MML minor #1

Метод опорных векторов. Ядра.

Повторение: задача классификации

$D = \{(x_i, y_i)\}_{i=1}^{N}$ — набор данных

$x_i \in \mathbb{R}^n$ — признаки

$y_i \in \{+1, -1\}$ — ответы

$a(x; \mu)$ — алгоритм классификации

μ — параметры алгоритма

$$L(D, \mu) = \frac{1}{N} \sum_{i=1}^{N} [a(x_i, \mu) \neq y_i]$$ — ошибка алгоритма

$$\hat{\mu} = \arg\min_{\mu} L(D, \mu)$$ — обучение алгоритма

Однозначно ли выбирается $\hat{\mu}$?

Повторение: линейный классификатор

$$h(x) = w^T x + b = w_1 x_1 + w_2 x_2 + \ldots + w_n x_n + b$$

$h(x) = 0$ — гиперплоскость

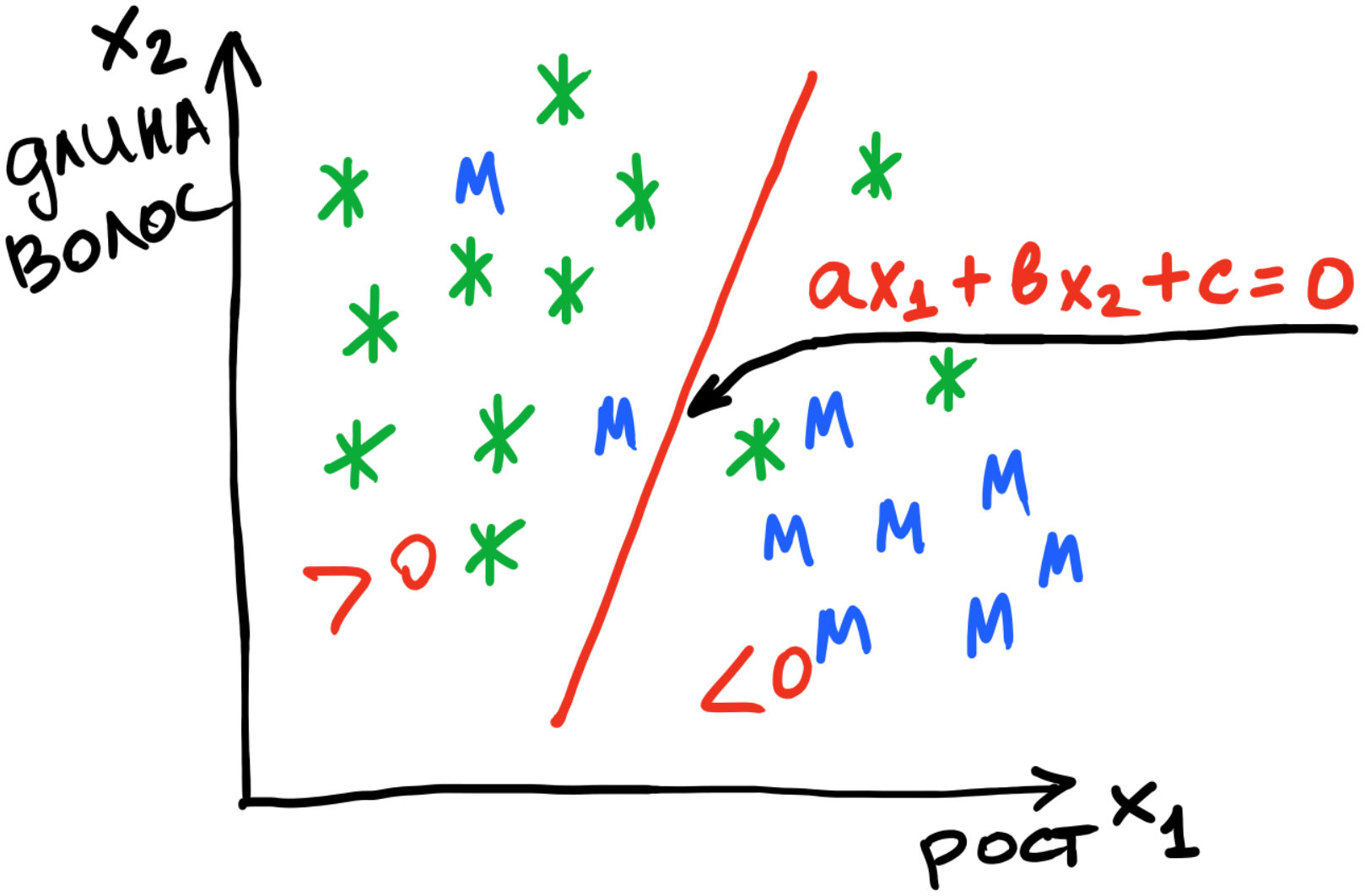
$$a(x;\mu) = \begin{cases} +1, & h(x) > 0; \\ -1, & h(x) < 0. \end{cases}$$

$\mu = (w, b)$

w — вектор весов

b — смещение (bias)

Повторение: линейный классификатор



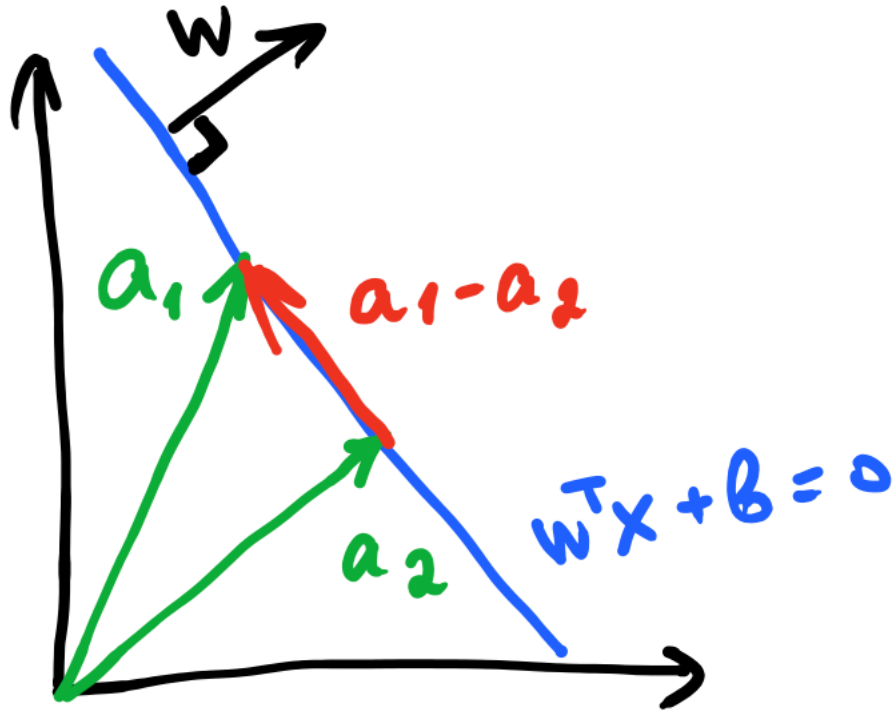
Нормаль гиперплоскости

a_1, a_2 — две произвольные точки на гиперплоскости $h(x) = w^T x + b$.

