

Машинное обучение. Support vector machine

Алексей Колесов

Белорусский государственный университет

5 ноября 2019 г.

Мотив

- линейные модели легко учить и интерпретировать
- пространства большой размерности «лучше» описывают объекты
- $\text{VCdim}(H) = d + 1$ 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

1 SVM

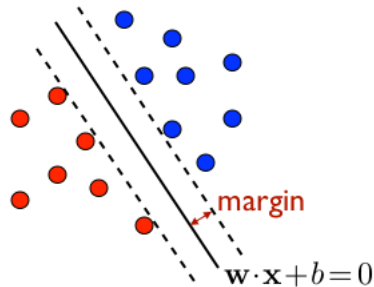
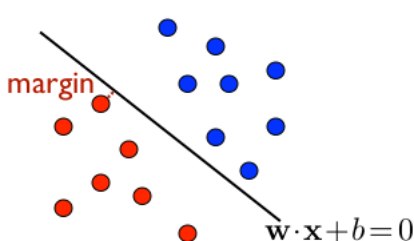
- Разделимая выборка (Hard-SVM)
- Неразделимая выборка (Soft-SVM)
- Анализ отступов

2 Kernel trick

Задача бинарной классификации

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

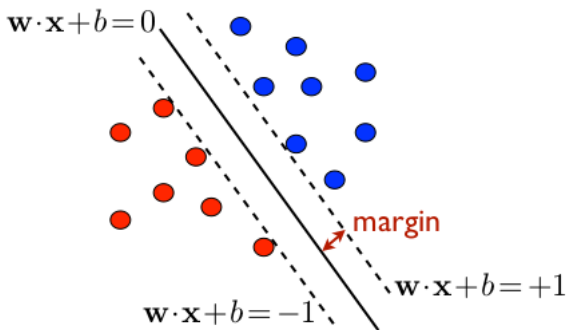
Линейная разделимость



Геометрическим отступом (geometric margin) называют

$$\text{величину } \rho = \min_{i \in [m]} \frac{|w \cdot x_i + b|}{\|w\|}$$

Оптимальная поверхность: максимальный отступ



Идея:

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

Оптимальная поверхность: максимальный отступ

$$\rho = \max_{w, b: y_i(w \cdot x_i + b) \geq 0} \min_{i \in [m]} \frac{|w \cdot x_i + b|}{\|w\|} \quad (1)$$

$$= \max_{w, b: y_i(w \cdot x_i + b) \geq 0 \wedge \min_{i \in [m]} |w \cdot x_i + b| = 1} \min_{i \in [m]} \frac{|w \cdot x_i + b|}{\|w\|} \quad (2)$$

$$= \max_{w, b: y_i(w \cdot x_i + b) \geq 0 \wedge \min_{i \in [m]} |w \cdot x_i + b| = 1} \frac{1}{\|w\|} \quad (3)$$

$$= \max_{w, b: y_i(w \cdot x_i + b) \geq 1} \frac{1}{\|w\|} \quad (4)$$

Задача оптимизации

$$\min_{w,b} \frac{1}{2} \|w\|^2$$

при условии $y_i(\langle w, x_i \rangle + b) \geq 1, \forall i \in [m]$

Замечания:

- задача выпуклой оптимизации
- решение единственно для разделимого случая

Необходимые условия Каруша-Куна-Такера

Необходимые условия Каруша-Куна-Такера

Пусть мы решаем задачу условной оптимизации:

$$\begin{aligned} \min_{x \in X} f(x), \\ x \geq 0, g_i(x) \leq 0 \quad \forall i \in [m] \end{aligned}$$

Тогда для лагранжиана $L(x) = f(x) + \sum_{i=1}^m \alpha_i g_i(x)$ найдётся такой вектор $A = (\alpha_1, \dots, \alpha_m) \in \mathbb{R}^m$, 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: $\alpha_i \geq 0$

Оптимальная разделяющая поверхность

Лангранжиан:

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

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

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

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

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

- $h(x) = \text{sign} \left(\sum_{i=1}^m \alpha_i y_i (x_i \cdot x) + b \right)$
- 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:

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

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

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

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

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:

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

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

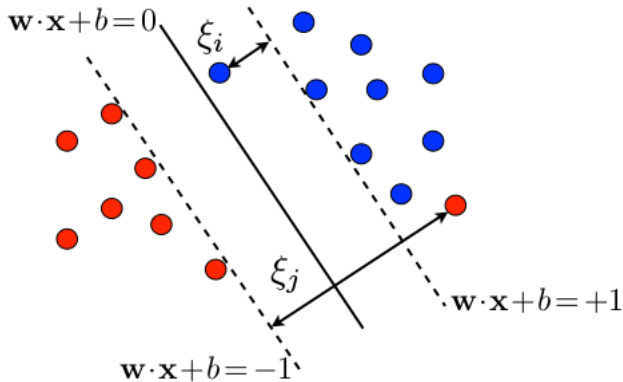
- данные зачастую неразделимы, т.е. для каждой пары (w, b) найдётся $x_i \in S$, что

$$y_i[w \cdot x_i + b] \not\geq 1$$

- идея из методов оптимизации: введём фиктивные переменные (slack variables) $\xi_i \geq 0$:

$$y_i[w \cdot x_i + b] \geq 1 - \xi_i$$

Soft-SVM



- разрешаем некоторым объектам находиться внутри «коридора»
- опорные вектора — у которых отступ равен 1 или меньше

