Лекция 6
Метод опорных векторов

Макаренко В.А., Габдуллин Р.А.

МГУ им. М.В. Ломоносова

14 февраля 2023

Задача классификации

X – множество объектов,
Y – множество ответов:

- $|Y| = 2$ – двухклассовая (binary) классификация.
- $|Y| = K$ – множественная (multiclass) классификация.

$y : X \to Y$ – неизвестная зависимость.

Дано:
$\{x_1, x_2, \dots, x_\ell\} \subset X$ – обучающая выборка,
$y_i = y(x_i),\ \ i = 1, \dots, \ell$ – известные ответы.

Найти:
$a : X \to Y$ – решающая функция, приближающая y на всём X.

Модель бинарной классификации

- Множество ответов:

$$Y = \{-1, 1\}.$$

- Семейство вещественных дискриминантных функций:

$$S = \{s(x, w) | w \in W\}.$$

- Семейство алгоритмов:

$$a(x, w) = \operatorname{sign} s(x, w).$$

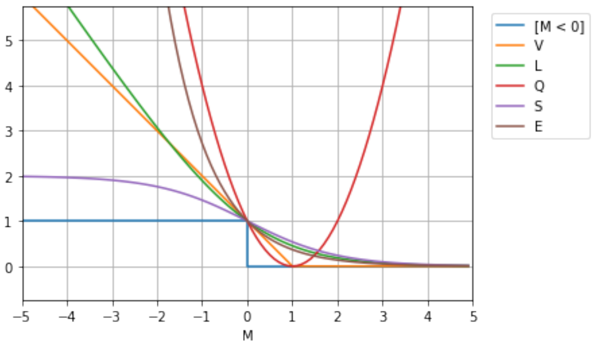
- Эмпирический риск:

$$Q(w, \mathbb{X}) = \sum_{i=1}^{\ell} [M(x_i, w) < 0] \equiv \sum_{i=1}^{\ell} [y_i \cdot s(x_i, w) < 0].$$

- Минимизация мажоранты эмпирического риска:

$$Q(w, \mathbb{X}) = \sum_{i=1}^{\ell} [M(x_i, w) < 0] \leqslant \sum_{i=1}^{\ell} \mathscr{L}(M(x_i, w)) \to \min_{w}.$$

Мажоранты эмпирического риска



Часто используемые функции потерь $\mathscr{L}$:

- $V(M) = (1 - M)_+$
- $L(M) = \log_2(1 + e^{-M})$
- $Q(M) = (1 - M)^2$
- $S(M) = 2(1 + e^{M})^{-1}$
- $E(M) = e^{-M}$

Метод опорных векторов (support vector machine)

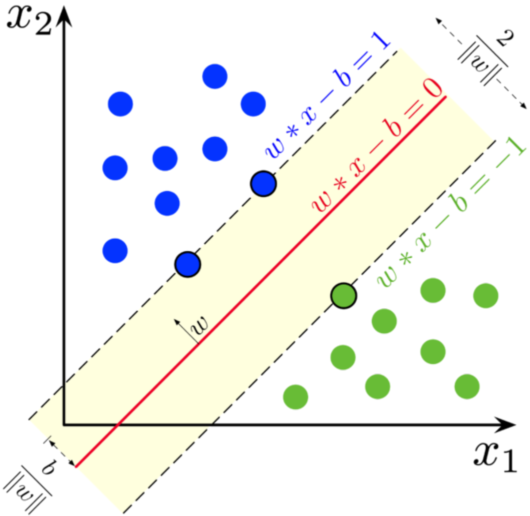


Рис.: Источник: neerc.ifmo.ru

SVM: линейно разделимый случай

- Обучающая выборка:

$$X^\ell = \{(x_i, y_i)\}_{i=1}^{\ell}, \quad x_i \in \mathbb{R}^n,\ y_i \in \{-1, 1\}.$$

- Семейство алгоритмов:

$$a(x, w, b) = \operatorname{sign}(\langle w, x\rangle - b).$$

- Отступ на i-м объекте:

$$M_i(w, b) = y_i(\langle w, x_i\rangle - b).$$

- Ориентированное расстояние от i-го объекта до гиперплоскости:

$$\frac{y_i(\langle w, x_i\rangle - b)}{\|w\|} = \frac{M_i(w, b)}{\|w\|}.$$

Видно, что значение не меняется при домножении w и b на одно и то же положительное число.

