Машинное обучение ФИВТ

DS-поток

Лекция 7

SVM
Метод опорных векторов

Бинарная классификация

SVM

Исходная задача:

решать сложно

$$\begin{cases} \frac{1}{2}\|\theta\|^2 + C\sum\limits_{i=1}^{n}\xi_i^+ \longrightarrow \min\limits_{\theta,\theta_0,\xi} \\ Y_i(\theta^T X_i + \theta_0) \geqslant 1 - \xi_i \quad i = 1, \ldots, n \end{cases}$$

Двойственная задача

решать легко

(*квадр. прогр.*)

$$\begin{cases} -\frac{1}{2}\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{n}\lambda_i\lambda_j Y_i Y_j \langle X_i, X_j\rangle + \sum\limits_{i=1}^{n}\lambda_i \longrightarrow \max\limits_{\lambda} \\ 0 \leqslant \lambda_i \leqslant C \quad i = 1, \ldots, n; \\ \sum\limits_{i=1}^{n}\lambda_i Y_i = 0 \end{cases}$$

Решение: $\widehat{\theta} = \sum\limits_{i=1}^{n}\lambda_i Y_i X_i; \quad \widehat{\theta}_0 = Y_i - \widehat{\theta}^T X_i$ где i т.ч. $\xi_i = 0$

$$\widehat{y}(x) = \left\langle \widehat{\theta}, x \right\rangle + \widehat{\theta}_0$$

SVM

Исходная задача

$$\begin{cases} \dfrac{1}{2}\|\theta\|^2 + C\sum\limits_{i=1}^{n}\xi_i^+ \longrightarrow \min\limits_{\theta,\theta_0,\xi} \\ Y_i(\theta^T X_i + \theta_0) \geqslant 1 - \xi_i \quad i = 1, \ldots, n \end{cases}$$

Преобразуем:

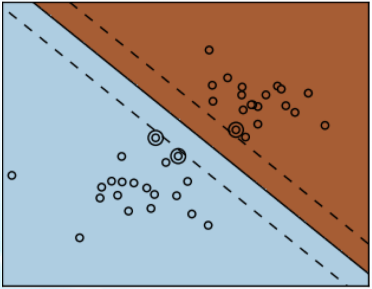
$$\sum_{i=1}^{n}(1 - M_i)^+ + \frac{1}{2C}\|\theta\|^2 \to \min_{\theta,\theta_0}$$

$$\text{где } M_i = Y_i(\theta^T X_i + \theta_0)$$

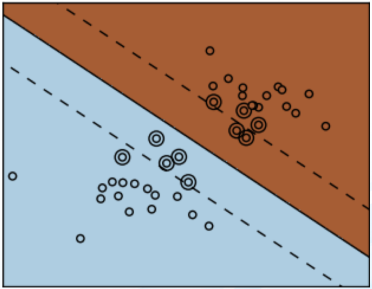
Вывод: SVM эквивалентен поиску линейного классификатора с ф-ией потерь $L(y, z) = (1 - yz)^+$ [*hinge loss*] и L_2-регуляризацией.

SVM: влияние константы C

Большой C
Слабая регуляризация



Малый C
Сильная регуляризация



На практике небольшие изменения в значении C не сильно влияют на вид классификатора.

$\Rightarrow$ C можно подбирать по достаточно грубой сетке.

SVM: ключевые моменты

- SVM — линейный классификатор с кусочно-линейной функцией потерь (hinge loss) и L2-регуляризацией.
- Придуман из соображений максим. зазора между классами.
- При линейно-разделимой выборке это означает максимизацию ширины разделющей полосы.
- При линейно неразделимой выборке добавляется возможность попадания объектов в полосу и штрафы за эти попадания.
- Объекты в глубине классов не влияют на разделяющую гиперплоскость.

Kernel trick

Недостаток:

Является линейной моделью $\Rightarrow$ разделяюшие полосы линейные.

Рассмотрим преобразование пространства объектов: $\psi : \mathscr{X} \to \mathscr{H}$,

где $\mathscr{X} = \mathbb{R}^d$ — исходное пр-во,

$\mathscr{H}$ — гильбертово пр-во, возможно, бесконечномерное.

Напоминание: в гильбертовом пр-ве есть скалярное произведение.

Обозначим $K(x_1, x_2) = \langle \psi(x_1), \psi(x_2) \rangle$

Наблюдения:

1. $\theta \in \mathscr{X} \Rightarrow$ к нему тоже применимо преобразование ψ.
2. используем только скалярные произведения.