$$h(a_1) = w^T a_1 + b = 0$$

$$h(a_2) = w^T a_2 + b = 0$$

$$w^T (a_1 - a_2) = 0$$



$(a_1 - a_2)$ — вектор в гиперплоскости

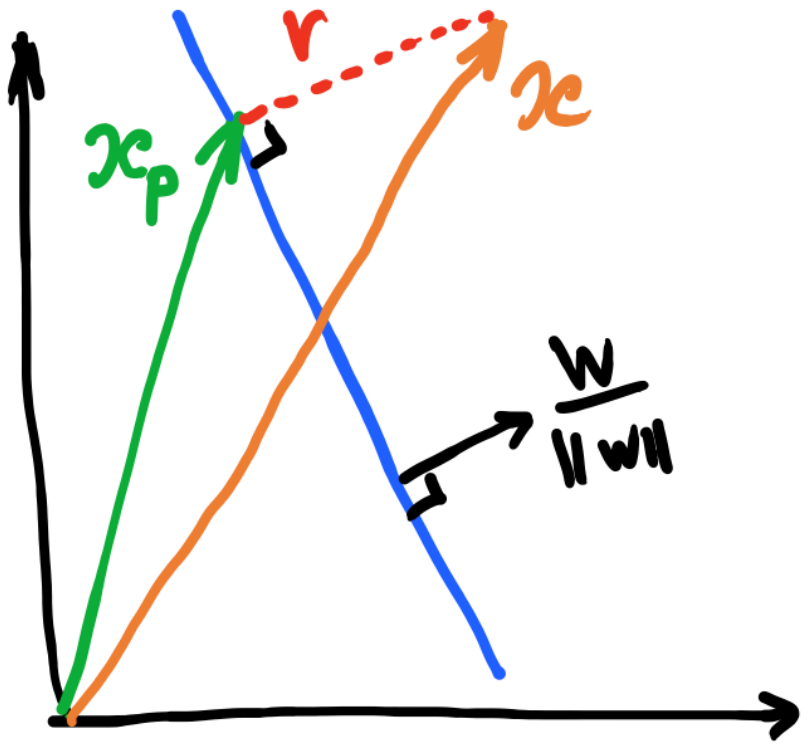
w — нормаль к гиперплоскости

Расстояние от гиперплоскости

Пусть $x \in \mathbb{R}^n$ — произвольная точка.

Обозначим x_P — ее проекция на гиперплоскость h,
r — расстояние от x до гиперплоскости (со знаком)

$$x = x_P + r\frac{w}{\|w\|}$$



Как выразить r через параметры h?

Расстояние от гиперплоскости

$$x = x_P + r\frac{w}{\|w\|}$$

$$h(x) = w^T\left(x_P + r\frac{w}{\|w\|}\right) + b = \boxed{w^T x_P + b} + r\frac{w^T w}{\|w\|} =$$

$$= h(x_P) + r\|w\| = r\|w\|$$

$$r = \frac{h(x)}{\|w\|} \qquad \boxed{|r| = yr = \frac{yh(x)}{\|w\|}}$$

Отступ классификатора

$D = \{(x_i, y_i)\}_{i=1}^{N}$ — набор данных

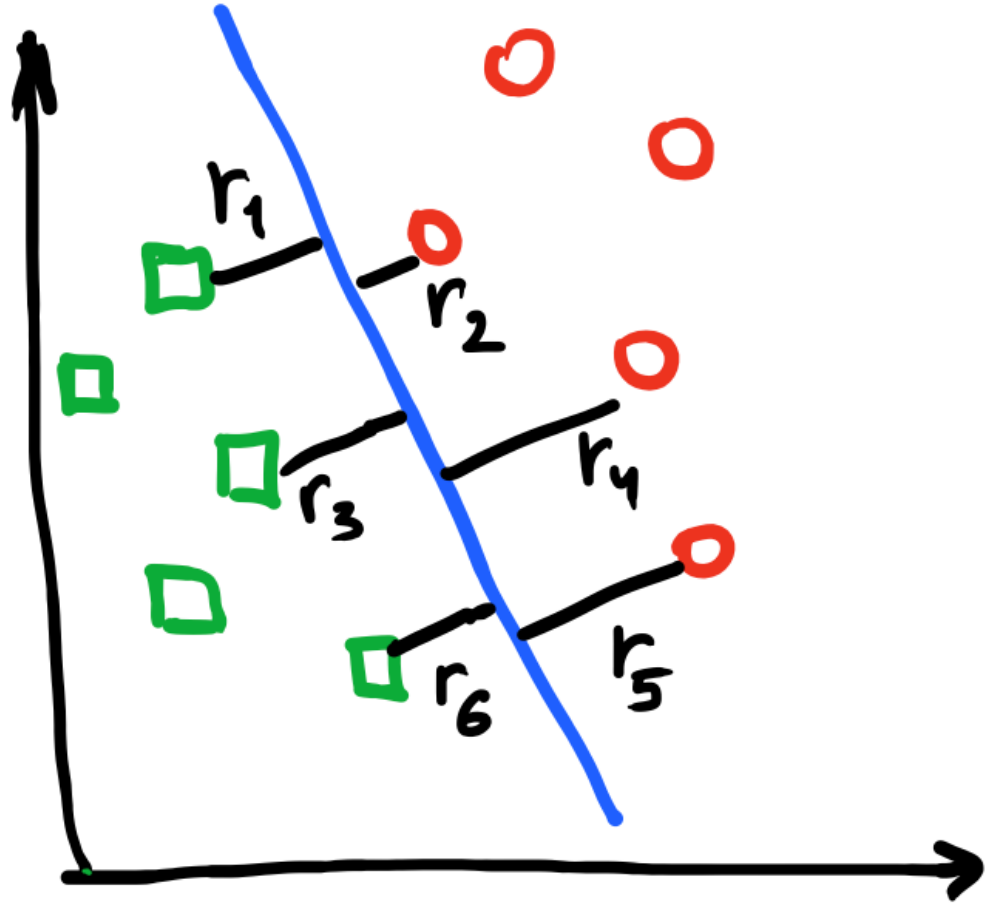
$$\delta^* = \min_{x_i} |r_i| = \min_{x_i} \frac{y_i h(x_i)}{\|w\|}$$

δ^* — отступ (margin) классификатора

Вектора, на которых достигается минимальное расстояние, называются опорными.

Зачем придумали отступ?

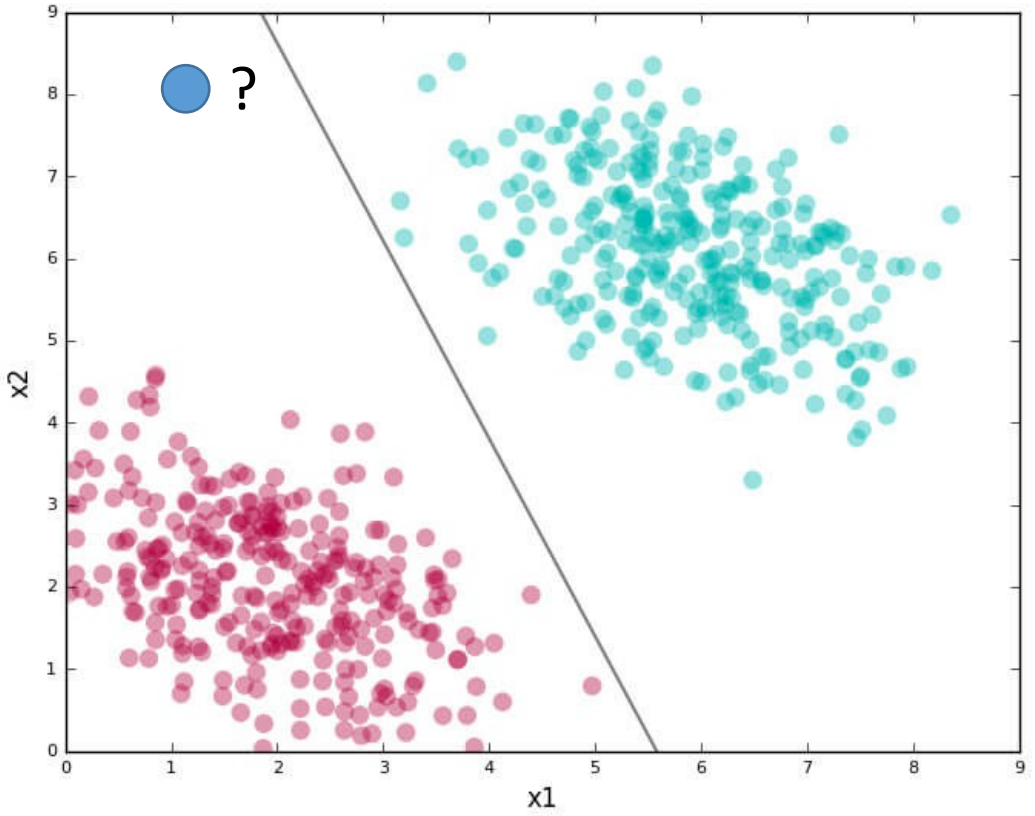
Зачем нам отступ?

