Машинное обучение. Support vector machine

Алексей Колесов

Белорусский государственный университет

25 октября 2017 г.

Мотив

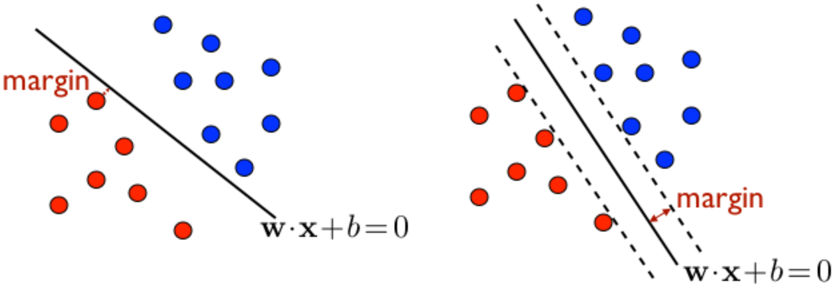
- линейные модели легко учить и интерпретировать
- пространства большой размерности «лучше» описывают объекты
- $\mathrm{VCdim}(H) = d + 1$ для моделей в $\mathbb{R}^d$
- SVM обладает рядом приятных свойств:
  - эффективно учится (даже в неразделимом сценарии)
  - выборочная сложность не зависит от d
  - обобщается с помощью kernel trick

Содержание

Задача бинарной классификации

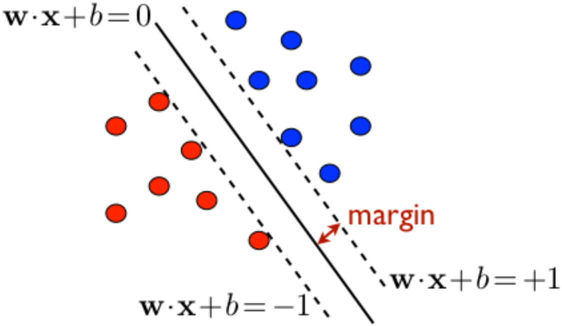
- выборка $S = ((x_1, y_1), \ldots, (x_m, y_m))$ из $X \times \{-1; +1\}$
- хотим найти $h : X \to \{-1; +1\}$ с низким $L_D(h)$
- будем искать в классе
  $H = \{x \mapsto \mathrm{sign}(\langle w, x\rangle + b) : w \in \mathbb{R}^d, b \in \mathbb{R}\}$

Линейная разделимость



Геометрическим отступом (geometric margin) называют величину $\rho = \min_{i\in[m]} \frac{|w \cdot x_i + b|}{||w||}$

Оптимальная поверхность: максимальный отступ



Идея:

- давайте максимизируем отступ
- т.е выберем $\rho = \max\limits_{w,b:y_i(w\cdot x_i+b)\geqslant 0} \min\limits_{i\in[m]} \frac{|w\cdot x_i + b|}{||w||}$

Оптимальная поверхность: максимальный отступ

$$\begin{aligned}
\rho &= \max_{w,b:y_i(w\cdot x_i+b)\geqslant 0} \min_{i\in[m]} \frac{|w\cdot x_i+b|}{||w||} && (1)\\
&= \max_{w,b:y_i(w\cdot x_i+b)\geqslant 0\wedge\min_{i\in[m]}|w\cdot x_i+b|=1} \min_{i\in[m]} \frac{|w\cdot x_i+b|}{||w||} && (2)\\
&= \max_{w,b:y_i(w\cdot x_i+b)\geqslant 0\wedge\min_{i\in[m]}|w\cdot x_i+b|=1} \frac{1}{||w||} && (3)\\
&= \max_{w,b:y_i(w\cdot x_i+b)\geqslant 1} \frac{1}{||w||} && (4)
\end{aligned}$$

Задача оптимизации

$$\min_{w,b} \frac{1}{2}||w||^2$$
$$\text{при условии } y_i(\langle w, x_i \rangle + b) \geqslant 1, \ \forall i \in [m]$$

