SVM-МЕТОД СТРОГОГО ЛИНЕЙНОГО ОТДЕЛЕНИЯ ДВУХ КОНЕЧНЫХ МНОЖЕСТВ*

В. Н. Малоземов А. В. Плоткин

v.malozemov@spbu.ru avplotkin@gmail.com

30 марта 2022 г.

1°. В пространстве $\mathbb{R}^n$ рассматриваются два конечных множества

$$P_1 = \{p_j\}_{j=1}^{s}, \qquad P_2 = \{p_j\}_{j=s+1}^{m},$$

где $s \in 1 : m-1$. Говорят, что множества P_1 и P_2 допускают строгое линейное отделение, если найдутся ненулевой вектор $w \in \mathbb{R}^n$ и число $\beta \in \mathbb{R}$, такие, что

$$\begin{aligned} \langle w, p_j \rangle + \beta &\geqslant 1 \quad \text{при} \quad j \in 1 : s, \\ \langle w, p_j \rangle + \beta &\leqslant -1 \quad \text{при} \quad j \in s+1 : m. \end{aligned} \tag{1}$$

Введем две гиперплоскости L_1 и L_2, определяемые уравнениями

$$\langle w, x \rangle + \beta = 1 \quad \text{и} \quad \langle w, x \rangle + \beta = -1.$$

Полоса H, ограниченная параллельными гиперплоскостями L_1 и L_2, называется *разделяющей полосой.* Ее ширина h определяется формулой

$$h = \min_{x_1 \in L_1,\, x_2 \in L_2} \|x_1 - x_2\|.$$

ЛЕММА. *Справедливо равенство* $h = \frac{2}{\|w\|}$.

Доказательство. Для любых $x_1 \in L_1$, $x_2 \in L_2$ имеем

$$2 = \langle w, x_1 - x_2 \rangle \leqslant \|w\| \cdot \|x_1 - x_2\|.$$

Отсюда следует, что

$$\|x_1 - x_2\| \geqslant \frac{2}{\|w\|}.$$

---

*Семинар по оптимизации, машинному обучению и искусственному интеллекту «O&ML» http://www.apmath.spbu.ru/oml/

Покажем, что это неравенство может выполняться как равенство. Возьмем любой элемент $x_2 \in L_2$ и положим

$$x_1 = x_2 + \frac{2}{\|w\|^2} w.$$

Так как

$$\langle w, x_1 \rangle + \beta = -1 + \left\langle w, \frac{2}{\|w\|^2} w \right\rangle = 1,$$

то $x_1 \in L_1$. при этом $\|x_1 - x_2\| = \frac{2}{\|w\|}$. Лемма доказана. □

Гиперплоскость L_0, определяемая уравнением $\langle w, x \rangle + \beta = 0$, называется *отделяющей* множество P_1 от P_2. Отметим, что расстояние от произвольной точки $x \in \mathbb{R}^n$ до L_0 равно следующей величине:

$$\frac{\big|\langle w, x \rangle + \beta\big|}{\|w\|}.$$

Это легко проверить. Обозначим $\alpha = \langle w, x \rangle + \beta$. Возьмем любую точку $x_0 \in L_0$. Запишем

$$|\alpha| = \big|\langle w, x - x_0 \rangle\big| \leqslant \|w\| \cdot \|x - x_0\|.$$

Отсюда следует, что

$$\|x - x_0\| \geqslant \frac{|\alpha|}{\|w\|}.$$

Для точки $x_0 = x - \frac{\alpha}{\|w\|^2} w$ имеем

$$\langle w, x_0 \rangle + \beta = \alpha - \left\langle w, \frac{\alpha}{\|w\|^2} w \right\rangle = 0.$$

Значит, $x_0 \in L_0$. При этом $\|x - x_0\| = \frac{|\alpha|}{\|w\|}$.

2°. Введем обозначения

$$g = \begin{pmatrix} w \\ \beta \end{pmatrix}; \quad a_j = \begin{pmatrix} p_j \\ 1 \end{pmatrix}, \ j \in 1:m; \quad \xi_j = \begin{cases} \ \ 1, & \text{при } j \in 1:s, \\ -1 & \text{при } j \in s+1:m. \end{cases}$$

С их помощью запись неравенств (1) можно упростить:

$$\langle \xi_j a_j, g \rangle \geqslant 1, \quad j \in 1:m. \tag{2}$$

Неравенства (2) представимы в компактном виде

$$A^T g \geqslant e, \tag{3}$$

где A — матрица со столбцами $\xi_1 a_1$, $\xi_2 a_2$, …, $\xi_m a_m$ и e — m-мерный вектор, все компоненты которого равны единице.

