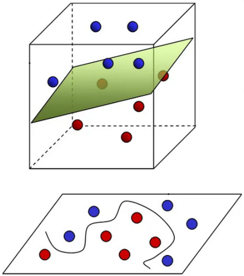
Ядерное обобщение методов

Виктор Китов

victorkitov.github.io



Курс поддержан фондом 'Интеллект'

Победитель конкурса VK среди курсов по IT

Содержание

Обобщение методов через ядра

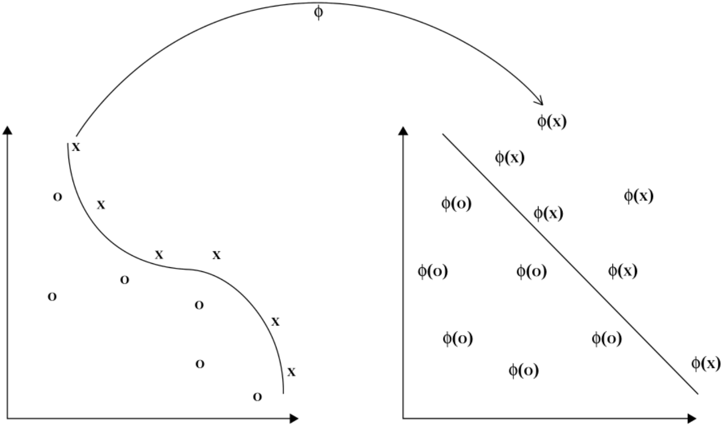
Обобщение методов через ядра (kernel trick)

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.

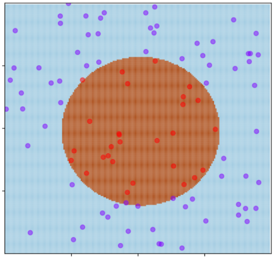
Спрямляющее пространство

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

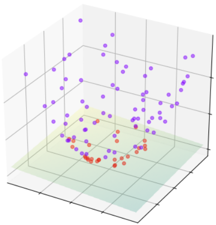


Важность преобразования признаков

- 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 $z=[z_1,z_2]$.

$$\begin{aligned}K(x,z) &= (x^Tz)^2=(x_1z_1+x_2z_2)^2=\\ &= x_1^2z_1^2+x_2^2z_2^2+2x_1x_2z_1z_2\\ &= \phi^T(x)\phi(z)\end{aligned}$$

для

$$\phi(x)=(x_1^2,x_2^2,\sqrt{2}x_1x_2)$$
$$\phi(z)=(z_1^2,z_2^2,\sqrt{2}z_1z_2)$$

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 $g: X \to \mathbb{R}$
   $$\int_X \int_X K(x,x')g(x)g(x')dxdx' \ge 0$$
   - 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 $K(x,x') \ge 0\ \forall x,x'$ 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

Содержание

Ядерное обобщение гребневой регрессии

$$\sum_{n=1}^{N}\left(x_n^T w - y_n\right)^2 + \lambda w^T w \to \min_{w}, \qquad \widehat{y}(x) = x^T w$$

Мы уже находили аналитическое решение[4]:

$$\widehat{w} = (X^T X + \lambda I)^{-1} X^T Y$$

в эквивалентном виде:

$$\begin{cases} \frac{1}{2}\left\|z\right\|^2 + \frac{1}{2}\lambda\left\|w\right\|^2 \to \min_{w,z} \\ z_i = x_i^T w - y_i & n = \overline{1, N} \end{cases}$$

Решение - методом множителей Лагранжа. Выпишем Лагранжиан:

$$L = \frac{1}{2} z^T z + \frac{1}{2}\lambda w^T w + \sum_i \alpha_i \left(x_i^T w - y_i - z_i\right)$$

[4] Рассчитайте вычислительную сложность обучения и построения прогноза.

Решение

Найдём экстремум Лагранжиана:

$$L = \frac{1}{2}z^T z + \frac{1}{2}\lambda w^T w + \sum_i \alpha_i \left(x_i^T w - y_i - z_i \right)$$

$$\frac{\partial L}{\partial w} = \lambda w + \sum_i \alpha_i x_i = 0$$

$$\frac{\partial L}{\partial z_i} = z_i - \alpha_i = 0$$

Отсюда $z_i = \alpha_i$ и $w = -\frac{1}{\lambda}\sum_i \alpha_i x_i$. Подставим в Лагранжиан и получим двойственную задачу (относительно двойственных переменных α).

