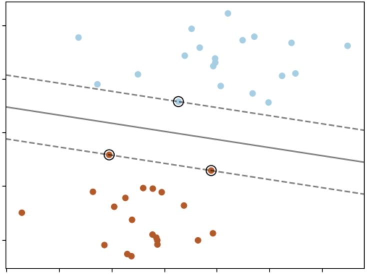
Метод опорных векторов

Виктор Китов

victorkitov.github.io

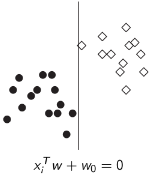


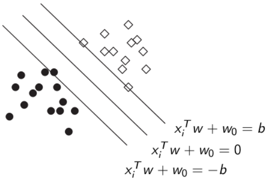
Курс поддержан фондом 'Интеллект'

Победитель конкурса VK среди курсов по IT

Содержание

Метод опорных векторов





Рассмотрим бинарную классификацию $y \in \{+1, -1\}$ линейно разделимой выборки.

Идея метода опорных векторов (support vector machines, SVM)

Выберем гиперплоскость, разделяющую классы с максимальным зазором.

Гиперплоскости $x_i^T w + w_0 = 0$, $x_i^T w + w_0 = b$, $x_i^T w + w_0 = -b$ поэтому величина зазора $\frac{2b}{\|w\|}$.

Метод опорных векторов

Объекты (x_i, y_i) отделены от разделяющей гиперплоскости $\geq \frac{b}{\|w\|}$, если

$$\begin{cases} x_i^T w + w_0 \geq b, & y_i = +1 \\ x_i^T w + w_0 \leq -b & y_i = -1 \end{cases} \quad i = 1, 2, \ldots N.$$

Это можно записать в виде

$$y_i(x_i^T w + w_0) \geq b, \quad i = 1, 2, \ldots N.$$

Максимизация зазора между классами:

$$2b / \|w\| \to \max_{w, w_0, b}$$

Оптимизационная задача

Оптимизационная задача:

$$\begin{cases} \frac{2b}{\|w\|} \to \max\limits_{w,w_0,b} \\ y_i(x_i^T w + w_0) \geq b, \quad i = 1, 2, ...N. \end{cases}$$

Если (w, w_0, b)-решение, то $(\alpha w, \alpha w_0, \alpha b)$ - тоже решение $\forall \alpha > 0$. Положим $b = 1$ $(\alpha = \frac{1}{b})$.

$$\begin{cases} \frac{2}{\|w\|} \to \max\limits_{w,w_0} \\ y_i(x_i^T w + w_0) \geq 1 \quad i = 1, 2, ...N. \end{cases}$$

Используя свойство $\arg\max \frac{2}{\|w\|} = \arg\min \frac{\|w\|}{2} = \arg\min \frac{\|w\|^2}{2}$:

$$\begin{cases} \frac{1}{2} w^T w \to \min\limits_{w,w_0} \\ y_i(x_i^T w + w_0) = M(x_i, y_i) \geq 1, \quad i = 1, 2, ...N. \end{cases}$$

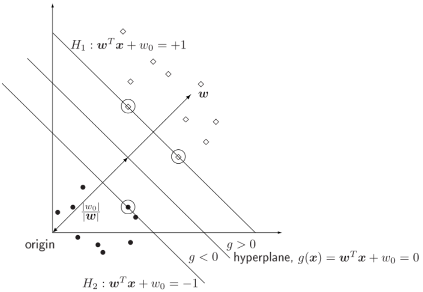
Типы объектов

Неинформативные объекты: $y_i(x_i^T w + w_0) > 1$

- не влияют на решение

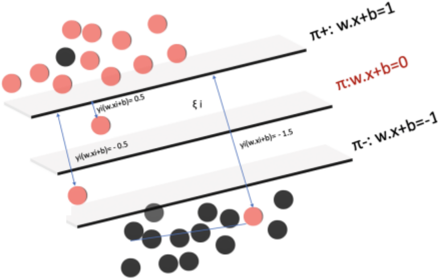
Опорные вектора: $y_i(x_i^T w + w_0) = 1$

- лежат на расстоянии $1/\|w\|$ к разделяющей гиперплоскости
- влияют на решение



Содержание

Линейно неразделимый случай



$$\begin{cases} \frac{1}{2} w^T w \to \min\limits_{w, w_0} \\ y_i(x_i^T w + w_0) = M(x_i, y_i) \geq 1, \quad i = 1, 2, \dots N. \end{cases}$$

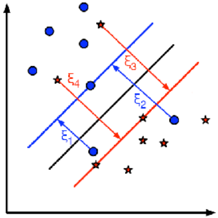
Ограничения становятся несовместными=>пустое множество решений.

