Ядерные методы

Виктор Китов
v.v.kitov@yandex.ru

МГУ им.Ломоносова, ф-т ВМиК, кафедра ММП.

I семестр 2015 г.

Содержание

Ridge регрессия

- Ridge регрессия - оптимизируемый критерий:

$$Q(\beta)=\sum_{n=1}^{N}\left(x_n^T\beta-y_n\right)^2+\lambda\sum_{d=1}^{D}\beta_d^2\to\min_{\beta}$$

- Условие стационарности:

$$\frac{dQ(\beta)}{d\beta}=2\sum_{n=1}^{N}\left(x_n^T\beta-y_n\right)x_n+2\lambda\beta=0 \tag{1}$$

- В векторной форме:

$$X^T\left(X\beta-Y\right)+\lambda\beta=0$$

Ridge регрессия

- Решение прямой задачи:

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

- Комментарий: $X^T X \succcurlyeq 0$ (неотрицательно определена), а $X^T X + \lambda I \succ 0$ (положительно определена), поэтому параметры ridge регрессии всегда однозначно идентифицируются.
- Сложность оценивания:
    - $X^T X + \lambda I$: $ND^2 + D$
    - $X^T Y$: DN
    - $(X^T X + \lambda I)^{-1}$: D^3
    - $(X^T X + \lambda I)^{-1} X^T Y$: D^2
    - Результирующая сложность оценивания: $O(ND^2 + D^3) = O(D^2(N + D))$.

Двойственное решение

Из векторная записи (1):

$$X^T(X\beta - Y) + \lambda\beta = 0$$

следует *двойственное решение* для вектора коэффициентов (линейная комбинация обучающих векторов):

$$\beta = \frac{1}{\lambda}X^T(Y - X\beta) = X^T\alpha \tag{2}$$

где

$$\alpha = \frac{1}{\lambda}(Y - X\beta) \tag{3}$$

называются *двойственными переменными*.

Прогнозирование:

$$\widehat{y}(x) = x^T\beta = x^T X^T\alpha = \sum_{i=1}^{N}\alpha_i\langle x, x_i\rangle$$

Двойственное решение

Для нахождения α подставим (2) в (3):

$$\alpha = \frac{1}{\lambda}(Y - X\beta) = \frac{1}{\lambda}(Y - XX^T\alpha)$$

$$\left(XX^T + \lambda I\right)\alpha = Y$$

$$\alpha = \left(XX^T + \lambda I\right)^{-1} Y$$

Сложность оценивания модели :
$XX^T + \lambda I$: $N^2D + N$
$\left(XX^T + \lambda I\right)^{-1}$: N^3
$\left(XX^T + \lambda I\right)^{-1} Y$: N^2
Итоговая сложность обучения: $O(N^2D + N^3) = O(N^2(D + N))$.
Сложность прогнозирования $\widehat{y}(x) = \langle x, \beta\rangle = \sum_{i=1}^{N} \alpha_i \langle x, x_i\rangle$: ND.

Преимущества двойственного решения

- Оптимальные α зависят только от скалярных произведений векторов (а не от полных признаковых представлений объектов):
$$\alpha = \left(XX^T + \lambda I\right)^{-1} Y = (G + \lambda I)^{-1} Y$$
где $G \in \mathbb{R}^{NxN}$ и $\{G\}_{ij} = \langle x_i, x_j \rangle$ - G называется *матрицей Грамма*.
- Прогноз также зависит только от скалярных произведений:
$$\widehat{y}(x) = \sum_{i=1}^{N} \alpha_i \langle x, x_i \rangle = \alpha^T v$$
где $v \in \mathbb{R}^N$ и $v_i = \langle x, x_i \rangle$.

Преимущества

- Оценка модели становится вычислительно проще при $D > N$
  - можно переходить в признаковые пространства большой размерности, даже бесконечномерные.

Преимущество двойственного решения

Не нужно представление объекта в том признаковом пространстве, в котором работает модель - в том пространстве нужна только возможность вычислять скалярные произведения для любой пары объектов.

Содержание

Kernel trick

Kernel trick

Определить не полное признаковое описание x, а только функцию вычисления скалярного произведения $K(x, x')$

- $\langle x, x' \rangle$ имеет сложность $O(D)$, а сложность вычисления $K(x, x')$ может быть меньше.

Комментарии

Kernel trick применим не только к ridge регрессии, но и к:

- методу K ближайших соседей
- K-средних, K-medoids (кластеризация)
- nearest medoid
- метод главных компонент (снижение размерности)
- метод опорных векторов
- многие другие

Когда исходное пространство объектов $\mathcal{X} = \mathbb{R}^M$, то всегда можно определить *линейное ядро*:

$$K(x, x') = \langle x, x' \rangle = \sum_{d=1}^{D} x_d x'_d$$

Характерные случаи применимости kernel trick

- признаковое пространство высокой размерности
  - все полиномы степени до M
  - Гауссово ядро $K(x,x') = e^{-\frac{1}{2\sigma^2}\left\|x-x'\right\|^2}$ соответствует признаковому пространству бесконечной размерности.
- сложно представить объекты векторами фиксированной длины
  - строки, множества, картинки, тексты, графы, 3D-структуры и т.д.
- существует естественные определения скалярного произведения
  - строки: число совместно встречающихся подстрок
  - множества: размер общего подмножества
    - пример: для множеств S_1 и S_2: $K(S_1,S_2) = 2^{|S_1 \cap S_2|}$ -ядро.
- скалярное произведение может быть эффективно посчитано

