Введение в искусственный интеллект. Машинное обучение

Тема семинара: Регрессия на основе опорных векторов — SVR

Бабин Д.Н., Иванов И.Е., Петюшко А.А.

кафедра Математической Теории Интеллектуальных Систем

План занятия

1. Постановка задачи SVR

План занятия

1. Постановка задачи SVR
2. Решение с помощью двойственной задачи

План занятия

1. Постановка задачи SVR
2. Решение с помощью двойственной задачи
3. Обобщение SVR с помощью ядрового трюка

Гребневая регрессия: напоминание

Рассмотрим задачу восстановления регрессии: $X \to Y$, $X = \mathbb{R}^n$, $Y = \mathbb{R}$ на обучающей выборке $X^m = (x_i, y_i)_{i=1}^m$.

Гребневая регрессия: напоминание

Рассмотрим задачу восстановления регрессии: $X \to Y$, $X = \mathbb{R}^n$, $Y = \mathbb{R}$ на обучающей выборке $X^m = (x_i, y_i)_{i=1}^m$.
Линейный алгоритм $a(x; w, w_0) = \langle w, x \rangle - w_0$.

Гребневая регрессия: напоминание

Рассмотрим задачу восстановления регрессии: $X \to Y$, $X = \mathbb{R}^n$, $Y = \mathbb{R}$ на обучающей выборке $X^m = (x_i, y_i)_{i=1}^m$.
Линейный алгоритм $a(x; w, w_0) = \langle w, x \rangle - w_0$.
Вспомним функцию потерь для гребневой регрессии ($u = (w, w_0), \hat{x} = (x, -1)$):

$$L(u, \hat{X}^m) = \sum_{i=1}^{m} (a(\hat{x}_i; u) - y_i)^2 + \frac{\alpha}{2} ||u||^2 \to \min$$

Гребневая регрессия: напоминание

Рассмотрим задачу восстановления регрессии: $X \to Y$, $X = \mathbb{R}^n$, $Y = \mathbb{R}$ на обучающей выборке $X^m = (x_i, y_i)_{i=1}^m$.
Линейный алгоритм $a(x; w, w_0) = \langle w, x \rangle - w_0$.
Вспомним функцию потерь для гребневой регрессии $(u = (w, w_0), \hat{x} = (x, -1))$:

$$L(u, \hat{X}^m) = \sum_{i=1}^{m} (a(\hat{x}_i; u) - y_i)^2 + \frac{\alpha}{2} ||u||^2 \to \min$$

Решение будет представляться в виде:

$$u = (\hat{X}^T \hat{X} + \alpha I_{n+1})^{-1} \cdot \hat{X}^T \cdot y$$

где $\hat{X}_{i,j} = \hat{x}_i^{(j)}$, $y = (y_1, \ldots, y_m)$, I_{n+1} — единичная матрица $(n+1) \times (n+1)$.

SVR: мотивация

Предположим:

- не обращаем внимание на те ошибки, которые по модулю меньше некоторого $\epsilon > 0$,
- ошибка — разность модуля обычной ошибки и этого значения ϵ,
- т.е. интересуют значения вне “коридора” шириной ϵ.

SVR: мотивация

Предположим:

- не обращаем внимание на те ошибки, которые по модулю меньше некоторого $\epsilon > 0$,
- ошибка — разность модуля обычной ошибки и этого значения ϵ,
- т.е. интересуют значения вне "коридора" шириной ϵ.

Обозначение:

$$ReLU(x) = \begin{cases} x, & x \geq 0, \\ 0, & x < 0 \end{cases}$$

Тогда функция потерь:

$$L(w, w_0; X^m) = \sum_{i=1}^{m} ReLU(|a(x_i; w, w_0) - y_i| - \epsilon) + \frac{1}{2C}||w||^2 \to \min$$

SVR: мотивация

Предположим:

- не обращаем внимание на те ошибки, которые по модулю меньше некоторого $\epsilon > 0$,
- ошибка — разность модуля обычной ошибки и этого значения ϵ,
- т.е. интересуют значения вне “коридора” шириной ϵ.

