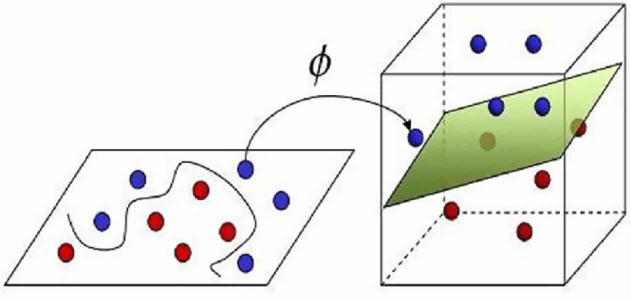
Ядерное обобщение методов

Виктор Китов

v.v.kitov@yandex.ru



Содержание

Построение прогнозов

1. Решим двойственную задачу, найдем оптимальные двойственные переменные α_i^*, $i = 1, 2, ...N$

$$\begin{cases} L_D = \sum_{i=1}^N \alpha_i - \frac{1}{2}\sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle \to \max_\alpha \\ \sum_{i=1}^N \alpha_i y_i = 0 \\ 0 \le \alpha_i \le C \end{cases}$$

2. Найдем оптимальное w_0:

$$w_0 = \frac{1}{|\tilde{SV}|}\left(\sum_{j \in \tilde{SV}} y_j - \sum_{j \in \tilde{SV}} \sum_{i \in \mathcal{SV}} \alpha_i^* y_i \langle x_i, x_j \rangle\right)$$

3. Построим прогноз для x:

$$\widehat{y} = \text{sign}[w^T x + w_0] = \text{sign}[\sum_{i \in \mathcal{SV}} \alpha_i^* y_i \langle x_i, x \rangle + w_0]$$

Обобщение решения через ядра

1. Решим двойственную задачу, найдем оптимальные двойственные переменные α_i^*, $i = 1, 2, ...N$

$$\begin{cases} L_D = \sum_{i=1}^N \alpha_i - \frac{1}{2}\sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j K(x_i, x_j) \to \max_\alpha \\ \sum_{i=1}^N \alpha_i y_i = 0 \\ 0 \le \alpha_i \le C \end{cases}$$

2. Найдем оптимальное w_0:

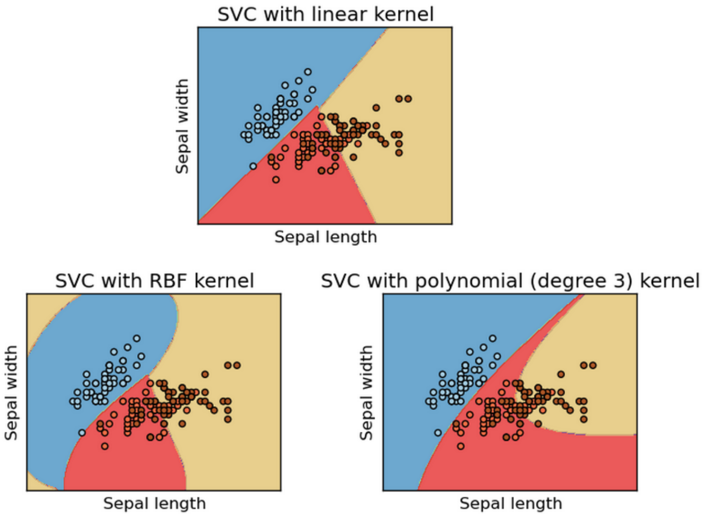
$$w_0 = \frac{1}{|SV|}\left(\sum_{j \in \tilde{SV}} y_j - \sum_{j \in \tilde{SV}} \sum_{i \in \mathcal{SV}} \alpha_i^* y_i K(x_i, x_j)\right)$$

3. Построим прогноз для x:

$$\widehat{y} = \text{sign}[w^T x + w_0] = \text{sign}[\sum_{i \in \mathcal{SV}} \alpha_i^* y_i K(x_i, x) + w_0]$$

- Замена $\langle x, x' \rangle \to K(x, x')$ для ядра $K(x, x') = \langle \phi(x), \phi(x') \rangle$