Преимущества kernel trick

- обобщение линейных методов на нелинейный случай
    - с сохранением вычислительной эффективности линейных методов
    - с сохранением преимуществ линейных методов
        - локальный оптимум является глобальным
        - нет локальных оптимумов=>меньше переобучение
- объекты для которых не существует векторных представлений фиксированной длины
- ускоренное вычисление скалярных произведений для высоких значений D

Содержание

Определение ядра

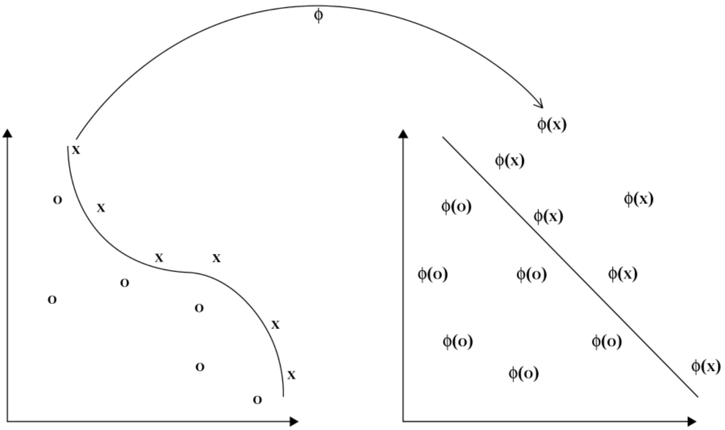
- $x \in \mathcal{X}$ заменяется признаковым описанием $\phi(x)$ размерности D.
  - $\mathcal{X}$ - пространство объектов (первичное представление, не обязательно $\mathbb{R}^M$)
  - Пример: $[x] \to [x, x^2, x^3]$

Ядро (Kernel)

Функция $K(x, x') : X \times X \to \mathbb{R}$ является ядром, если найдется такое признаковое описание $\phi(x) : \mathcal{X} \to X$ что $K(x, x') = \langle \phi(x), \phi(x') \rangle$.

- $\langle x, x' \rangle$ определяется как $\langle \phi(x), \phi(x') \rangle = K(x, x')$
- Частные разновидности ядер:
  - $K(x, x') = K(x - x')$ - стационарные ядра (инвариантны к параллельным переносам)
  - $K(x, x') = K(\|x - x'\|)$ - радиальные базисные функции

Иллюстрация



Полиномиальное ядро

- Пример 1: пусть $D=2$.

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

где $\phi(x) = (x_1^2, x_2^2, \sqrt{2} x_1 x_2)$

- Пример 2: пусть $D=2$.

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

где $\phi(x) = (1,\ x_1^2,\ x_2^2,\ \sqrt{2} x_1,\ \sqrt{2} x_2,\ \sqrt{2} x_1 x_2)$

- В общем, для $D \geq 1$ $K(x,z) = (x^T z)^M$ соответствует расширению признакового пространства до всех мономов степени M, а $(1 + x^T z)^M$ - до всех мономов степени не выше M.

Свойства ядер

Теорема (Мерсер, упрощенная формулировка): Функция $K(x, x')$ является ядром тогда и только тогда, когда

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

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

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

Построение новых ядер

- Обучение ядер (kernel learning) - отдельная область.
- Сложно доказывать неотрицательную определенность $K(x, x')$ в каждом новом случае.
- Новые ядра обычно строят из известных ядер, применяя преобразования сохраняющие свойство ядра:
- Известные ядра:
    - обычное скалярное определение $\langle x, x' \rangle$
    - константа $K(x, x') \equiv 1$
    - $x^T A x$ для любой $A \succcurlyeq 0$

Преобразования, сохраняющие свойство ядра

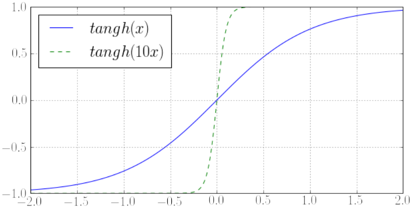
Если $K_1(x,x')$, $K_2(x,x')$ - произвольные ядра, $c>0$ - константа, $q(\cdot)$ - полином с неотрицательными коэффициентами, $h(x)$ и $\varphi(x)$ - произвольные функции, отображающие $\mathcal{X}\to\mathbb{R}$ и $\mathcal{X}\to\mathbb{R}^M$ соответственно, то ядрами также являются:

1. $K(x,x')=cK_1(x,x')$
2. $K(x,x')=K_1(x,x')K_2(x,x')$
3. $K(x,x')=K_1(x,x')+K_2(x,x')$
4. $K(x,x')=K_1(\varphi(x),\varphi(x'))$
5. $K(x,x')=h(x)K_1(x,x')h(x')$
6. $K(x,x')=e^{K_1(x,x')}$

Часто используемые функции $K(x, x')$

$K(x, x')$	**определение**
линейная	$\langle x, x' \rangle$
полиномиальная	$(\gamma \langle x, x' \rangle + r)^d$
RBF	$\exp(-\gamma \left\| x - x' \right\|^2)$
сигмоидальное	$\mathrm{tangh}(\gamma \langle x, y \rangle + r)$

- Линейная, полиномиальная и RBF функция являются ядрами Мерсера, а сигмоидальная - нет.



