Лекция 4

Метод опорных векторов

Машинное обучение

Сергей Муравьёв / Андрей Фильченков

25.09.2020

План лекции

- Линейно разделимый случай
- Линейно неразделимый случай
- Ядерный трюк
- Выбор и синтез ядер
- Регуляризация для метода опорных векторов

- В презентации используются материалы курса «Машинное обучение» К.В. Воронцова
- Слайды доступны: **shorturl.at/hjyAX**
- Видео доступны: **shorturl.at/ltVZ3**

План лекции

- Линейно разделимый случай
- Линейно неразделимый случай
- Ядерный трюк
- Выбор и синтез ядер
- Регуляризация для метода опорных векторов

Основная идея

Если мы предполагаем, что классификатор должен быть линейным, как лучше всего его определить?

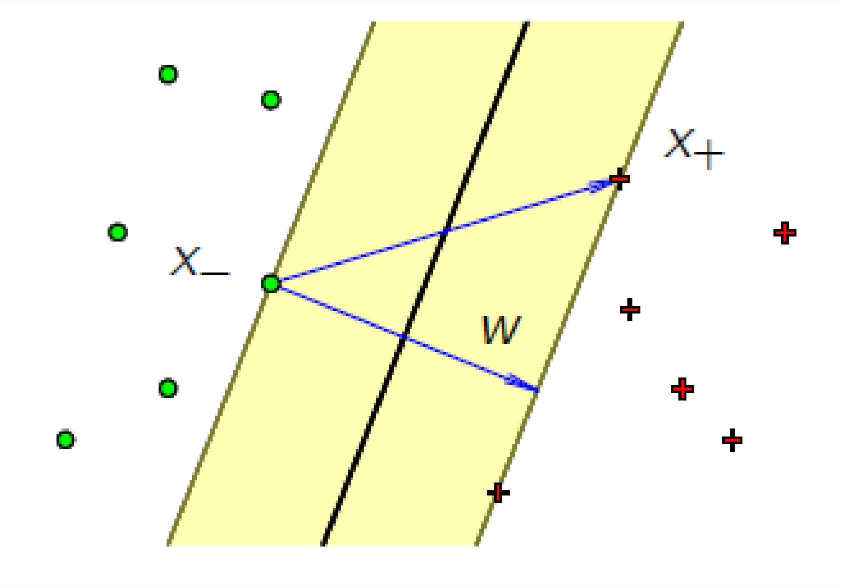
Основная идея: поиск поверхности, наиболее удаленной от классов (классификация с большим запасом).

Линейно разделимый случай

Основная гипотеза: выборка является линейно разделимой:

$$\exists w, w_0 \colon M_i(w, w_0) = y_i(\langle w, x_i \rangle - w_0) > 0, i = 1, \ldots, |\mathcal{D}|.$$

Может существовать несколько разделительных гиперплоскостей, поэтому среди них можно выделить ту, которая имеет максимальное расстояние от обоих классов.



Разделяющая полоса

Нормализуем величину отступа:

$$\min_i M_i(w, w_0) = 1.$$

Уравнение разделяющей полосы:

$$\{x: -1 \leq \langle w, x \rangle - w_0 \leq 1\}.$$

Ширина полосы:

$$\frac{\langle x_+ - x_-, w \rangle}{||w||} = \frac{(\langle x_+, w \rangle - w_0) - (\langle x_-, w \rangle - w_0)}{||w||} = \frac{2}{||w||}.$$

Формализуем задачу оптимизации:

$$\begin{cases} ||w||^2 \to \min_{w, w_0}; \\ M_i(w, w_0) \geq 1, i = 1, \dots, |\mathcal{D}|. \end{cases}$$

План лекции

- Линейно разделимый случай
- Линейно неразделимый случай
- Ядерный трюк
- Выбор и синтез ядер
- Регуляризация для метода опорных векторов

Линейно неразделимый случай

Основная гипотеза: выборка не является линейно разделимой:

$$\forall w, w_0\ \exists x_d\colon M_d(w, w_0) = y_d(\langle w, x_d\rangle - w_0) < 0$$

Такой разделительной гиперплоскости не существует.