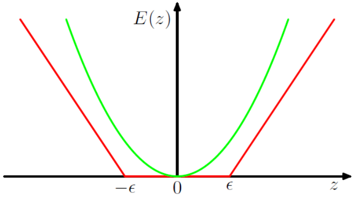
Обозначение:

$$ReLU(x) = \begin{cases} x, & x \geq 0, \\ 0, & x < 0 \end{cases}$$

Тогда функция потерь:

$$L(w, w_0; X^m) = \sum_{i=1}^{m} ReLU(|a(x_i; w, w_0) - y_i| - \epsilon) + \frac{1}{2C}||w||^2$$

Сравнение функции потерь $ReLU(|a(x_i; w, w_0) - y_i| - \epsilon)$ и $(a(x_i; w, w_0) - y_i)^2$:



SVR: обозначения

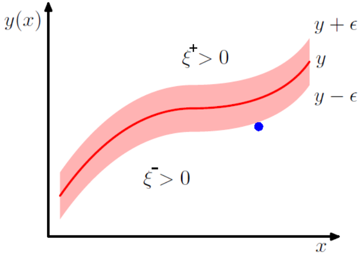
По аналогии с SVM введем два типа дополнительных переменных $\xi_i^- \geq 0$ и $\xi_i^+ \geq 0$, которые будут отвечать за выход за "коридор" шириной ϵ:

$$\begin{cases} \xi_i^+ = & ReLU(\langle w, x_i \rangle - w_0 - y_i - \epsilon), \\ \xi_i^- = & ReLU(-\langle w, x_i \rangle + w_0 + y_i - \epsilon) \end{cases}$$

SVR: обозначения

По аналогии с SVM введем два типа дополнительных переменных $\xi_i^- \geq 0$ и $\xi_i^+ \geq 0$, которые будут отвечать за выход за "коридор" шириной ϵ:

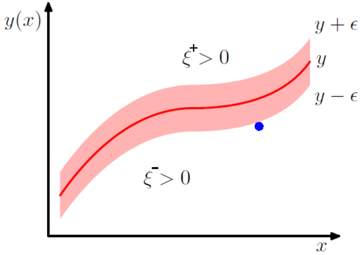
$$\begin{cases} \xi_i^+ = & ReLU(\langle w, x_i \rangle - w_0 - y_i - \epsilon), \\ \xi_i^- = & ReLU(-\langle w, x_i \rangle + w_0 + y_i - \epsilon) \end{cases}$$



SVR: обозначения

По аналогии с SVM введем два типа дополнительных переменных $\xi_i^- \geq 0$ и $\xi_i^+ \geq 0$, которые будут отвечать за выход за “коридор” шириной ϵ:

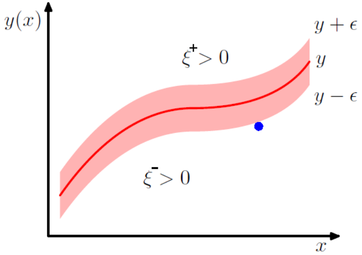
$$\begin{cases} \xi_i^+ = & ReLU(\langle w, x_i \rangle - w_0 - y_i - \epsilon), \\ \xi_i^- = & ReLU(-\langle w, x_i \rangle + w_0 + y_i - \epsilon) \end{cases}$$



Таким образом, получим, что:

- Если предсказание алгоритма $a(x_i; w, w_0)$ находится внутри ϵ-“коридора”, то есть $y_i - \epsilon \leq \langle w, x_i \rangle - w_0 \leq y_i + \epsilon$, то $\xi_i^+ = \xi_i^- = 0$, По аналогии с SVM введем два типа дополнительных переменных $\xi_i^- \geq 0$ и $\xi_i^+ \geq 0$, которые будут отвечать за выход за "коридор" шириной ϵ:

$$\begin{cases} \xi_i^+ = & ReLU(\langle w, x_i \rangle - w_0 - y_i - \epsilon), \\ \xi_i^- = & ReLU(-\langle w, x_i \rangle + w_0 + y_i - \epsilon) \end{cases}$$



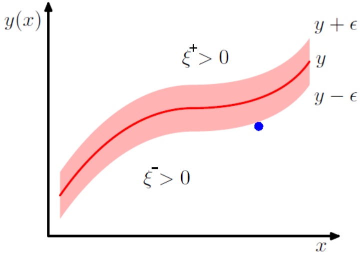
Таким образом, получим, что:

- Если предсказание алгоритма $a(x_i; w, w_0)$ находится внутри ϵ-"коридора", то есть $y_i - \epsilon \leq \langle w, x_i \rangle - w_0 \leq y_i + \epsilon$, то $\xi_i^+ = \xi_i^- = 0$,
- Если $\langle w, x_i \rangle - w_0 > y_i + \epsilon$, то $\xi_i^+ > 0, \xi_i^- = 0$,