SVM

Исходная задача:

решать сложно

$$\begin{cases} \frac{1}{2}\langle\theta,\theta\rangle + C\sum_{i=1}^{n}\xi_i^{+} \longrightarrow \min\limits_{\theta,\theta_0,\xi} \\ Y_i(\langle\theta, X_i\rangle + \theta_0) \geqslant 1-\xi_i \quad i=1,\dots,n \end{cases}$$

Двойственная задача

решать легко

(*квадр. прогр.*)

$$\begin{cases} -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j Y_i Y_j \langle X_i, X_j\rangle + \sum_{i=1}^{n}\lambda_i \longrightarrow \max\limits_{\lambda} \\ 0 \leqslant \lambda_i \leqslant C \quad i=1,\dots,n; \\ \sum_{i=1}^{n}\lambda_i Y_i = 0 \end{cases}$$

Решение: $\widehat{\theta} = \sum_{i=1}^{n}\lambda_i Y_i X_i$; $\quad \widehat{\theta}_0 = Y_i - \langle\theta, X_i\rangle$ где i т.ч. $\xi_i = 0$

$$\widehat{y}(x) = \left\langle \widehat{\theta}, x \right\rangle + \widehat{\theta}_0$$

SVM

Исходная задача:

решать сложно

$$\begin{cases} \frac{1}{2}\langle\theta,\theta\rangle + C\sum_{i=1}^{n}\xi_i^{+} \longrightarrow \min\limits_{\theta,\theta_0,\xi} \\ Y_i(\langle\theta, X_i\rangle + \theta_0) \geqslant 1 - \xi_i \quad i = 1,\dots,n \end{cases}$$

Двойственная задача

решать легко

(квадр. прогр.)

$$\begin{cases} -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j Y_i Y_j\,\langle X_i, X_j\rangle + \sum_{i=1}^{n}\lambda_i \longrightarrow \max\limits_{\lambda} \\ 0 \leqslant \lambda_i \leqslant C \quad i = 1,\dots,n; \\ \sum_{i=1}^{n}\lambda_i Y_i = 0 \end{cases}$$

Решение: $\cancel{\widehat{\theta} = \sum_{i=1}^{n}\lambda_i Y_i X_i; \quad \widehat{\theta}_0 = Y_i - \langle\theta, X_i\rangle \text{ где } i \text{ т.ч. } \xi_i = 0}$

$$\widehat{y}(x) = \sum_{i=1}^{n}\lambda_i Y_i\,\langle X_i, x\rangle + \left(Y_\ell - \sum_{i=1}^{n}\lambda_i Y_i\langle X_i, X_\ell\rangle\right), \text{ т.ч. } X_\ell \text{ на границе.}$$

SVM

Исходная задача:

решать сложно

$$\begin{cases} \frac{1}{2}K(\theta,\theta) + C\sum_{i=1}^{n}\xi_i^{+} \longrightarrow \min\limits_{\theta,\theta_0,\xi} \\ Y_i(K(\theta,X_i)+\theta_0) \geqslant 1-\xi_i \quad i=1,\ldots,n \end{cases}$$

Двойственная задача

решать легко

(*квадр. прогр.*)

$$\begin{cases} -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j Y_i Y_j K(X_i,X_j) + \sum_{i=1}^{n}\lambda_i \longrightarrow \max\limits_{\lambda} \\ 0 \leqslant \lambda_i \leqslant C \quad i=1,\ldots,n; \\ \sum_{i=1}^{n}\lambda_i Y_i = 0 \end{cases}$$

Решение: $\cancel{\widehat{\theta} = \sum_{i=1}^{n}\lambda_i Y_i X_i; \quad \widehat{\theta}_0 = Y_i - \langle\theta, X_i\rangle \text{ где } i \text{ т.ч. } \xi_i = 0}$

$$\widehat{y}(x) = \sum_{i=1}^{n}\lambda_i Y_i K(X_i,x) + \left(Y_\ell - \sum_{i=1}^{n}\lambda_i Y_i K(X_i,X_\ell)\right), \text{ т.ч. } X_\ell \text{ на}$$

границе

Kernel trick

Наблюдение: не нужно знать саму ψ.

Определение:

$K(x_1, x_2)$ — *ядро*, если $\exists \psi : \mathscr{X} \to \mathscr{H}$ т.ч. $K(x_1, x_2) = \langle \psi(x_1), \psi(x_2) \rangle$,

где $\mathscr{H}$ — некоторое гильбертово пространство.

Замечание. Не путать с ядром из KDE.

