Введение в искусственный интеллект. Машинное обучение

Лекция 6. Машины опорных векторов – SVM

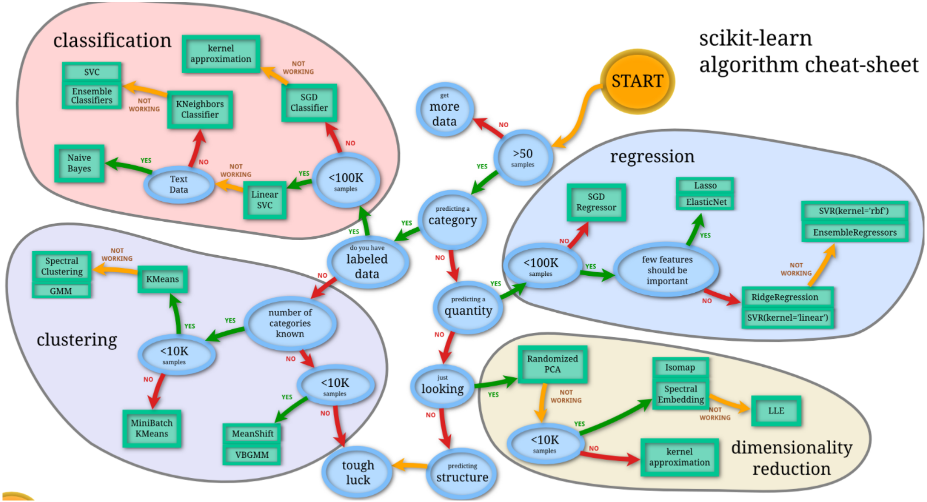
МаТИС

29 марта 2019г.

План лекции

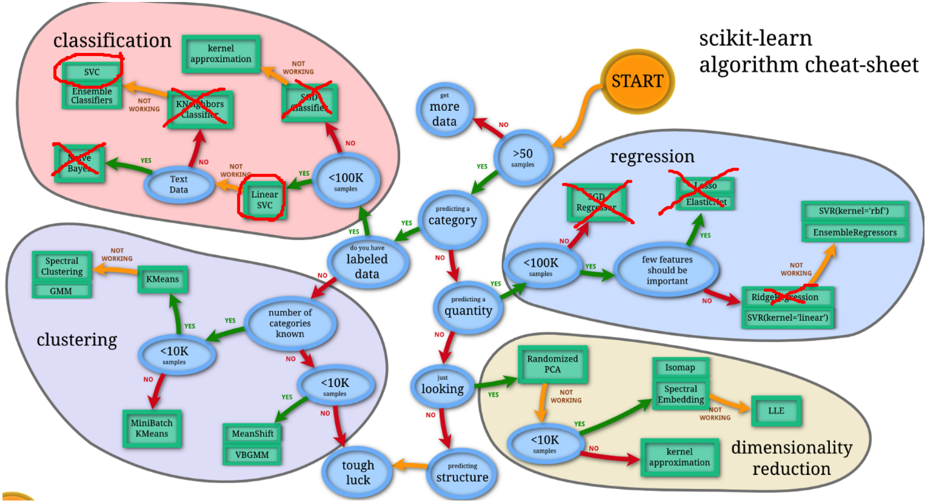
1. Случай линейной разделимости
2. Случай линейной неразделимости
3. Решение с помощью двойственной задачи
4. Обобщение SVM с помощью ядрового трюка
5. Пример расчетов для SVM

Дорожная карта Scikit-Learn[1]



[1] https://scikit-learn.org/stable/tutorial/machine_learning_map/

Дорожная карта Scikit-Learn[1]



[1] https://scikit-learn.org/stable/tutorial/machine_learning_map/

Вспомним прошлую лекцию

Рассмотрим задачу бинарной классификации: $X \to Y$, $X = \mathbb{R}^n$, $Y = \{+1, -1\}$ на обучающей выборке $X^m = (x_i, y_i)_{i=1}^m$.
Линейный классификатор $a(x; w, w_0) = \text{sign}(\langle w, x\rangle - w_0)$.
Минимизация эмпирического риска в данном случае:

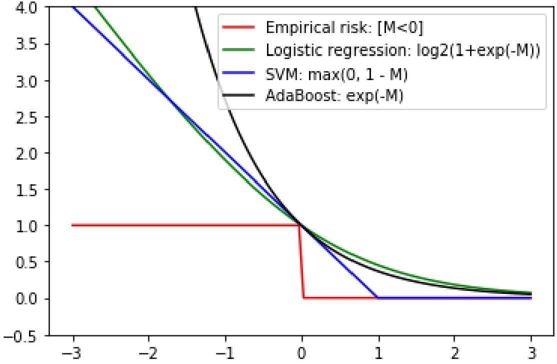
$$R(w, w_0, X^m) = \sum_{i=1}^{m}[a(x_i; w, w_0) \neq y_i] = \sum_{i=1}^{m}[M_i(w, w_0) < 0],$$

где отступ объекта x_i: $M_i = (\langle w, x_i\rangle - w_0)y_i$.
Добавим аппроксимацию и L_2 регуляризацию:

$$R(w, w_0, X^m) \leq \sum_{i=1}^{m}\max(0, 1 - M_i(w, w_0)) + \frac{1}{2C}||w||^2$$

Аппроксимация и регуляризация

- Аппроксимация штрафует за приближение к границе классов: $M_i = 1$
- Регуляризация штрафует неустойчивые решения



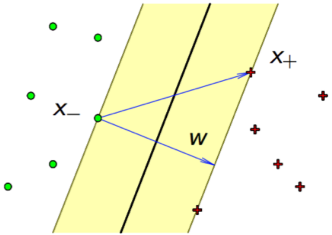
Случай линейной разделимости

Линейная разделимость

$\exists w, w_0$ т.ч. $M_i(w, w_0) > 0$ для всех $i = 1, \ldots m$.