Замечания:

- задача выпуклой оптимизации
- решение единственно для разделимого случая
- можно (неэквивалентно) перейти к однородной задаче

Необходимые условия Каруша-Куна-Такера

Необходимые условия Каруша-Куна-Такера

Пусть мы решаем задачу условной оптимизации:

$$\min_{x \in X} f(x),$$
$$x \geqslant 0,\ g_i(x) \leqslant 0\ \forall i \in [m]$$

Тогда для лангранжиана $L(x) = f(x) + \sum_{i=1}^{m} \alpha_i g_i(x)$ найдётся такой вектор $\lambda \in \mathbb{R}^m$, что для $\hat{x} \in \operatorname{argmin} f$ выполняются следующие условия:

- **стационарности**: $\min_{x} L(x) = L(\hat{x})$
- **дополняющей нежёсткости**: $\alpha_i g_i(\hat{x}) = 0,\ \forall i \in [m]$
- **неотрицательности**: $\alpha_i \geqslant 0$

Оптимальная разделяющая поверхность

Ланграножиан:

$$L(w,b,\alpha) = \frac{1}{2}||w||^2 - \sum_{i=1}^{m} \alpha_i [y_i(w \cdot x_i + b) - 1]$$

- $\nabla_w L = w - \sum_{i=1}^{m} \alpha_i y_i x_i = 0 \Leftrightarrow w = \sum_{i=1}^{m} \alpha_i y_i x_i$
- $\nabla_b L = -\sum_{i=1}^{m} \alpha_i y_i = 0 \Leftrightarrow \sum_{i=1}^{m} \alpha_i y_i = 0$
- $\alpha_i [y_i(w \cdot x_i + b) - 1] = 0$

Опорные вектора

- из условия о дополняющей нежёсткости:
  $\alpha_i[y_i(w \cdot x_i + b) - 1] = 0 \Rightarrow \alpha_i = 0 \vee [y_i(w \cdot x_i + b) - 1] = 0$
- опорными называются вектора, для которых
  $\alpha_i \neq 0 \wedge y_i(w \cdot x_i + b) = 1$
- опорные вектора можно выбрать не единственным образом

Двойственная задача

Подставим $w = \sum_{i=1}^{m} \alpha_i y_i x_i$ в L:

$$L = \frac{1}{2}\left\|\sum_{i=1}^{m} \alpha_i y_i x_i\right\|^2 - \sum_{i,j\in[m]} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) - \sum_{i=1}^{m} \alpha_i y_i b + \sum_{i=1}^{m} \alpha_i$$

$$L = -\frac{1}{2}\sum_{i,j\in[m]} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) + \sum_{i=1}^{m} \alpha_i$$

Решение

$$\min_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j \in [m]} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j)$$

$$\text{при условии: } \alpha_i \geqslant 0,\ \sum_{i=1}^{m} \alpha_i y_i = 0$$

- $h(x) = \operatorname{sign}\left(\sum_{i=1}^{m} \alpha_i y_i (x_i \cdot x) + b \right)$
- b можно найти из условия опорности вектора для любого $\alpha_i \neq 0$

Leave one out error

- Leave one out (LOO) error определяется следующим образом:

$$L_{loo}(L) = \frac{1}{m}\sum_{i=1}^{m} 1_{h_{S\setminus\{x_i\}}(x_i)=y_i}$$

- L_{loo} — несмещённая оценка ошибки обобщения для гипотезы на выборке размера $m-1$:

$$\underset{S\sim D^m}{\mathbb{E}}[L_{loo}] = \frac{1}{m}\sum_{i=1}^{m}\underset{S}{\mathbb{E}}[1_{h_{S\setminus\{x_i\}}(x_i)=y_i}] \qquad = \underset{S}{\mathbb{E}}[1_{h_{S\setminus\{x\}}(x)=y}] \tag{5}$$

$$= \underset{S'\sim D^{m-1}}{\mathbb{E}}[\underset{x\sim D}{\mathbb{E}} 1_{h_{S'}(x)=y}] \quad = \underset{S'\sim D^{m-1}}{\mathbb{E}}[L_D(h_{S'})] \tag{6}$$