Теорема:

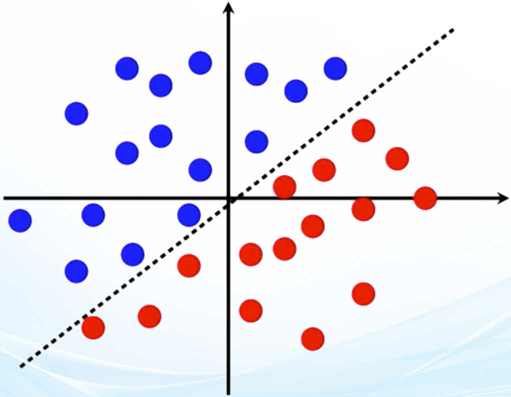
Функция $K(x_1, x_2)$ является ядром $\iff$

- $K(x_1, x_2)$ симметрична: $K(x_1, x_2) = K(x_2, x_1)$;
- $K(x_1, x_2)$ неотрицательно определена:

  т.е. для любых $x_1, ..., x_n \in \mathscr{X}$ матрица $(K(x_i, x_j))_{ij}$ неотр. определена.

Kernel trick

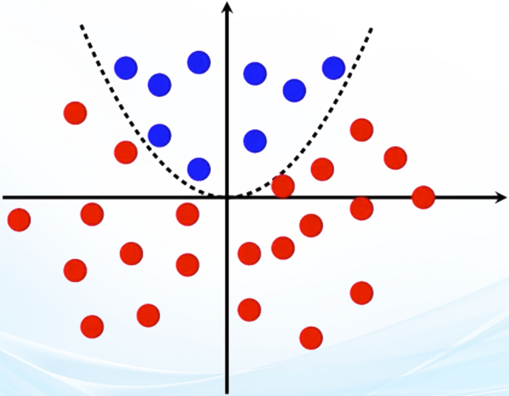
Линейное ядро



$$K(x_1, x_2) = \langle x_1, x_2 \rangle$$

Kernel trick

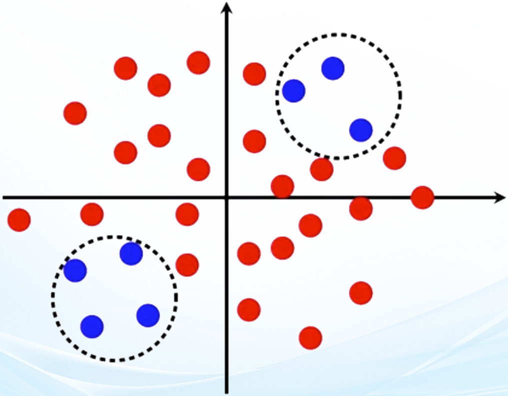
Полиномиальное ядро



$$K(x_1, x_2) = (\gamma\langle x_1, x_2\rangle + r)^d$$

Kernel trick

Радиальное ядро



$$K(x_1, x_2) = e^{-\gamma \langle x_1 - x_2 \rangle^2}$$

Kernel trick

Почему это ядра?

Рассмотрим пример квадратичного ядра, т.е. $K(x,z) = \langle x, z\rangle^2$.

Пусть $\mathscr{X} = \mathbb{R}^2$, т.е. $x = (x_1, x_2),\ z = (z_1, z_2)$

Разложим $K(x,z)$:

$$K(x,z) = \ \langle x, z\rangle^2 \ = \ \langle (x_1, x_2), (z_1, z_2)\rangle^2 \ =$$

$$= (x_1 z_1 + x_2 z_2)^2 = x_1^2 z_1^2 + x_2^2 z_2^2 + 2 x_1 z_1 x_2 z_2 =$$

$$= \ \left\langle \left(x_1^2, x_2^2, \sqrt{2} x_1 x_2 \right)\ ,\ \left(z_1^2, z_2^2, \sqrt{2} z_1 z_2 \right) \right\rangle$$

Таким образом, $\mathscr{H} = \mathbb{R}^3,\ \ \psi(x) = \left(x_1^2, x_2^2, \sqrt{2} x_1 x_2 \right)$

Линейная поверхность в $\mathscr{H}$ соответствует квадратичной в $\mathscr{X}$.

Пространство $\mathscr{H}$ называется **спрямляющим**.

Недостатки SVM

- Нужно подбирать C.
- Нет общих рекомендаций для выбора $K(x, z)$.
- Работает только для бинарной классификации.

SVM
Метод опорных векторов

Многоклассовый случай

Многоклассовый SVM

Теперь $Y_i \in \{1, ..., K\}$.

Пусть в выборке имеется константный признак
т.е. не нужно явно указывать сдвиг θ_0.

Идея: отделяем класс k от всех остальных по знач. $\langle \theta_k, x \rangle$.
где θ_k — вектор параметров.