Дополнение

- Другие алгоритмы, допускающие обобщение через ядра: K-ближайших соседей, K-средних, K-medoids, nearest medoid, метод главных компонент, метод опорных векторов и многие другие..
- Определение расстояния через ядро:

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

- Скалярное определение нормализованных векторов:

$$\left\langle \frac{\phi(x)}{\|\phi(x)\|}, \frac{\phi(x')}{\|\phi(x')\|}\right\rangle = \frac{\langle \phi(x), \phi(x')\rangle}{\sqrt{\langle \phi(x),\phi(x)\rangle}, \sqrt{\langle \phi(x'),\phi(x')\rangle}} = \frac{K(x,x')}{\sqrt{K(x,x)K(x',x')}}$$

Содержание

Линейный метод опорных векторов

- Решение для весовых коэффициентов:

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

- Дискриминантная функция

$$g(x) = \sum_{i \in \mathcal{SV}} \alpha_i y_i < x_i, x > + w_0$$

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

где $SV = \{i : y_i(x_i^T w + w_0 \leq 1)\}$ - индексы всех опорных векторов, а $\tilde{SV} = \{i : y_i(x_i^T w + w_0 = 1\}$ - индексы граничных опорных векторов.

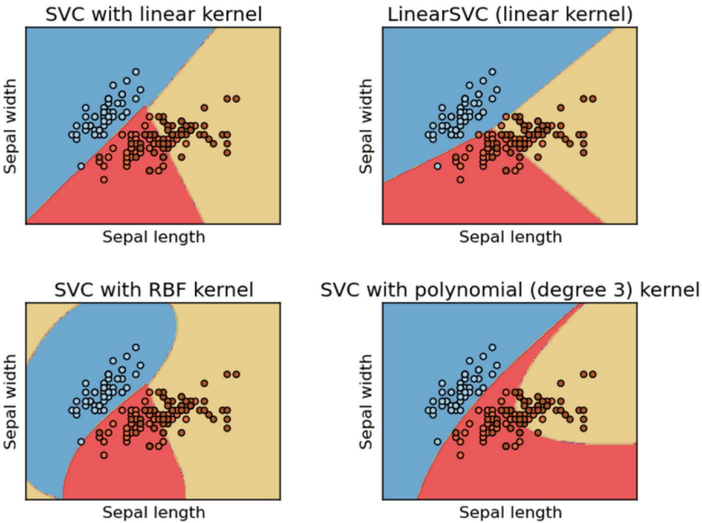
Ядерное обобщение

Дискриминантная функция:

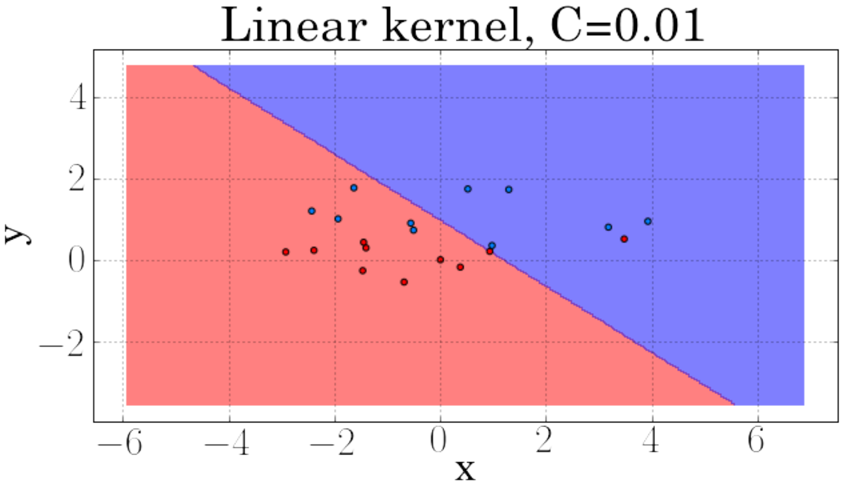
$$g(x) = \sum_{i \in \mathcal{SV}} \alpha_i y_i K(x_i, x) + w_0$$