SVM: линейно разделимый случай

- Цель – сделать ориентированное расстояние от разделяющей гиперплоскости до ближайшего к ней объекта как можно больше:

$$\min_{1\leqslant i\leqslant \ell} \frac{y_i(\langle w, x_i\rangle - b)}{\|w\|} = \min_{1\leqslant i\leqslant \ell} \frac{M_i(w,b)}{\|w\|} \to \max_{w,b}.$$

- Задача оптимизации (за счет возможности нормировки):

$$\begin{cases} \frac{1}{\|w\|} \to \max\limits_{w,b}, \\ \min\limits_{1\leqslant i\leqslant \ell} M_i(w,b) = 1. \end{cases} \iff \begin{cases} \frac{1}{2}\|w\|^2 \to \min\limits_{w,b}, \\ M_i(w,b) \geqslant 1, \quad i = \overline{1,\ell}. \end{cases}$$

- В результате:
  - Минимальный отступ: $\min\limits_{1\leqslant i\leqslant \ell} M_i(w,b) = 1$.
  - Расстояние до блажайшего объекта: $\frac{1}{\|w\|}$.
  - Расстояние до начала координат: $\frac{|b|}{\|w\|}$.
  - Ширина полосы: $\frac{2}{\|w\|}$.

SVM: линейно неразделимый случай

- Введем штрафы за попадание в разделяющую полосу или на территорию другого класса.
- Задача оптимизации:

$$\begin{cases} M_i(w,b) \geqslant 1-\xi_i, & 1 \leqslant i \leqslant \ell, \\ \xi_i \geqslant 0, & 1 \leqslant i \leqslant \ell, \\ \frac{1}{2}\|w\|^2 + C\sum\limits_{i=1}^{\ell}\xi_i \to \min, \end{cases}$$

  где $C > 0$ – гиперпараметр.
- Эквивалентная задача безусловной оптимизации (hinge loss):

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{\ell}(1 - M_i(w,b))_+ \to \min$$

Условия Каруша–Куна–Таккера

Задача нелинейного программирования:

$$\begin{cases} f(x) \to \min\limits_{x \in X}, \\ g_i(x) \leqslant 0, \qquad 1 \leqslant i \leqslant m, \\ h_j(x) = 0, \qquad 1 \leqslant j \leqslant k. \end{cases}$$

Если x – точка локального минимума, то существуют такие множители $\mu_i, \lambda_j \, (1 \leqslant i \leqslant m, 1 \leqslant j \leqslant k)$, что для функции Лагранжа

$$L(x, \mu, \lambda) = f(x) + \sum_{i=1}^{m} \mu_i g_i(x) + \sum_{j=1}^{k} \lambda_j h_j(x)$$

выполняются условия

$$\begin{cases} \frac{\partial L}{\partial x} = 0, & \\ g_i(x) \leqslant 0, h_j(x) = 0, & \text{(исходные ограничения)} \\ \mu_i \geqslant 0, & \text{(двойственные ограничения)} \\ \mu_i g_i(x) = 0 & \text{(условия дополняющей нежесткости)} \end{cases}$$

Условия Каруша–Куна–Таккера в SVM

Задача оптимизации:

$$\begin{cases} M_i(w,b) \geqslant 1 - \xi_i, & 1 \leqslant i \leqslant \ell, \\ \xi_i \geqslant 0, & 1 \leqslant i \leqslant \ell, \\ \frac{1}{2}\|w\|^2 + C\sum\limits_{i=1}^{\ell} \xi_i \to \min. \end{cases}$$

Функция Лагранжа:

$$L(w,b,\xi,\lambda,\eta) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{\ell} \lambda_i (M_i(w,b) - 1) - \sum_{i=1}^{\ell} \xi_i (\lambda_i + \eta_i - C)$$

Условия Каруша–Куна–Таккера:

$$\begin{cases} \frac{\partial L}{\partial w} = 0, \quad \frac{\partial L}{\partial b} = 0, \quad \frac{\partial L}{\partial \xi} = 0, \\ \xi_i \geqslant 0, \quad \lambda_i \geqslant 0, \quad \eta_i \geqslant 0, \quad M_i(w,b) \geqslant 1 - \xi_i, & 1 \leqslant i \leqslant \ell, \\ \lambda_i = 0 \quad \text{либо} \quad M_i(w,b) = 1 - \xi_i, & 1 \leqslant i \leqslant \ell, \\ \eta_i = 0 \quad \text{либо} \quad \xi_i = 0, & 1 \leqslant i \leqslant \ell. \end{cases}$$