Запуск с разными ядрами



Содержание

Обобщение методов через ядра

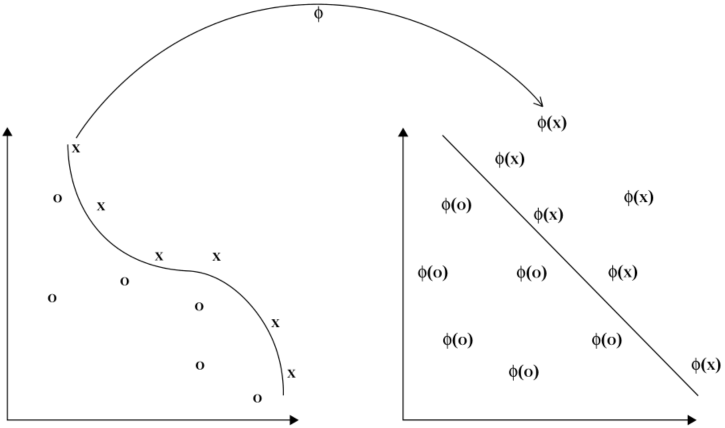
Обобщение методов через ядра (kernel trick)

Если прогноз зависит только от объектов через попарные скалярные произведения, заменим $\langle x, z\rangle \longrightarrow K(x, z)$.

- $K(x, z)$ должно соответствовать скалярному произведению $\langle\phi(x), \phi(z)\rangle$ после некоторого преобразования признаков $x \to \phi(x)$.
- $K(x, z)$ называется ядром Мерсера (Mercer kernel).
    - не каждая $K(x, z)$ будет ядром Мерсера
- Интуитивно $K(x, z)$ измеряет близость объектов.
- Допускают ядерное обобщение: классификатор опорных векторов, регрессия опорных векторов, гребневая регрессия, K-NN, K-средних, метод главных компонент и др.

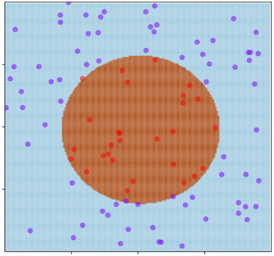
Спрямляющее пространство

Пространство $\phi(x)$ называется спрямляющим пространством.

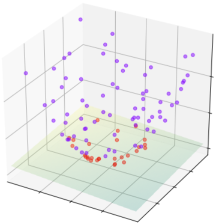


Важность преобразования признаков

- Рассмотрим линейную классификацию.
- Исходное пространство $x = [x^1, x^2]$ не может разделить классы.
- $\phi(x) = [x^1, x^2, (x^1)^2 + (x^2)^2]$ - может.
    - соответствующее ядро: $K(x, z) = \langle x, z \rangle + \|x\|^2 \|z\|^2$



x



ϕ(x)

Обобщение расстояния через ядра[1]

- Евклидово расстояние в исходном пространстве x:

$$\rho(x,z)^2 = \langle x-z, x-z\rangle$$

- Евклидово расстояние в спрямляющем пространстве $\phi(x)$:

$$\begin{aligned}
\rho(x,z)^2 &= \langle \phi(x)-\phi(z), \phi(x)-\phi(z)\rangle \\
&= \langle \phi(x), \phi(x)\rangle + \langle \phi(z), \phi(z)\rangle - 2\langle \phi(x), \phi(z)\rangle \\
&= K(x,x) + K(z,z) - 2K(x,z)
\end{aligned}$$

- Таким способом можем обобщить через ядра все метрические методы: метод ближайших центроидов, K-ближайших соседей, K-средних, метод главных компонент и т.д.

[1] Как в терминах ядер можем посчитать расстояние между $\phi(x)/\left\|\phi(x)\right\|$ и $\phi(z)/\left\|\phi(z)\right\|$?