Линейно неразделимый случай

Разрешим частичные нарушения ограничений на величины нарушений ξ_i (slack variables):

$$\begin{cases} \frac{1}{2}w^T w + C\sum_{i=1}^{N}\xi_i \to \min_{w,w_0,\xi} \\ y_i(w^T x_i + w_0) = M(x_i, y_i) \geq 1 - \xi_i, \; i = 1, 2, \dots N \\ \xi_i \geq 0, \; i = 1, 2, \dots N \end{cases}$$

- Штраф за нарушение C контролирует точность модели (в противовес простоте).
- Подбирается по сетке на валидации.
- Другие штрафы возможны, например $C\sum_i \xi_i^2$.



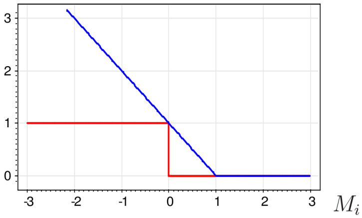
Типы объектов

- **Неинформативные объекты**:
  - $y_i(w^T x_i + w_0) > 1$
- **Опорные вектора** SV:
  - $y_i(w^T x_i + w_0) \leq 1$
  - **пограничные** $\widetilde{SV}$:
    - $y_i(w^T x_i + w_0) = 1$
  - **объекты-нарушители**:
    - $y_i(w^T x_i + w_0) > 0$: нарушитель корректно классифицирован
    - $y_i(w^T x_i + w_0) < 0$: нарушитель некорректно классифицирован

Безусловная оптимизация

Оптимизационная задача:

$$\begin{cases}\frac{1}{2}w^T w + C\sum_{i=1}^N \xi_i \to \min_{w,w_0,\xi} \\ y_i(w^T x_i + w_0) = M_i(w,w_0) \geq 1-\xi_i, \\ \xi_i \geq 0,\ i=1,2,...N\end{cases}$$



может быть переписана как

$$\frac{1}{2C}\|w\|_2^2 + \sum_{i=1}^N [1-M_i(w,w_0)]_+ \to \min_{w,w_0}$$

Таким образом, метод - линейный классификатор с функцией потерь $\mathcal{L}(M)=[1-M]_+$ и L_2 регуляризацией (обобщается на другие).

Разреженность решения

- Решение зависит только от опорных векторов.
- Это видно из условия $\mathcal{L}(M) = 0$ для $M \geq 1$.
    - хорошо классифицированные объекты с $M\geq 1$ не влияют на решение
- Разреженность решения - метод менее устойчив к выбросам
    - выбросы - всегда опорные объекты

Использование SVM

```
...
from sklearn.svm import SVC
from sklearn.metrics import accuracy_score

X_train, X_test, Y_train, Y_test =
    get_demo_classification_data()
model = SVC(C=1)              # инициализация модели
model.fit(X_train, Y_train)   # обучение модели
Y_hat = model.predict(X_test) # построение прогнозов
print(f'Точность прогнозов: \
{100*accuracy_score(Y_test, Y_hat):.1f}%')

print(f'Число опорных векторов к каждом классе: \
{model.n_support_}')
# первые 5 опорных векторов:
print(model.support_vectors_[:5])
```

- 1/C - вес при регуляризаторе.
- Больше информации. Полный код.

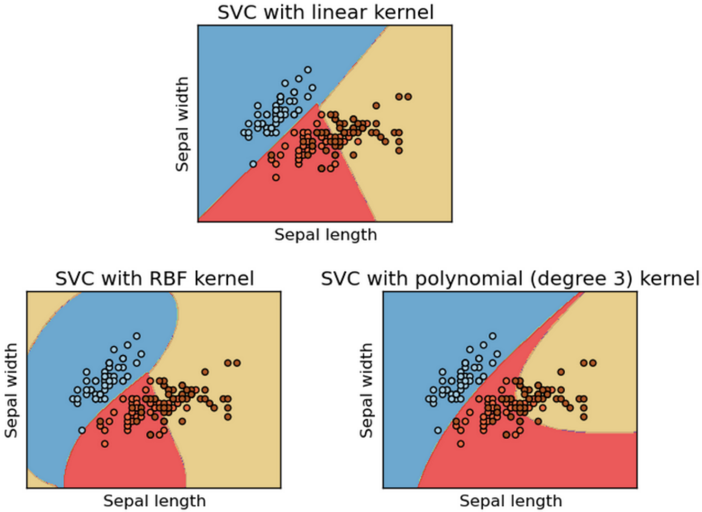
Ядерное обобщение

- Решение исходной задачи с ограничениями

$$\begin{cases} \frac{1}{2}w^T w + C\sum_{i=1}^N \xi_i \to \min_{w,w_0,\xi} \\ y_i(w^T x_i + w_0) = M_i(w, w_0) \geq 1 - \xi_i, \\ \xi_i \geq 0, \ i = 1, 2, ...N \end{cases}$$

  вычислительно сложнее (нужно использовать условия Каруша-Куна-Такеера), зато позволяет получить решение, зависящее только от $\langle x', x'' \rangle$.
- Замена $\langle x', x'' \rangle \to K(x', x'')$ позволяет обобщить метод и сделать его нелинейным!
  - это ядерное обобщение (kernel trick).