SVR: обозначения

По аналогии с SVM введем два типа дополнительных переменных $\xi_i^- \geq 0$ и $\xi_i^+ \geq 0$, которые будут отвечать за выход за "коридор" шириной ϵ:

$$\begin{cases} \xi_i^+ = & ReLU(\langle w, x_i \rangle - w_0 - y_i - \epsilon), \\ \xi_i^- = & ReLU(-\langle w, x_i \rangle + w_0 + y_i - \epsilon) \end{cases}$$



Таким образом, получим, что:

- Если предсказание алгоритма $a(x_i; w, w_0)$ находится внутри ϵ-"коридора", то есть $y_i - \epsilon \leq \langle w, x_i \rangle - w_0 \leq y_i + \epsilon$, то $\xi_i^+ = \xi_i^- = 0$,
- Если $\langle w, x_i \rangle - w_0 > y_i + \epsilon$, то $\xi_i^+ > 0, \xi_i^- = 0$,
- Если $\langle w, x_i \rangle - w_0 < y_i - \epsilon$, то $\xi_i^+ = 0, \xi_i^- > 0$.

SVR: формулировка

Тогда оптимизационную задачу можно переписать как (домножив функцию потерь на C и учитывая $ReLU(x) \geq x$):

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

Условия Каруша-Куна-Таккера (ККТ)[1,2]

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

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

[1] W. Karush (1939). Minima of Functions of Several Variables with Inequalities as Side Constraints.
[2] Kuhn, H. W.; Tucker, A. W. (1951). Nonlinear programming.

Условия Каруша-Куна-Таккера (ККТ)[1,2]

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

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

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

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

[1] W. Karush (1939). Minima of Functions of Several Variables with Inequalities as Side Constraints.
[2] Kuhn, H. W.; Tucker, A. W. (1951). Nonlinear programming.

Условия ККТ и SVR

Функция Лагранжа для SVR

$L(w, w_0, \xi^+, \xi^-; \lambda^+, \lambda^-, \eta^+, \eta^-) = \frac{1}{2}||w||^2 + C\sum_{i=1}^m(\xi_i^+ + \xi_i^-) + \sum_{i=1}^m \lambda_i^+(\langle w, x_i\rangle - w_0 - y_i - \epsilon - \xi_i^+) + \sum_{i=1}^m \lambda_i^-(-\langle w, x_i\rangle + w_0 + y_i - \epsilon - \xi_i^-) - \sum_{i=1}^m(\xi_i^+\eta_i^+ + \xi_i^-\eta_i^-)$, где:

$\lambda_i^{\pm}$ – переменные, двойственные к ограничениям $\pm\langle w, x_i\rangle \mp w_0 \mp y_i - \epsilon - \xi_i^{\pm} \leq 0$,

$\eta_i^{\pm}$ – переменные, двойственные к ограничениям $-\xi_i^{\pm} \leq 0$.

Условия ККТ и SVR

Функция Лагранжа для SVR

$L(w, w_0, \xi^+, \xi^-; \lambda^+, \lambda^-, \eta^+, \eta^-) = \frac{1}{2}||w||^2 + C\sum_{i=1}^m(\xi_i^+ + \xi_i^-) + \sum_{i=1}^m \lambda_i^+(\langle w, x_i\rangle - w_0 - y_i - \epsilon - \xi_i^+) + \sum_{i=1}^m \lambda_i^-(-\langle w, x_i\rangle + w_0 + y_i - \epsilon - \xi_i^-) - \sum_{i=1}^m(\xi_i^+\eta_i^+ + \xi_i^-\eta_i^-)$, где:

$\lambda_i^\pm$ – переменные, двойственные к ограничениям $\pm\langle w, x_i\rangle \mp w_0 \mp y_i - \epsilon - \xi_i^\pm \leq 0$,

$\eta_i^\pm$ – переменные, двойственные к ограничениям $-\xi_i^\pm \leq 0$.