LOO для Hard-SVM

Sparcity bound для Hard-SVM

Пусть h_S оптимальная гиперплоскость для S и $N_{SV}(S)$ — количество опорных векторов в определении h_S. Тогда:

$$\mathop{\mathbb{E}}_{S \sim D^m}[L_D(h_S)] \leqslant \mathop{\mathbb{E}}_{S \sim D^{m-1}}\left[\frac{N_{SV}(S)}{m+1}\right]$$

Идея доказательства: если $h_{S \setminus \{x\}}$ ошибается на x, то x — опорный вектор для h_S

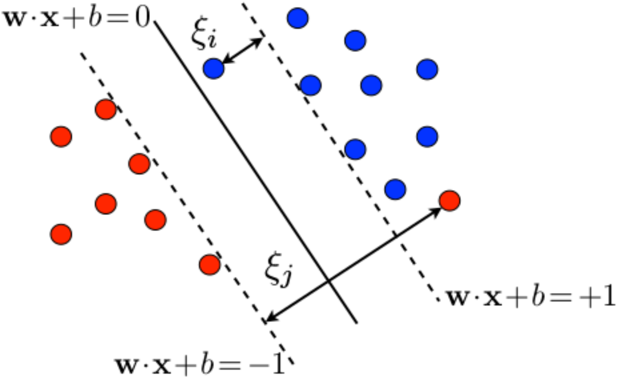
Замечания

- алгоритм опубликован Вапником и Червоненкисом в 1965-м году
- разделимые выборки встречаются редко
- для SVM есть более сильные оценки, чем sparsity bound

Soft-SVM

- данные зачастую неразделимы, т.е. для каждой пары (w, b) найдётся $x_i \in S$, что
$$y_i[w \cdot x_i + b] \not\geqslant 1$$
- идея из методов оптимизации: введём фиктивные переменные (slack variables) $\xi_i \geqslant 0$:
$$y_i[w \cdot x_i + b] \geqslant 1 - \xi_i$$

Soft-SVM



- разрешаем некоторых объектом находится внутри «коридора»
- опорные вектора — у которых отступ равен 1 или меньше

Задача оптимизации для Soft-SVM

Задача оптимизации:

$$\min_{w,b,\xi} \frac{1}{2}||w||^2 + C\sum_{i=1}^{m} \xi_i$$

$$\text{при условии: } y_i(w \cdot x_i + b) \geqslant 1 - \xi_i,\ \xi_i \geqslant 0\ \forall i \in [m]$$

Замечания:

- $C \geqslant 0$ — tradeoff-переменная
- задача выпуклой оптимизации
- имеет единственное решение
- попадает под RLM-сценарий!

Замечания

- как выбирать C?
- определить, правда ли данная гиперплоскость минимизирует функцию потерь на выборке — NP-полная задача (как функция от d)
- можно оптимизировать не сумму фиктивных переменных — сумма и сумма квадратов дают удобную форму

Сведение к RLM

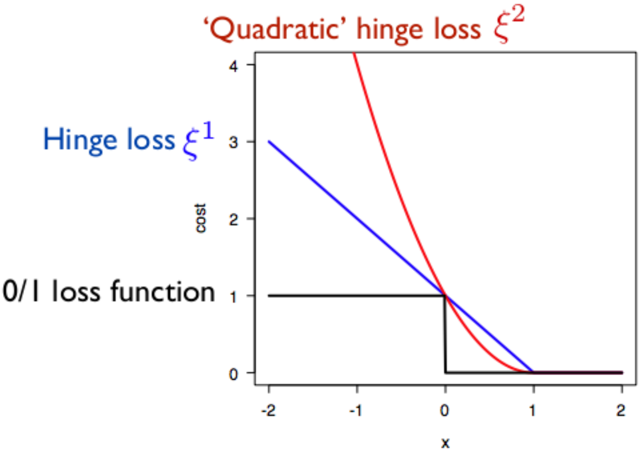
Задачу оптимизации можно переписать:

$$\min_{w,b} \frac{1}{2}||w||^2 + C\sum_{i=1}^{m}(1 - y_i(w \cdot x_i + b))_+$$

Функции потерь:

- hinge-loss: $l(x, y) = (1 - yh(x))_+$
- квадратичный hinge-loss: $l(x, y) = (1 - yh(x))_+^2$

hinge-loss


'Quadratic' hinge loss ξ^2
Hinge loss ξ^1
0/1 loss function
cost
x
4
3
2
1
0
-2
-1
0
1
2

SGD для Soft-SVM

- перепишем задачу оптимизации в нотации прошлой лекции:
$$\min_w \left(\frac{\lambda}{2}||w||^2 + \frac{1}{m}\sum_{i=1}^{m} \max\{0, 1 - y_i\langle w, x_i\rangle\} \right)$$
- для RLM с регуляризацией Тихонова можем записать: $w^{(t+1)} = -\frac{1}{\lambda t}\sum_{j=1}^{t} v_j$, где v_j — субградиент функции потерь для случайного элемента в точке $w^{(j)}$
- субградиент для hinge-loss либо ноль, либо $-y_i x_i$

SGD для Soft-SVM

Алгоритм 1 SGD для Soft-SVM

Вход: $T > 0$

$\theta^{(1)} = 0$
for $t = 1, \dots T$ **do**
    $w^{(t)} = \frac{1}{\lambda t}\theta^{(t)}$
    выбрать i равновероятно из $[m]$
    if $y_i \langle w^{(t)}, x_i \rangle < 1$ **then**
        $\theta^{(t+1)} = \theta^{(t)} + y_i x_i$
    else
        $\theta^{(t+1)} = \theta^{(t)}$
    end if
end for
return $\overline{w} = \sum_{t=1}^{T} w^{(t)}$

Оптимальная разделяющая поверхность

Лангранжиан ($\alpha_i \geqslant 0$, $\beta_i \geqslant 0$):

$$L(w, b, \alpha, \beta) =$$
$$\frac{1}{2}||w||^2 + C\sum_{i=1}^{m} \xi_i - \sum_{i=1}^{m} \alpha_i[y_i(w \cdot x_i + b) - 1 + \xi_i] - \sum_{i=1}^{m} \beta_i \xi_i$$

- $\nabla_w L = w - \sum_{i=1}^{m} \alpha_i y_i x_i = 0 \Leftrightarrow w = \sum_{i=1}^{m} \alpha_i y_i x_i$
- $\nabla_b L = -\sum_{i=1}^{m} \alpha_i y_i = 0 \Leftrightarrow \sum_{i=1}^{m} \alpha_i y_i = 0$
- $\nabla_{\xi_i} L = C - \alpha_i - \beta_i = 0 \Leftrightarrow \alpha_i + \beta_i = C$
- $\alpha_i[y_i(w \cdot x_i + b) - 1 + \xi_i] = 0$
- $\beta_i \xi_i = 0$

Опорные вектора

- из условия о дополняющей нежёсткости:
$$\alpha_i[y_i(w \cdot x_i + b) - 1 + \xi_i] = 0 \Rightarrow$$
$$\alpha_i = 0 \vee [y_i(w \cdot x_i + b) - 1 + \xi_i] = 0$$
- опорными называются вектора, для которых $\alpha_i \neq 0 \wedge y_i(w \cdot x_i + b) = 1 - \xi_i$
- опорные вектора можно выбрать не единственным образом