Условия Каруша–Куна–Таккера в SVM

Функция Лагранжа:

$$L(w, b, \xi, \lambda, \eta) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{\ell} \lambda_i (M_i(w, b) - 1) - \sum_{i=1}^{\ell} \xi_i (\lambda_i + \eta_i - C)$$

Продифференцируем и приравняем производные к нулю:

$$\frac{\partial L}{\partial w} = w - \sum_{i=1}^{\ell} \lambda_i y_i x_i = 0 \quad \Longleftrightarrow \quad w = \sum_{i=1}^{\ell} \lambda_i y_i x_i,$$

$$\frac{\partial L}{\partial b} = -\sum_{i=1}^{\ell} \lambda_i y_i = 0 \quad \Longleftrightarrow \quad \sum_{i=1}^{\ell} \lambda_i y_i = 0,$$

$$\frac{\partial L}{\partial \xi_i} = -\lambda_i - \eta_i + C = 0 \quad \Longleftrightarrow \quad \lambda_i + \eta_i = C, \quad 1 \leqslant i \leqslant \ell.$$

Условия Каруша–Куна–Таккера в SVM

Условия Каруша–Куна–Таккера:

$$
\begin{cases}
w = \sum\limits_{i=1}^{\ell} \lambda_i y_i x_i, \quad \sum\limits_{i=1}^{\ell} \lambda_i y_i = 0, \quad \lambda_i + \eta_i = C, & 1 \leqslant i \leqslant \ell \\
\xi_i \geqslant 0, \quad \lambda_i \geqslant 0, \quad \eta_i \geqslant 0, \quad M_i(w, b) \geqslant 1 - \xi_i, & 1 \leqslant i \leqslant \ell, \\
\lambda_i = 0 \quad \text{либо} \quad M_i(w, b) = 1 - \xi_i, & 1 \leqslant i \leqslant \ell, \\
\eta_i = 0 \quad \text{либо} \quad \xi_i = 0, & 1 \leqslant i \leqslant \ell.
\end{cases}
$$

- Объект x_i называется опорным, если $\lambda_i \neq 0$.
- Можем разделить объекты на три типа:
  1. $\lambda_i = 0 \;\Rightarrow\; \eta_i = C, \xi_i = 0, M_i \geqslant 1$ – периферийные объекты,
  2. $0 < \lambda_i < C \;\Rightarrow\; 0 < \eta_i < C,\ \xi_i = 0,\ M_i = 1$ – опорные граничные объекты,
  3. $\lambda_i = C \;\Rightarrow\; \eta_i = 0,\ \xi_i > 0, M_i < 1$ – опорные объекты-нарушители.

Двойственная задача

Подставляем в функцию Лагранжа полученные ограничения и приходим к двойственной задаче:

$$\begin{cases} -L(\lambda) = -\sum\limits_{i=1}^{\ell} \lambda_i + \frac{1}{2} \sum\limits_{i=1}^{\ell} \sum\limits_{j=1}^{\ell} \lambda_i \lambda_j y_i y_j \langle x_i, x_j \rangle \to \min\limits_{\lambda}, \\ 0 \leqslant \lambda \leqslant C, \quad 1 \leqslant i \leqslant \ell, \\ \sum\limits_{i=1}^{\ell} \lambda_i y_i = 0. \end{cases}$$

Решение прямой задачи выражается через решение двойственной:

$$\begin{cases} w = \sum\limits_{i=1}^{\ell} \lambda_i y_i x_i, \\ b = \langle w, x_i \rangle - y_i, \quad \forall i : M_i = 1. \end{cases}$$

Линейный классификатор принимает вид:

$$a(x) = \operatorname{sign}\left(\sum_{i=1}^{\ell} \lambda_i y_i \langle x, x_i \rangle - b \right).$$

Нелинейное обобщение, ядерный переход

Отобразим объекты в пространство более высокой размерности с помощью функции $\psi : X \to H$. Будем считать, что пространство H обладает скалярным произведением, тогда

$$a(x) = \operatorname{sign}\left(\sum_{i=1}^{\ell} \lambda_i y_i \langle \psi(x), \psi(x_i)\rangle - b\right) =$$

$$= \operatorname{sign}\left(\sum_{i=1}^{\ell} \lambda_i y_i K(x, x_i) - b\right),$$

где $K(x, x_i) = \langle \psi(x), \psi(x_i)\rangle$

Определение

Функция $K : X \times X \to H$ называется ядром, если она представима в виде

$$K(x, x') = \langle \psi(x), \psi(x')\rangle$$

при некотором отображении $\psi : X \to H$, где H – пространство со скалярным произведением.