Итоговое предсказание имеет вид

$$\widehat{y}(x) = \operatorname*{arg\,max}_{k \in \{1,...,K\}} \langle \theta_k, x \rangle.$$

Двуклассовый случай

$$\begin{cases} \dfrac{1}{2} \|\theta_1\|^2 \to \min\limits_{\theta_1} \\ Y_i \langle \theta_1, X_i \rangle \geqslant 1, \quad i = 1, \ldots, n \end{cases}$$

Многоклассовый

$$\begin{cases} \dfrac{1}{2} \sum\limits_{k=1}^{K} \|\theta_k\|^2 \to \min\limits_{\theta} \\ \langle \theta_{Y_i}, X_i \rangle - \langle \theta_k, X_i \rangle \geqslant 1, \quad i = 1..n; \ k \in \{1..K\} \setminus \{Y_i\} \end{cases}$$

Многоклассовый SVM

Общий случай

Двуклассовый:

$$\begin{cases} \frac{1}{2}\|\theta\|^2 + C\sum_{i=1}^{n}\xi_i^+ \longrightarrow \min\limits_{\theta,\theta_0,\xi} \\ Y_i(\theta^T X_i + \theta_0) \geqslant 1 - \xi_i \quad i = 1, \dots, n \end{cases}$$

Многоклассовый:

$$\begin{cases} \frac{1}{2}\sum_{k=1}^{K}\|\theta_k\|^2 + C\sum_{i=1}^{n}\xi_i^+ \longrightarrow \min\limits_{\theta,\xi} \\ \langle\theta_{Y_i}, X_i\rangle - \langle\theta_k, X_i\rangle \geqslant 1 - \xi_i, \quad i = 1..n;\ k \in \{1..K\} \setminus \{Y_i\} \end{cases}$$

Эквивалентная функция потерь

Рассмотрим следующую **функцию потерь**:

$$\mathcal{L}(X_i) = \max_k \Big\{ \langle \theta_k, X_i \rangle + 1 - I\{k = Y_i\} \Big\} - \langle \theta_{Y_i}, X_i \rangle$$

Выражение, по которому берется max:

- Если $k = Y_i$, то оно равно $\langle \theta_k, X_i \rangle$, иначе оно равно $\langle \theta_k, X_i \rangle + 1$

Получаем

- Если оценка за верный класс больше оценок за остальные классы хотя бы на единицу, т.е. $\forall k : \langle \theta_{Y_i}, X_i \rangle > \langle \theta_k, X_i \rangle + 1$,

  $\Rightarrow$ Максимум достигается на $k = Y_i$

  $\Rightarrow \mathcal{L}(X_i) = \langle \theta_{Y_i}, X_i \rangle - \langle \theta_{Y_i}, X_i \rangle = 0$

- Иначе, т.е. $\exists k : \langle \theta_{Y_i}, X_i \rangle < \langle \theta_k, X_i \rangle + 1$

  $\Rightarrow$ Максимум достигается на $k \neq Y_i$

  $\Rightarrow \mathcal{L}(X_i) = \langle \theta_k, X_i \rangle + 1 - \langle \theta_{Y_i}, X_i \rangle > 0$

Вывод: Штрафуем как за неверный ответ на объекте, так и за попадание в разделяющую полосу

SVM
Метод опорных векторов

История

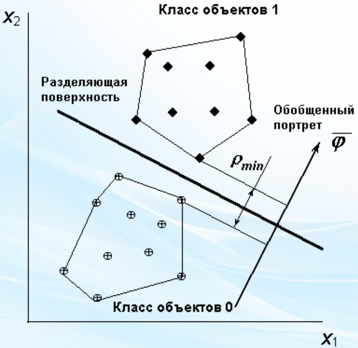
Метод обобщенного портрета

Предложен в 1960-е Вапником и Червоненкисом.

Первая версия: Отделение отделение гиперплоскостью точек одного класса на сфере от всего остального.

Вторая версия: Разделение гиперплоскостью двух классов на сфере.

Третья версия: Разделение гиперплоскостью двух классов в пр-ве.

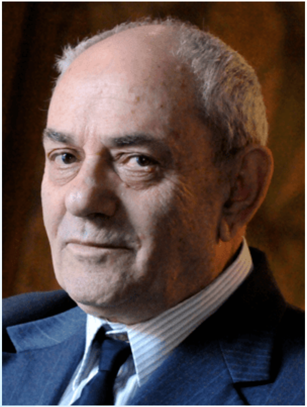


SVM: предложен Вапником в 1990-е

Ключевые отличия:

1. Ошибки ξ_i
2. Ядра.

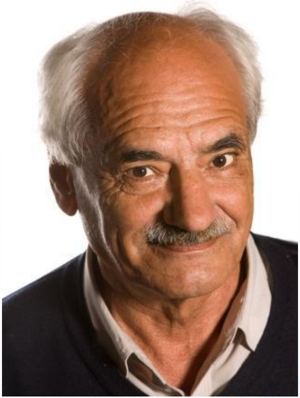
В середине 2000-х был популярным.



Вапник Владимир Наумович

род. 1936

учился в Узбекистане



Червоненкис Алексей Яковлевич

1938-2014

выпускник МФТИ, 1961