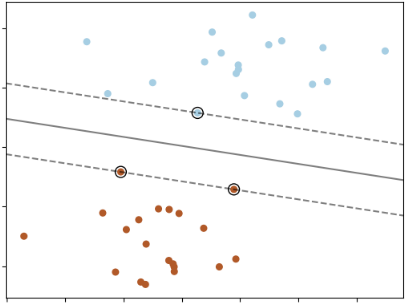
Метод опорных векторов

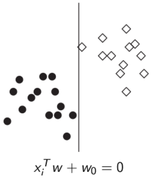
Виктор Китов

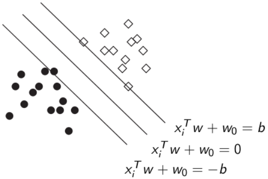
v.v.kitov@yandex.ru



Содержание

Метод опорных векторов





Рассмотрим бинарную классификацию $y \in \{+1, -1\}$ линейно разделимой выборки.

Идея метода опорных векторов (support vector machines, SVM)

Выберем гиперплоскость, разделяющую классы с максимальным зазором.

Гиперплоскости $x_i^T w + w_0 = 0$, $x_i^T w + w_0 = b$, $x_i^T w + w_0 = -b$ поэтому величина зазора $\frac{2b}{\|w\|}$.

Метод опорных векторов

Объекты (x_i, y_i) отделены от разделяющей гиперплоскости $\geq \frac{b}{\|w\|}$, если

$$\begin{cases} x_i^T w + w_0 \geq b, & y_i = +1 \\ x_i^T w + w_0 \leq -b & y_i = -1 \end{cases} \qquad i = 1, 2, \ldots N.$$

Это можно записать в виде

$$y_i(x_i^T w + w_0) \geq b, \quad i = 1, 2, \ldots N.$$

Максимизация зазора между классами:

$$2b/\left\|w\right\| \to \max_{w, w_0, b}$$

Оптимизационная задача

Оптимизационная задача:

$$\begin{cases} \frac{2b}{\|w\|} \to \max\limits_{w,w_0,b} \\ y_i(x_i^T w + w_0) \ge b, \quad i = 1, 2, ...N. \end{cases}$$

Если (w, w_0, b)-решение, то $(\alpha w, \alpha w_0, \alpha b)$ - тоже решение $\forall \alpha > 0$. Положим $b = 1$ $(\alpha = \frac{1}{b})$.

$$\begin{cases} \frac{2}{\|w\|} \to \max\limits_{w,w_0} \\ y_i(x_i^T w + w_0) \ge 1 \quad i = 1, 2, ...N. \end{cases}$$

Используя свойство $\arg\max \frac{2}{\|w\|} = \arg\min \frac{\|w\|}{2} = \arg\min \frac{\|w\|^2}{2}$:

$$\begin{cases} \frac{1}{2} w^T w \to \min\limits_{w,w_0} \\ y_i(x_i^T w + w_0) = M(x_i, y_i) \ge 1, \quad i = 1, 2, ...N. \end{cases}$$

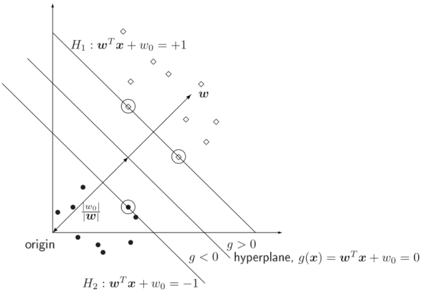
Типы объектов

Неинформативные объекты: $y_i(x_i^T w + w_0) > 1$

- не влияют на решение

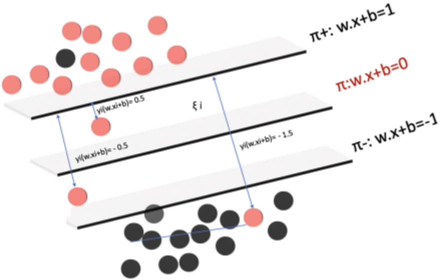
Опорные вектора: $y_i(x_i^T w + w_0) = 1$