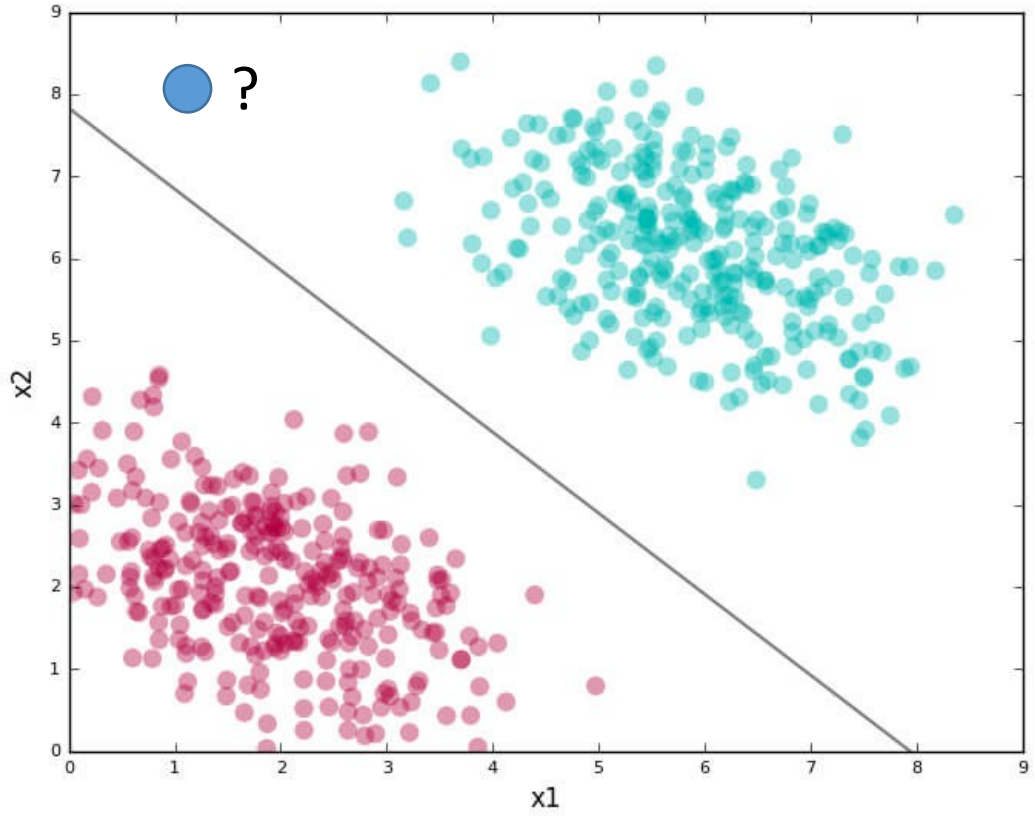


$$\delta^* = \min_{x_i} |r_i| = \min_{x_i} \frac{y_i h(x_i)}{\|w\|}$$

$$\delta^* \to \max_{w}$$

Почему хотим большой отступ





Здесь отступ больше.
Меньше ошибка на тесте!

Каноническая гиперплоскость

$$h(x) = w^T x + b \qquad h'(x) = cw^T x + cb$$

Уравнения задают одну и ту же гиперплоскость

Пусть x_k — опорный вектор. Выберем w, b так, чтобы

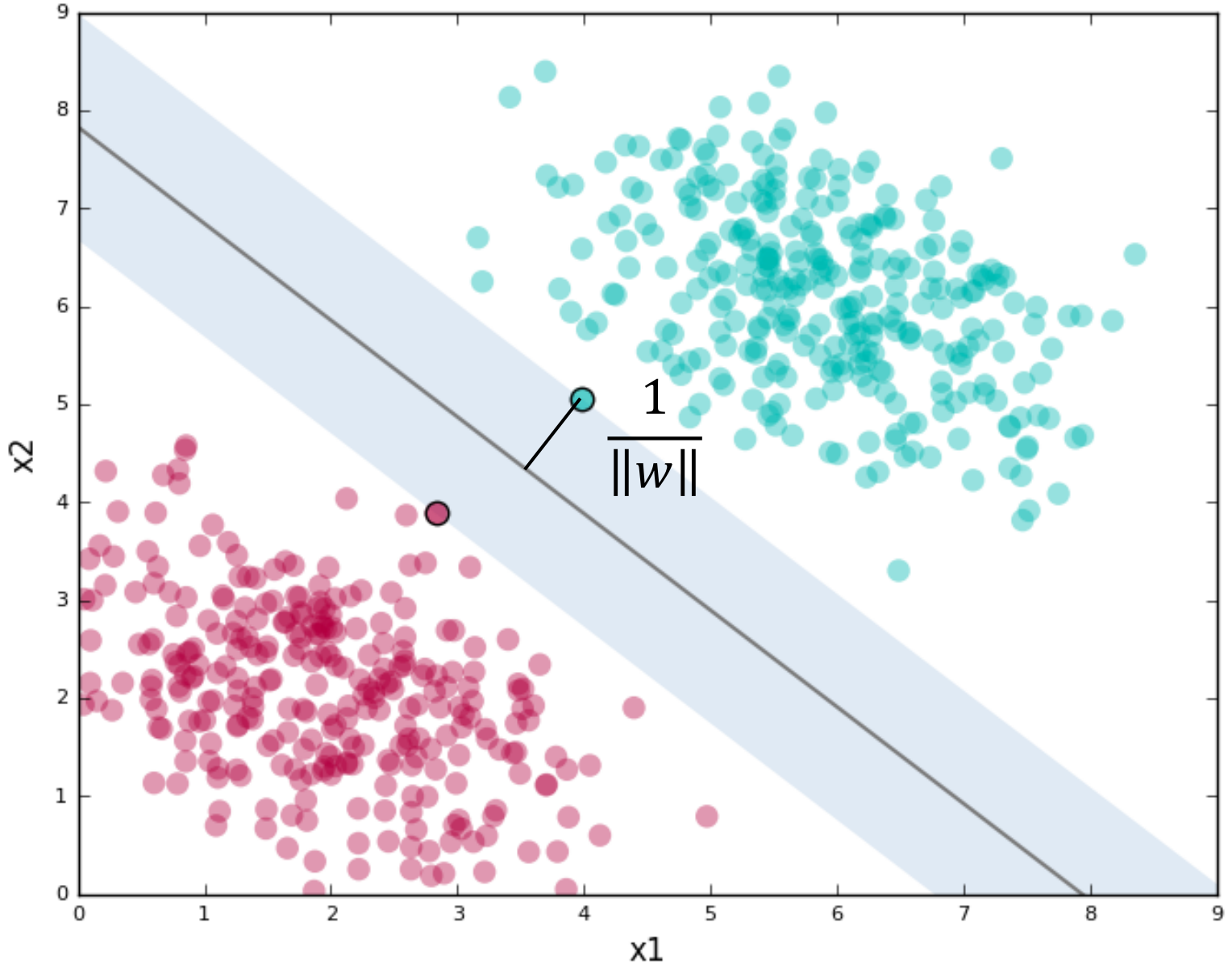
$$y_k h(x_k) = 1$$

Тогда

$$r_k = \frac{y_k h(x_k)}{\|w\|} = \frac{1}{\|w\|}$$

(для всех опорных векторов)

Опорные вектора и отступ



SVM (линейная разделимость)

$D = \{(x_i, y_i)\}_{i=1}^{N}$ — обучающая выборка (линейно разделимая)

