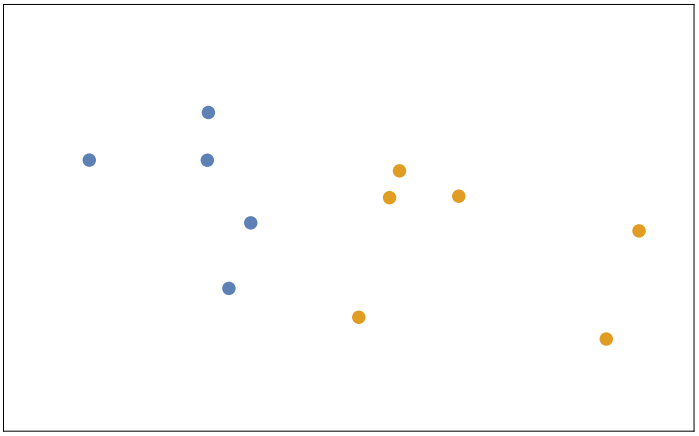
Метод опорных векторов

Рассмотрим задачу классификации на два класса $y \in \{-1, +1\}$, в которой объекты описываются n-мерными вещественными векторами: $x = \mathbb{R}^n$, $y \in \{-1, +1\}$. Будем использовать линейный классификатор вида:

$$a(x) = \text{sign}(\langle \theta, x \rangle + \theta_0). \tag{1}$$

Тогда уравнение $\langle \theta, x \rangle + \theta_0 = 0$ описывает гиперплоскость, разделяющую объекты на классы.

Рассмотрим пример линейно разделимой выборки.

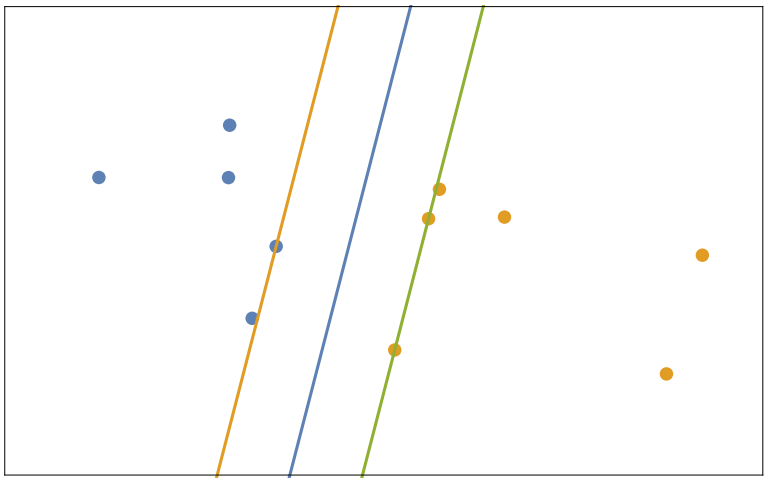


В данном случае разделяющая гиперплоскость не единственна. Функционал качества

$$Q(a(x), y) = \sum_{i=1}^{l} [y_i(\langle \theta, x_i \rangle + \theta_0) \le 0] \tag{2}$$

принимает значение 0 при различных расположениях разделяющей гиперплоскости.

Для выбора положения разделяющей гиперплоскости будем отталкиваться от идеи, что она должна находиться как можно дальше от объектов двух различных классов. Таким образом возникает понятие разделяющей полосы, или зазора между классами.



Заметим, что параметры линейного порогового классификатора определены с точностью до нормировки: алгоритм $a(x)$ не изменится, если его параметры умножить на одну и ту же положительную константу. Удобно выбрать эту константу таким образом, чтобы выполнялось условие

$$\min_{i=1,\dots,l} y_i(\langle \theta, x_i \rangle + \theta_0) = 1. \tag{3}$$

Тогда расстояние между ближайшими двумя объектами разных классов должно быть как

можно больше:

$$\frac{\langle \theta, x_+ \rangle}{\|\theta\|} - \frac{\langle \theta, x_- \rangle}{\|\theta\|} = \frac{(1-\theta_0)-(-1-\theta_0)}{\|\theta\|} = \frac{2}{\|\theta\|} \to \max.$$

Ширина разделяющей полосы максимальна, когда норма вектора θ минимальна. Таким образом можно составить задачу квадратичного программирования для построения оптимальной разделяющей гиперплоскости:

$$\frac{1}{2}\|\theta\|^2 \to \min, \tag{4}$$

$$y_i(\langle \theta, x_i \rangle + \theta_0) \geq 1, \ \ i = 1, \ \ldots, \ l. \tag{5}$$

Выражение $M_i = y_i(\langle \theta, x_i \rangle + \theta_0)$ является отступом объекта i от разделяющей гиперплоскости. Если величина отступа $M_i > 0$, то объект классифицирован верно, если же $M_i < 0$, то на объекте i модель допускает ошибку.