- лежат на расстоянии $1/\|w\|$ к разделяющей гиперплоскости
- влияют на решение



Содержание

Линейно неразделимый случай



$$\begin{cases} \frac{1}{2} w^T w \to \min\limits_{w, w_0} \\ y_i(x_i^T w + w_0) = M(x_i, y_i) \geq 1, \quad i = 1, 2, ... N. \end{cases}$$

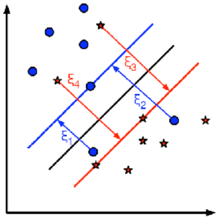
Ограничения становятся несовместными=>пустое множество решений.

Линейно неразделимый случай

Разрешим частичные нарушения ограничений на величины нарушений ξ_i (slack variables):

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

- Штраф за нарушение C контролирует точность модели (в противовес простоте).
- Подбирается по сетке на валидации.
- Другие штрафы возможны, например $C\sum_i \xi_i^2$.



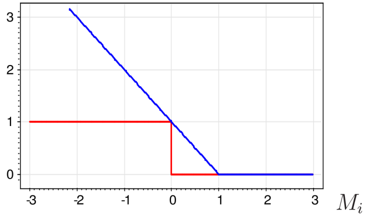
Типы объектов

- **Неинформативные объекты**:
  - $y_i(w^T x_i + w_0) > 1$
- **Опорные вектора** SV**:**
  - $y_i(w^T x_i + w_0) \leq 1$
  - **пограничные** $\widetilde{SV}$:
    - $y_i(w^T x_i + w_0) = 1$
  - **объекты-нарушители**:
    - $y_i(w^T x_i + w_0) > 0$: нарушитель корректно классифицирован
    - $y_i(w^T x_i + w_0) < 0$: нарушитель некорректно классифицирован

Безусловная оптимизация

Оптимизационная задача:

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



может быть переписана как

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

Таким образом, метод - линейный классификатор с функцией потерь $\mathcal{L}(M) = [1 - M]_+$ и L_2 регуляризацией (обобщается на другие).

Разреженность решения

- Решение зависит только от опорных векторов.
- Это видно из условия $\mathcal{L}(M) = 0$ для $M \geq 1$.
    - хорошо классифицированные объекты с $M \geq 1$ не влияют на решение
- Разреженность решения - метод менее устойчив к выбросам
    - выбросы - всегда опорные объекты

Многоклассовый метод опорных векторов

C дискриминантных ф-ций строятся одновременно:

$$g_c(x) = (\mathbf{w}^c)^T x + w_0^c, \qquad c = \overline{1, C}.$$

Линейно разделимый случай:

$$\begin{cases} \sum_{c=1}^{C} (\mathbf{w}^c)^T \mathbf{w}^c \to \min_{\mathbf{w}} \\ (\mathbf{w}^{y_n})^T x_n + w_0^{y_n} - (\mathbf{w}^c)^T x - w_0^c \geq 1 \quad \forall c \neq y_n, \\ n = \overline{1, N}. \end{cases}$$

Линейно неразделимый случай:

$$\begin{cases} \sum_{c=1}^{C} (\mathbf{w}^c)^T \mathbf{w}^c + C \sum_{n=1}^{N} \xi_n \to \min_{w} \\ (\mathbf{w}^{y_n})^T x + w_0^{y_n} - (\mathbf{w}^c)^T x - w_0^c \geq 1 - \xi_n \quad \forall c \neq y_n, \\ \xi_n \geq 0, \quad n = \overline{1, N}. \end{cases}$$

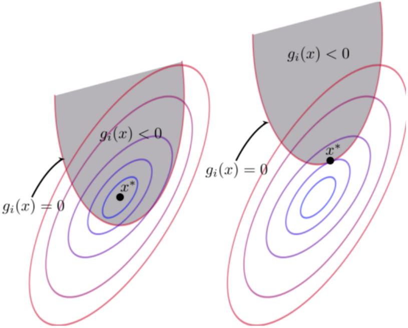
Настраивается медленнее, по точности сравним с бинарным обобщением через один-против-всех и один-против-одного.