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

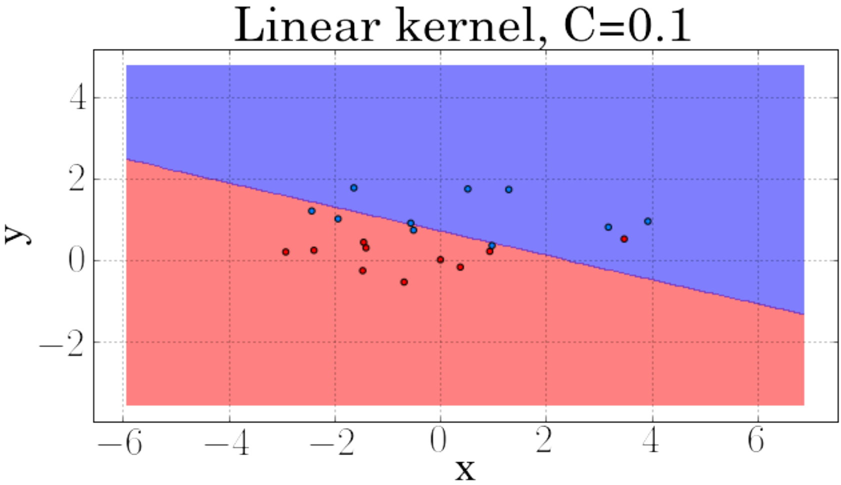
Области классов для ядерного обобщения



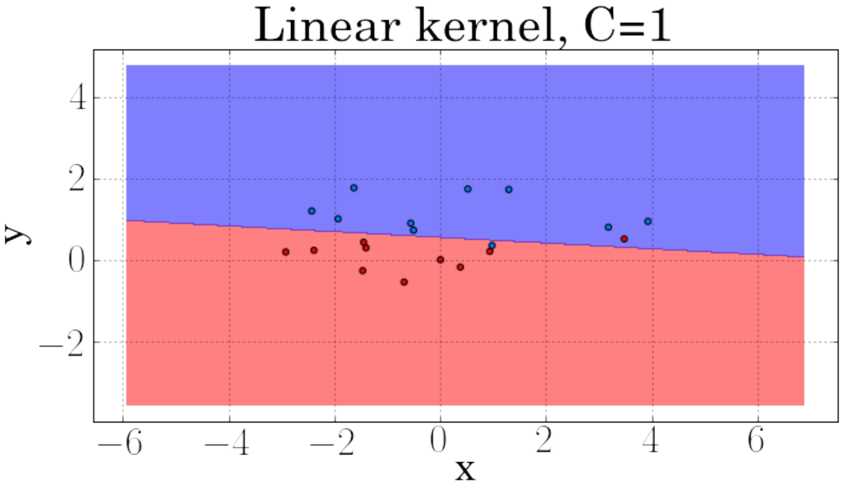
Линейное ядро - изменяемое C



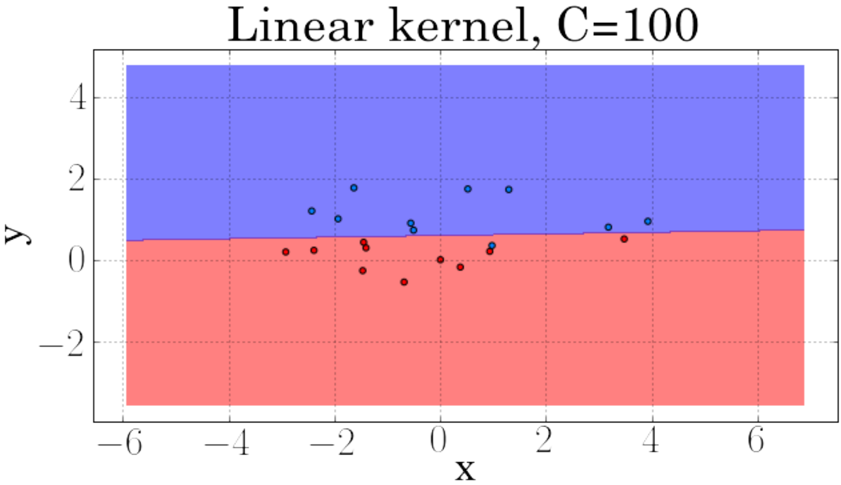
Линейное ядро - изменяемое C



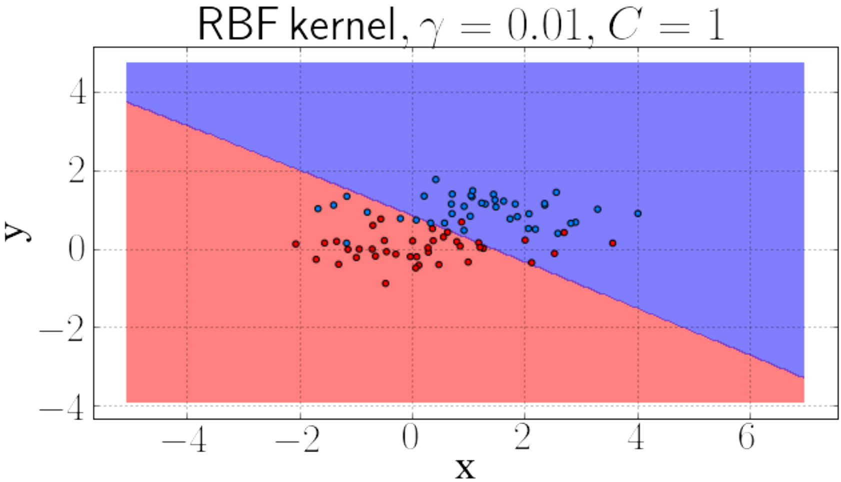