Таким образом, модель (4)-(5) позволяет найти оптимальную разделяющую гиперплоскость только в том случае, если выборка линейно разделима. В противном случае решения не существует, поскольку $\exists\ (x_i, y_i): y_i(\langle \theta, x_i \rangle + \theta_0) < 0$ и ограничение (5) не выполняется.

Ослабим ограничения (5):

$$\frac{1}{2}\|\theta\|^2 + C\sum_{i=1}^{l}\xi_i \to \min, \tag{6}$$

$$y_i(\langle \theta, x_i \rangle + \theta_0) \geq 1 - \xi_i, \ \ i = 1, \ \ldots, \ l, \tag{7}$$

$$\xi_i \geq 0, \ \ i = 1, \ \ldots, \ l. \tag{8}$$

Здесь C – параметр регуляризации.

> **Упражнение**
>
> От задачи (6) перейти к задаче безусловной оптимизации.

*Hinge loss

Двойственная задача

Для нахождения решения задачи нелинейного программирования воспользуемся условиями Куна-Таккера существования экстремума.

Рассмотрим задачу оптимизации следующего вида:

$$f(x) \to \min, \tag{9}$$

$$g_i(x) \leq 0, \ i = 1, \ \ldots, \ m, \tag{10}$$

$$h_j(x) = 0, \ j = 1, \ \ldots, \ k. \tag{11}$$

Выпишем необходимые условия Куна-Таккера. Если x – точка локального минимума, то найдутся множители λ_i, μ_i такие, что

$$\mathcal{L}(x; \mu, \lambda) = f(x) + \sum_{i=1}^{m}\lambda_i g_i(x) + \sum_{j=1}^{k}\mu_j h_j(x) \to \max, \tag{12}$$

$$\frac{\partial \mathcal{L}}{\partial x} = 0, \tag{13}$$

$$g_i(x) \leq 0, \tag{14}$$

$$h_j(x) = 0, \tag{15}$$

$$\lambda_i \geq 0,\ \mu_j \geq 0, \tag{16}$$

$$\lambda_i\, g_i(x) = 0, \tag{17}$$

$$\mu_j\, h_j(x) = 0. \tag{18}$$

где (14)-(15) – исходные ограничения, (17) – условие дополняющей нежесткости.

Составим функцию Лагранжа и ограничения для задачи (6)-(8).

$$\mathcal{L}(x;\, \mu,\, \lambda) = \frac{1}{2}\,\|\theta\|^2 + C\sum_{i=1}^{l}\xi_i + \sum_{i=1}^{l}\lambda_i[1 - \xi_i - y_i(\langle\theta,\, x_i\rangle + \theta_0)] - \sum_{i=1}^{l}\mu_i\,\xi_i \to \max, \tag{19}$$

$$\frac{\partial\mathcal{L}}{\partial\theta} = 0, \tag{20}$$

$$\frac{\partial\mathcal{L}}{\partial\theta_0} = 0, \tag{21}$$

$$\frac{\partial\mathcal{L}}{\partial\xi_i} = 0, \tag{22}$$

$$y_i(\langle\theta,\, x_i\rangle + \theta_0) \geq 1 - \xi_i,\ \ i = 1,\ \ldots,\ l, \tag{23}$$

$$\xi_i \geq 0,\ \ i = 1,\ \ldots,\ l. \tag{24}$$

$$\lambda_i \geq 0,\ \mu_i \geq 0, \tag{25}$$

$$\lambda_i[1 - \xi_i - y_i(\langle\theta,\, x_i\rangle + \theta_0)] = 0 \tag{26}$$

$$\mu_j\,\xi_j = 0. \tag{27}$$

Из (20): $\theta - \sum_{i=1}^{l}\lambda_i\, y_i\, x_i = 0 \Rightarrow \theta = \sum_{i=1}^{l}\lambda_i\, y_i\, x_i$.

Из (21): $-\sum_{i=1}^{l}\lambda_i\, y_i = 0\ \Rightarrow \sum_{i=1}^{l}\lambda_i\, y_i = 0$.

Из (22): $C - \lambda_i - \mu_i = 0$.