Классификатор:

$$a(x; w, b) = \begin{cases} +1, & w^T x + b \geqslant 0; \\ -1, & w^T x + b < 0. \end{cases}$$

Задача

$$\max_{w,b} \frac{1}{\|w\|} \quad \text{при условии} \quad y_i(w^T x_i + b) \geqslant 1.$$

$$i = 1, \ldots, N.$$

SVM (линейная разделимость)

$D = \{(x_i, y_i)\}_{i=1}^{N}$ — обучающая выборка (линейно разделимая)

Классификатор:

$$a(x; w, b) = \begin{cases} +1, & w^T x + b \geqslant 0; \\ -1, & w^T x + b < 0. \end{cases}$$

Задача

$$\min_{w,b} \|w\|^2 \quad \text{при условии} \quad y_i(w^T x_i + b) \geqslant 1.$$

$$i = 1, \ldots, N.$$

Отсутствие линейной разделимости

Не существует решений для

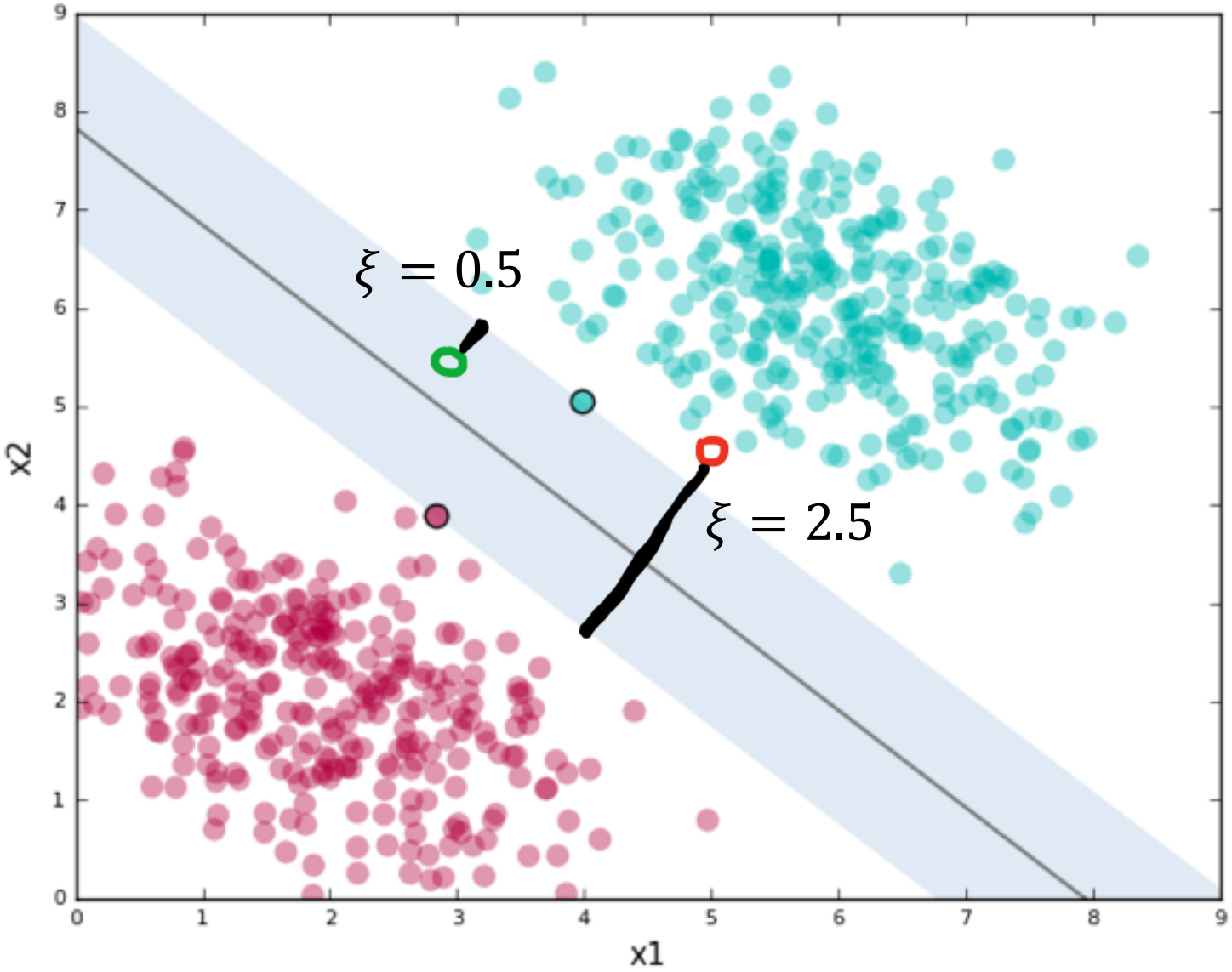
$$y_i(w^T x_i + b) \geqslant 1, \quad i = 1, \ldots, N.$$

Разрешим некоторым объектам нарушать условие

$$y_i(w^T x_i + b) \geqslant 1 - \xi_i, \qquad \xi_i \geqslant 0,$$
$$i = 1, \ldots, N.$$

- $\xi_i = 0$ — обычный объект
- $0 < \xi_i \leqslant 1$ — объект попадает в отступ, но классифицируется верно
- $\xi_i > 1$ — объект классифицируется неверно

Штрафы ξ



SVM (Общий случай)

Первая попытка:

$$\min_{w,b} \|w\|^2 \quad \text{при условии}$$

$$y_i(w^T x_i + b) \geqslant 1 - \xi_i, \quad \xi_i \geqslant 0.$$

$$i = 1, \ldots, N.$$

Проблема: будет большая ошибка классификации.

SVM (Общий случай)

$$\min_{w,b} \|w\|^2 + C \sum_{i=1}^{N} \xi_i \quad \text{при условии}$$

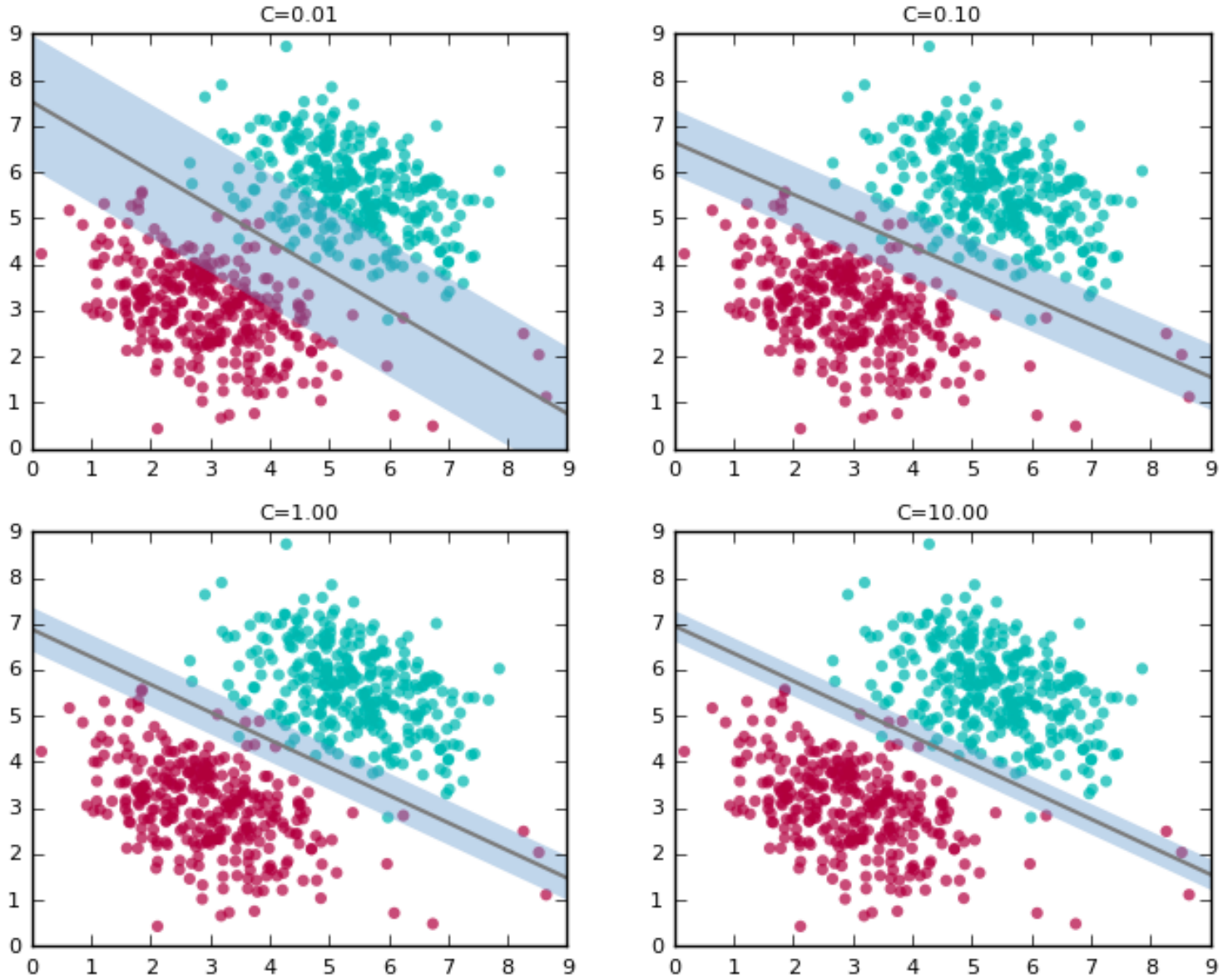
$$y_i(w^T x_i + b) \geqslant 1 - \xi_i, \quad \xi_i \geqslant 0.$$