Очевидно, что можно перенормировать вектор w, т.ч. $\min_i M_i(w, w_0) = 1$.

Разделяющая полоса: $-1 \leq \langle w, x_i \rangle - w_0 \leq +1$.
Разделяющая гиперплоскость (посередине):
$\langle w, x_i \rangle - w_0 = 0$.
Можем добиться того, что существует по крайней мере одна точка на каждой из границ
(**Упражнение**: доказать): $\exists x_\pm : \langle w, x_\pm \rangle - w_0 = \pm 1$.
Ширина полосы: $\frac{\langle x_+ - x_-, w \rangle}{||w||} = \frac{2}{||w||} \to \max_w$



Случай линейной разделимости – вывод

Т.о., в случае линейной разделимости можно оптимизационную задачу записать как:

$$\begin{cases} \frac{1}{2}||w||^2 \to \min_w, \\ y_i(\langle w, x_i \rangle - w_0) \geq 1, \quad i = 1, \ldots, m \end{cases}$$

Случай линейной неразделимости

Обобщим[2] задачу на этот случай: алгоритм может допускать ошибки на обучающих объектах.

[2] Вапник, Червоненкис, 1963

Случай линейной неразделимости

Обобщим[2] задачу на этот случай: алгоритм может допускать ошибки на обучающих объектах.
Ограничение: таких ошибок должно быть поменьше.

[2]Вапник, Червоненкис, 1963

Случай линейной неразделимости

Обобщим[2] задачу на этот случай: алгоритм может допускать ошибки на обучающих объектах.
Ограничение: таких ошибок должно быть поменьше.
Решение: введение дополнительных переменных $\xi_i \geq 0$, характеризующих величину ошибки (уменьшение отступа) на объектах x_i.

[2] Вапник, Червоненкис, 1963

Случай линейной неразделимости

Обобщим[2] задачу на этот случай: алгоритм может допускать ошибки на обучающих объектах.
Ограничение: таких ошибок должно быть поменьше.
Решение: введение дополнительных переменных $\xi_i \geq 0$, характеризующих величину ошибки (уменьшение отступа) на объектах x_i.

$$\begin{cases} \frac{1}{2}||w||^2 + C\sum_{i=1}^{m} \xi_i \to \min_{w, w_0, \xi}, & \\ y_i(\langle w, x_i \rangle - w_0) \geq 1 - \xi_i, & i = 1, \dots, m, \\ \xi_i \geq 0 & i = 1, \dots, m. \end{cases}$$

[2] Вапник, Червоненкис, 1963

Случай линейной неразделимости

Обобщим[2] задачу на этот случай: алгоритм может допускать ошибки на обучающих объектах.

Ограничение: таких ошибок должно быть поменьше.

Решение: введение дополнительных переменных $\xi_i \geq 0$, характеризующих величину ошибки (уменьшение отступа) на объектах x_i.

$$\begin{cases} \frac{1}{2}||w||^2 + C\sum_{i=1}^m \xi_i \to \min_{w,w_0,\xi}, & \\ y_i(\langle w, x_i\rangle - w_0) \geq 1 - \xi_i, & i = 1, \ldots, m, \\ \xi_i \geq 0 & i = 1, \ldots, m. \end{cases}$$

Замечание. Положительная константа C определяет компромисс между максимизацией ширины разделяющей полосы и минимизацией суммарной ошибки.

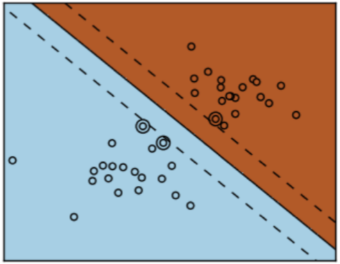
[2] Вапник, Червоненкис, 1963

О константе C

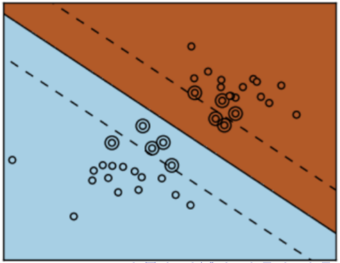
Поставленная выше оптимизационная задача эквивалентна минимизации аппроксимированного э.р. с регуляризатором:

$$\sum_{i=1}^{m} \max(0, 1 - y_i(\langle w, x_i \rangle - w_0)) + \frac{1}{2C}||w||^2 \to min_{w,w_0}$$

Большое значение C: узкая полоса, мало ошибок



Маленькое значение C: широкая полоса, много ошибок



Условия Каруша-Куна-Таккера (ККТ)

Условия ККТ – это **необходимые** условия решения задачи нелинейного программирования (обобщение метода множителей Лагранжа).
Задача нелинейного программирования:

$$\begin{cases} f(x) \to \min_x, & \\ g_i(x) \le 0, & i = 1, \ldots, k, \\ h_j(x) = 0, & j = 1, \ldots, \ell. \end{cases}$$

Условия Каруша-Куна-Таккера (ККТ)

Условия ККТ – это **необходимые** условия решения задачи нелинейного программирования (обобщение метода множителей Лагранжа).
Задача нелинейного программирования:

$$\begin{cases} f(x) \to \min_x, & \\ g_i(x) \leq 0, & i = 1, \dots, k, \\ h_j(x) = 0, & j = 1, \dots, \ell. \end{cases}$$

Необходимые условия: если x - точка локального минимума, то существуют множители μ_i, λ_j, т.ч.