Линейное ядро - изменяемое C



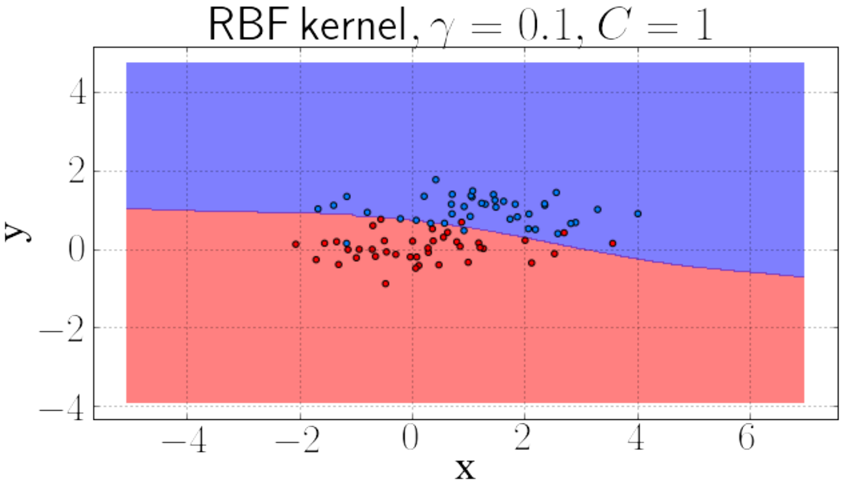
Линейное ядро - изменяемое C



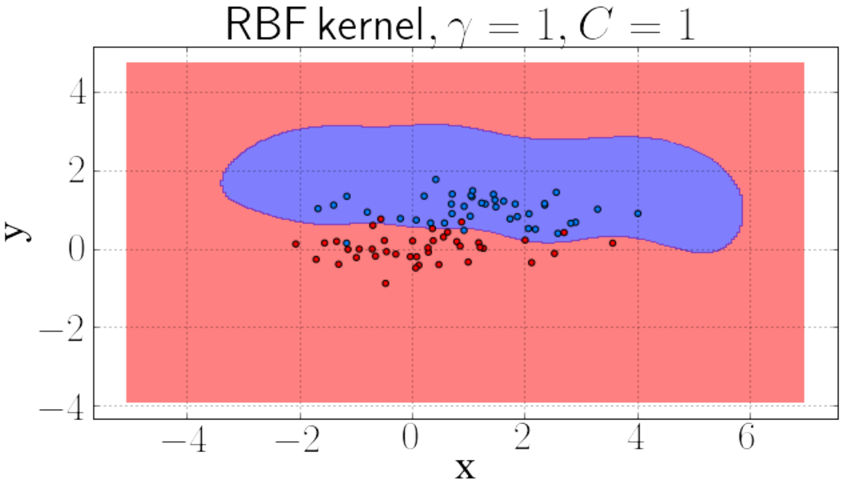
RBF ядро - изменяемое γ



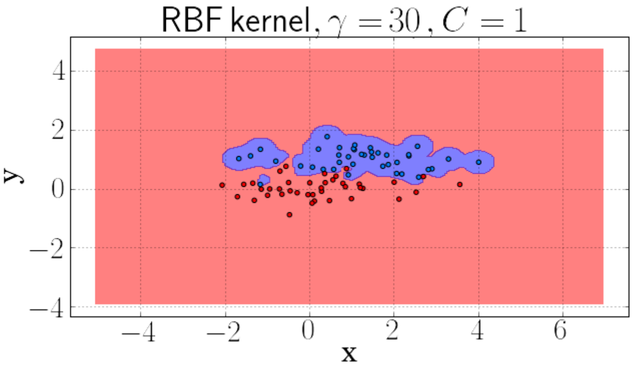
RBF ядро - изменяемое γ



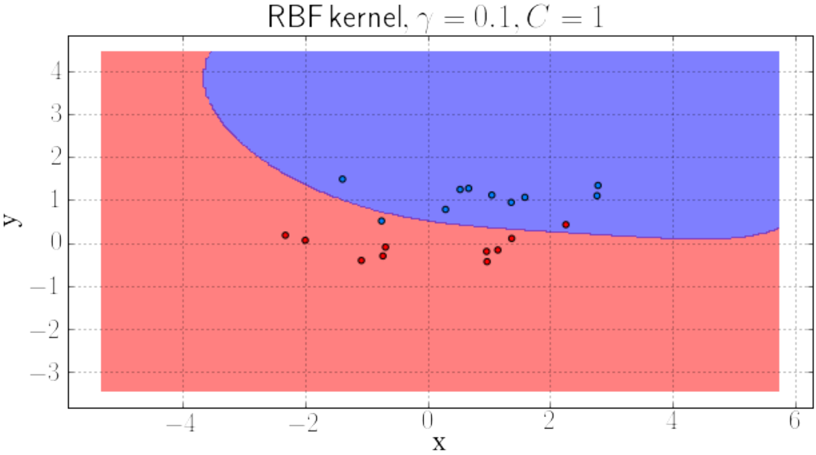
RBF ядро - изменяемое γ



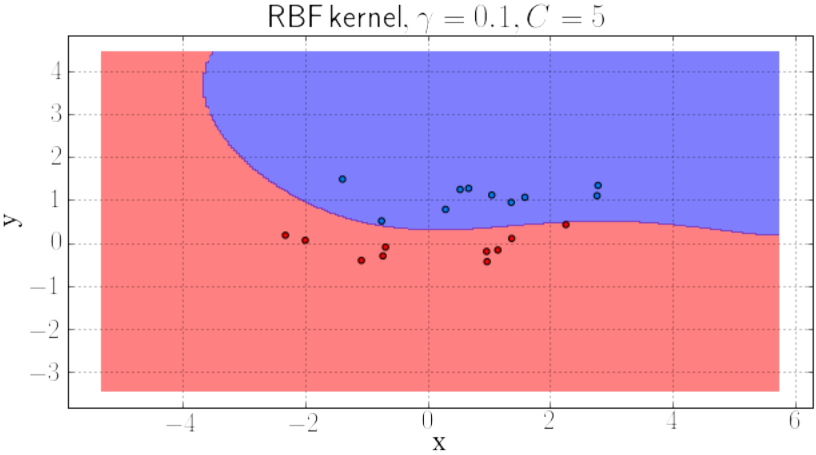