$$i = 1, \ldots, N.$$

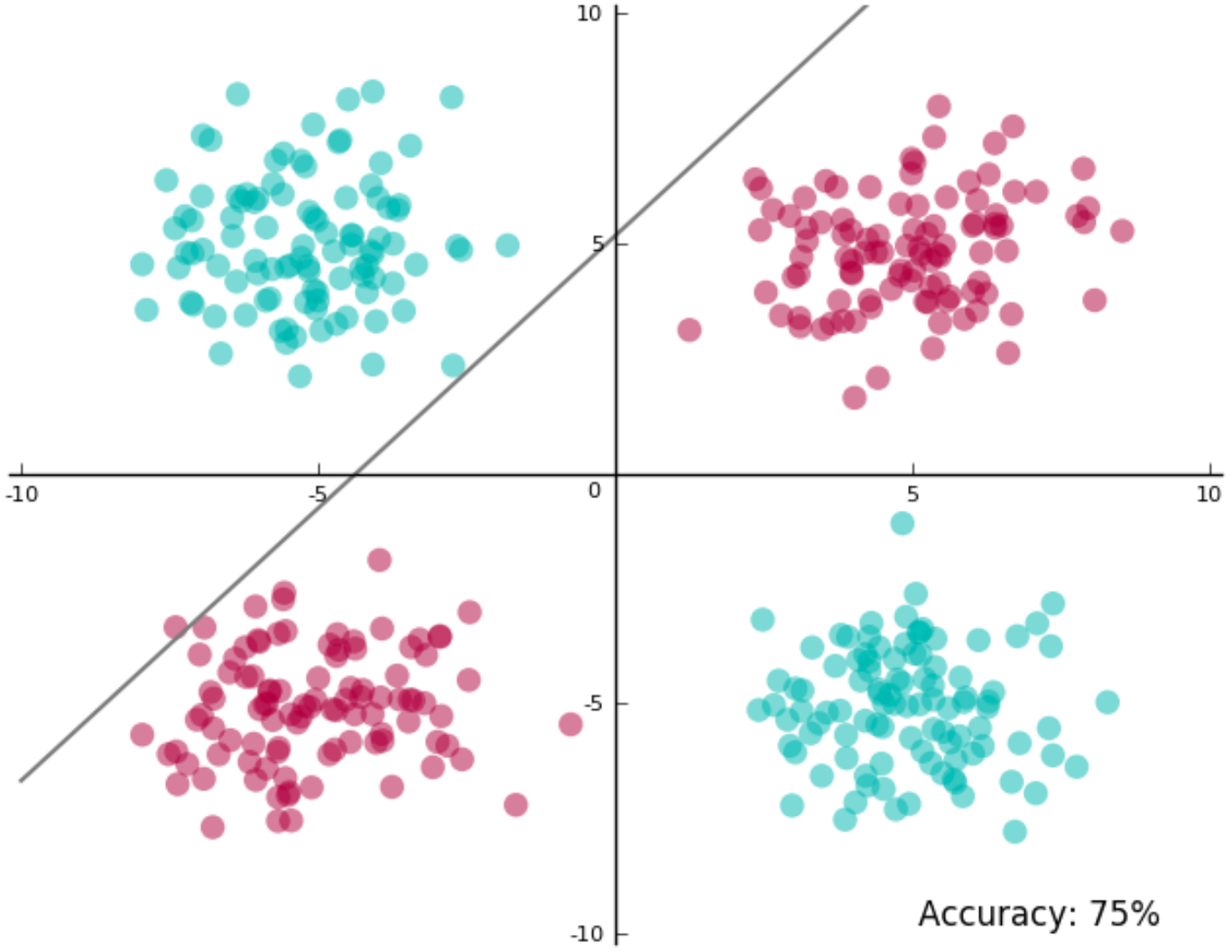
C — параметр регуляризации

- $C \to 0$ — сильная регуляризация, слабо учитываются данные
- $C \to \infty$ — слабая регуляризация, настройка на данные

Эффект регуляризации



А как быть тут?



Как добиться линейной разделимости?

Расширение признаков

Как добиться линейной разделимости?
Ответ: нужно добавить нелинейные признаки.

x — исходные признаки
$\varphi(x)$ — расширенные признаки.

Примеры для $x = (x_1, x_2)$:

- $\varphi(x) = (x_1, x_2, x_1^2, x_2^2, x_1 x_2)$
- $\varphi(x) = (x_1, x_2, \ln x_1, \ln x_2)$

Проблемы при большом числе признаков:

- Вычислительная сложность
- Проклятие размерности

Преобразуем задачу SVM

Исходная задача

$$\min_{w,b} \|w\|^2 + C \sum_{i=1}^{N} \xi_i$$

$$y_i(w^T x_i + b) \geqslant 1 - \xi_i, \quad i = 1, \ldots, N$$

$$\xi_i \geqslant 0, \quad i = 1, \ldots, N$$

Исходная решающая функция

$$h(z) = w^T z + b$$

Двойственная задача

- Будем искать решение в виде $$\mathbf{w} = \sum_{j=1}^{N} \alpha_j y_j \mathbf{x}_j$$
- Подставим в решающую функцию

$$f(x) = \left(\sum_{j=1}^{N} \alpha_j y_j \mathbf{x}_j \right)^{\top} \mathbf{x} + b = \sum_{j=1}^{N} \alpha_j y_j \left(\mathbf{x}_j^{\top} \mathbf{x} \right) + b$$

- Выразим $\|w\|^2$

$$||\mathbf{w}||^2 = \left\{ \sum_j \alpha_j y_j \mathbf{x}_j \right\}^{\top} \left\{ \sum_k \alpha_k y_k \mathbf{x}_k \right\} = \sum_{jk} \alpha_j \alpha_k y_j y_k (\mathbf{x}_j^{\top} \mathbf{x}_k)$$