Содержание

Примеры ядер

- $K(x,z) = \langle x, z \rangle$ - ядро с $\phi(x) = x$
- $K(x,z) \equiv 1$ - ядро с $\phi(x) = 1$
- $x^T Az,\ A \succ 0$ - ядро с $\phi(x) = Ux$ с верхней треугольной матрицей U из разложения Холецкого

$$A = U^T U$$
$$x^T Az = x^T U^T Uz = (Ux)^T Uz = \langle Ux, Uz \rangle$$

Полиномиальное ядро

Рассмотрим объекты с двумя признаками $x = [x_1, x_2]$ и $z = [z_1, z_2]$.

$$
\begin{aligned}
K(x, z) &= (x^T z)^2 = (x_1 z_1 + x_2 z_2)^2 = \\
&= x_1^2 z_1^2 + x_2^2 z_2^2 + 2 x_1 x_2 z_1 z_2 \\
&= \phi^T(x) \phi(z)
\end{aligned}
$$

для

$$
\phi(x) = (x_1^2, x_2^2, \sqrt{2} x_1 x_2)
$$
$$
\phi(z) = (z_1^2, z_2^2, \sqrt{2} z_1 z_2)
$$

Полиномиальное ядро со смещением

Рассмотрим объекты с двумя признаками $x=[x_1,x_2]$ и $z=[z_1,z_2]$.

$$\begin{aligned} K(x,z) &= (1+x^T z)^2 = (1+x_1 z_1 + x_2 z_2)^2 = \\ &= 1 + x_1^2 z_1^2 + x_2^2 z_2^2 + 2x_1 z_1 + 2x_2 z_2 + 2x_1 x_2 z_1 z_2 \\ &= \phi^T(x)\phi(z) \end{aligned}$$

для

$$\begin{aligned} \phi(x) =& (1,\ x_1^2,\ x_2^2,\ \sqrt{2}x_1,\ \sqrt{2}x_2,\ \sqrt{2}x_1 x_2) \\ \phi(z) =& (1,\ z_1^2,\ z_2^2,\ \sqrt{2}z_1,\ \sqrt{2}z_2,\ \sqrt{2}z_1 z_2) \end{aligned}$$

Популярные ядра

Ядро	**Формула**	**Параметры**
линейное	$\langle x, z\rangle$	-
полиномиальное	$(\alpha\langle x, z\rangle + \beta)^M$	$\alpha > 0, \beta \geq 0, M = 1, 2, \ldots$
Гауссово (RBF)	$e^{-\gamma\|x-z\|^2}$	$\gamma > 0$

- **Гауссово ядро**: переход в бесконечномерное $\phi(x)$.
- **Полиномиальное ядро с** $\beta = 0$: $\phi(x)$ из всех полиномиальных признаков порядка M.
  - C_{M+D-1}^{D-1} признаков
  - расчет $\langle\phi(x), \phi(z)\rangle$: $O\left(C_{M+D-1}^{D-1}\right)$, расчет $K(x, z)$: $O(D)$
- **Полиномиальное ядро с** $\beta > 0$[2]: $\phi(x)$ из всех полиномиальных признаков порядка $\leq M$.
  - C_{M+D}^{D} признаков ($(D+1)$-й признак $\equiv 1$)
  - расчет $\langle\phi(x), \phi(z)\rangle$: $O\left(C_{M+D}^{D}\right)$, расчет $K(x, z)$: $O(D)$

[2] Сравните сложность построения прогнозов методом опорных векторов напрямую и через ядра.

Число мономов из D признаков порядка M

- Число мономов из D признаков порядка M равно C_{M+D-1}^{D-1}.
- Пусть звезда обозначает увеличение степени, а черта отделяет признаки.
    - M звезд, $D-1$ черточек.

$$
\begin{array}{rcl}
x_1^3 & \longleftrightarrow & \star\star\star\,|\,| \\
x_1^2x_2 & \longleftrightarrow & \star\star\,|\star| \\
x_1^2x_3 & \longleftrightarrow & \star\star\,|\,|\star \\
x_1x_2^2 & \longleftrightarrow & \star\,|\star\star| \\
x_1x_2x_3 & \longleftrightarrow & \star\,|\star\,|\star \\
x_1x_3^2 & \longleftrightarrow & \star\,|\,|\star\star \\
x_2^3 & \longleftrightarrow & |\star\star\star| \\
x_2^2x_3 & \longleftrightarrow & |\star\star|\star \\
x_2x_3^2 & \longleftrightarrow & |\star\,|\star\star \\
x_3^3 & \longleftrightarrow & |\,|\star\star\star
\end{array}
$$

Вычисление числа мономов степени 3 от 3х признаков.