Двойственная задача

$$L = \frac{1}{2}\alpha^T\alpha + \frac{1}{2}\lambda\frac{1}{\lambda^2}\sum_{i,j}\alpha_i\alpha_j x_i^T x_j + \sum_i \alpha_i \sum_j \alpha_j x_j^T x_i \left(-\frac{1}{\lambda}\right) - \sum_i \alpha_i (y_i + \alpha_i)$$

$$= \frac{1}{2\lambda}\sum_{i,j}\alpha_i\alpha_j x_i^T x_j - \frac{1}{\lambda}\sum_{i,j}\alpha_i\alpha_j x_i^T x_j - \frac{1}{2}\alpha^T\alpha - \sum_i \alpha_i y_i$$

$$= -\frac{1}{2\lambda}\sum_{i,j}\alpha_i\alpha_j x_i^T x_j - \sum_i \alpha_i y_i - \frac{1}{2}\alpha^T\alpha \to extr_\alpha$$

Изменим знак (не влияющий на точку экстремума):

$$\frac{1}{2\lambda}\sum_{i,j}\alpha_i\alpha_j x_i^T x_j + \frac{1}{2}\alpha^T\alpha + \sum_i \alpha_i y_i \to extr_\alpha$$

Решение

Определим матрицу Грамма $M \in \mathbb{R}^{N \times N}$ с элементами $\{M\}_{i,j} = x_i^T x_j$. Тогда можем переписать задачу на экстремум в матричном виде:

$$Q = \frac{1}{2\lambda}\alpha^T M\alpha + \frac{1}{2}\alpha^T\alpha + \alpha^T y \to extr_\alpha$$

$$\frac{dQ}{d\alpha} = \frac{1}{\lambda}M\alpha + \alpha + y = 0,$$

что эквивалентно

$$\left(\frac{1}{\lambda}M + I\right)\alpha = -y \quad \Longrightarrow \quad \alpha = -\left(\frac{1}{\lambda}M + I\right)^{-1} y$$

Прогноз[5]:

$$\widehat{y}(x) = x^T w = -\frac{1}{\lambda}\sum_i \alpha_i x_i^T x$$

[5] Рассчитайте вычислительную сложность обучения и построения прогноза.

Ядерное обобщение

$$\alpha = -\left(\frac{1}{\lambda}M + I\right)^{-1} y, \qquad \{M\}_{i,j} = \langle x_i, x_j \rangle \to K(x_i, x_j)$$

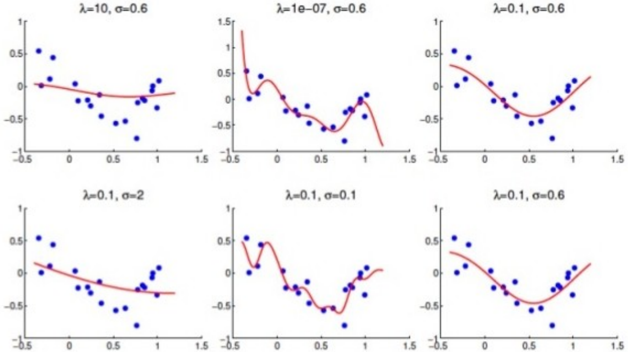
$$\widehat{y}(x) = \langle w, x \rangle = -\frac{1}{\lambda}\sum_i \alpha_i \langle x_i, x \rangle \to -\frac{1}{\lambda}\sum_i \alpha_i K(x_i, x)$$

Применим к Гауссову ядру:

$$K\left(x, x'\right) = e^{-\gamma\|x - x'\|_2^2} = e^{\|x - x'\|^2/\left(2\sigma^2\right)}$$

Ядерное обобщение[6]

$$\widehat{y}(x) = -\frac{1}{\lambda}\sum_i \alpha_i K(x_i, x), \qquad K\left(x, x'\right) = e^{\|x-x'\|^2/\left(2\sigma^2\right)}$$



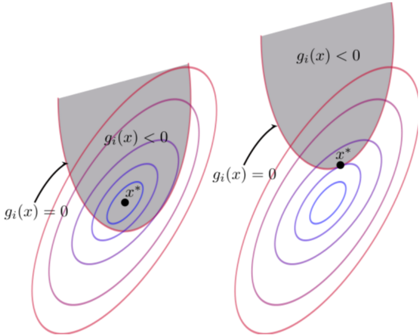
[6] Оцените влияние λ, σ на сложность модели.

Содержание

Условия Каруша-Куна-Таккера

Рассмотрим оптимизационную задачу:

$$\begin{cases} f(x) \to \min_x \\ g_m(x) \leq 0 \qquad m = \overline{1, M} \end{cases} \tag{1}$$



Необходимые условия оптимальности

Определим Лагранжиан

$$L(x,\lambda)=f(x)+\sum_{m=1}^{M}\lambda_m g_m(x)$$

Теорема (необходимые условия оптимальности):

- Пусть x^* - решение (1),
- $f(x^*)$ и $g_m(x^*)$, $m=1,2,...M$ - непрерывно-дифференцируемы в x^*.
- Выполнены условия регулярности Слейтера: $\exists x:\ g_m(x)<0\,\forall m$.