Задача оптимизации для Soft-SVM

Задача оптимизации:

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

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

Замечания:

- $C \geq 0$ — tradeoff-переменная
- задача выпуклой оптимизации
- имеет единственное решение

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

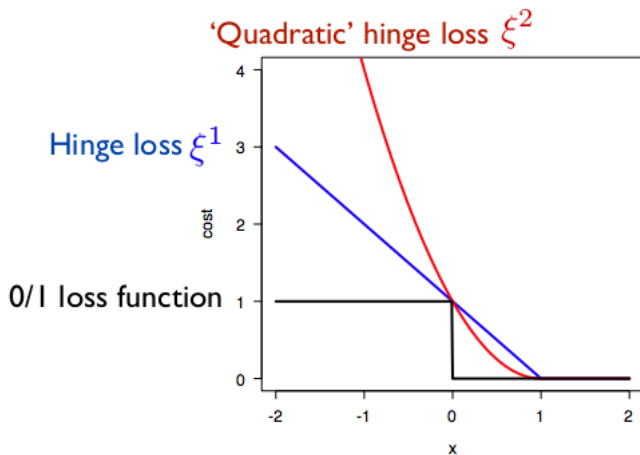
Задачу оптимизации можно переписать:

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

Функции потерь:

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

hinge-loss



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: $T > 0$

```
1:  $\theta^{(1)} = 0$ 
2: for  $t = 1, \dots, T$  do
3:    $w^{(t)} = \frac{1}{\lambda t} \theta^{(t)}$ 
4:   выбрать  $i$  равномерно из  $[m]$ 
5:   if  $y_i \langle w^{(t)}, x_i \rangle < 1$  then
6:      $\theta^{(t+1)} = \theta^{(t)} + y_i x_i$ 
7:   else
8:      $\theta^{(t+1)} = \theta^{(t)}$ 
9:   end if
10: end for
11: return  $\bar{w} = \sum_{t=1}^T w^{(t)}$ 
```

Оптимальная разделяющая поверхность

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

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

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

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

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

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

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

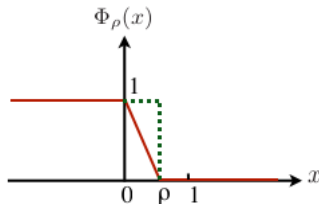
$$\max_{\alpha} \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j \in [m]} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j)$$

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

- $h(x) = \text{sign} \left(\sum_{i=1}^m \alpha_i y_i (x_i \cdot x) + b \right)$
- 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:

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

Margin Bound для линейных классификаторов

Margin Bound для линейных классификаторов

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

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

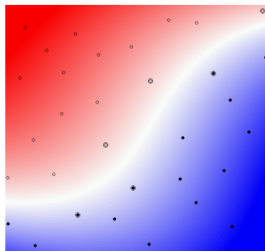
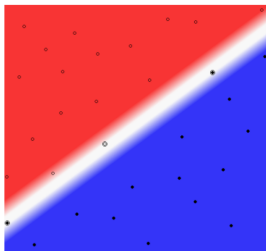
Содержание

1 SVM

- Разделимая выборка (Hard-SVM)
- Неразделимая выборка (Soft-SVM)
- Анализ отступов

2 Kernel trick

Нелинейная разделимость



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

Ядерные методы

- **идея:**

- зададим $K : X \times X \rightarrow \mathbb{R}$ (**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, \dots, 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 $c \in \mathbb{R}^m$ $c^T K c \geq 0$

Пример ядра

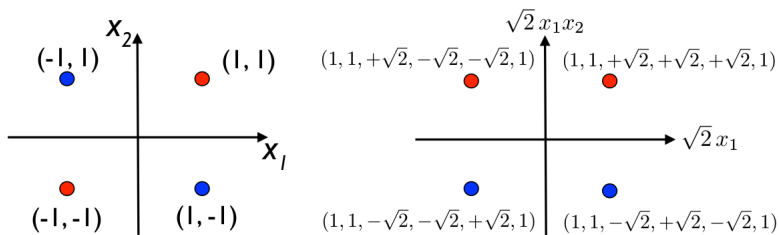
Следующая функция называется полиномиальным ядром:

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

Например, при $N = 2$ и $d = 2$:

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

XOR-problem



Ещё примеры

- гауссово ядро

$$K(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right), \sigma \neq 0$$

- сигмоидальное ядро

$$K(x, y) = \tanh(a(x \cdot y) + b), a, b \geq 0$$

SVM с kernel trick

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

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

- $h(x) = \text{sign} \left(\sum_{i=1}^m \alpha_i y_i K(x_i, x) + b \right)$
- 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

1 SVM

- Разделимая выборка (Hard-SVM)
- Неразделимая выборка (Soft-SVM)
- Анализ отступов

2 Kernel trick

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)