RBF ядро - изменяемое γ



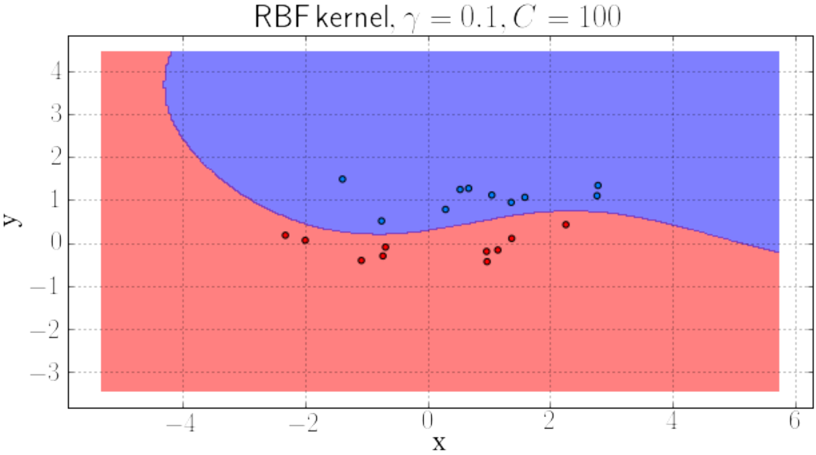
RBF ядро - изменяемое C



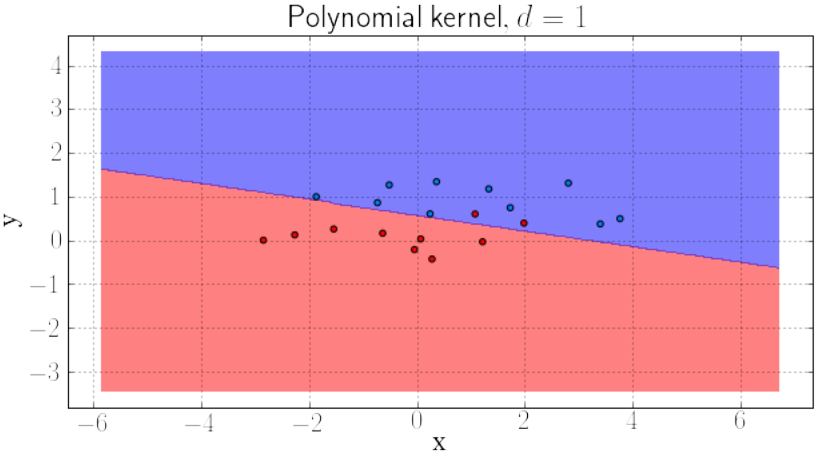
RBF ядро - изменяемое C



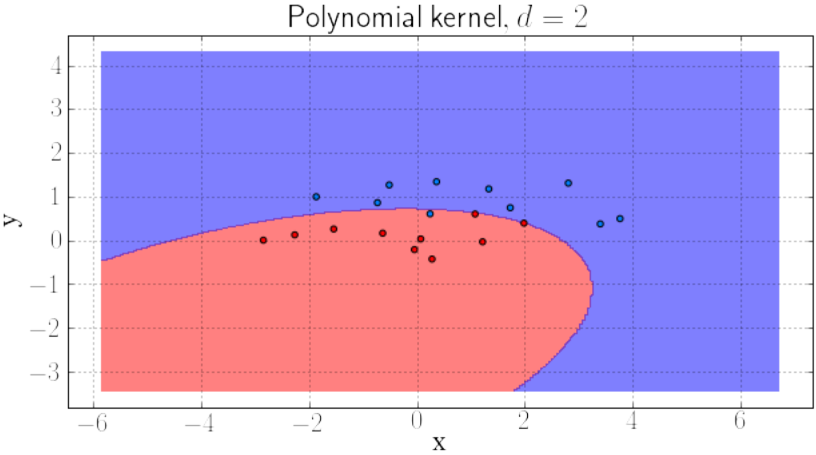
RBF ядро - изменяемое C



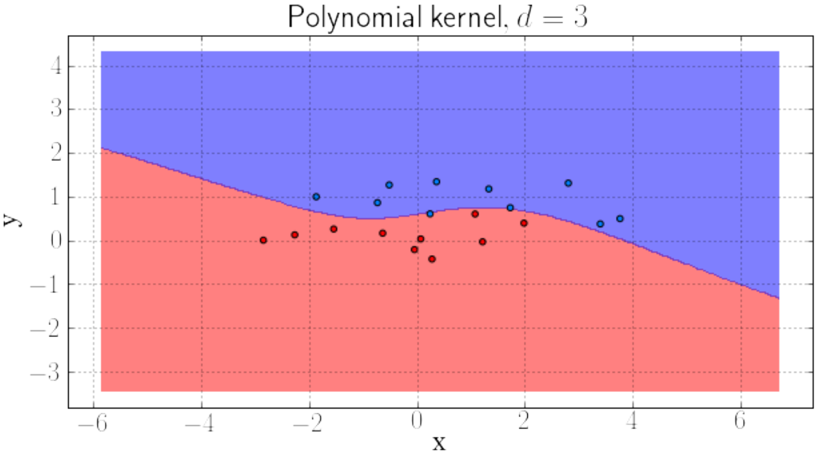
Полиномиальное ядро - изменяемое d