Необходимые условия примут вид:

$$\begin{cases} \frac{\partial L}{\partial w} = 0; \frac{\partial L}{\partial w_0} = 0; \frac{\partial L}{\partial \xi^\pm} = 0 & \\ \xi_i^\pm \geq 0, \lambda_i^\pm \geq 0, \eta_i^\pm \geq 0 & i = 1, \dots, m \\ \lambda_i^\pm = 0 \text{ либо } \pm\langle w, x_i\rangle \mp w_0 \mp y_i - \epsilon - \xi_i^\pm = 0 & i = 1, \dots, m \\ \eta_i^\pm = 0 \text{ либо } \xi_i^\pm = 0 & i = 1, \dots, m \end{cases}$$

Дифференцируем функцию Лагранжа

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

Дифференцируем функцию Лагранжа

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

1. Одновременно $\lambda_i^+ > 0, \lambda_i^- > 0$ не бывает (следствие из условий дополняющей нежесткости)

Дифференцируем функцию Лагранжа

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

1. Одновременно $\lambda_i^+ > 0, \lambda_i^- > 0$ не бывает (следствие из условий дополняющей нежесткости)
2. $\lambda_i^+ = \lambda_i^- = 0 \Rightarrow \eta_i^+ = \eta_i^- = C \Rightarrow \xi_i^+ = \xi_i^- = 0$, и вектор x_i внутри ϵ-“коридора”

Дифференцируем функцию Лагранжа

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

1. Одновременно $\lambda_i^+ > 0, \lambda_i^- > 0$ не бывает (следствие из условий дополняющей нежесткости)
2. $\lambda_i^+ = \lambda_i^- = 0 \Rightarrow \eta_i^+ = \eta_i^- = C \Rightarrow \xi_i^+ = \xi_i^- = 0$, и вектор x_i внутри ϵ-“коридора”
3. $\lambda_i^+ > 0$ либо $\lambda_i^- > 0$ соответствуют т.н. “опорным” векторам x_i

Дифференцируем функцию Лагранжа

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

1. Одновременно $\lambda_i^+ > 0, \lambda_i^- > 0$ не бывает (следствие из условий дополняющей нежесткости)
2. $\lambda_i^+ = \lambda_i^- = 0 \Rightarrow \eta_i^+ = \eta_i^- = C \Rightarrow \xi_i^+ = \xi_i^- = 0$, и вектор x_i внутри ϵ-“коридора”
3. $\lambda_i^+ > 0$ либо $\lambda_i^- > 0$ соответствуют т.н. “опорным” векторам x_i
   - $0 < \lambda_i^\pm < C (\Rightarrow \eta_i^\pm > 0 \Rightarrow \xi_i^\pm = 0)$ соответствует вектору x_i на границе $\pm\langle w, x_i\rangle \mp w_0 \mp y_i - \epsilon = 0$

Дифференцируем функцию Лагранжа

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

1. Одновременно $\lambda_i^+ > 0, \lambda_i^- > 0$ не бывает (следствие из условий дополняющей нежесткости)
2. $\lambda_i^+ = \lambda_i^- = 0 \Rightarrow \eta_i^+ = \eta_i^- = C \Rightarrow \xi_i^+ = \xi_i^- = 0$, и вектор x_i внутри ϵ-“коридора”
3. $\lambda_i^+ > 0$ либо $\lambda_i^- > 0$ соответствуют т.н. “опорным” векторам x_i
   - $0 < \lambda_i^\pm < C (\Rightarrow \eta_i^\pm > 0 \Rightarrow \xi_i^\pm = 0)$ соответствует вектору x_i на границе $\pm\langle w, x_i\rangle \mp w_0 \mp y_i - \epsilon = 0$
   - $\lambda_i^\pm = C$ соответствует вектору x_i вне ϵ-“коридора” при $\xi_i^\pm > 0$, либо лежащему на границе $\pm\langle w, x_i\rangle \mp w_0 \mp y_i - \epsilon = 0$ при $\xi_i^\pm = 0$.

О двойственных задачах

Прямая задача:

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

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

[3] Wolfe, P. (1961). A duality theorem for non-linear programming.

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: $Q(\mu, \lambda) = \min_x L(x, \mu, \lambda)$.

[3] Wolfe, P. (1961). A duality theorem for non-linear programming.

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:

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

[3]Wolfe, P. (1961). A duality theorem for non-linear programming.

О двойственных задачах

Прямая задача:

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

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:

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

Теорема (Дуальность Вулфа[3])

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

[3] Wolfe, P. (1961). A duality theorem for non-linear programming.