- Еще несколько шагов…

Двойственная задача

Преобразованная задача

$$\max_{\alpha} \left(\sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j x_i^T x_j \right)$$

$$0 \leqslant \alpha_i \leqslant C, \quad i = 1, \ldots, N$$

$$\sum_{i=1}^{N} \alpha_i y_i = 0$$

Преобразованная решающая функция

$$h(z) = \sum_{\alpha_i > 0} \alpha_i y_i x_i^T z + \underset{i, 0 < \alpha_i < C}{\text{average}} \left(y_i - \sum_{\alpha_j > 0} \alpha_j y_j x_j^T x_i \right)$$

Двойственная задача

Преобразованная задача

$$\max_{\alpha} \left(\sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle \right)$$

$$0 \leqslant \alpha_i \leqslant C, \quad i = 1, \ldots, N$$

$$\sum_{i=1}^{N} \alpha_i y_i = 0$$

Преобразованная решающая функция

$$h(z) = \sum_{\alpha_i > 0} \alpha_i y_i \langle x_i, z \rangle + \underset{i, 0 < \alpha_i < C}{\text{average}} \left(y_i - \sum_{\alpha_j > 0} \alpha_j y_j \langle x_j, x_i \rangle \right)$$

Трюк с ядром

Преобразованная задача

$$\max_{\alpha} \left(\sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \right)$$

$$0 \leqslant \alpha_i \leqslant C, \quad i = 1, \ldots, N$$

$$\sum_{i=1}^{N} \alpha_i y_i = 0$$

Преобразованная решающая функция

$$h(z) = \sum_{\alpha_i > 0} \alpha_i y_i K(x_i, z) + \underset{i, 0 < \alpha_i < C}{\text{average}} \left(y_i - \sum_{\alpha_j > 0} \alpha_j y_j K(x_j, x_i) \right)$$

Расширение признаков и ядра

x — исходные признаки

$\varphi(x)$ — расширенные признаки.

$K(x, y) = \langle \varphi(x), \varphi(y) \rangle$ — ядро.

Норма

$$\|\varphi(x)\|^2 = \langle \varphi(x), \varphi(x) \rangle = K(x, x)$$

Расстояние

$$\begin{aligned}\rho(\varphi(x), \varphi(y))^2 = \|\varphi(x) - \varphi(y)\| = \\ = \langle \varphi(x) - \varphi(y), \varphi(x) - \varphi(y) \rangle = \\ = \langle \varphi(x), \varphi(x) \rangle + \langle \varphi(y), \varphi(y) \rangle - 2\langle \varphi(x), \varphi(y) \rangle = \\ = K(x, x) + K(y, y) - 2K(x, y).\end{aligned}$$

Будем работать только с ядрами

Идея: работать только с ядрами.

Пример

$$x = (x_1, x_2)$$

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

$$K(x, y) = \langle \varphi(x), \varphi(y) \rangle =$$
$$= x_1^2 y_1^2 + x_2^2 y_2^2 + 2 x_1 y_1 x_2 y_2 = (x_1 y_1 + x_2 y_2)^2 = \langle x, y \rangle^2$$

А нужно ли в явном виде задавать $\varphi(x)$?

Скорость работы

Ядра работают быстрее, чем честный расчет новых признаков $\varphi(x)$ и вычисление скалярного произведения на них

Другие ядра

Другие ядра

- $K(x, y) = \langle x, y \rangle^d$ полиномиальное ядро
- $K(x, y) = (\langle x, y \rangle + 1)^d$ полиномиальное ядро
- $K(x, y) = e^{-\|x-y\|^2}$ гауссовское ядро

Как составлять ядра

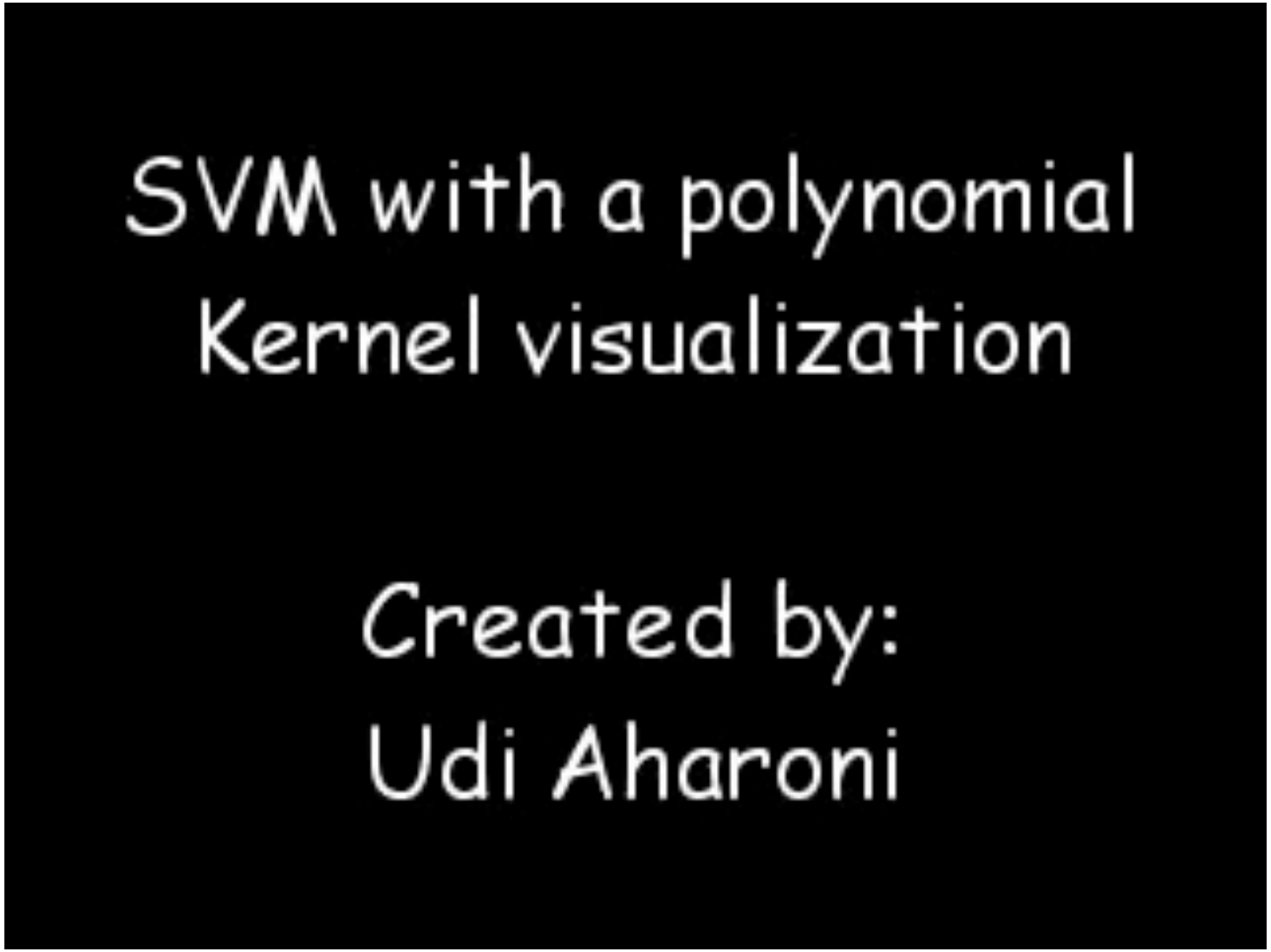
Стандартные ядра

- $K(x, y) = 1$
- $K(x, y) = \langle x, y \rangle$
- $K(x, y) = e^{-\|x-y\|^2}$
- $K(x, y) = e^{-\|x-y\|}$