$$\begin{cases} \frac{\partial L}{\partial x} = 0; L(x, \mu, \lambda) = f(x) + \sum_{i=1}^k \mu_i g_i(x) + \sum_{j=1}^\ell \lambda_j h_j(x) & \text{(функция Лагранжа)} \\ g_i(x) \leq 0, h_j(x) = 0 & \text{(исходные ограничения)} \\ \mu_i \geq 0 & \text{(двойственные ограничения)} \\ \mu_i g_i(x) = 0 & \text{(дополняющая нежесткость)} \end{cases}$$

Условия ККТ и SVM

Функция Лагранжа для SVM

$L(w, w_0, \xi; \lambda, \eta) = \frac{1}{2}||w||^2 - \sum_{i=1}^m \lambda_i(M_i(w, w_0) - 1) - \sum_{i=1}^m \xi_i(\lambda_i + \eta_i - C)$, где:
λ_i – переменные, двойственные к ограничениям $M_i(w, w_0) \geq 1 - \xi_i$,
η_i – переменные, двойственные к ограничениям $\xi_i \geq 0$.

Условия ККТ и SVM

Функция Лагранжа для SVM

$L(w, w_0, \xi; \lambda, \eta) = \frac{1}{2}||w||^2 - \sum_{i=1}^m \lambda_i (M_i(w, w_0) - 1) - \sum_{i=1}^m \xi_i (\lambda_i + \eta_i - C)$, где:
λ_i – переменные, двойственные к ограничениям $M_i(w, w_0) \geq 1 - \xi_i$,
η_i – переменные, двойственные к ограничениям $\xi_i \geq 0$.

Необходимые условия примут вид:

$$\begin{cases} \frac{\partial L}{\partial w} = 0; \frac{\partial L}{\partial w_0} = 0; \frac{\partial L}{\partial \xi} = 0 & \\ \xi_i \geq 0, \lambda_i \geq 0, \eta_i \geq 0 & i = 1, \dots, m \\ \lambda_i = 0 \text{ либо } M_i(w, w_0) = 1 - \xi_i & i = 1, \dots, m \\ \eta_i = 0 \text{ либо } \xi_i = 0 & i = 1, \dots, m \end{cases}$$

Дифференцируем функцию Лагранжа

$$L(w, w_0, \xi; \lambda, \eta) = \frac{1}{2}||w||^2 - \sum_{i=1}^m \lambda_i(M_i(w, w_0) - 1) - \sum_{i=1}^m \xi_i(\lambda_i + \eta_i - C).$$

$$\begin{cases} \frac{\partial L}{\partial w} = w - \sum_{i=1}^m \lambda_i y_i x_i = 0 & \Rightarrow w = \sum_{i=1}^m \lambda_i y_i x_i \\ \frac{\partial L}{\partial w_0} = -\sum_{i=1}^m \lambda_i y_i = 0 & \Rightarrow \sum_{i=1}^m \lambda_i y_i = 0 \\ \frac{\partial L}{\partial \xi_i} = -\lambda_i - \eta_i + C = 0 & \Rightarrow \lambda_i + \eta_i = C \end{cases}$$

Дифференцируем функцию Лагранжа

$$L(w, w_0, \xi; \lambda, \eta) = \frac{1}{2}||w||^2 - \sum_{i=1}^m \lambda_i (M_i(w, w_0) - 1) - \sum_{i=1}^m \xi_i (\lambda_i + \eta_i - C).$$

$$\begin{cases} \frac{\partial L}{\partial w} = w - \sum_{i=1}^m \lambda_i y_i x_i = 0 & \Rightarrow w = \sum_{i=1}^m \lambda_i y_i x_i \\ \frac{\partial L}{\partial w_0} = - \sum_{i=1}^m \lambda_i y_i = 0 & \Rightarrow \sum_{i=1}^m \lambda_i y_i = 0 \\ \frac{\partial L}{\partial \xi_i} = -\lambda_i - \eta_i + C = 0 & \Rightarrow \lambda_i + \eta_i = C \end{cases}$$

1. $\lambda_i = 0, \eta_i = C, \xi_i = 0, M_i \geq 1$: периферийные объекты
2. $0 < \lambda_i < C, 0 < \eta_i < C, \xi_i = 0, M_i = 1$: опорные объекты на границе
3. $\lambda_i = C, \eta_i = 0, \xi_i > 0, M_i < 1$: опорные объекты-ошибки

О двойственных задачах

Прямая задача:

$$\begin{cases} f(x) \to \min_x, & \\ g_i(x) \le 0, & i = 1, \dots, k, \\ h_j(x) = 0, & j = 1, \dots, \ell. \end{cases}$$

Функция Лагранжа: $L(x, \mu, \lambda) = f(x) + \sum_{i=1}^{k} \mu_i g_i(x) + \sum_{j=1}^{\ell} \lambda_j h_j(x) \to \min_x$.

О двойственных задачах

Прямая задача:

$$\begin{cases} f(x) \to \min_x, & \\ g_i(x) \leq 0, & i = 1, \ldots, k, \\ h_j(x) = 0, & j = 1, \ldots, \ell. \end{cases}$$

Функция Лагранжа: $L(x, \mu, \lambda) = f(x) + \sum_{i=1}^{k} \mu_i g_i(x) + \sum_{j=1}^{\ell} \lambda_j h_j(x) \to \min_x$.
Двойственная функция: $Q(\mu, \lambda) = \min_x L(x, \mu, \lambda)$.