Тогда $\exists \lambda_1^*,\lambda_2^*,...\lambda_M^*$, что x^* удовлетворяет условию:

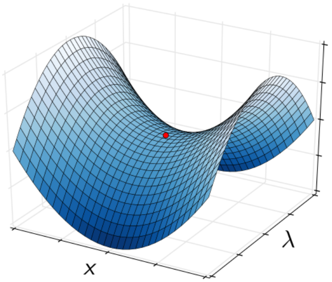
$$\begin{cases}\nabla_x f(x^*)+\sum_{i=1}^{M}\lambda_i^*\nabla_x g_i(x^*)=0 & \text{стационарность}\\ g_m(x^*)\le 0,\ \ m=\overline{1,M} & \text{достижимость}\\ \lambda_m^*\ge 0,\ \ m=\overline{1,M} & \text{неотрицательность}\\ \lambda_m^* g_m(x^*)=0,\ \ m=\overline{1,M} & \text{дополняющая нежесткость}\end{cases}$$

Условия Каруша-Куна-Таккера (KKT)

Предположим $f(x)$ и $g_m(x)$, $m = \overline{1, M}$ выпуклы. Тогда

1. Условия Каруша-Куна-Таккера (2) становятся **достаточными**, чтобы x^* было решением (1).
2. (x^*, λ^*) являются седловой точкой Лагранжиана:

$$L(x^*, \lambda) \leq L(x^*, \lambda^*) \leq L(x, \lambda^*) \quad \forall x \, \forall \lambda \in \mathbb{R}_+^M$$



Седловая точка

Двойственная задача

Из условия $\nabla_x L(x^*, \lambda^*) = 0$ можем найти $x^* = x(\lambda^*)$.
Поскольку (x^*, λ^*) - седловая точка $L(x, \lambda)$ можем найти λ^* из двойственной задачи:

$$\begin{cases} L(x(\lambda), \lambda) \to \max_\lambda & \\ g_m(x(\lambda)) \leq 0 & m = \overline{1, M} \\ \lambda_m \geq 0 & m = \overline{1, M} \\ \lambda_m g_m(x(\lambda)) = 0 & m = \overline{1, M} \end{cases}$$

В целом, выпуклость $f(x)$ и $g_m(x)$, $m = \overline{1, M}$ обеспечивает:

- все локальные минимумы являются глобальными
- множество минимумов выпукло
- если $f(x)$ - строго выпукла и минимум существует, то он единственный.

Содержание

Линейно неразделимый случай

$$\begin{cases} \frac{1}{2}w^T w + C\sum_{i=1}^N \xi_i \to \min_{w,w_0,\xi} \\ y_i(w^T x_i + w_0) = M(x_i, y_i) \ge 1 - \xi_i,\ i = 1, 2, ...N \\ \xi_i \ge 0,\ i = 1, 2, ...N \end{cases}$$

Лагранжиан:

$$L = \frac{1}{2}w^T w + C\sum_{i=1}^N \xi_i - \sum_{i=1}^N \alpha_i(y_i(w^T x_i + w_0) - 1 + \xi_i) - \sum_{i=1}^N \beta_i \xi_i$$

Условия ККТ:

$$\begin{cases} \frac{\partial L_P}{\partial w} = \overrightarrow{0},\ \frac{\partial L_P}{\partial w_0} = 0,\ \frac{\partial L_P}{\partial \xi_i} = 0 & \text{стационарность} \\ y_i(x_i^T w + w_0) \ge 1 - \xi_i,\ \ \xi_i \ge 0 & \text{достижимость} \\ \alpha_i \ge 0,\ \beta_i \ge 0 & \text{двойственные переменные} \ge 0 \\ \alpha_i(y_i(w^T x_i + w_0) - 1 + \xi_i) = 0 & \text{дополняющая нежесткость} \\ \beta_i \xi_i = 0, \quad i = 1, 2, ...N & \text{дополняющая нежесткость} \end{cases}$$

Решение условий ККТ

$$\frac{\partial L}{\partial w}=\overrightarrow{0}:\ w=\sum_{i=1}^{N}\alpha_i y_i x_i \tag{3}$$

$$\frac{\partial L}{\partial w_0}=0:\ \sum_{i=1}^{N}\alpha_i y_i=0$$

$$\frac{\partial L}{\partial \xi_i}=0:\ C-\alpha_i-\beta_i=0 \tag{4}$$

Подставляя эти ограничения в L, получим *двойственную задачу*[7]:

$$\begin{cases} L_D=\sum_{i=1}^{N}\alpha_i-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j x_i^T x_j\to\max_\alpha \\ \sum_{i=1}^{N}\alpha_i y_i=0 \\ 0\le\alpha_i\le C\quad(\text{используя (4) и } \alpha_i\ge 0,\ r_i\ge 0) \end{cases} \tag{5}$$

[7] Седловая точка лагранжиана, min для w, w_0, ξ_i и max для α_i, β_i.