Мотивация применения ядер

Мотивация применения ядер:

- Пространство x недостаточно, чтобы разделить классы, а $\phi(x)$-достаточно.
- $\langle\phi(x), \phi(z)\rangle$ напрямую вычисляются долго, а $K(x, z)$-быстро.
    - для метода опорных векторов с полиномиальным ядром, сложность прогнозов напрямую $O(C_{M+D}^{D})$, а через ядра $O\left(|SV| \cdot D\right)$
- $\phi(x)$ соответствует переходу в бесконечномерное пространство
    - например, для Гауссова ядра
- $x \notin \mathbb{R}^D$, зато определено скалярное произведение $K(x, z)$:
    - строки произвольной длины ($K()$ зависит от общей подстроки)
    - множества ($K()$ зависит от общего подмножества)
    - графы ($K()$ зависит от общего подграфа)

Является ли $K(x,z)$ ядром Мерсера?

Теорема (Мерсер): Функция $K(x,x')$ - ядро Мерсера тогда и только тогда, когда

1. она симметрична: $K(x,x') = K(x',x)$
2. и неотрицательно определена:
   - определение 1: для любой функции $g: X \to \mathbb{R}$
   $$\int_X \int_X K(x,x')g(x)g(x')dxdx' \geq 0$$
   - определение 2 (эквивалентное): для любого M и набора $x_1, x_2, ...x_M$ матрица Грамма неотрицательно определена $\{K(x_i,x_j)\}_{i,j=1}^M \succeq 0$

Является ли $K(x,z)$ ядром Мерсера?

Теорема (Мерсер): Функция $K(x,x')$ - ядро Мерсера тогда и только тогда, когда

1. она симметрична: $K(x,x') = K(x',x)$
2. и неотрицательно определена:
   - определение 1: для любой функции $g: X \to \mathbb{R}$

   $$\int_X \int_X K(x,x')g(x)g(x')dxdx' \geq 0$$

   - определение 2 (эквивалентное): для любого M и набора $x_1, x_2, ...x_M$ матрица Грамма неотрицательно определена $\{K(x_i,x_j)\}_{i,j=1}^M \succeq 0$

Комментарий: Неотрицательности $K(x,x') \geq 0 \; \forall x, x'$ недостаточно, например

$$\begin{pmatrix} 1 & -1 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ -1 \end{pmatrix} = \begin{pmatrix} 1 & -1 \end{pmatrix} \begin{pmatrix} -1 \\ 1 \end{pmatrix} = -1-1 < 0$$

Преобразования, не выводящие из класса ядер

Новые ядра можно строить проще из других ядер, используя преобразования, не выводящие из класса ядер[3]:

1. $K(x,z) = cK_1(x,z),\ \forall c > 0$
2. $K(x,z) = K_1(x,z)K_2(x,z)$
3. $K(x,z) = K_1(x,z) + K_2(x,z)$
4. $K(x,z) = K_1(\varphi(x),\varphi(z)),\ \forall\varphi : \mathcal{X} \to \mathbb{R}^M$
5. $K(x,z) = h(x)K_1(x,z)h(z),\ \forall h : \mathcal{X} \to \mathbb{R}$
6. $K(x,z) = e^{K_1(x,z)}$

[3] Докажите некоторые из свойств

Заключение

- Метод обобщается через ядра, если прогноз зависит только от $\langle x, x' \rangle$.
- Обобщение через ядра $\langle x, x' \rangle \longrightarrow K(x, x')$
- Подходит не каждая $K(x, x')$, а удовлетворяющая условию $K(x, x') = \langle \phi(x), \phi(x') \rangle$
- Ядра можно строить из других ядер.
- Ядра можно настраивать под задачу (kernel learning).
    - также как $\rho(x, x')$ в метрических методах.
- Популярные алгоритмы с ядерных обобщением:
    - классификатор опорных векторов
    - регрессия опорных векторов
    - гребневая регрессия