О двойственных задачах

Прямая задача:

$$\begin{cases} f(x) \to \min_x, & \\ g_i(x) \le 0, & i = 1, \dots, k, \\ h_j(x) = 0, & j = 1, \dots, \ell. \end{cases}$$

Функция Лагранжа: $L(x, \mu, \lambda) = f(x) + \sum_{i=1}^{k} \mu_i g_i(x) + \sum_{j=1}^{\ell} \lambda_j h_j(x) \to \min_x$.
Двойственная функция: $Q(\mu, \lambda) = \min_x L(x, \mu, \lambda)$.
Двойственная задача:

$$\begin{cases} Q(\mu, \lambda) \to \max_{\mu, \lambda}, & \\ \mu_i \le 0, & i = 1, \dots, k \end{cases}$$

Теорема

Если $f(x), g_i(x), h_j(x)$ – выпуклые функции, x^* – решение прямой задачи, а (μ^*, λ^*) – решение двойственной задачи, то $Q(\mu^*, \lambda^*) = f(x^*)$.

Двойственная задача для SVM

Подставим решения прямой задачи $w = \sum_{i=1}^{m} \lambda_i y_i x_i, \sum_{i=1}^{m} \lambda_i y_i = 0, \lambda_i + \eta_i = C$ в функцию Лагранжа
$L(w, w_0, \xi; \lambda, \eta) = \frac{1}{2}||w||^2 - \sum_{i=1}^{m} \lambda_i (y_i(\langle w, x_i \rangle - w_0) - 1) - \sum_{i=1}^{m} \xi_i(\lambda_i + \eta_i - C)$. Получим:

Двойственная задача для SVM

Подставим решения прямой задачи $w = \sum_{i=1}^{m} \lambda_i y_i x_i, \sum_{i=1}^{m} \lambda_i y_i = 0, \lambda_i + \eta_i = C$ в функцию Лагранжа
$L(w, w_0, \xi; \lambda, \eta) = \frac{1}{2}||w||^2 - \sum_{i=1}^{m} \lambda_i (y_i(\langle w, x_i \rangle - w_0) - 1) - \sum_{i=1}^{m} \xi_i(\lambda_i + \eta_i - C)$. Получим:

$$Q(\lambda) = \frac{1}{2}\sum_{i=1}^{m} \lambda_i y_i x_i \sum_{j=1}^{m} \lambda_j y_j x_j - \sum_{i=1}^{m} \lambda_i y_i \langle \sum_{j=1}^{m} \lambda_j y_j x_j, x_i \rangle + w_0 \sum_{i=1}^{m} \lambda_i y_i + \sum_{i=1}^{m} \lambda_i - \sum_{i=1}^{m} \xi_i (C - C) =$$

$$= -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \lambda_i \lambda_j y_i y_j \langle x_i, x_j \rangle + \sum_{i=1}^{m} \lambda_i$$

.

Двойственная задача для SVM

Объединяя, получаем формулировку двойственной задачи:

$$\begin{cases} -Q(\lambda) = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\lambda_i\lambda_j y_i y_j\langle x_i, x_j\rangle - \sum_{i=1}^{m}\lambda_i \to \min_{\lambda}, \\ 0 \le \lambda_i \le C, \\ \sum_{i=1}^{m}\lambda_i y_i = 0. \end{cases}$$

[3] https://en.wikipedia.org/wiki/Quadratic_programming

Двойственная задача для SVM

Объединяя, получаем формулировку двойственной задачи:

$$\begin{cases} -Q(\lambda) = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\lambda_i\lambda_j y_i y_j \langle x_i, x_j\rangle - \sum_{i=1}^{m}\lambda_i \to \min_{\lambda}, \\ 0 \le \lambda_i \le C, \\ \sum_{i=1}^{m}\lambda_i y_i = 0. \end{cases}$$

Минимизируется квадратичный функционал $Q(\lambda)$, имеющий неотрицательно определённую квадратичную форму $\Rightarrow$ этот функционал – выпуклый.

[3] https://en.wikipedia.org/wiki/Quadratic_programming

Двойственная задача для SVM

Объединяя, получаем формулировку двойственной задачи:

$$\begin{cases} -Q(\lambda) = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\lambda_i\lambda_j y_i y_j \langle x_i, x_j\rangle - \sum_{i=1}^{m}\lambda_i \to \min_{\lambda}, \\ 0 \le \lambda_i \le C, \\ \sum_{i=1}^{m}\lambda_i y_i = 0. \end{cases}$$

Минимизируется квадратичный функционал $Q(\lambda)$, имеющий неотрицательно определённую квадратичную форму $\Rightarrow$ этот функционал – выпуклый.
Область, определяемая линейными ограничениями, также выпуклая.

[3] https://en.wikipedia.org/wiki/Quadratic_programming

Объединяя, получаем формулировку двойственной задачи:

$$\begin{cases} -Q(\lambda) = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\lambda_i\lambda_j y_i y_j \langle x_i, x_j\rangle - \sum_{i=1}^{m}\lambda_i \to \min_{\lambda}, \\ 0 \le \lambda_i \le C, \\ \sum_{i=1}^{m}\lambda_i y_i = 0. \end{cases}$$

Минимизируется квадратичный функционал $Q(\lambda)$, имеющий неотрицательно определённую квадратичную форму $\Rightarrow$ этот функционал – выпуклый. Область, определяемая линейными ограничениями, также выпуклая. Следовательно, данная двойственная задача имеет **единственное** решение.

[3] https://en.wikipedia.org/wiki/Quadratic_programming

Двойственная задача для SVM

Объединяя, получаем формулировку двойственной задачи:

$$\begin{cases} -Q(\lambda) = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\lambda_i\lambda_j y_i y_j \langle x_i, x_j\rangle - \sum_{i=1}^{m}\lambda_i \to \min_\lambda, \\ 0 \leq \lambda_i \leq C, \\ \sum_{i=1}^{m}\lambda_i y_i = 0. \end{cases}$$

Минимизируется квадратичный функционал $Q(\lambda)$, имеющий неотрицательно определённую квадратичную форму $\Rightarrow$ этот функционал – выпуклый.
Область, определяемая линейными ограничениями, также выпуклая.
Следовательно, данная двойственная задача имеет **единственное** решение.
Способ решения – методами **квадратичного** программирования (например, можно использовать метод внутренней точки[3]).