Мы все еще можем попытаться найти гиперплоскость с наименьшими значениями отступов для каждого объекта.

Линейно неразделимый случай

В случае линейной неразделимости заданной выборки:

$$\begin{cases} \frac{1}{2}||w||^2 + C\sum_{i=1}^{|\mathcal{D}|}\xi_i \to \min_{w,w_0,\xi}; \\ M_i(w,w_0) \geq 1 - \xi_i, i = 1, \dots, |\mathcal{D}|; \\ \xi_i \geq 0, \qquad i = 1, \dots, |\mathcal{D}|. \end{cases}$$

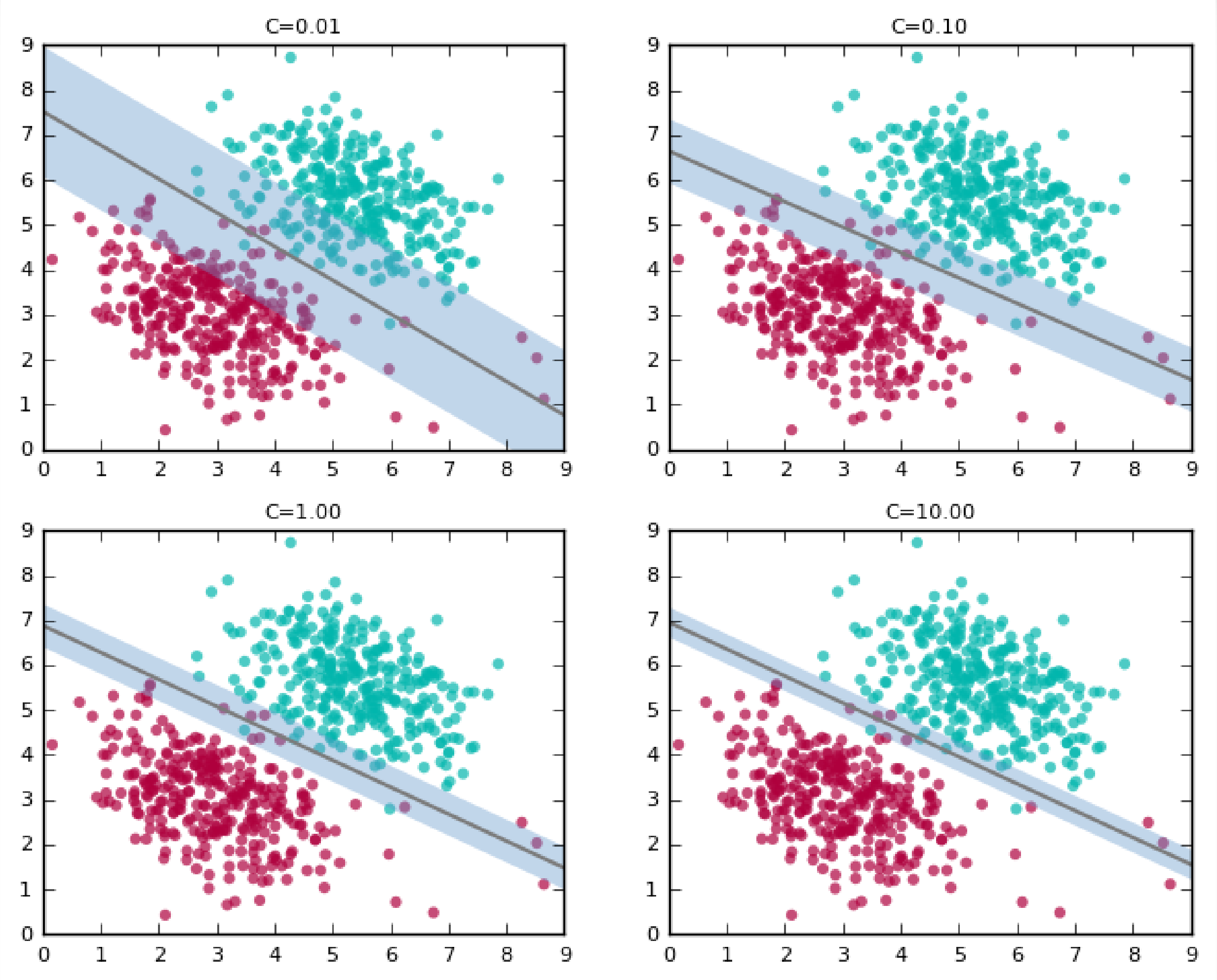
Эквивалентная задача безусловной оптимизации:

$$\sum_{i=1}^{|\mathcal{D}|}\big(1 - M_i(w,w_0)\big)_+ + \frac{1}{2C}||w||^2 \to \min_{w,w_0},$$

где $(x)_+ = (x + |x|)/2$.