Преобразование ядер

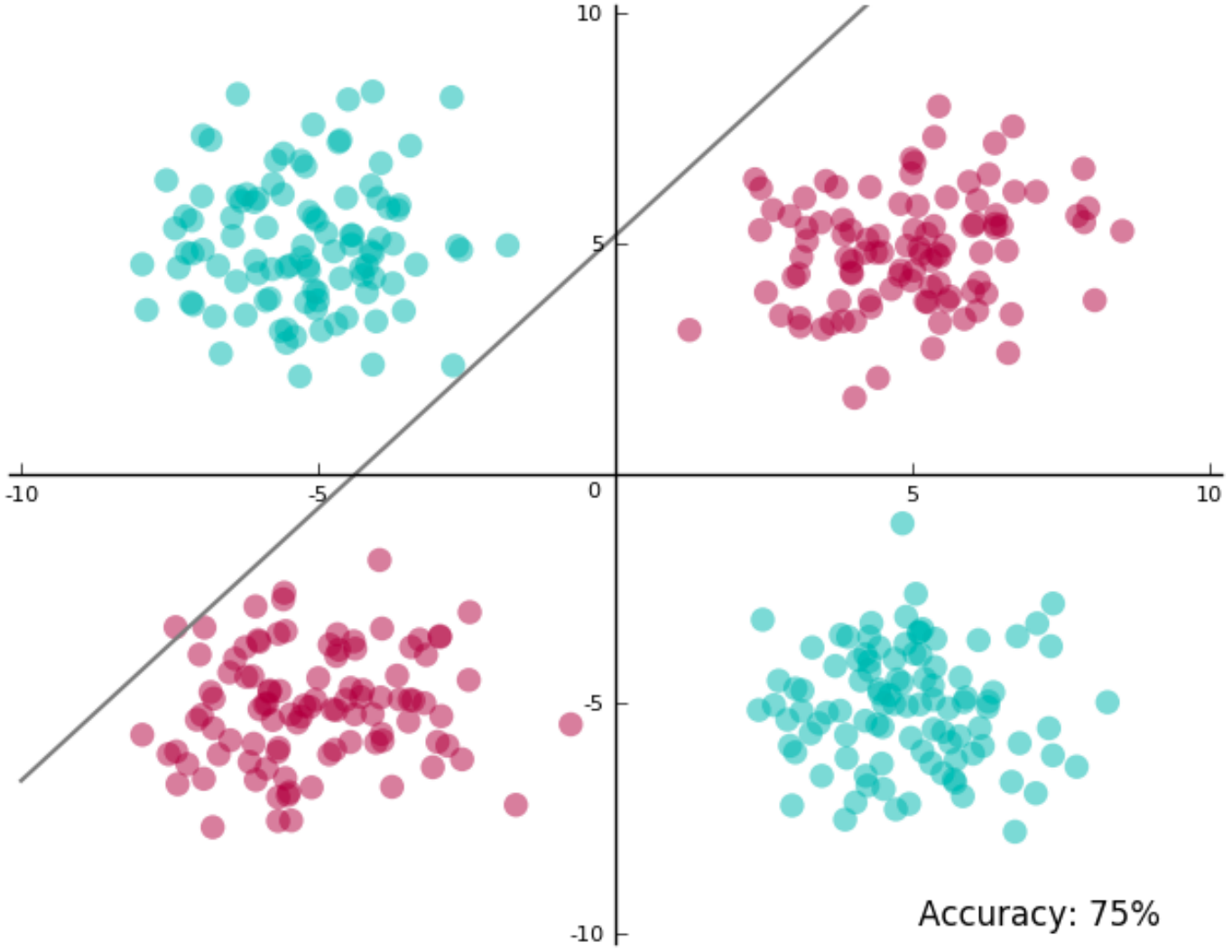
- $K(x, y) = K_1(x, y)K_2(x, y)$
- $K(x, y) = C_1K_1(x, y) + C_2K_2(x, y)$