[3] https://en.wikipedia.org/wiki/Quadratic_programming

Решение прямой задачи для SVM

Пусть единственное решение двойственной задачи $(\lambda_i)_{i=1}^m$. Тогда решение прямой задачи выражается через решение двойственной как:

$$\begin{cases} w = \sum_{\lambda_i \neq 0} \lambda_i y_i x_i, & \text{суммируем только по опорным векторам } \lambda_i \neq 0 \\ w_0 = \langle w, x_j \rangle - y_j & \text{для опорного вектора на границе } 0 < \lambda_j < C \end{cases}$$

Решение прямой задачи для SVM

Пусть единственное решение двойственной задачи $(\lambda_i)_{i=1}^m$. Тогда решение прямой задачи выражается через решение двойственной как:

$$\begin{cases} w = \sum_{\lambda_i \neq 0} \lambda_i y_i x_i, & \text{суммируем только по опорным векторам } \lambda_i \neq 0 \\ w_0 = \langle w, x_j \rangle - y_j & \text{для опорного вектора на границе } 0 < \lambda_j < C \end{cases}$$

При этом сам линейный классификатор примет вид

$$a(x) = \operatorname{sign}\left(\sum_{i=1}^{m} \lambda_i y_i \langle x_i, x \rangle - w_0\right)$$

что можно понимать как линейность в пространстве $\mathbb{R}^m$ с признаками $f_i = \langle x_i, x \rangle$.

Обобщение SVM

Если в исходном пространстве сложно разделить выборку, то попробуем перейти в пространство большей размерности[4] $\varphi : X \to H$.

[4] Босер, Гийон, Вапник, 1992

Обобщение SVM

Если в исходном пространстве сложно разделить выборку, то попробуем перейти в пространство большей размерности[4] $\varphi : X \to H$.
Ядро – функция $K : X \times X \to \mathbb{R}$, т.ч. $K(x_1, x_2) = \langle \varphi(x_1), \varphi(x_2) \rangle$ при некотором $\varphi : X \to H$, где H – гильбертово пространство.

[4] Босер, Гийон, Вапник, 1992

Обобщение SVM

Если в исходном пространстве сложно разделить выборку, то попробуем перейти в пространство большей размерности[4] $\varphi : X \to H$.
Ядро – функция $K : X \times X \to \mathbb{R}$, т.ч. $K(x_1, x_2) = \langle \varphi(x_1), \varphi(x_2) \rangle$ при некотором $\varphi : X \to H$, где H – гильбертово пространство.

Теорема Мерсера

Функция $K(x_1, x_2)$ является ядром $\Leftrightarrow$ 1) Она симметрична $K(x_1, x_2) = K(x_2, x_1)$ и 2) Неотрицательно определена $\int_X \int_X K(x_1, x_2) f(x_1) f(x_2) dx_1 dx_2 \geq 0$ для любой функции $f : X \to \mathbb{R}$.

Замечание. Если K не удовлетворяет указанным выше условиям, то минимизируемый функционал для классификатора уже не будет выпуклым, и решение может оказаться не единственным!

[4] Босер, Гийон, Вапник, 1992

SVM с другими ядрами

Изначально наша двойственная задача была сформулирована с терминах линейного ядра:

$$\begin{cases} -Q(\lambda) = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\lambda_i\lambda_j y_i y_j\langle x_i, x_j\rangle - \sum_{i=1}^{m}\lambda_i \to \min_\lambda, \\ 0 \le \lambda_i \le C, \\ \sum_{i=1}^{m}\lambda_i y_i = 0. \end{cases}$$

Когда мы меняем ядро с $\langle x_i, x_j\rangle$ на $K(x_i, x_j)$, задача преобразуется в:

$$\begin{cases} -Q(\lambda) = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\lambda_i\lambda_j y_i y_j K(x_i, x_j) - \sum_{i=1}^{m}\lambda_i \to \min_\lambda, \\ 0 \le \lambda_i \le C, \\ \sum_{i=1}^{m}\lambda_i y_i = 0. \end{cases}$$

При этом сам линейный классификатор принимает вид (x_j - опорный вектор на границе):

$$a(x) = \mathrm{sign}(\sum_{i=1}^{m}\lambda_i y_i K(x_i, x) - w_0), w_0 = \sum_{i=1}^{m}\lambda_i y_i K(x_i, x_j) - y_j$$

Пример нахождения пространства H

Пусть $X = \mathbb{R}^2, K(u, v) = \langle u, v \rangle^2$ при $u = (u_1, u_2), v = (v_1, v_2)$.
Хотим найти H и $\varphi : X \to H$, т.ч. $K(x_1, x_2) = \langle \varphi(x_1), \varphi(x_2) \rangle$.
Сделаем эквивалентные преобразования:
$K(u, v) = \langle u, v \rangle^2 = \langle (u_1, u_2), (v_1, v_2) \rangle^2 = (u_1 v_1 + u_2 v_2)^2 = u_1^2 v_1^2 + u_2^2 v_2^2 + 2 u_1 v_1 u_2 v_2$