Пример ядерно-обобщённых прогнозов



Многоклассовый метод опорных векторов

C дискриминантных ф-ций строятся одновременно:

$$g_c(x) = (\mathrm{w}^c)^T x + w_0^c, \qquad c = \overline{1, C}.$$

Линейно разделимый случай:

$$\begin{cases} \sum_{c=1}^{C} (\mathrm{w}^c)^T \mathrm{w}^c \to \min_{\mathrm{w}} \\ (\mathrm{w}^{y_n})^T x_n + w_0^{y_n} - (\mathrm{w}^c)^T x - w_0^c \geq 1 \quad \forall c \neq y_n, \\ n = \overline{1, N}. \end{cases}$$

Линейно неразделимый случай:

$$\begin{cases} \sum_{c=1}^{C} (\mathrm{w}^c)^T \mathrm{w}^c + C \sum_{n=1}^{N} \xi_n \to \min_{w} \\ (\mathrm{w}^{y_n})^T x + w_0^{y_n} - (\mathrm{w}^c)^T x - w_0^c \geq 1 - \xi_n \quad \forall c \neq y_n, \\ \xi_n \geq 0, \quad n = \overline{1, N}. \end{cases}$$

Настраивается медленнее, по точности сравним с бинарным обобщением через один-против-всех и один-против-одного.

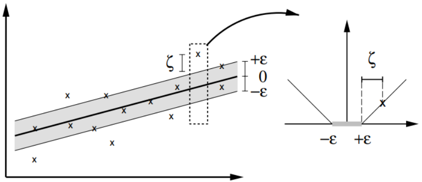
Регрессия опорных векторов

Эквивалентная формулировка (без ограничений неравенства):

$$\frac{1}{2}\|\beta\|_2^2 + C\sum_{n=1}^{N}\mathcal{L}(x_n^T\beta+\beta_0-y_n) \to \min_{\beta\in\mathbb{R}^D}$$

$$\mathcal{L}(u) = \begin{cases} 0, & \text{если } |u| \le \varepsilon \\ |u|-\varepsilon & \text{иначе} \end{cases}$$

ε – нечувствительная ф-ция потерь



Решение будет зависеть только от объектов, где $|\text{ошибка}| \ge \varepsilon$, называемых опорными векторами.

Регрессия опорных векторов

Идея: допускаем небольшие $\pm\varepsilon$ отклонения, L_2 регуляризация.

$$\begin{cases} \frac{1}{2}\left\|\beta\right\|_2^2 \to \min_{\beta\in\mathbb{R}^D} & \text{(смещение } \beta_0 \text{ пишем явно)} \\ x_n^T\beta + \beta_0 - y_n \le \varepsilon & n = \overline{1,N} \\ y_n - x_n^T\beta - \beta_0 \le \varepsilon & n = \overline{1,N} \end{cases}$$

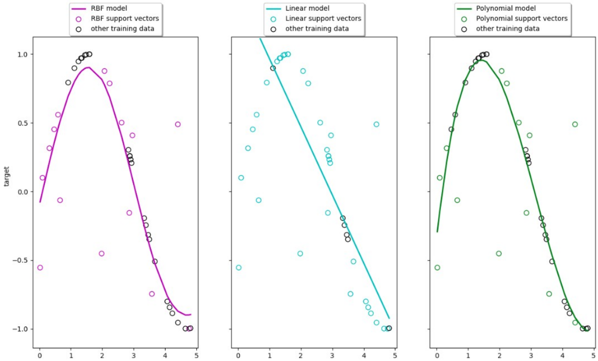
Если невозможно вписать все ошибки в интервал $[-\varepsilon, \varepsilon]$, воспользуемся методом общего вида:

$$\begin{cases} \frac{1}{2}\left\|\beta\right\|_2^2 + C\sum_{n=1}^{N}(\xi_n + \xi_n^*) \to \min_{\beta\in\mathbb{R}^D;\xi_n,\xi_n^*\in\mathbb{R}^N} & \\ x_n^T\beta + \beta_0 - y_n \le \varepsilon + \xi_n, \quad \xi_n \ge 0 & n = \overline{1,N} \\ y_n - x_n^T\beta - \beta_0 \le \varepsilon + \xi_n^*, \quad \xi_n^* \ge 0 & n = \overline{1,N} \end{cases}$$

$C \ge 0$ - гиперпараметр, контролирующий противоречие между точностью и простотой модели.

Ядерное обобщение

- Решение через задачу с ограничениями приводит к решению, зависящему только от $\langle x', x'' \rangle$.
- Ядерное обобщение (kernel trick)[1]: $\langle x', x'' \rangle \to K(x', x'')$



[1] Источник.

Заключение

- Метод опорных векторов - линейный классификатор с L_2 регуляризацией и функцией потерь hinge.
- Геометрически метод максимизирует зазор между классами.
- Решение зависит только от опорных векторов с $M \leq 1$.
- Регрессия опорных векторов - линейная регрессия с L_2 регуляризацией и ε-нечувствительной функцией потерь.
- Методы допускают ядерное обобщение
    - и становятся нелинейными