Данная формула является аппроксимацией эмпирического риска.

Анализ константы *C*



Задача нелинейного программирования

Задача математического программирования:

$$\begin{cases} f(x) \to \min_x \\ g_i(x) \leq 0, \\ h_j(x) = 0. \end{cases} \qquad i = 1, \dots, m; j = 1, \dots, k.$$

Лагранжиан:

$$\mathcal{L}(x; \mu, \lambda) = f(x) + \sum_{i=1}^{m} \mu_i g_i(x) + \sum_{j=1}^{k} \lambda_j h_j(x)$$

Условие Каруша—Куна—Таккера:

$$\frac{\delta \mathcal{L}}{\delta x}(x^*; \mu, \lambda) = 0.$$

$$\begin{cases} g_i(x^*) \leq 0; \\ h_j(x^*) = 0; \\ \mu_i \geq 0; \\ \mu_i g_i(x^*) = 0. \end{cases} \qquad i = 1, \dots, m; j = 1, \dots, k.$$

Условия ККТ в задаче опорных векторов

Лагранжиан

$$\boldsymbol{\mathcal{L}}(w, w_0, \xi; \alpha, \beta) = \frac{1}{2}||w||^2 - \sum_{i=1}^{|\mathcal{D}|} \alpha_i (M_i(w, w_0) - 1) - \sum_{j=1}^{|\mathcal{D}|} \xi_j(\alpha_j + \beta_j - C)$$

α_i – переменные, двойственные для ограничений $M_i \geq 1 - \xi_i$;

β_i – переменные, двойственные для ограничений $\xi_i \geq 0$.

Условия минимума:

$$\begin{cases} \dfrac{\delta\boldsymbol{\mathcal{L}}}{\delta w} = 0; \dfrac{\delta\boldsymbol{\mathcal{L}}}{\delta w_0} = 0; \dfrac{\delta\boldsymbol{\mathcal{L}}}{\delta \xi} = 0; \\ \xi_i \geq 0; \alpha_i \geq 0; \beta_i \geq 0; \\ \alpha_i = 0 \text{ или } M_i(w, w_0) = 1 - \xi_i; \\ \beta_i = 0 \text{ или } \xi_i = 0; \end{cases}$$

$i = 1, \ldots, |\mathcal{D}|$.

Опорные вектора

Типы объектов:

1. $\alpha_i = 0; \beta_i = C; \xi_i = 0; M_i > 1$

периферийные объекты.

2. $0 < \alpha_i < C; 0 < \beta_i < C; \xi_i = 0; M_i = 1$

опорные пограничные объекты.

3. $\alpha_i = C; \beta_i = 0; \xi_i > 0; M_i < 1$

опорные нарушители.

Объект x_i – **опорные объекты,** если $\alpha_i \neq 0$.

Задача нелинейного программирования

$$-\mathcal{L}(\alpha) = -\sum_{i=1}^{|\mathcal{D}|} \alpha_i + \frac{1}{2}\sum_{i=1}^{|\mathcal{D}|}\sum_{j=1}^{|\mathcal{D}|} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle \to \min_\alpha$$

$$\begin{cases} 0 \le \alpha_i \le C; \\ \sum_{j=1}^{\ell} \alpha_i y_i = 0. \end{cases}$$

Решение задачи может быть выражено следующим образом:

$$\begin{cases} w = \sum_{i=1}^{|\mathcal{D}|} \alpha_i y_i x_i\,; \\ w_0 = \langle w, x_i \rangle - y_i. \end{cases} \quad \forall i\colon\ \alpha_i > 0, M_i = 1.$$