Пример нахождения пространства H

Пусть $X = \mathbb{R}^2, K(u, v) = \langle u, v\rangle^2$ при $u = (u_1, u_2), v = (v_1, v_2)$.
Хотим найти H и $\varphi : X \to H$, т.ч. $K(x_1, x_2) = \langle \varphi(x_1), \varphi(x_2)\rangle$.
Сделаем эквивалентные преобразования:

$$K(u, v) = \langle u, v\rangle^2 = \langle (u_1, u_2), (v_1, v_2)\rangle^2 = (u_1 v_1 + u_2 v_2)^2 = u_1^2 v_1^2 + u_2^2 v_2^2 + 2u_1 v_1 u_2 v_2$$

$$K(u, v) = \langle (u_1^2, u_2^2, \sqrt{2}u_1 u_2), (v_1^2, v_2^2, \sqrt{2}v_1 v_2)\rangle$$

Пример нахождения пространства H

Пусть $X = \mathbb{R}^2, K(u,v) = \langle u, v\rangle^2$ при $u = (u_1, u_2), v = (v_1, v_2)$.
Хотим найти H и $\varphi : X \to H$, т.ч. $K(x_1, x_2) = \langle \varphi(x_1), \varphi(x_2)\rangle$.
Сделаем эквивалентные преобразования:
$K(u,v) = \langle u, v\rangle^2 = \langle (u_1, u_2), (v_1, v_2)\rangle^2 = (u_1 v_1 + u_2 v_2)^2 = u_1^2 v_1^2 + u_2^2 v_2^2 + 2u_1 v_1 u_2 v_2$
$K(u,v) = \langle (u_1^2, u_2^2, \sqrt{2} u_1 u_2), (v_1^2, v_2^2, \sqrt{2} v_1 v_2)\rangle$
Т.о., $H = \mathbb{R}^3$ и $\varphi : (u_1, u_2) \mapsto (u_1^2, u_2^2, \sqrt{2} u_1 u_2)$.

Пример нахождения пространства H

Пусть $X = \mathbb{R}^2, K(u, v) = \langle u, v\rangle^2$ при $u = (u_1, u_2), v = (v_1, v_2)$.
Хотим найти H и $\varphi : X \to H$, т.ч. $K(x_1, x_2) = \langle\varphi(x_1), \varphi(x_2)\rangle$.
Сделаем эквивалентные преобразования:
$K(u, v) = \langle u, v\rangle^2 = \langle(u_1, u_2), (v_1, v_2)\rangle^2 = (u_1v_1 + u_2v_2)^2 = u_1^2v_1^2 + u_2^2v_2^2 + 2u_1v_1u_2v_2$
$K(u, v) = \langle(u_1^2, u_2^2, \sqrt{2}u_1u_2), (v_1^2, v_2^2, \sqrt{2}v_1v_2)\rangle$
Т.о., $H = \mathbb{R}^3$ и $\varphi : (u_1, u_2) \mapsto (u_1^2, u_2^2, \sqrt{2}u_1u_2)$.
Линейной поверхности в H будет соответствовать квадратичная поверхность в X.

Операции над ядрами

- Скалярное произведение $K(x_1, x_2) = \langle x_1, x_2 \rangle$ – ядро
- Константа $K(x_1, x_2) = c$ – ядро
- Произведение ядер $K(x_1, x_2) = K_1(x_1, x_2)K_2(x_1, x_2)$ – ядро
- Для любой $\varphi : X \to \mathbb{R}$ сепарабельная $K(x_1, x_2) = \varphi(x_1)\varphi(x_2)$ – ядро
- Линейная $K(x_1, x_2) = \alpha_1 K_1(x_1, x_2) + \alpha_2 K_2(x_1, x_2)$ – ядро при $\alpha_1, \alpha_2 > 0$, K_1, K_2 - ядрах
- Для любой $\varphi : X \to X$ подстановка $K(x_1, x_2) = K_1(\varphi(x_1), \varphi(x_2))$ – ядро при K_1 - ядро

Примеры ядер

- Полиномиальное ядро с мономами степени d: $K(x_1, x_2) = \langle x_1, x_2 \rangle^d$

Примеры ядер

- Полиномиальное ядро с мономами степени d: $K(x_1, x_2) = \langle x_1, x_2 \rangle^d$
- Полиномиальное ядро с мономами степени $\leq d$: $K(x_1, x_2) = (\langle x_1, x_2 \rangle + 1)^d$

Примеры ядер

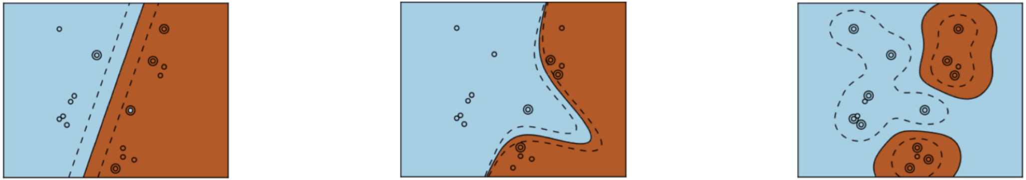
- Полиномиальное ядро с мономами степени d: $K(x_1, x_2) = \langle x_1, x_2 \rangle^d$
- Полиномиальное ядро с мономами степени $\leq d$: $K(x_1, x_2) = (\langle x_1, x_2 \rangle + 1)^d$
- Радиальное ядро (RBF): $K(x_1, x_2) = \exp(-\gamma ||x_1 - x_2||^2)$ (наиболее универсальное)

- Полиномиальное ядро с мономами степени d: $K(x_1, x_2) = \langle x_1, x_2 \rangle^d$
- Полиномиальное ядро с мономами степени $\leq d$: $K(x_1, x_2) = (\langle x_1, x_2 \rangle + 1)^d$
- Радиальное ядро (RBF): $K(x_1, x_2) = \exp(-\gamma ||x_1 - x_2||^2)$ (наиболее универсальное)

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

Полиномиальное ядро
$K(x_1, x_2) = (\langle x_1, x_2 \rangle + 1)^3$

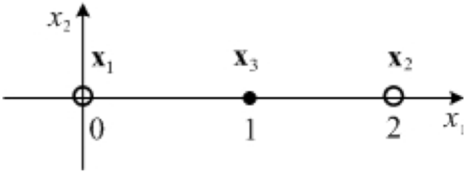
Радиальное ядро
$K(x_1, x_2) = \exp(-||x_1 - x_2||^2)$



Пример расчётов для SVM 1

Задача

Методом опорных векторов разделить классы $A = \{x_1, x_2\}$ и $B = \{x_3\}$, если $x_1 = (0, 0), x_2 = (2, 0), x_3 = (1, 0)$. Полагаем, что $y_1 = y_2 = +1, y_3 = -1$.