Определение типа объектов

- **неинформативные объекты**: $y_i(w^T x_i + w_0) > 1 <=> \xi_i = 0$, $y_i(w^T x_i + w_0) - 1 + \xi_i > 0 => \alpha_i = 0$
    - опорные вектора SV будут имет $\alpha_i > 0$.
- **опорные объекты нарушители** $SV \setminus \tilde{SV}$: $y_i(w^T x_i + w_0) < 1 <=> \xi_i > 0 => \beta_i = 0 <=> \alpha_i = C$.
- **опорные пограничные объекты** $\widetilde{\mathcal{SV}}$: $y_i(w^T x_i + w_0) = 1 =>$
    - $\xi_i = 0 => \beta_i > 0 => \alpha_i < C$ в общем случае
    - $y_i(w^T x_i + w_0) - 1 + \xi_i = 0 => \alpha_i > 0$ в общем случае

В общем случае $\alpha_i \in (0, C)$, ($\alpha_i = 0, C$ - частный случай).

Решение

1. Решим (5), чтобы найти α_i^*
2. Используя (3) и условие $\alpha_i^* = 0$ для неинформативный объектов, получим оптимальную w

$$w = \sum_{i \in \mathcal{SV}} \alpha_i^* y_i x_i$$

3. w_0 может быть найдено из условия пограничности объекта:

$$y_i(x_i^T w + w_0) = 1, \forall i \in \widetilde{\mathcal{SV}} \quad (6)$$

Решение для w_0

Домножая (6) на y_i, получим

$$x_i^T w + w_0 = y_i \quad \forall i \in \widetilde{\mathcal{SV}} \tag{7}$$

Вычислительно более устойчивое решение: просуммируем 7 по всем $i \in \widetilde{\mathcal{SV}}$:

$$n_{\tilde{SV}} w_0 = \sum_{j \in \tilde{SV}} \left(y_j - x_j^T w \right) = \sum_{j \in \tilde{SV}} y_j - \sum_{j \in \tilde{SV}} x_j^T w,$$

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

Если нет пограничных объектов, можно найти w_0 одномерной оптимизацией.

Построение прогнозов

1. Решаем двойственную задачу, что найти α_i^*, $i = 1, 2, \dots N$

$$\begin{cases} \sum_{i=1}^N \alpha_i - \frac{1}{2}\sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle \to \max_\alpha \\ \sum_{i=1}^N \alpha_i y_i = 0 \\ 0 \le \alpha_i \le C \quad (\text{используя (4) и } \alpha_i \ge 0,\ r_i \ge 0) \end{cases}$$

2. Находим w_0:

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

3. Строим прогноз для нового x:

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

Построение прогнозов

1. Решаем двойственную задачу, что найти α_i^*, $i = 1, 2, \dots N$

$$\begin{cases} L_D = \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle \to \max_\alpha \\ \sum_{i=1}^N \alpha_i y_i = 0 \\ 0 \le \alpha_i \le C \quad \text{(используя (4) и } \alpha_i \ge 0,\ r_i \ge 0) \end{cases}$$

2. Находим w_0:

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

3. Строим прогноз для нового x:

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

- На всех этапах нам нужно знать не x, а скалярные произведения $\langle x, x' \rangle$!

Обобщение через ядра

1. Решаем двойственную задачу, что найти α_i^*, $i = 1, 2, \dots N$

$$\begin{cases} L_D = \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j K(x_i, x_j) \to \max_\alpha \\ \sum_{i=1}^N \alpha_i y_i = 0 \\ 0 \le \alpha_i \le C \quad \text{(используя (4) и } \alpha_i \ge 0,\ r_i \ge 0) \end{cases}$$

2. Находим w_0:

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

3. Строим прогноз для нового x:

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

- Заменили $\langle x, x' \rangle \to K(x, x')$ для $K(x, x') = \langle \phi(x), \phi(x') \rangle$ с некоторым преобразованием признаков $\phi(\cdot)$.

Обобщение

Ядерно-обобщенный метод x:

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

$K(x, z) = \langle \phi(x), \phi(z) \rangle$ - ядро Мерсера.