Будем искать решение неравенства (3), порождающее разделяющую полосу H наибольшей ширины h. С учетом леммы приходим к экстремальной задаче

$$\begin{gathered} \tfrac{1}{2}\left\|w\right\|^2 \to \min, \\ A^T g \geqslant e. \end{gathered} \tag{4}$$

Введем матрицу

$$D = \left(\begin{array}{ccc:c} & & & 0 \\ & E_n & & \vdots \\ & & & 0 \\ \hdashline 0 & \dots & 0 & 0 \end{array}\right).$$

Очевидно, что $\left\|w\right\|^2 = \langle Dg, g\rangle$. Задача (4) принимает вид стандартной задачи квадратичного программирования

$$\begin{gathered} \tfrac{1}{2}\langle Dg, g\rangle \to \min, \\ A^T g \geqslant e. \end{gathered} \tag{5}$$

Запишем двойственную задачу с помощью таблицы данных задачи (5) (см. [1]).

Таблица 1. Таблица данных задачи (5)

	$0,\dots,0$	0	
v	$-E_n$	0 $\vdots$ 0	$-\frac{1}{2}v$
λ	$0,\dots,0$	0	0
u_1	$\xi_1 p_1^{\mathsf{T}}$	ξ_1	1
u_2	$\xi_2 p_2^{\mathsf{T}}$	ξ_2	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$
u_m	$\xi_m p_m^{\mathsf{T}}$	ξ_m	1

Получим

$$\begin{gathered}
-\tfrac{1}{2}\langle v, v\rangle + \sum_{i=1}^{m} u_i \to \max,\\
-v^T + \sum_{i=1}^{m} u_i \xi_i p_i^T = \mathbb{O},\\
\sum_{i=1}^{m} u_i \xi_i = 0,\\
u_i \geqslant 0 \ \text{ при } \ i \in 1:m.
\end{gathered} \tag{6}$$

В частности,

$$v = \sum_{i=1}^{m} u_i \xi_i p_i. \tag{7}$$

Обозначим через A_0 матрицу со столбцами $\xi_1 p_1$, $\xi_2 p_2$, …, $\xi_m p_m$. Тогда формулу (7) можно записать так:

$$v = A_0 u.$$

Переменную v в задаче (6) можно исключить. Поменяв знак у целевой функции, придем к следующей задаче:

$$\begin{gathered}
\tfrac{1}{2}\langle A_0^T A_0 u, u\rangle - \sum_{i=1}^{m} u_i \to \min,\\
\sum_{i=1}^{m} \xi_i u_i = 0,\\
u_i \geqslant 0 \ \text{ при } \ i \in 1:m.
\end{gathered} \tag{8}$$

Именно эту задачу и нужно решать.

3°. Допустим, что мы нашли решение u^* задачи (8). Решением двойственной задач (6) будет пара (z^*, u^*), где $z^* = \begin{pmatrix} v^* \\ \lambda^* \end{pmatrix}$,

$$v^* = \sum_{i=1}^{m} u_i^* \xi_i p_i$$

и λ^* — произвольное вещественное число. По первой теореме двойственности существует решение $g_* = \left(\begin{smallmatrix} w_* \\ \beta_* \end{smallmatrix}\right)$ у прямой задачи (5). По второй теореме двойственности справедливы соотношения

$$D g_* = D z^*, \tag{9}$$

$$u_i^* \Big(\xi_i \big(\langle w_*, p_i\rangle + \beta_*\big) - 1\Big) = 0, \quad i \in 1:m. \tag{10}$$

Согласно (9) и определению матрицы D получаем $w_* = v^*$, так что

$$w_* = \sum_{i=1}^{m} u_i^* \xi_i p_i. \tag{11}$$

Обозначим $I^* = \{i \in 1:m \mid u_i^* > 0\}$. Из (10) следуют равенства

$$\xi_i\big(\langle w_*, p_i\rangle + \beta_*\big) = 1 \quad \text{при} \quad i \in I^*. \tag{12}$$

Отметим, что $I^* \neq \varnothing$. В противном случае $u^* = \mathbb{O}$, $w_* = \mathbb{O}$. Формула (2) при $g = g_*$ принимает вид $\xi_j \beta_* \geqslant 1$, $j \in 1:m$. Это невозможно, поскольку по определению ξ_j существуют как $\xi_j = 1$, так и $\xi_j = -1$.

Векторы p_i при $i \in I^*$ называется *опорным* для решения $g_* = \binom{w_*}{\beta_*}$. На них выполняются равенства (12). В представлении (11) для w_* можно оставить только опорные векторы,

$$w_* = \sum_{i \in I^*} u_i^* \xi_i p_i.$$

Константа β_* определяется из (12) при любом $i \in I^*$.