SVM с полиномиальным ядром

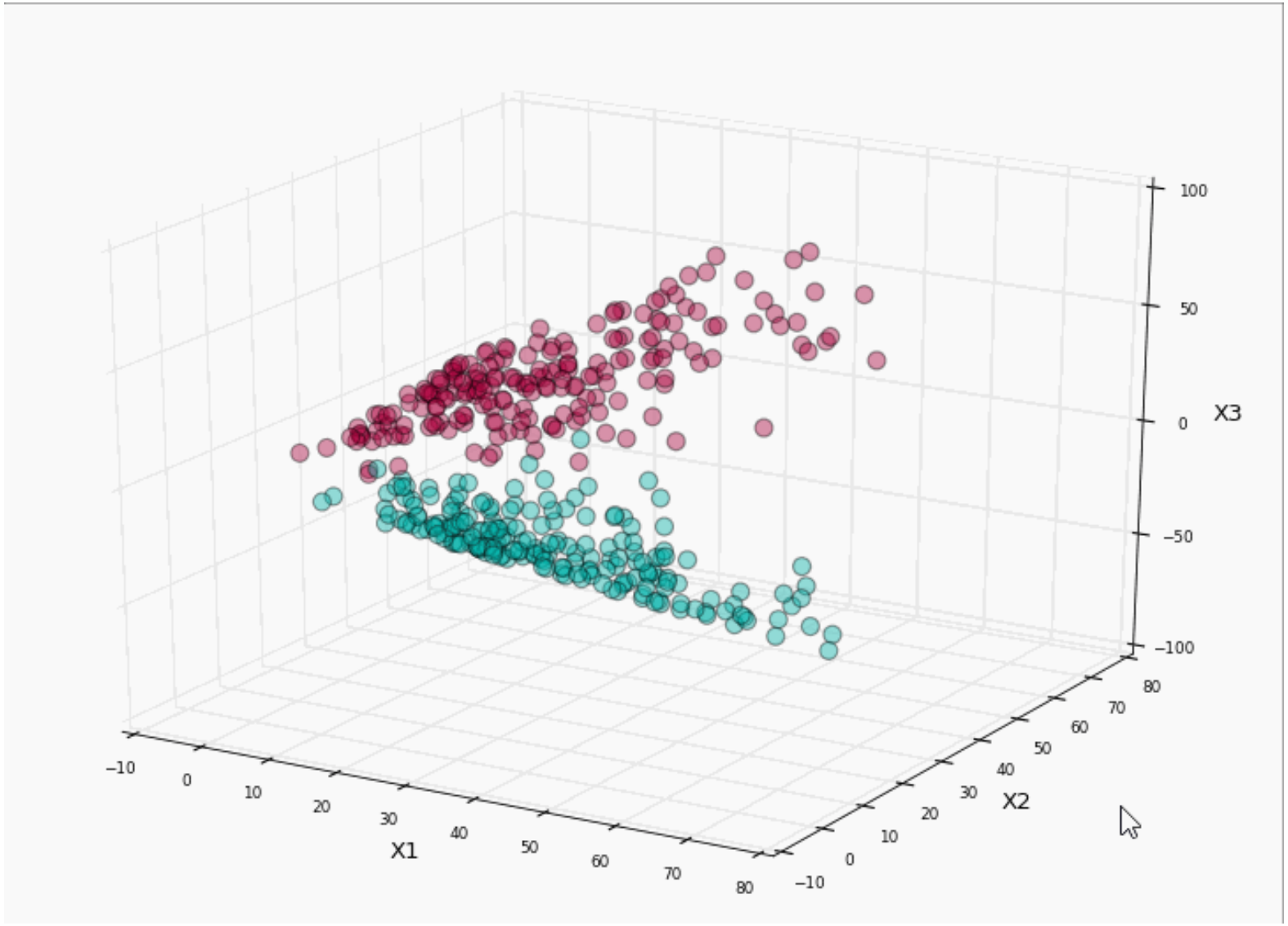


$$\varphi(x) = (x_1, x_2, x_1^2 + x_2^2)$$

Вернемся к примеру



Применим ядровой SVM



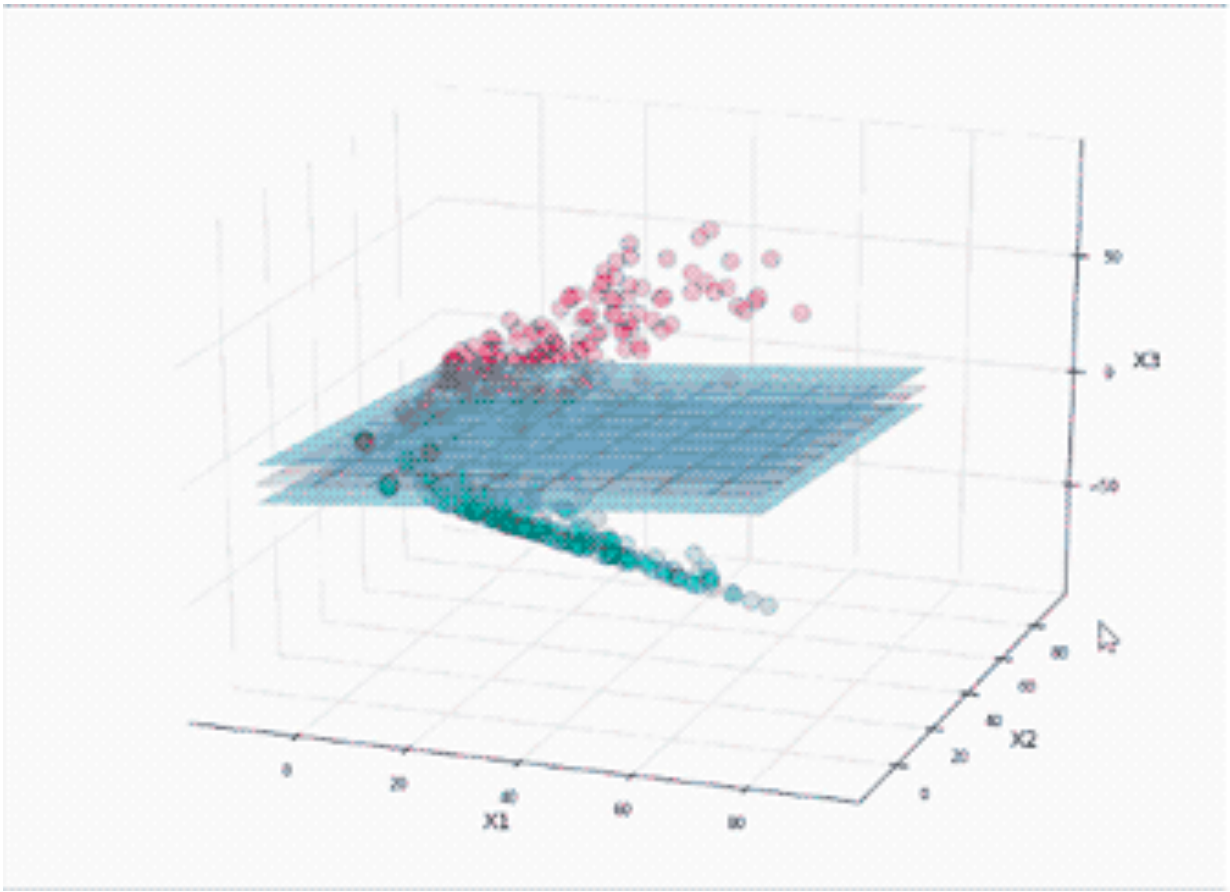
$$X_1 = x_1^2$$

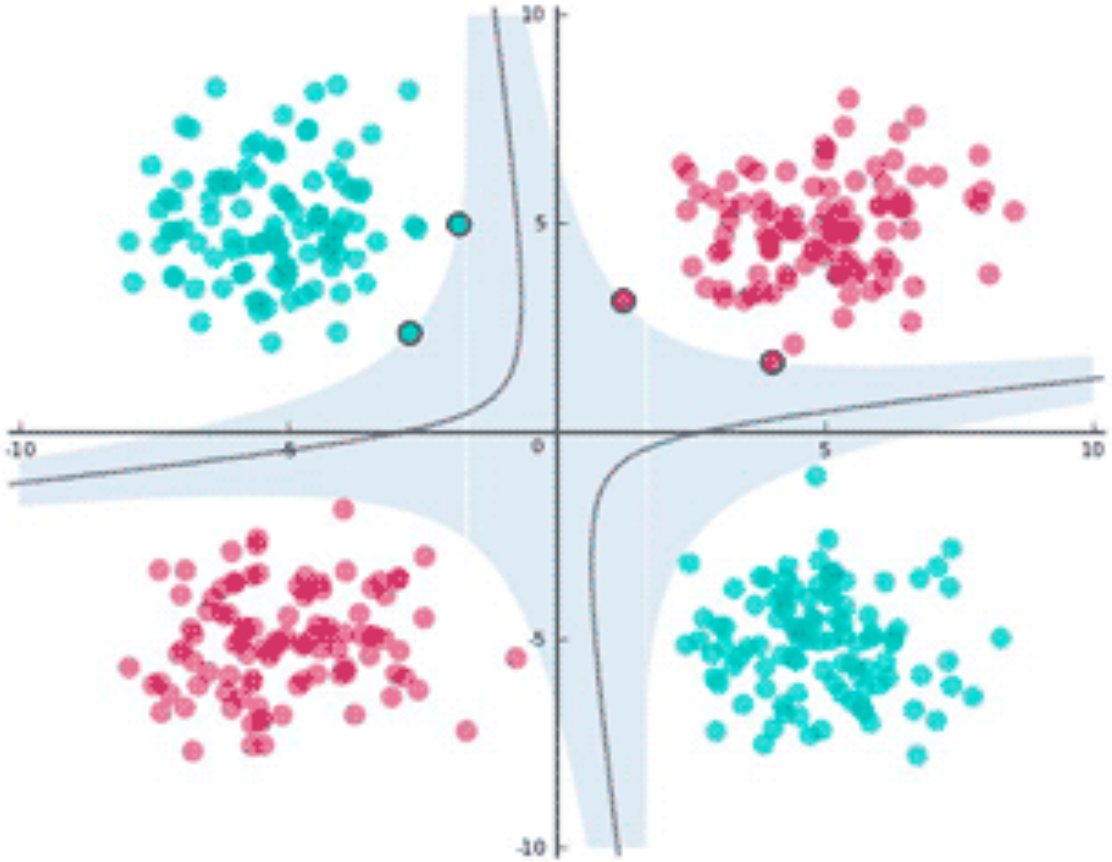
$$X_2 = x_2^2$$

$$X_3 = \sqrt{2}x_1x_2$$

$$\mathrm{K}(x, y) = < x, y >^2$$

Результат





Ядра для строк

S — некоторая последовательность «слов»
(например, текст или последовательность ДНК)

Возможные признаки
(например, $S =$ «dog and cat and cow»)

- Сколько раз встретилось каждое слово
  (dog: 1, and: 2, cat: 1, cow: 1)

- Сколько раз встретились последовательности слов длины d
  («dog and»: 1, «and cat»: 1, ..., «cow and»: 0, ...)

Ядра для строк

- Ядро: скалярное произведение на счетчиках последовательностей длины 2

$$A = \text{«dog and cat and cow»}$$
$$B = \text{«cat and cat and cat»}$$

- $K(A, B) = (\text{“}and\ cat\text{”})\ 1 \times 2 + (\text{”}cat\ and\text{”})\ 1 \times 2 = 4$

Преимущества и недостатки SVM

Преимущества

- Достаточно эффективное решение
- Высокая обобщающая способность

Недостатки

- Не очень высокая устойчивость к шуму
  Опорные вектора могут быть шумовыми
- Непонятно, как выбирать C
  Обычно выбирается кросс-валидацией
- Нет рецепта как выбирать ядра под задачу

Ссылки

- https://github.com/esokolov/ml-course-hse/blob/master/2016-spring/lecture-notes/lecture16-kernels.pdf Вывод двойственной задачи SVM в §2.2
- http://www.robots.ox.ac.uk/~az/lectures/ml/lect3.pdf Хорошие слайды про SVM
- https://blog.statsbot.co/support-vector-machines-tutorial-c1618e635e93 Блог пост про SVM с анимациями
- https://www.youtube.com/watch?v=3liCbRZPrZA Видео про полиномиальное ядро