Указание [доказать]

Любые $m + 1$ векторов могут быть разделены на любые два класса с помощью мономиального отображения степени не больше m.

Пример расчётов для SVM 2

Поэтому отображение ищем в виде $\varphi : X \to \{x_1^{i_1} x_2^{i_2} \ldots x_n^{i_n}\}_{i_1+i_2+\dots+i_n \le m}$ для $X = \mathbb{R}^n, n = 2, m = 2$.

Пример расчётов для SVM 2

Поэтому отображение ищем в виде $\varphi : X \to \{x_1^{i_1} x_2^{i_2} \dots x_n^{i_n}\}_{i_1+i_2+\dots+i_n \leq m}$ для $X = \mathbb{R}^n, n = 2, m = 2$.
Ядро, соответствующее этому отображению: $K(x_1, x_2) = (\langle x_1, x_2 \rangle + 1)^2$.

Пример расчётов для SVM 2

Поэтому отображение ищем в виде $\varphi : X \to \{x_1^{i_1} x_2^{i_2} \dots x_n^{i_n}\}_{i_1+i_2+\dots+i_n \le m}$ для $X = \mathbb{R}^n, n = 2, m = 2$.
Ядро, соответствующее этому отображению: $K(x_1, x_2) = (\langle x_1, x_2\rangle + 1)^2$.
Будем решать задачу

$$\begin{cases} -Q(\lambda) = \frac{1}{2}\sum_{i=1}^m \sum_{j=1}^m \lambda_i \lambda_j y_i y_j K(x_i, x_j) - \sum_{i=1}^m \lambda_i \to \min_\lambda, \\ 0 \le \lambda_i \le C, \\ \sum_{i=1}^m \lambda_i y_i = 0. \end{cases}$$

Пример расчётов для SVM 3

$$-Q(\lambda) = \frac{1}{2}\sum_{i=1}^{3}\sum_{j=1}^{3}\lambda_i\lambda_j y_i y_j(\langle x_i, x_j\rangle + 1)^2 - \sum_{i=1}^{3}\lambda_i =$$
$$\frac{1}{2}(2\lambda_1\lambda_2 - 2\lambda_1\lambda_3 + \lambda_1^2 - 18\lambda_2\lambda_3 + 25\lambda_2^2 + 4\lambda_3^2) - (\lambda_1 + \lambda_2 + \lambda_3).$$

Пример расчётов для SVM 3

$-Q(\lambda) = \frac{1}{2}\sum_{i=1}^{3}\sum_{j=1}^{3}\lambda_i\lambda_j y_i y_j(\langle x_i, x_j\rangle + 1)^2 - \sum_{i=1}^{3}\lambda_i =$
$\frac{1}{2}(2\lambda_1\lambda_2 - 2\lambda_1\lambda_3 + \lambda_1^2 - 18\lambda_2\lambda_3 + 25\lambda_2^2 + 4\lambda_3^2) - (\lambda_1 + \lambda_2 + \lambda_3).$
Из условия $\sum_{i=1}^{3}\lambda_i y_i = 0$ имеем $\lambda_3 = \lambda_1 + \lambda_2$.

Пример расчётов для SVM 3

$-Q(\lambda) = \frac{1}{2}\sum_{i=1}^{3}\sum_{j=1}^{3}\lambda_i\lambda_j y_i y_j(\langle x_i, x_j\rangle + 1)^2 - \sum_{i=1}^{3}\lambda_i =$
$\frac{1}{2}(2\lambda_1\lambda_2 - 2\lambda_1\lambda_3 + \lambda_1^2 - 18\lambda_2\lambda_3 + 25\lambda_2^2 + 4\lambda_3^2) - (\lambda_1 + \lambda_2 + \lambda_3).$
Из условия $\sum_{i=1}^{3}\lambda_i y_i = 0$ имеем $\lambda_3 = \lambda_1 + \lambda_2$.
Подставим в выражение для $-Q(\lambda) = \frac{1}{2}(3\lambda_1^2 + 11\lambda_2^2 - 10\lambda_1\lambda_2) - 2(\lambda_1 + \lambda_2)$.