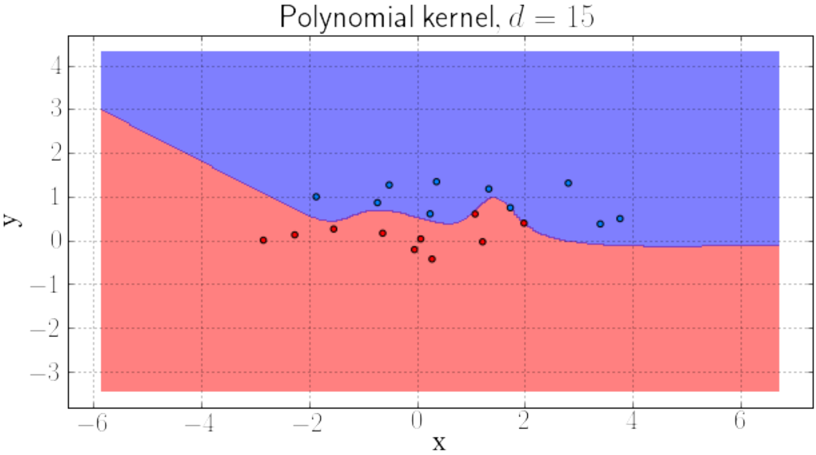
Полиномиальное ядро - изменяемое d



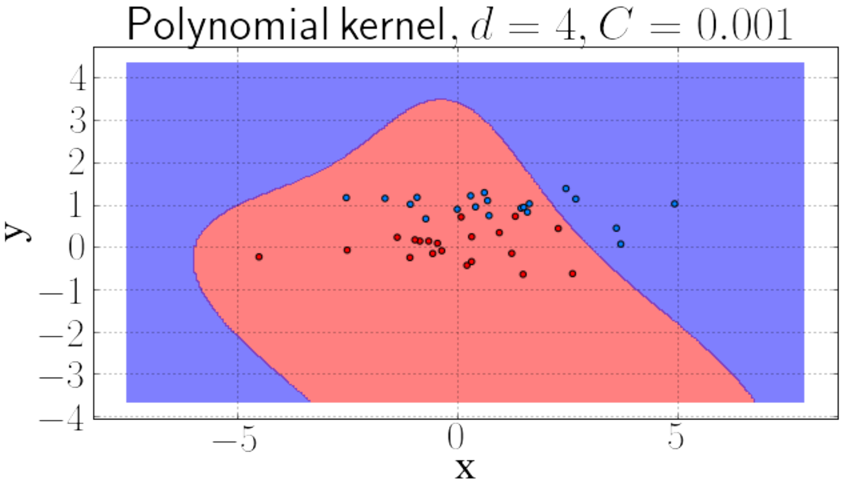
Полиномиальное ядро - изменяемое d



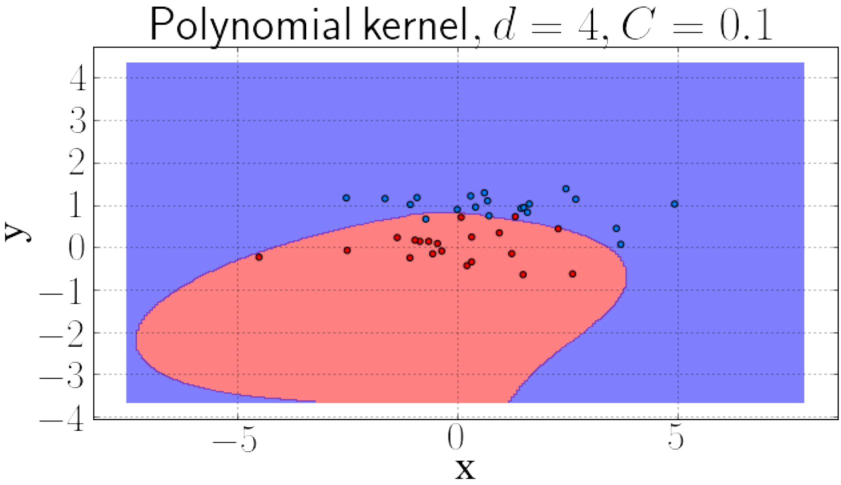
Полиномиальное ядро - изменяемое d



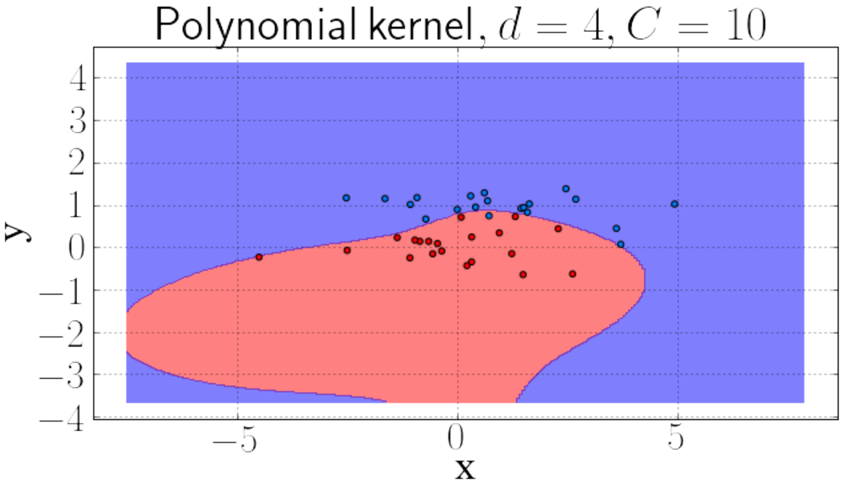
Полиномиальное ядро - изменяемое C