Содержание

Условия Каруша-Куна-Таккера

Рассмотрим оптимизационную задачу:

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



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:

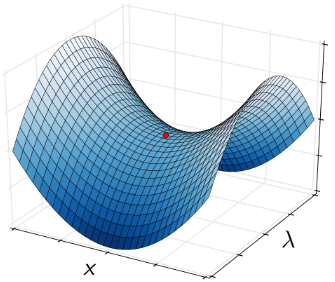
$$\begin{cases} \nabla_x f(x^*) + \sum_{i=1}^{M} \lambda_i^* \nabla_x g_i(x^*) = 0 & \text{стационарность} \\ g_m(x^*) \le 0, \;\; m = \overline{1,M} & \text{достижимость} \\ \lambda_m^* \ge 0, \;\; m = \overline{1,M} & \text{неотрицательность} \\ \lambda_m^* g_m(x^*) = 0, \;\; m = \overline{1,M} & \text{дополняющая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

Условия Каруша-Куна-Таккера (ККТ)

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:

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

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

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

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:

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

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*[1]:

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

[1] 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:

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

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:

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

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:

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

- 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^*, $i = 1, 2, \ldots N$

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

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

5 Визуализация работы SVM с ядрами
- SVM - линейное ядро
- SVM - полиномиальное ядро
- SVM - Гауссово ядро

Параметр C

Условная оптимизация:

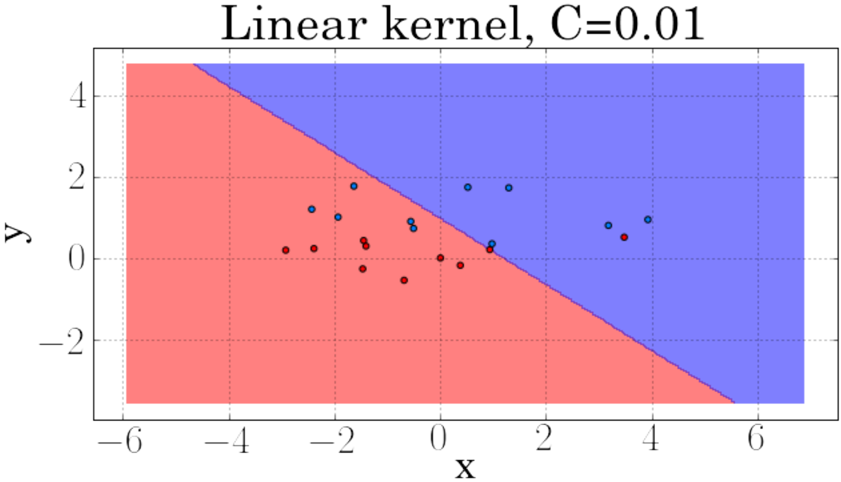
$$\begin{cases} \frac{1}{2}w^T w + C\sum_{i=1}^{N}\xi_i \to \min_{w,w_0,\xi} \\ y_i(w^T x_i + w_0) = M(x_i, y_i) \geq 1 - \xi_i, \ i = 1, 2, \dots N \\ \xi_i \geq 0, \ i = 1, 2, \dots N \end{cases}$$

Безусловная оптимизация:

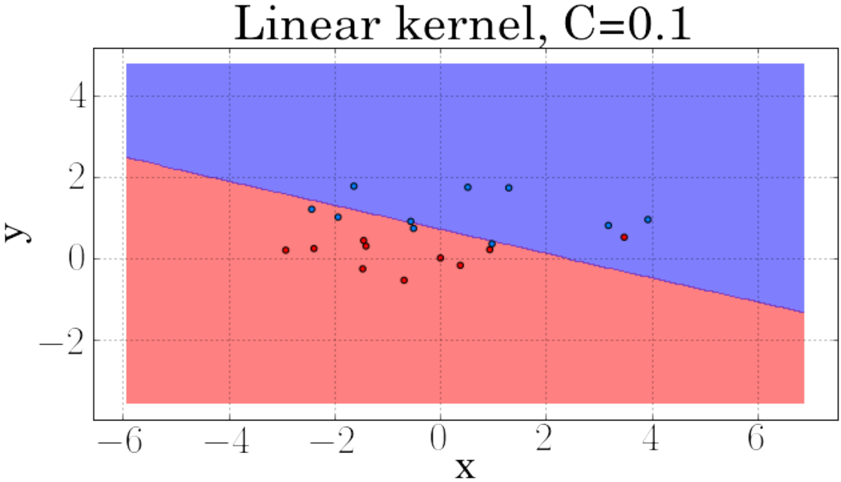
$$\frac{1}{2C}\|w\|_2^2 + \sum_{i=1}^{N}[1 - M_i(w, w_0)]_+ \to \min_{w,w_0}$$