Ядро	$K(x, z)$
линейное	$\langle x, z \rangle$
полиномиальное	$(a\langle x, z \rangle + b)^d, \; a > 0, \; b \geq 0, \; d = 1, 2, \ldots$
RBF (Гауссово)	$e^{-\gamma \lVert x - z \rVert^2}, \; \gamma > 0$

Содержание

6. Визуализация работы SVM с ядрами
   - SVM - линейное ядро
   - SVM - полиномиальное ядро
   - SVM - Гауссово ядро

Параметр C

Условная оптимизация:

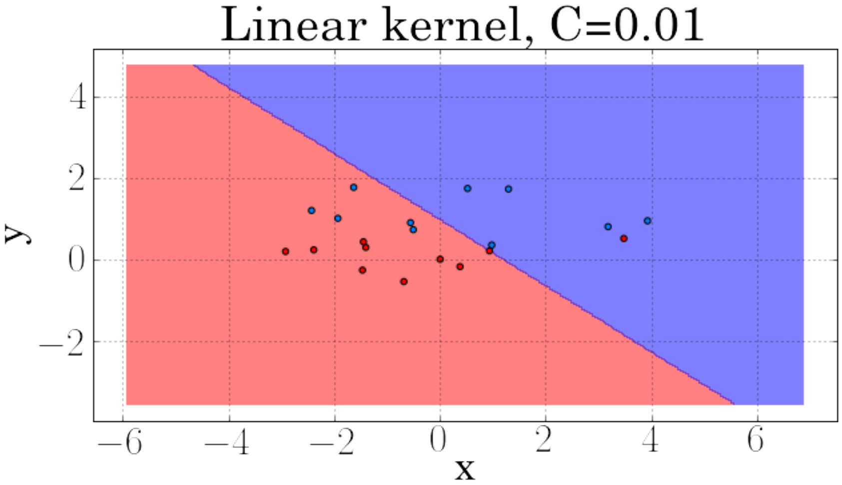
$$\begin{cases} \frac{1}{2}w^T w + C\sum_{i=1}^{N}\xi_i \to \min_{w,w_0,\xi} \\ y_i(w^T x_i + w_0) = M(x_i, y_i) \geq 1 - \xi_i,\ i = 1, 2, \ldots N \\ \xi_i \geq 0,\ i = 1, 2, \ldots N \end{cases}$$

Безусловная оптимизация:

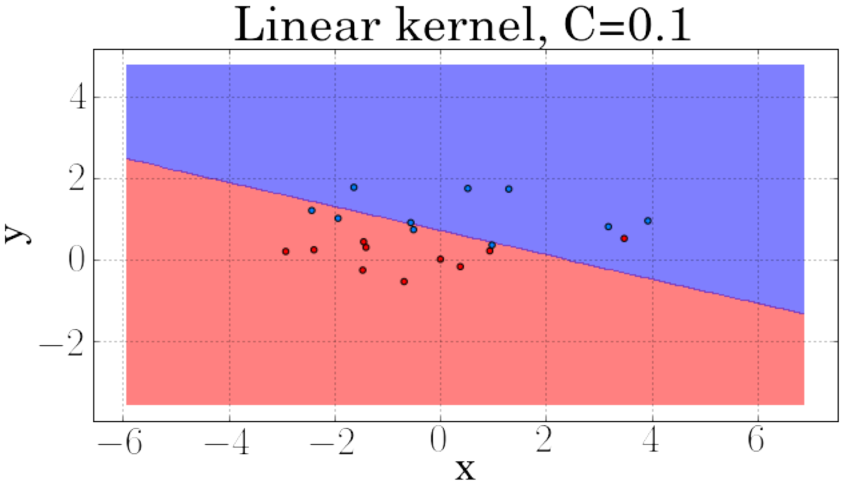
$$\frac{1}{2C}\|w\|_2^2 + \sum_{i=1}^{N}[1 - M_i(w, w_0)]_+ \to \min_{w,w_0}$$

Параметр C контролирует противоречие: простота$\leftrightarrow$точность.