Рассмотренный метод строго линейного отделения двух конечных множеств называется методом опорных векторов или SVM-методом (сокращенно от Support Vector Machine) [2].

4°. Задача (8) не имеет решения, когда ее целевая функция неограничена снизу. В этом случае и двойственная задача (6) не имеет решения. По первой теореме двойственности отсутствует решение у прямой задачи (5). Поскольку целевая функция задачи (5) ограничена снизу нулем, то по теореме существования решения множество планов задачи (5) пусто. Это означает, что система неравенств (1) несовместна, то есть что множества P_1 и P_2 не допускают строго линейного отделения.

5°. В докладе [3] рассматривается задача, аналогичная задаче (4): найти решение неравенства (3), порождающее разделяющую полосу наибольшей ℓ_1-ширины. Эта задача формализуется следующим образом:

$$\begin{aligned} \|w\|_\infty &\to \min, \\ A^T g &\geqslant e, \end{aligned} \tag{13}$$

где $g = \binom{w}{\beta}$. Напомним, что задача (4) эквивалентна задача квадратичного программирования (5), в то время как задача (13) сводится к задаче линейного программирования

$$\begin{aligned} t &\to \min, \\ A^T g &\geqslant e, \\ -w + te &\geqslant 0, \\ w - te &\geqslant 0. \end{aligned} \tag{14}$$

При этом обе задачи ориентированы на расширение разделяющей полосы.

6°. Очевидно, что задача (5) более трудоемкая, чем задача (14). Мы провели численные эксперименты, чтобы сравнить качество линейного отделения двух конечных множеств при двух описанных подходах.

На рисунках 1 и 2 представлены результаты SVM-отделения. В первом случае имеем четыре $(3+1)$ опорные точки (они выделены кружками). И во втором случае получены четыре $(2+2)$ опорные точки.

На рисунках 3 и 4 сравниваются SVM-отделение и отделение, основанное на решении задачи (13) (которая сводится к задаче линейного программирования (14)). Обозначим через h ширину разделяющей полосы при SVM-отделении, а через h_1 — ширину разделяющей полосы в случае решения задачи (14). В ситуации, представленной на рис. 3, имеем $h = 0.213$, $h_1 = 0.202$. На рис. 4 разделяющие плоскости отличаются существенно, но значения h и h_1 близки: $h = 0.130$, $h_1 = 0.095$.

ЛИТЕРАТУРА

1. Малоземов В. Н., Плоткин А. В. *Двойственность в квадратичном программировании. Задача Сильвестра.* // Семинар «CNSA & NDO». Избранные доклады. 8 декабря 2021 г. (`http://apmath.spbu.ru/cnsa/reps21.shtml#1208`)

2. Воронцов К. В. *Лекции по методу опорных векторов.* (`http://www.ccas.ru/voron/download/SVM.pdf`)

3. Малоземов В. Н., Плоткин А. В. *Строгое линейное отделение двух конечных множеств и линейное программирование* // Семинар «O & ML». Избранные доклады. 24 февраля 2022 г. (`http://apmath.spbu.ru/oml/reps22.shtml#0224`)

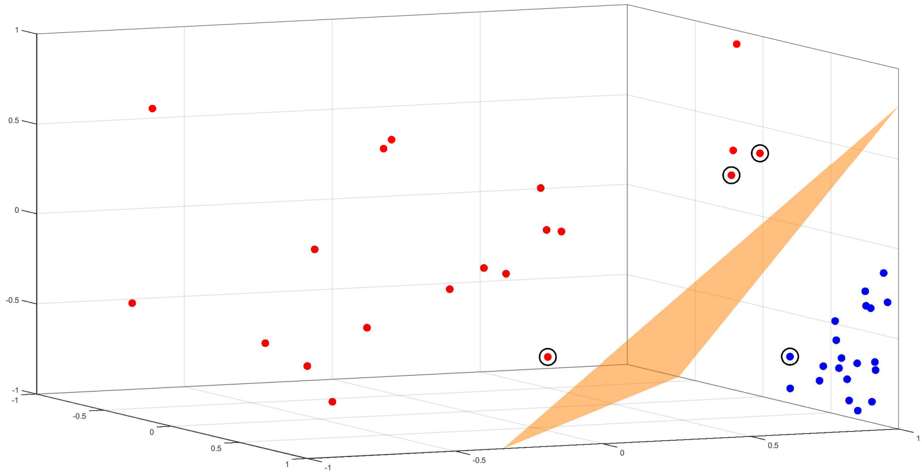

Рис. 1. $s = 20$, $m = 40$; четыре $(3+1)$ опорные точки.