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

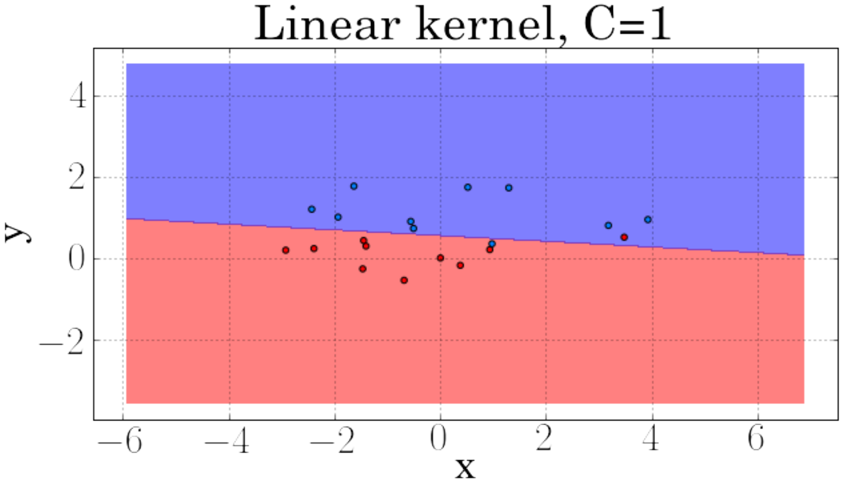
Линейное ядро, влияние C



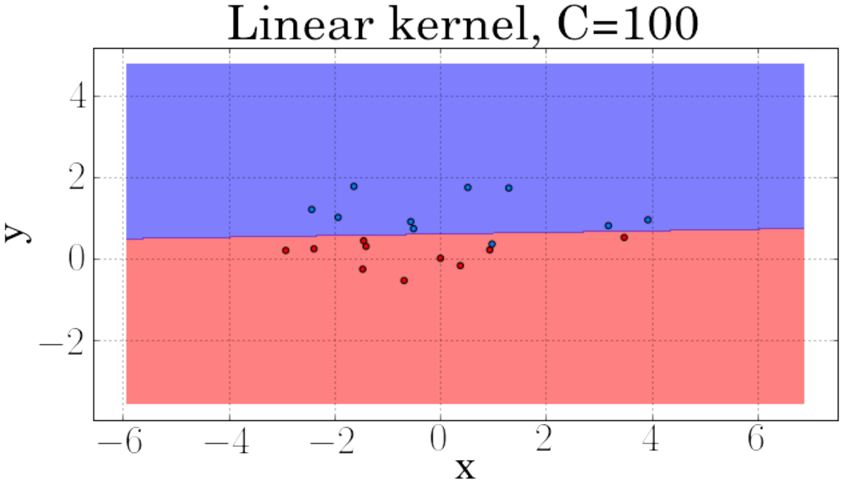
Линейное ядро, влияние C



Линейное ядро, влияние C



Линейное ядро, влияние C



5 Визуализация работы SVM с ядрами

- SVM - линейное ядро
- SVM - полиномиальное ядро
- SVM - Гауссово ядро

Полиномиальное ядро

Полиномиальное ядро:

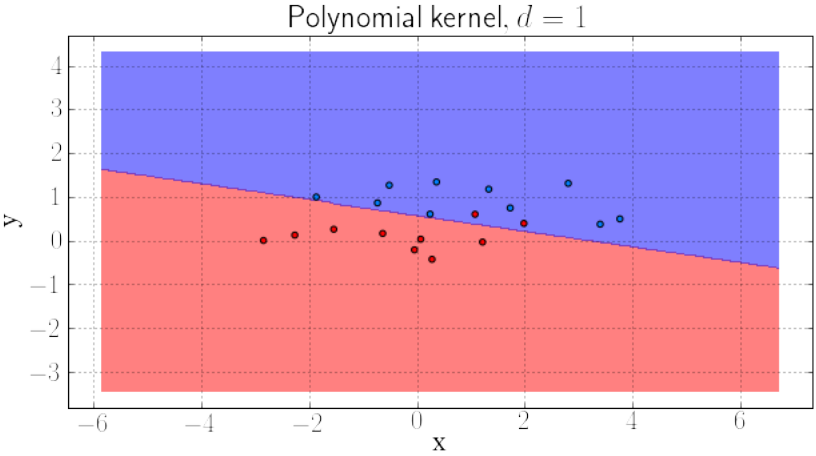
$$K(x,z)=\left(a\langle x,z\rangle+b\right)^{d},\;\; a>0,\;\; b\geq 0,\;\; d=1,2,...$$

Прогноз

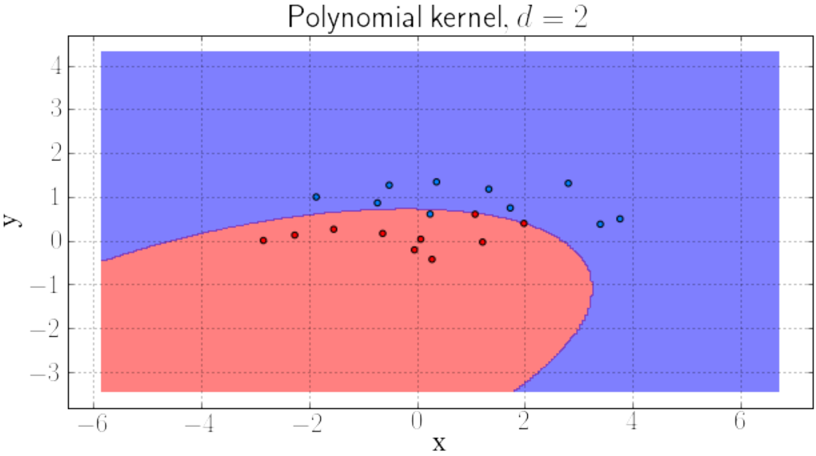
$$\widehat{y}(x)=\operatorname{sign}\left(\sum_{i\in\mathcal{SV}}\alpha_i^* y_i K(x_i,x)+w_0\right)=$$
$$=\operatorname{sign}\left(\sum_{i\in\mathcal{SV}}\alpha_i^* y_i\left(a\langle x,x_i\rangle+b\right)^{d}+w_0\right)$$

Граница между классами - полиномиальная поверхность порядка d.