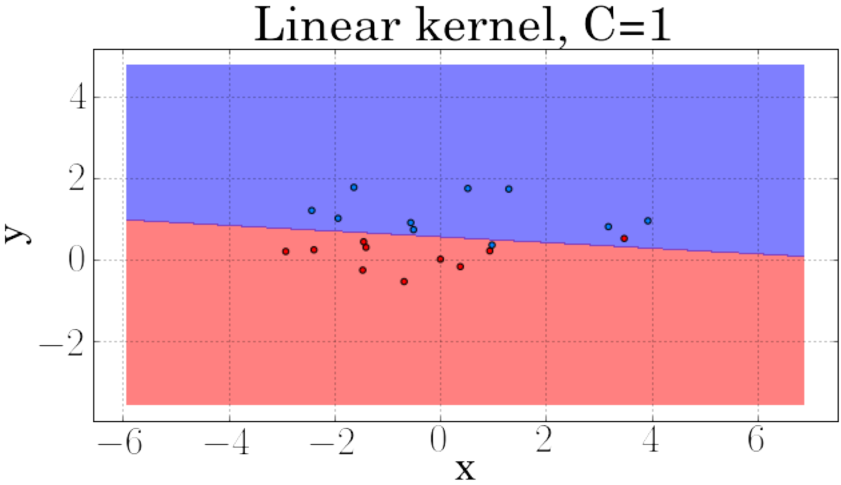
Линейное ядро, влияние C



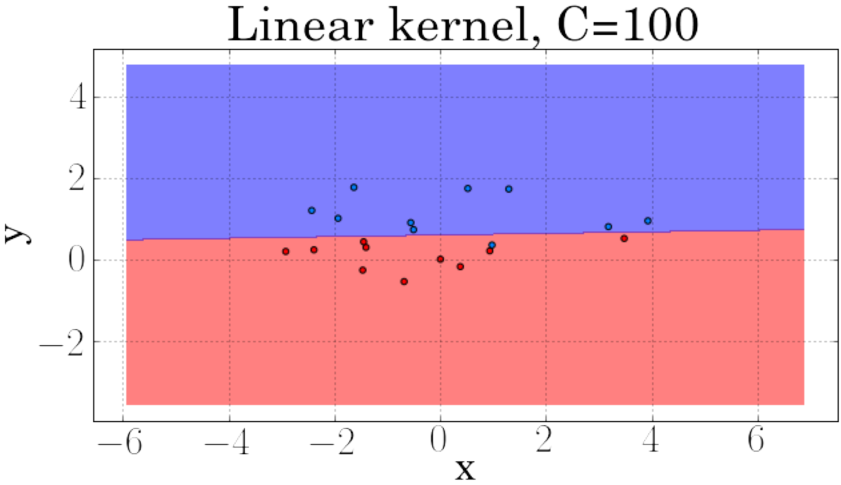
Линейное ядро, влияние C



Линейное ядро, влияние C



Линейное ядро, влияние C



6. Визуализация работы SVM с ядрами
   - SVM - линейное ядро
   - **SVM - полиномиальное ядро**
   - SVM - Гауссово ядро

Полиномиальное ядро

Полиномиальное ядро:

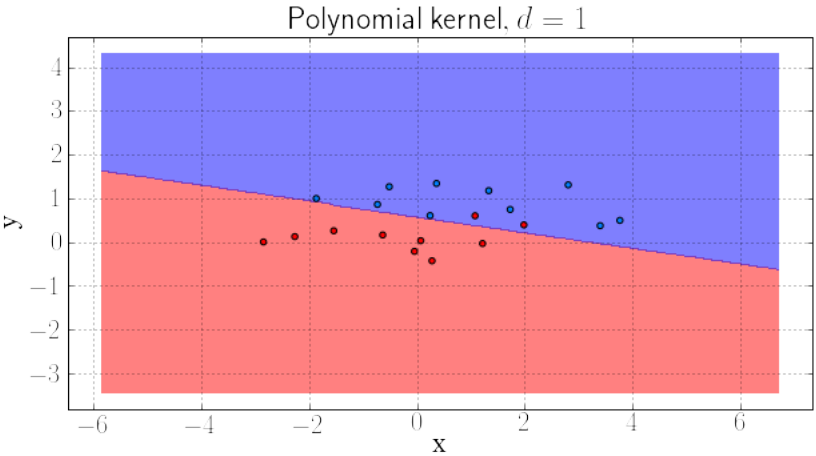
$$K(x,z) = (a\langle x,z\rangle + b)^{d}, \;\; a>0,\;\; b\geq 0,\;\; d=1,2,...$$

Прогноз

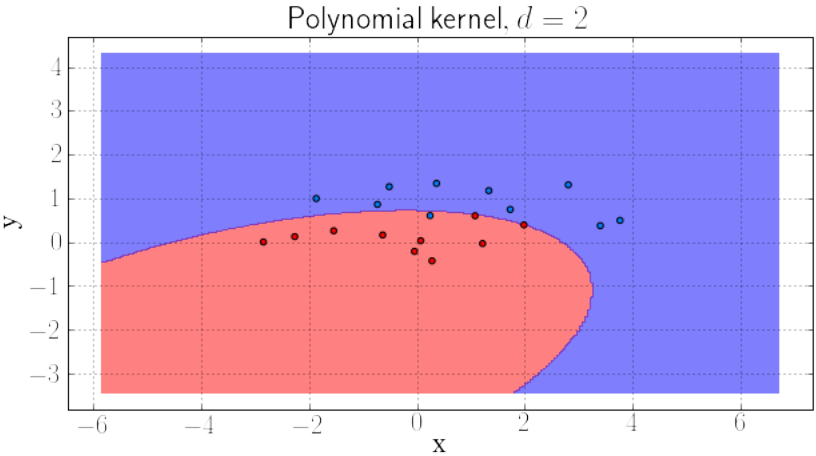
$$\widehat{y}(x) = \text{sign}\left(\sum_{i\in\mathcal{SV}} \alpha_i^* y_i K(x_i,x) + w_0\right) =$$
$$= \text{sign}\left(\sum_{i\in\mathcal{SV}} \alpha_i^* y_i \left(a\langle x,x_i\rangle + b\right)^{d} + w_0\right)$$

Граница между классами - полиномиальная поверхность порядка d.