Критерий ядерной функции

Теорема (Мерсер)

Функция $K(x, x')$ является ядром тогда и только тогда, когда выполнены два условия:

- *Симметричность:*

$$K(x, x') = K(x', x).$$

- *Неотрицательная определенность:*

$$\int_X \int_X K(x, x')g(x)g(x')\,dx\,dx' \geqslant 0, \quad \forall g : X \to \mathbb{R}.$$

Конструирование ядер

- Скалярное произведение $< x, x' >$ является ядром.
- Константа $K(x, x') = 1$ является ядром.
- Произведение ядер $K(x, x') = K_1(x, x')K_2(x, x')$ является ядром.
- Для любой функции $\psi : X \to \mathbb{R}$ произведение $K(x, x') = \psi(x)\psi(x')$ является ядром.
- Линейная комбинация ядер с неотрицательными коэффициентами $K(x, x') = \alpha K_1(x, x') + \beta K_2(x, x')$ является ядром.
- Композиция произвольной функции $\psi : X \to X$ и произвольного ядра $K(x, x') = K_0(\psi(x), \psi(x'))$ является ядром.
- Композиция произвольного ядра и произвольной функции $f : \mathbb{R} \to \mathbb{R}$, представимой в виде сходящегося степенного ряда с неотрицательными коэффициентами $K(x, x') = f(K_0(x, x'))$ является ядром. В частности, функции $f(z) = e^z$ и $f(z) = \frac{1}{1-z}$ от ядра являются ядрами.

Примеры ядер

Следующие функции являются ядрами:

- $K(x, x') = \langle x, x' \rangle$ – линейное ядро
- $K(x, x') = (\gamma \langle x, x' \rangle + r)^d$ – полиномиальное ядро.
- $K(x, x') = \exp(-\gamma \|x - x'\|^2)$ – сеть радиальных базисных функций.
- $K(x, x,')= \tan(\gamma \langle x, x' \rangle + r)$ – сигмоидная.

Связь SVM с нейронными сетями

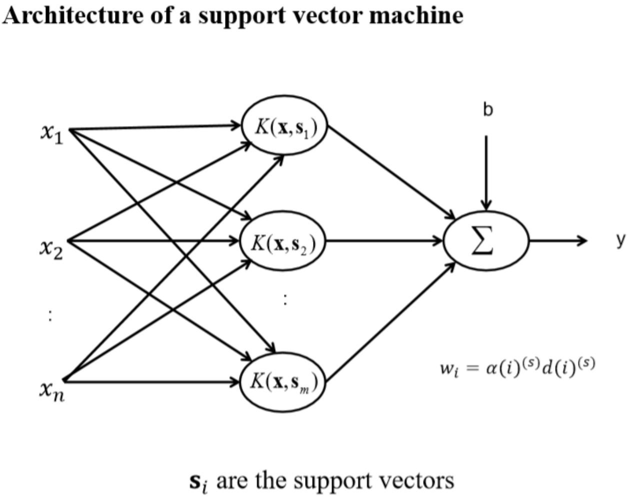


Рис.: Источник: stackexchange.com

Метод наименьших квадратов с L_2-регуляризатором:

$$Q(a, \mathbb{X}) = \sum_{i=1}^{\ell} (\langle w, x_i \rangle + w_0 - y_i)^2 + \tau \|w\|^2 \to \min_{w, w_0} .$$

Функция потерь SVM в задаче регрессии:

$$Q(a, \mathbb{X})_\varepsilon = \sum_{i=1}^{\ell} \left| \langle w, x_i \rangle + w_0 - y_i \right|_\varepsilon + \tau \|w\|^2 \to \min_{w, w_0},$$

где

$$|z|_\varepsilon = \max\{0, |z| - \varepsilon\}.$$