Двойственная задача

Рассмотрим L как функцию от ξ:

$$L(\xi) = c + C\sum_{i=1}^{m}\xi_i - \sum_{i=1}^{m}\alpha_i\xi_i = c + \sum_{i=1}^{m}\xi_i(C-\alpha_i) = \text{[при оптимальности]} = d$$

Подставим $w = \sum_{i=1}^{m}\alpha_i y_i x_i$ в L:

$$L = \frac{1}{2}\left\|\sum_{i=1}^{m}\alpha_i y_i x_i\right\|^2 - \sum_{i,j\in[m]}\alpha_i\alpha_j y_i y_j(x_i\cdot x_j) - \sum_{i=1}^{m}\alpha_i y_i b + \sum_{i=1}^{m}\alpha_i$$

$$L = -\frac{1}{2}\sum_{i,j\in[m]}\alpha_i\alpha_j y_i y_j(x_i\cdot x_j) + \sum_{i=1}^{m}\alpha_i$$

Решение

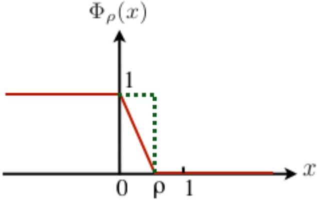
$$\min_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j \in [m]} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j)$$

$$\text{при условии: } 0 \leqslant \alpha_i \leqslant C, \ \sum_{i=1}^{m} \alpha_i y_i = 0$$

- $h(x) = \operatorname{sign}\left(\sum_{i=1}^{m} \alpha_i y_i (x_i \cdot x) + b\right)$
- b можно найти из условия опорности вектора для любого $\alpha_i \neq 0$

Отсуп как уверенность модели

- пусть ρ — заданный уровень «уверенности». Тогда функция ρ-margin задаётся вот так:



- для выборки S эмпирическая margin-функция потерь определяется как:

$$L_\rho(h) = \frac{1}{m}\sum_{i=1}^{m}\Phi(y_i h(x_i)) \leqslant \frac{1}{m}\sum_{i=1}^{m}1_{y_i h(x_i)\leqslant\rho}$$

Margin Bound для линейных классификаторов

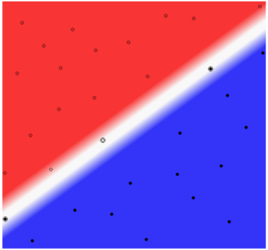
Margin Bound для линейных классификаторов

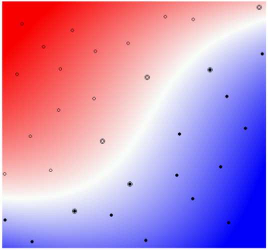
Пусть $\rho > 0$ и H — линейные классификаторы в $\mathbb{R}^d$. Пусть $||x|| \leqslant R$ и $||w|| \leqslant B$. Тогда $\forall \delta > 0$ с вероятностью не меньше $1 - \delta$ для любой гипотезы $h \in H$:

$$L_D(h) \leqslant L_\rho(h) + 2\sqrt{\frac{R^2 B^2 / \rho}{m}} + 3\sqrt{\frac{\log \frac{2}{\delta}}{2m}}$$

Содержание

Нелинейная разделимость





- линейная разделимость почти не встречается
- **идея**: переведём объекты в высокомерное пространство $X \to \Phi(X)$
- обобщающая способность не зависит от $\dim \Phi(X)$, лишь от m и ρ

Ядерные методы

- **идея**:
  - зададим $K : X \times X \to \mathbb{R}$ (**ядро** или kernel), так что:
  $$\Phi(x) \cdot \Phi(y) = K(x, y)$$
  - K можно понимать как меру близости
- **преимущества**:
  - **эффективность**: не надо переводить в пространство и там считать скалярное произведение
  - **гибкость**: класс ядерных функций очень велик

Характеризация ядер

Характеризация ядер

Функция K является ядром, если для любых $x_1, \ldots, x_m$ матрица $A = [K(x_i, x_j)]_{i,j}$ является симметричной положительно полуопределённой. Т.е. A симметрична, и выполняются любое из двух эквивалентных условий:

- все собственные значения A неотрицательны
- для любого $c \in \mathbb{R}^m$ $c^T K c \geqslant 0$

Пример ядра

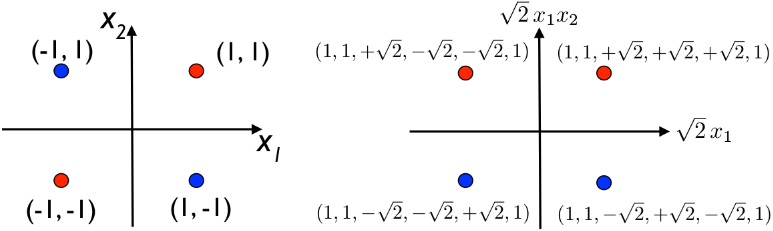
Следующая функция называется полиномиальным ядром:

$$\forall x, y \in \mathbb{R}^N \; K(x, y) = (x \cdot y + c)^d, \; c > 0$$

Например, при $N = 2$ и $d = 2$:

$$K(x, y) = (x_1 y_1 + x_2 y_2 + c)^2$$
$$= \begin{bmatrix} x_1^2 \\ x_2^2 \\ \sqrt{2} x_1 x_2 \\ \sqrt{2c} x_1 \\ \sqrt{2c} x_2 \\ c \end{bmatrix} \cdot \begin{bmatrix} y_1^2 \\ y_2^2 \\ \sqrt{2} y_1 y_2 \\ \sqrt{2c} y_1 \\ \sqrt{2c} y_2 \\ c \end{bmatrix}$$

XOR-problem


x2
(-1, 1)
(1, 1)
x1
(-1, -1)
(1, -1)
$\sqrt{2}\,x_1x_2$
$(1, 1, +\sqrt{2}, -\sqrt{2}, -\sqrt{2}, 1)$
$(1, 1, +\sqrt{2}, +\sqrt{2}, +\sqrt{2}, 1)$
$\sqrt{2}\,x_1$
$(1, 1, -\sqrt{2}, -\sqrt{2}, +\sqrt{2}, 1)$
$(1, 1, -\sqrt{2}, +\sqrt{2}, -\sqrt{2}, 1)$

Ещё примеры

- гауссово ядро
$$K(x,y) = \exp\left(-\frac{||x-y||^2}{2\sigma^2}\right),\ \sigma \neq 0$$
- сигмоидальное ядро
$$K(x,y) = \tanh(a(x\cdot y)+b),\ a,b \geqslant 0$$

SVM с kernel trick

$$\min_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j \in [m]} \alpha_i \alpha_j y_i y_j K(x_i, x_j)$$

$$\text{при условии: } 0 \leqslant \alpha_i \leqslant C, \; \sum_{i=1}^{m} \alpha_i y_i = 0$$

- $h(x) = \operatorname{sign}\left(\sum_{i=1}^{m} \alpha_i y_i K(x_i, x) + b\right)$
- b можно найти из условия опорности вектора для любого $\alpha_i \neq 0$

Следующие функции являются ядрами

- константа
- сумма ядер
- произведение ядер
- поточечный предел ядер
- композиция со степенными рядами с неотрицательными коэффициентами (например, $\exp(K)$ — ядро, для любого ядра K)

Итоги

Kernel method:

- эффективный
- гибкий
- позволяет внести индуктивный bias

Содержание

Итоги

- рассмотрели варианты SVM для разделимой и неразделимой выборки
- привели алгоритм для нахождения оптимальной гипотезы
- разобрали margin bound для линейных классификаторов
- изучили kernel trick

Литература

- Mohry — Foundations of Machine Learning (главы SVM и Kernel methods)
- Shai Shalev-Shwartz and Shai Ben-David — Understanding Machine Learning: From theory to algorithms (главы 15-16)