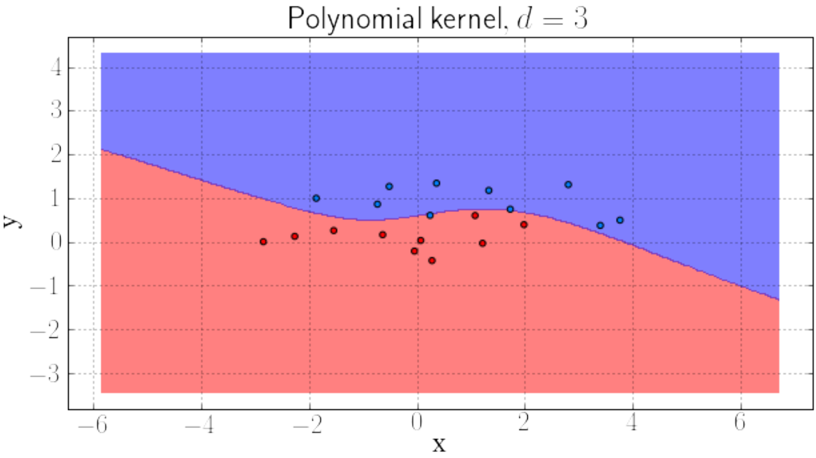
Полиномиальное ядро, влияние d



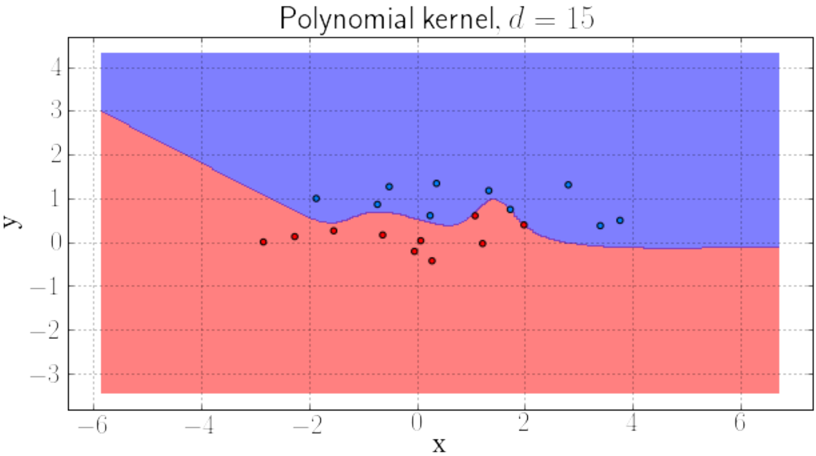
Полиномиальное ядро, влияние d



Полиномиальное ядро, влияние d



Полиномиальное ядро, влияние d



6. Визуализация работы SVM с ядрами
   - SVM - линейное ядро
   - SVM - полиномиальное ядро
   - SVM - Гауссово ядро

Гауссово ядро

Гауссово ядро:

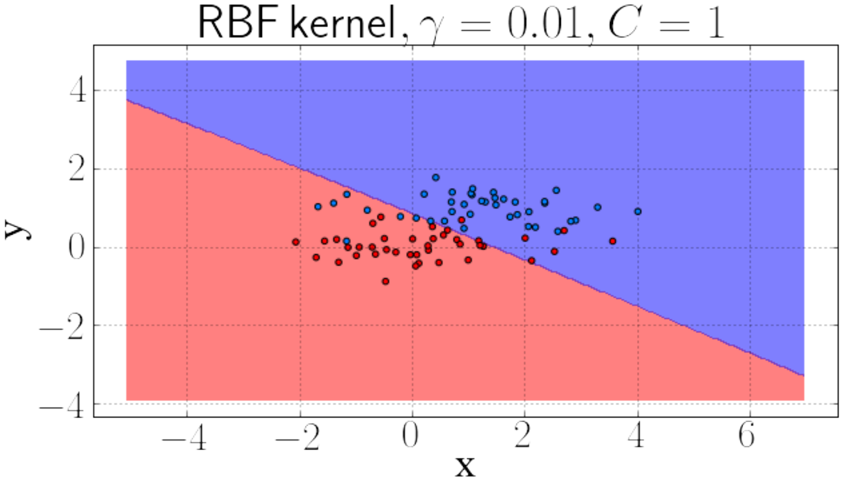
$$K(x, z) = e^{-\gamma \|x-z\|^2}, \ \gamma > 0$$