Пример расчётов для SVM 3

$-Q(\lambda) = \frac{1}{2}\sum_{i=1}^{3}\sum_{j=1}^{3}\lambda_i\lambda_j y_i y_j(\langle x_i, x_j\rangle + 1)^2 - \sum_{i=1}^{3}\lambda_i =$
$\frac{1}{2}(2\lambda_1\lambda_2 - 2\lambda_1\lambda_3 + \lambda_1^2 - 18\lambda_2\lambda_3 + 25\lambda_2^2 + 4\lambda_3^2) - (\lambda_1 + \lambda_2 + \lambda_3).$
Из условия $\sum_{i=1}^{3}\lambda_i y_i = 0$ имеем $\lambda_3 = \lambda_1 + \lambda_2$.
Подставим в выражение для $-Q(\lambda) = \frac{1}{2}(3\lambda_1^2 + 11\lambda_2^2 - 10\lambda_1\lambda_2) - 2(\lambda_1 + \lambda_2)$.
Дифференцируем Q по λ:

$$\begin{cases} \frac{\partial Q}{\partial \lambda_1} = 0 & \Leftrightarrow 3\lambda_1 - 5\lambda_2 = 2, \\ \frac{\partial Q}{\partial \lambda_2} = 0 & \Leftrightarrow -5\lambda_1 + 11\lambda_2 = 2 \end{cases}$$

Пример расчётов для SVM 3

$-Q(\lambda) = \frac{1}{2}\sum_{i=1}^{3}\sum_{j=1}^{3}\lambda_i\lambda_j y_i y_j(\langle x_i, x_j\rangle + 1)^2 - \sum_{i=1}^{3}\lambda_i =$
$\frac{1}{2}(2\lambda_1\lambda_2 - 2\lambda_1\lambda_3 + \lambda_1^2 - 18\lambda_2\lambda_3 + 25\lambda_2^2 + 4\lambda_3^2) - (\lambda_1 + \lambda_2 + \lambda_3).$
Из условия $\sum_{i=1}^{3}\lambda_i y_i = 0$ имеем $\lambda_3 = \lambda_1 + \lambda_2$.
Подставим в выражение для $-Q(\lambda) = \frac{1}{2}(3\lambda_1^2 + 11\lambda_2^2 - 10\lambda_1\lambda_2) - 2(\lambda_1 + \lambda_2)$.
Дифференцируем Q по λ:

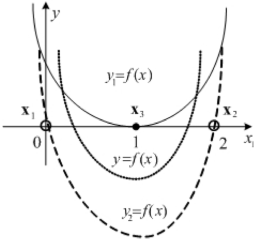
$$\begin{cases} \frac{\partial Q}{\partial \lambda_1} = 0 & \Leftrightarrow 3\lambda_1 - 5\lambda_2 = 2, \\ \frac{\partial Q}{\partial \lambda_2} = 0 & \Leftrightarrow -5\lambda_1 + 11\lambda_2 = 2 \end{cases}$$

Решая линейную систему уравнений, получаем $(\lambda_1, \lambda_2, \lambda_3) = (4, 2, 6)$.

Пример расчётов для SVM 4

Найдем разделяющую поверхность в виде: $f(x) = \sum_{i=1}^{3} \lambda_i y_i K(x_i, x) - w_0$,
$w_0 = \sum_{i=1}^{3} \lambda_i y_i K(x_i, x_j) - y_j$.
Поскольку все $\lambda > 0$, то можем взять в качестве опорного вектора x_1.
Получим $w_0 = 4K(x_1, x_1) + 2K(x_2, x_1) - 6K(x_3, x_1) - 1 = -1$.
Теперь рассчитаем основную часть: $\sum_{i=1}^{3} \lambda_i y_i K(x_i, x) = 4 \cdot 1 + 2(2x_1 + 1)^2 - 6(x_1 + 1)^2$.
Объединяя, получаем разделяющую поверхность: $f(x) = 2x_1^2 - 4x_1 + 1$.

Нули разделителя: $f(x) = 0 \Leftrightarrow x_1 = 1 \pm \frac{\sqrt{2}}{2}$.
Первый край полосы:
$f_1(x) = f(x) + 1 = 2(x_1 - 1)^2$. Нули: $x_1 = 1$.
Второй край полосы:
$f_2(x) = f(x) - 1 = 2x_1(x_1 - 2)$. Нули: $x_1 = 0, x_1 = 2$.



Многоклассовый SVM

Предположим, что нужно построить классификатор методом опорных векторов для задачи классификации с количеством классов $N > 2$. Тогда возможны 2 варианта:

Многоклассовый SVM

Предположим, что нужно построить классификатор методом опорных векторов для задачи классификации с количеством классов $N > 2$. Тогда возможны 2 варианта:

Стратегия “один-против-всех”

Обучаем N бинарных SVM-классификаторов, каждый из которых отделяет некоторый класс от остальных $N-1$ классов.
Затем в качестве класса берем $a(x, w, w_0) = \arg\max_{c \in Y}(\langle w^c, x \rangle - w_0^c)$

Многоклассовый SVM

Предположим, что нужно построить классификатор методом опорных векторов для задачи классификации с количеством классов $N > 2$. Тогда возможны 2 варианта:

Стратегия "один-против-всех"

Обучаем N бинарных SVM-классификаторов, каждый из которых отделяет некоторый класс от остальных $N-1$ классов.
Затем в качестве класса берем $a(x, w, w_0) = \arg\max_{c \in Y}(\langle w^c, x\rangle - w_0^c)$

Стратегия "каждый-против-каждого"

Обучаем $\frac{N(N-1)}{2}$ бинарных SVM-классификаторов, каждый из которых отделяет между собой некоторую пару классов.
В результате применения классификаторов получаем $\frac{N(N-1)}{2}$ доминирующих классов.
Итоговый класс выбирается большинством голосов.

Плюсы и минусы SVM

Плюсы

- Наглядная оптимизационная модель
- Задача имеет единственное решение
- Легко обобщается для нелинейной классификации

Плюсы и минусы SVM

Плюсы

- Наглядная оптимизационная модель
- Задача имеет единственное решение
- Легко обобщается для нелинейной классификации

Минусы

- Непонятно, как подбирать ядро в конкретном случае
- Нет встроенного отбора признаков (как в регрессии LASSO, например)
- Подбор константы C
- Решение задачи квадратичного программирования, особенно с экзотическими ядрами, может занять много времени

Источники

На основе материалов сайта http://www.machinelearning.ru.