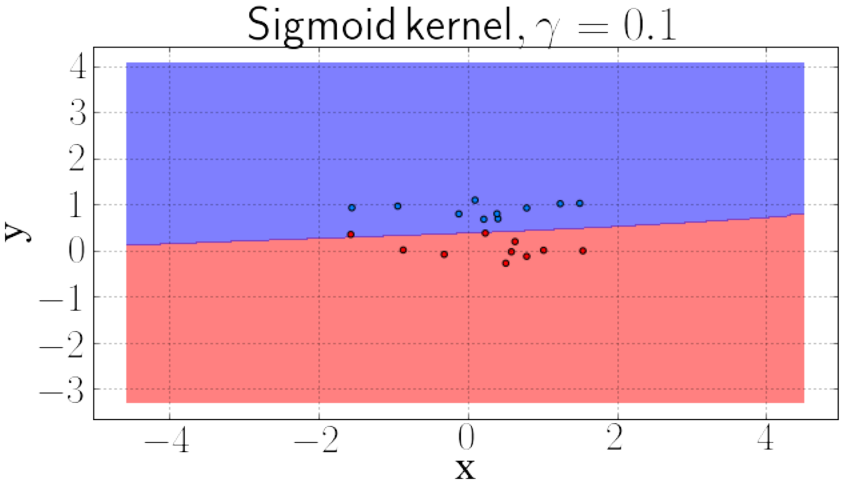
Полиномиальное ядро - изменяемое C



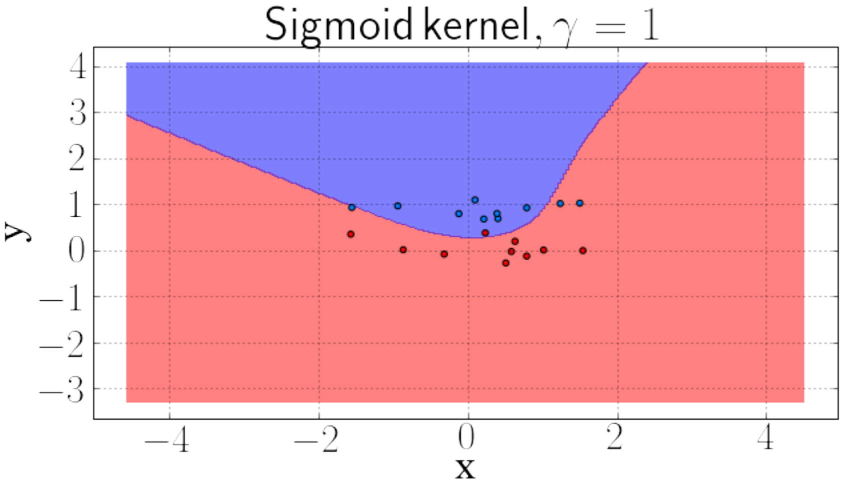
Полиномиальное ядро - изменяемое C



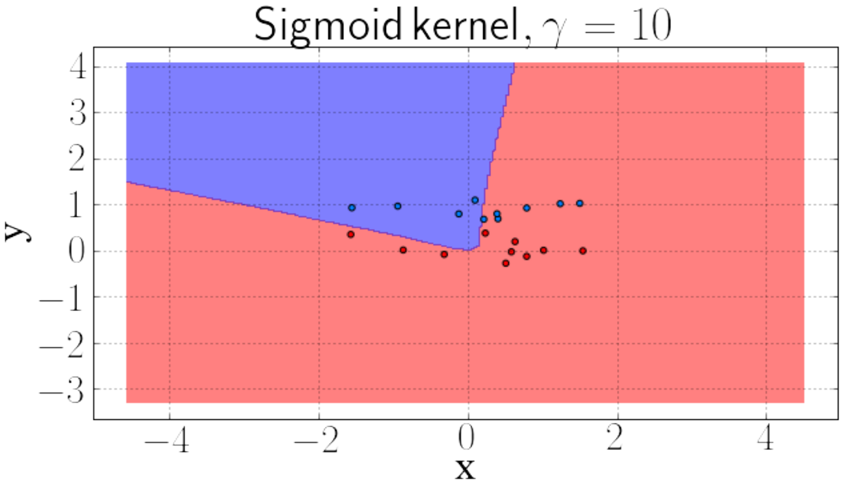
Сигмоидальное ядро - изменяемое γ



Сигмоидальное ядро - изменяемое γ



Сигмоидальное ядро - изменяемое γ



Сигмоидальное ядро - изменяемое C