Прогноз

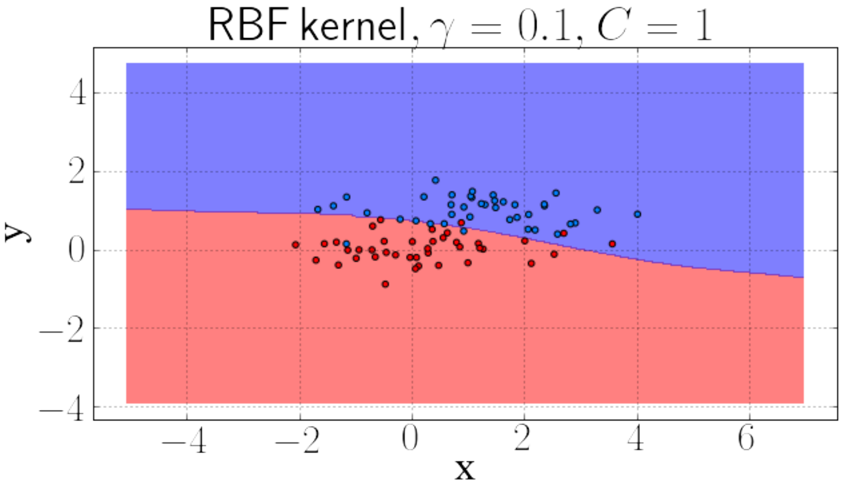
$$\begin{aligned}\widehat{y}(x) &= \text{sign}\left(\sum_{i \in \mathcal{SV}} \alpha_i^* y_i K(x_i, x) + w_0\right) \\ &= \text{sign}\left(\sum_{i \in \mathcal{SV}} \alpha_i^* y_i e^{-\gamma \|x-x_i\|^2} + w_0\right)\end{aligned}$$

Классификация на основе близости x к опорным объектам в весами α_i^* (их важность).

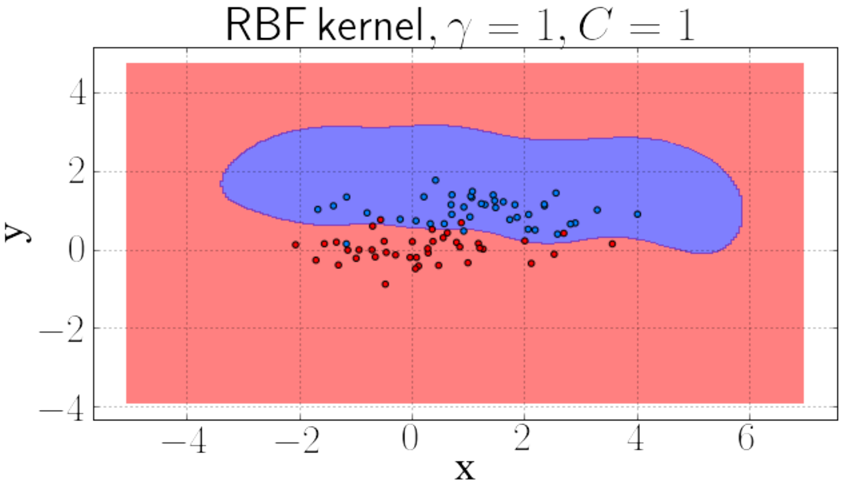
Гауссово ядро, влияние γ



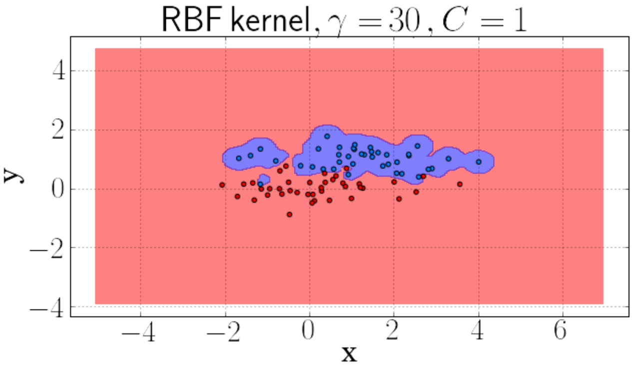
Гауссово ядро, влияние γ



Гауссово ядро, влияние γ



Гауссово ядро, влияние γ



Использование SVM с Гауссовым ядром

```
...
from sklearn.svm import SVC
from sklearn.metrics import accuracy_score

X_train, X_test, Y_train, Y_test =
    get_demo_classification_data()

model = SVC(kernel='rbf', gamma=1)    # инициализация
model.fit(X_train, Y_train)           # обучение модели
Y_hat = model.predict(X_test)         # построение
    прогнозов
print(f'Точность прогнозов: \
{100*accuracy_score(Y_test, Y_hat):.1f}%')
```

- 1/C - вес при регуляризаторе.
- Больше информации. Полный код.

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