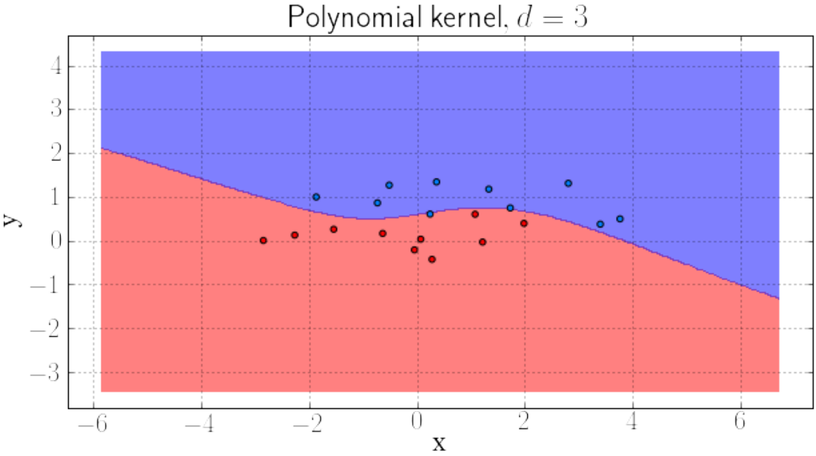
Полиномиальное ядро, влияние d



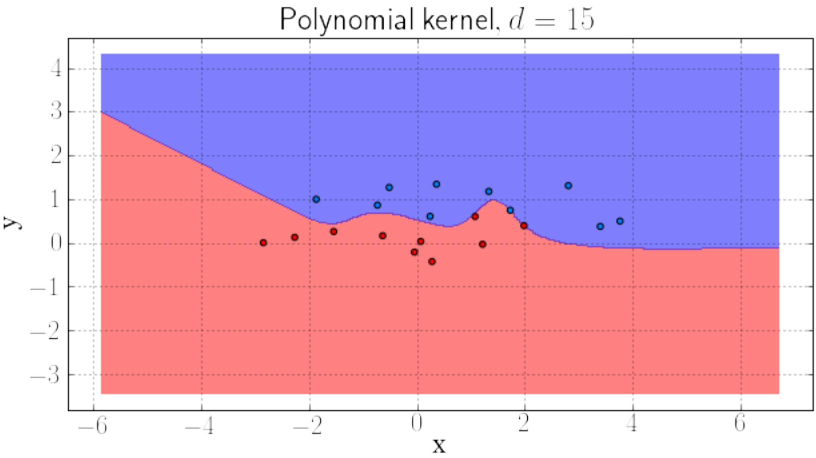
Полиномиальное ядро, влияние d



Полиномиальное ядро, влияние d



Полиномиальное ядро, влияние d



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

Гауссово ядро

Гауссово ядро:

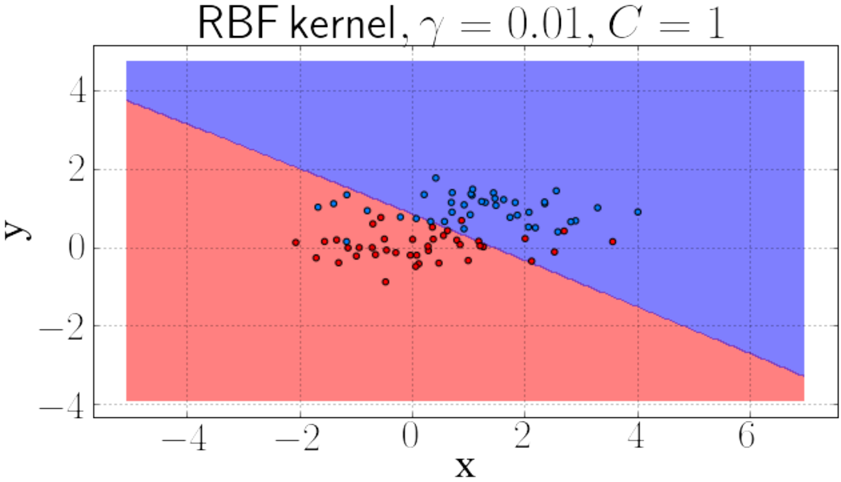
$$K(x,z)=e^{-\gamma\|x-z\|^2},\ \gamma>0$$