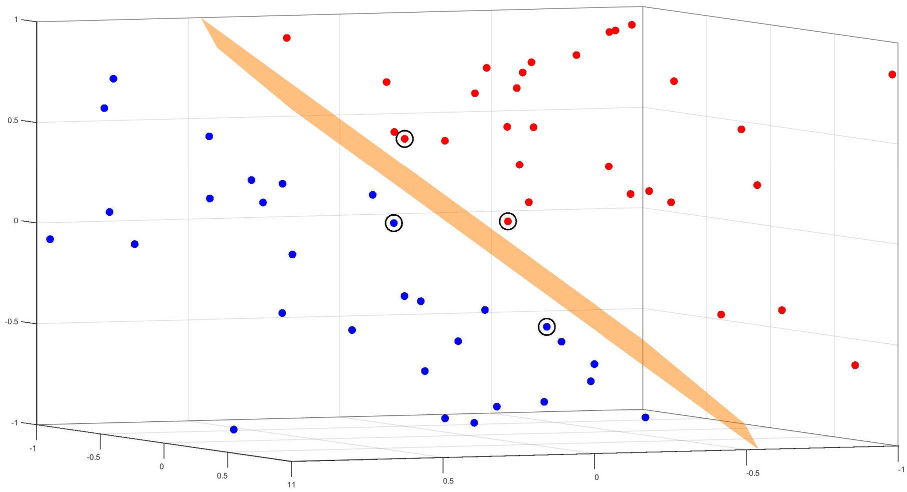

Рис. 2. $s = 30$, $m = 60$; четыре $(2+2)$ опорные точки.

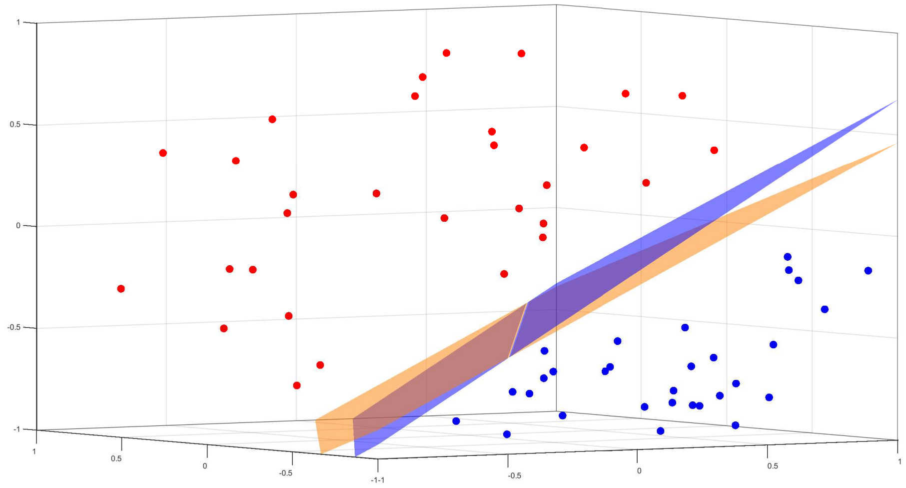

Рис. 3. $s = 30$, $m = 60$; $h = 0.213$, $h_1 = 0.202$.

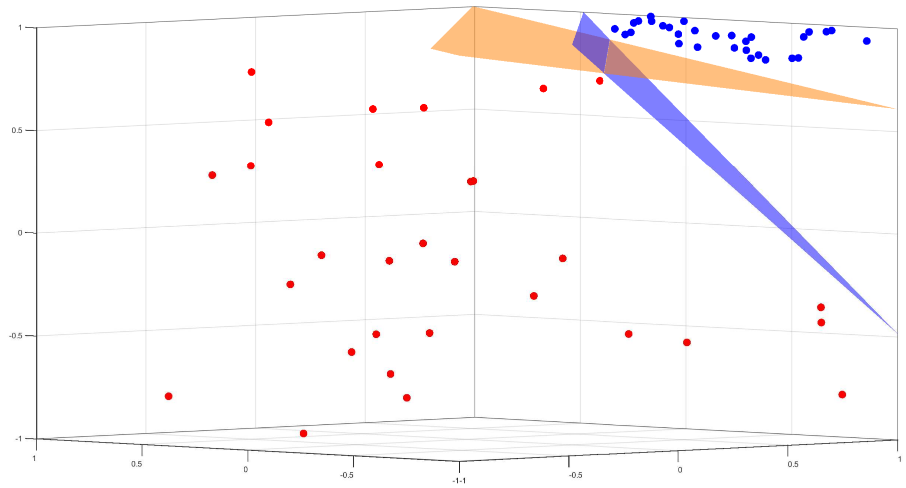

Рис. 4. $s = 30$, $m = 60$; $h = 0.130$, $h_1 = 0.095$.