Из (26): либо $\lambda_i = 0$, либо $y_i(\langle\theta,\, x_i\rangle + \theta_0) = 1 - \xi_i$.

Из (27): либо $\mu_j = 0$, либо $\xi_i = 0$.

Из (20) следует, что искомый вектор весов θ является линейной комбинацией векторов обучающей выборки $(x_i,\, y_i)$, причём только тех, для которых $\lambda_i > 0$. Если $\lambda_i > 0$, то объект обучающей выборки x_i называется **опорным** вектором.

Т.к. $\mu_i \geq 0$, то из (22): $0 \leq \lambda_i \leq C$.

Упростим функцию Лагранжа, пользуясь полученными формулами:

$$\begin{aligned}\mathcal{L}(x;\, \mu,\, \lambda) = \frac{1}{2}\left\|\sum_{i=1}^{l}\lambda_i\, y_i\, x_i\right\|^2 + \sum_{i=1}^{l}\lambda_i\left[1 - y_i\left(\left\langle\sum_{i=1}^{l}\lambda_i\, y_i\, x_i,\ x_i\right\rangle + \theta_0\right)\right] + \sum_{i=1}^{l}\xi_i(C - \lambda_i - \mu_i) = \\ = \sum_{i=1}^{l}\lambda_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\lambda_i\,\lambda_j\, y_i\, y_j\,\langle x_i,\, x_j\rangle \to \max,\end{aligned} \tag{28}$$

при ограничениях:

$$0 \leq \lambda_i \leq C,\ \ i = 1,\ \ldots,\ l, \tag{29}$$

$$\sum_{i=1}^{l}\lambda_i\, y_i = 0. \tag{30}$$

Возможны три допустимых сочетания значений переменных ξ_i, λ_i, μ_i и отступов M_i.

1. $\lambda_i = 0,\ \mu_i = C,\ \xi_i = 0,\ M_i \geq 1$: объект $(x_i,\, y_i)$ классифицируется правильно и не влияет на

решение θ. Такие объекты называют периферийными или неинформативными.

2. $0 < \lambda_i < C,\ 0 < \mu_i < C,\ \xi_i = 0,\ M_i = 1$: объект (x_i, y_i) классифицируется правильно и лежит в точности на границе разделяющей полосы. Такие объекты называются **опорными граничными**.

3. $\lambda_i = C,\ \mu_i = 0,\ \xi_i > 0,\ M_i < 1$: объект (x_i, y_i) либо лежит внутри разделяющей полосы, но классифицируется правильно (если $0 < \xi_i < 1,\ 0 < M_i < 1$), либо попадает на границу классов ($\xi_i = 1,\ M_i = 0$), либо лежит не в своем классе (если $\xi_i > 1,\ M_i < 0$). Такой объект называют **опорным нарушителем**.

Для получения готового классификатора вычислим θ по формуле (20). Для определения θ_0 достаточно взять произвольный опорный граничный вектор и выразить θ_0 из равенства:

$$\begin{gathered} y_i(\langle \theta,\ x_i \rangle + \theta_0) = 1, \\ \text{т.к. } y_i \in \{-1,\ +1\} : \langle \theta,\ x_i \rangle + \theta_0 = y_i \ \Rightarrow\ \theta_0 = y_i - \langle \theta,\ x_i \rangle. \end{gathered} \tag{31}$$

Для повышения численной устойчивости рекомендуется брать медиану множества значений θ_0, вычисленных по всем граничным опорным векторам:

$$\theta_0 = \operatorname{med} \{ y_i - \langle \theta,\ x_i \rangle :\ \lambda_i > 0,\ M_i = 1,\ i = 1,\ \ldots,\ l \}. \tag{32}$$

Таким образом, алгоритм классификации представляется в следующем виде:

$$a(x) = \operatorname{sign}\left(\left\langle \sum_{i=1}^{l} \lambda_i\, y_i\, x_i,\ x \right\rangle + \theta_0 \right) = \operatorname{sign}\left(\sum_{i=1}^{l} \lambda_i\, y_i\, \langle x_i,\ x \rangle + \theta_0 \right) \tag{33}$$

Обратим внимание, что суммирование идёт не по всей выборке, а только по опорным векторам, для которых $\lambda_i \neq 0$. Классификатор $a(x)$ не изменится, если все остальные объекты исключить из выборки. Ненулевыми λ_i обладают не только граничные опорные объекты, но и объекты-нарушители. Это говорит о недостаточной устойчивости к шуму метода опорных векторов.