Двойственная задача для SVR

Подставим решения прямой задачи

$$w = \sum_{i=1}^{m}(\lambda_i^- - \lambda_i^+)x_i, \sum_{i=1}^{m}(\lambda_i^- - \lambda_i^+) = 0, \lambda_i^\pm + \eta_i^\pm = C$$

в функцию Лагранжа

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

Двойственная задача для SVR

Подставим решения прямой задачи

$$w = \sum_{i=1}^{m}(\lambda_i^- - \lambda_i^+)x_i, \sum_{i=1}^{m}(\lambda_i^- - \lambda_i^+) = 0, \lambda_i^\pm + \eta_i^\pm = C$$

в функцию Лагранжа

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

Получим:

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

Двойственная задача для SVR

Подставим решения прямой задачи

$$w = \sum_{i=1}^{m} (\lambda_i^- - \lambda_i^+) x_i, \sum_{i=1}^{m} (\lambda_i^- - \lambda_i^+) = 0, \lambda_i^\pm + \eta_i^\pm = C$$

в функцию Лагранжа

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

Получим:

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

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

Двойственная задача для SVR

Объединяя, получаем формулировку двойственной задачи:

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

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

Двойственная задача для SVR

Объединяя, получаем формулировку двойственной задачи:

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

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

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

Объединяя, получаем формулировку двойственной задачи:

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

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

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

Двойственная задача для SVR

Объединяя, получаем формулировку двойственной задачи:

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

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

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

Двойственная задача для SVR

Объединяя, получаем формулировку двойственной задачи:

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

Минимизируется квадратичный функционал $-Q(\lambda^{\pm})$, 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[4]).

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

Решение прямой задачи для SVR

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

$$\begin{cases} w = \sum_{i=1}^m (\lambda_i^- - \lambda_i^+) x_i, & \text{суммируем только по опорным векторам } \lambda_i^{\pm} \neq 0 \\ w_0 = \sum_{i=1}^m (\lambda_i^- - \lambda_i^+) \langle x_i, x_j \rangle - y_j - \epsilon & \text{для опорного вектора на верхней границе} \\ 0 < \lambda_j^+ < C, & \langle w, x_j \rangle - w_0 - y_j - \epsilon = 0 \end{cases}$$

Решение прямой задачи для SVR

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

$$\begin{cases} w = \sum_{i=1}^m (\lambda_i^- - \lambda_i^+) x_i, & \text{суммируем только по опорным векторам } \lambda_i^{\pm} \neq 0 \\ w_0 = \sum_{i=1}^m (\lambda_i^- - \lambda_i^+) \langle x_i, x_j \rangle - y_j - \epsilon & \text{для опорного вектора на верхней границе} \\ 0 < \lambda_j^+ < C, & \langle w, x_j \rangle - w_0 - y_j - \epsilon = 0 \end{cases}$$

При этом сам линейный алгоритм примет вид

$$a(x) = \sum_{i=1}^m (\lambda_i^- - \lambda_i^+) \langle x_i, x \rangle - w_0$$

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

SVR и нелинейность

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

SVR и нелинейность

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

Определение ядра

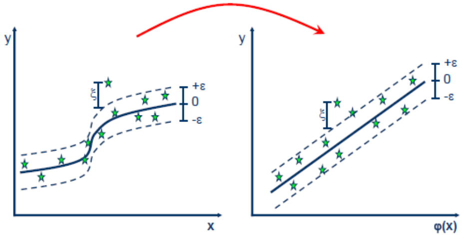
Ядро – функция $K : X \times X \to \mathbb{R}$, т.ч. $K(x_1, x_2) = \langle \varphi(x_1), \varphi(x_2) \rangle$ при некотором $\varphi : X \to H$, где H – гильбертово пространство.

SVR и нелинейность

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

Определение ядра

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.



SVR с другими ядрами

Изначально наша двойственная задача была сформулирована в терминах линейного ядра:

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

SVR с другими ядрами

Изначально наша двойственная задача была сформулирована в терминах линейного ядра:

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

Когда мы переходим в другое пространство $\varphi : X \to H$, то соответственно меняем ядро с $\langle x_i, x_j\rangle$ на $K(x_i, x_j)$. При этом задача преобразуется в:

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

SVR с другими ядрами

А линейный алгоритм примет вид (x_j - опорный вектор на верхней границе):

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

где линейная часть $w = \sum_{i=1}^{m} (\lambda_i^- - \lambda_i^+) \varphi(x_i)$

Плюсы и минусы SVR

Плюсы

- Наглядная оптимизационная модель
- Задача имеет единственное решение
- Легко обобщается для нелинейной регрессии
- Отличный пример использования теории обобщающей способности

Плюсы и минусы SVR

Плюсы

- Наглядная оптимизационная модель
- Задача имеет единственное решение
- Легко обобщается для нелинейной регрессии
- Отличный пример использования теории обобщающей способности

Минусы

- Непонятно, как подбирать ядро в конкретном случае
- Подбор константы C
- Решение задачи квадратичного программирования, особенно с экзотическими ядрами, может занять много времени

Время для вопросов