Прогноз

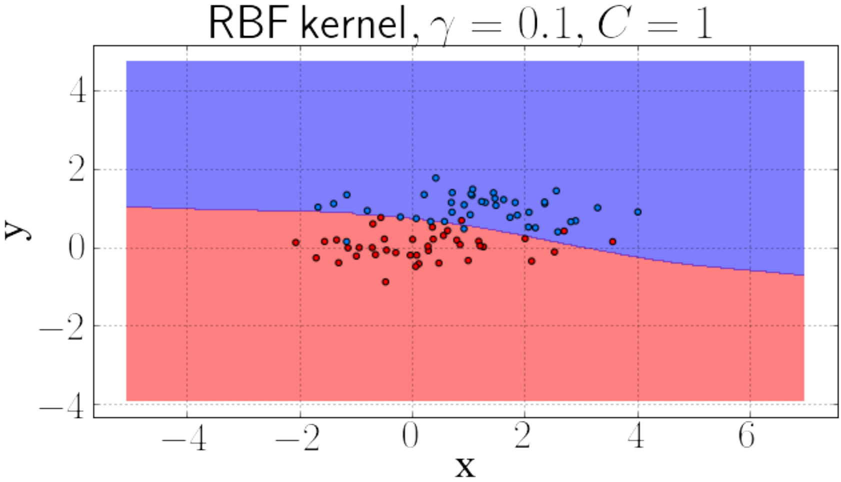
$$\begin{aligned}\widehat{y}(x)&=\text{sign}\left(\sum_{i\in\mathcal{SV}}\alpha_i^*y_iK(x_i,x)+w_0\right)\\&=\text{sign}\left(\sum_{i\in\mathcal{SV}}\alpha_i^*y_ie^{-\gamma\|x-x_i\|^2}+w_0\right)\end{aligned}$$

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

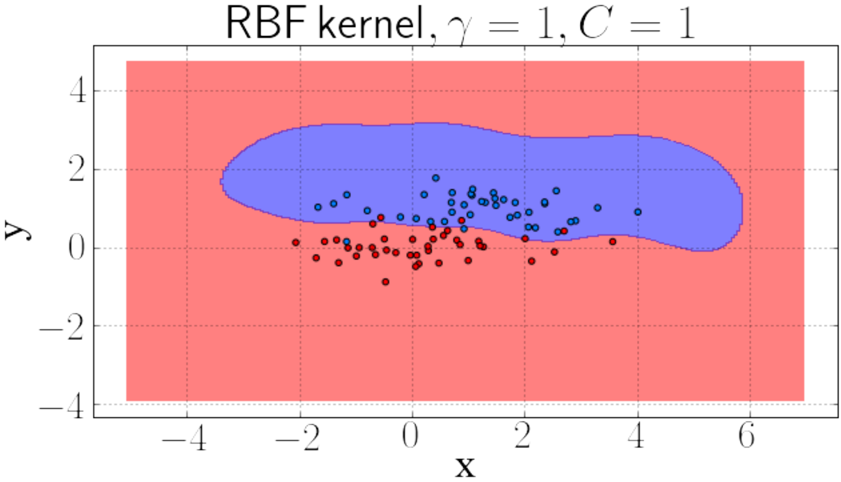
Гауссово ядро, влияние γ



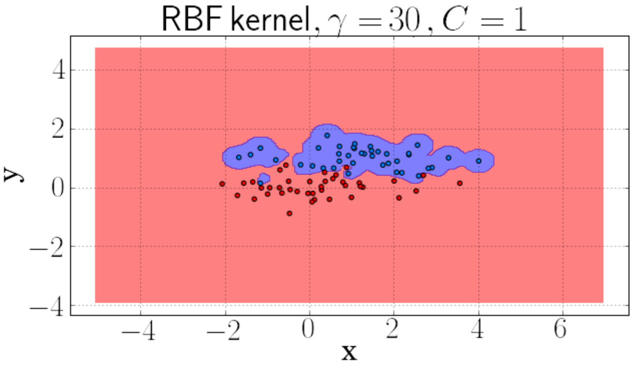
Гауссово ядро, влияние γ



Гауссово ядро, влияние γ



Гауссово ядро, влияние γ



Заключение

- Метод опорных векторов - линейный классификатор с L_2 регуляризацией и функцией потерь hinge.
- Геометрически метод максимизирует зазор между классами.
- Решение зависит только от опорных векторов с $M \leq 1$.
- Решение через двойственную задачу зависит не от x, а от $\langle x_i, x_j \rangle$
  - допускает обобщение $\langle x_i, x_j \rangle \longrightarrow K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$
  - линейное ядро - без обобщения
  - полиномиальное ядро - полиномиальная граница
  - Гауссово ядро - метрический метод