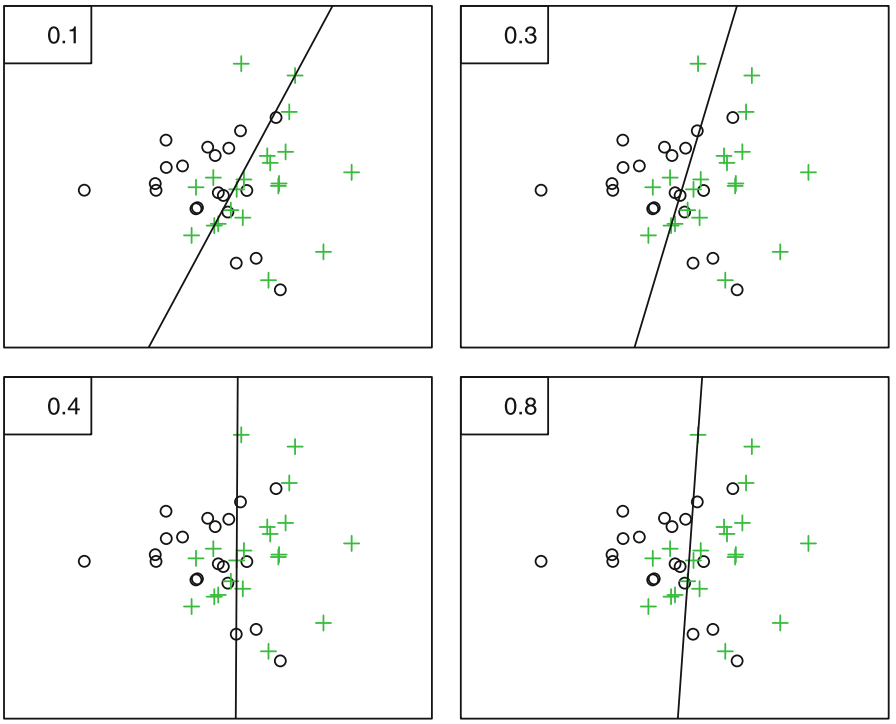

Fig. 1 Optimal separating lines for different values of p

3 Application to approximation

Theorem 3 can also be applied to an approximation problem, in which a set X is given, and a hyperplane $H(u, \beta)$ is sought minimizing the sum of concave functions f_x of the distances from the data points x to $H(u, \beta)$,

$$\begin{array}{ll} \min & \displaystyle\sum_{x \in X} f_x(d(x, H(u, \beta))) \\ \text{s.t.} & u \neq 0,\ \beta \in \mathbb{R}, \end{array} \tag{17}$$

which can also be written as

$$\begin{array}{ll} \min & \displaystyle\sum_{x \in X} f_x\left(\frac{(-\langle u\ ;\ x\rangle + \beta)^+}{\gamma^\circ(u)}\right) + \sum_{x \in X} f_x\left(\frac{(\langle u\ ;\ x\rangle - \beta)^+}{\gamma^\circ(u)}\right) \\ \text{s.t.} & u \neq 0,\ \beta \in \mathbb{R}. \end{array} \tag{18}$$

Indeed, taking $A = B = X$ in Theorem 3 one immediately obtains

Corollary 4 *Problem* (17) *has an optimal solution* (u, β) *such that* (u, β) *contains at least* d *affinely independent points of* X.

References

Carrizosa E, Plastria F (2008) Optimal expected distance separating halfspace. Math Oper Res 33:662–677
Carrizosa E, Rodríguez-Chía A (1997) Weber problems with alternative transportation systems. Eur J Oper Res 97:87–93
Chen C, Mangasarian OL (1996) Hybrid misclassification minimization. Adv Comput Math 5:127–136
Karam A, Caporossi G, Hansen P (2007) Arbitrary-norm hyperplane separation by variable neighbourhood search. IMA J Manag Math 18:173–189
Mangasarian OL (1994) Misclassification minimization. J Glob Optim 5:309–323
Mangasarian OL (1999) Arbitrary-norm separating plane. Oper Res Lett 24:15–23
Plastria F (2009) Asymmetric distances, semidirected networks and majority in Fermat–Weber problems. Ann Oper Res 167:121–155
Plastria F, Carrizosa E (2001) Gauge-distances and median hyperplanes. J Optim Theory Appl 110:173–182
Plastria F, Carrizosa E (2012) Minmax-distance approximation and separation problems: geometrical properties. Math Program 123:153–177
Plastria FS, De Bruyne S, Carrizosa E (2010) Alternating local search based VNS for linear classification. Ann Oper Res 174:121–134
Stam A (1997) Nontraditional approaches to statistical classification: some perspectives on L_p-norm methods. Ann Oper Res 74:1–36