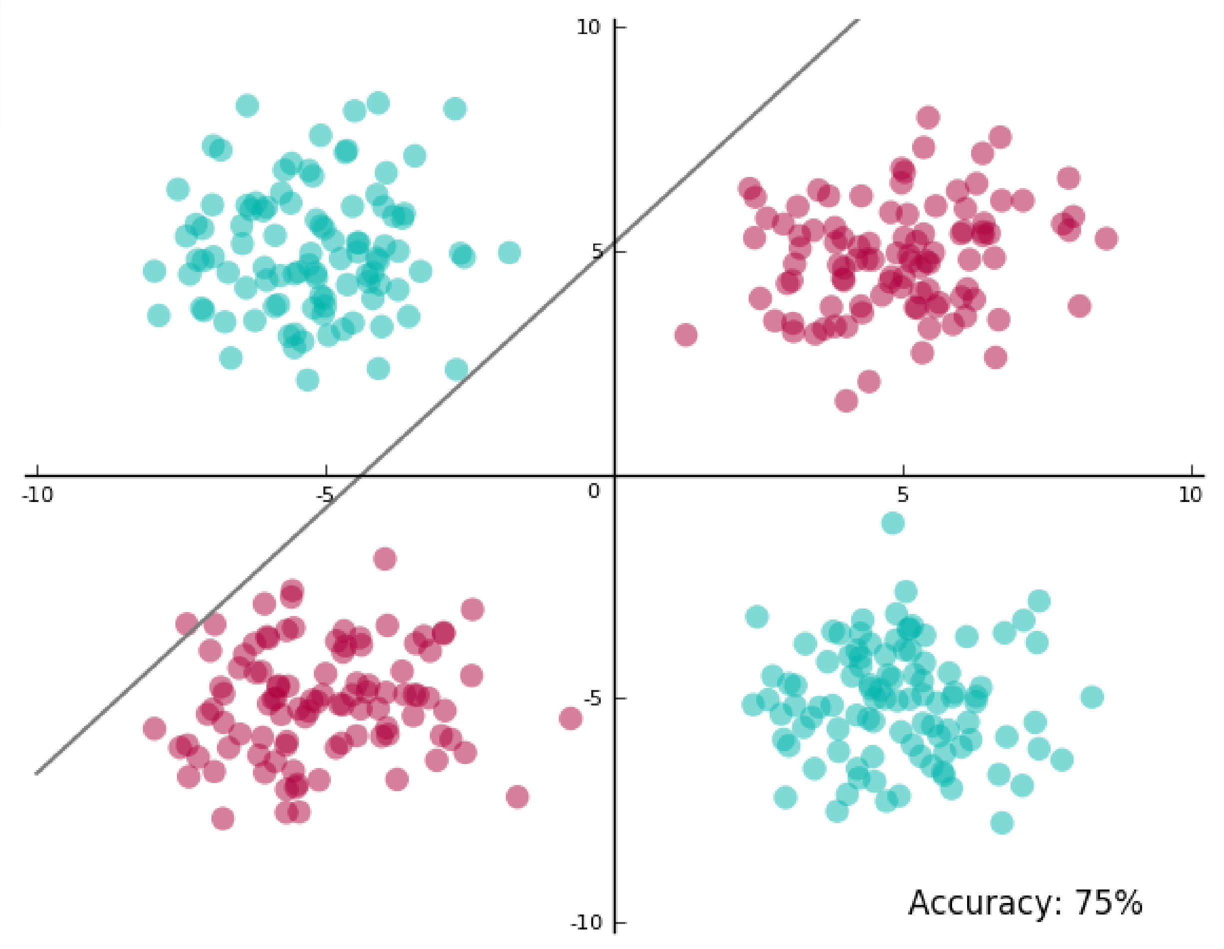
Линейный классификатор:

$$a(x) = \operatorname{sign}\left(\sum_{i=1}^{|\mathcal{D}|} \alpha_i y_i \langle x_i, x \rangle - w_0\right).$$

План лекции

- Линейно разделимый случай
- Линейно неразделимый случай
- Ядерный трюк
- Выбор и синтез ядер
- Регуляризация для метода опорных векторов

Плохой случай линейной неразделимости

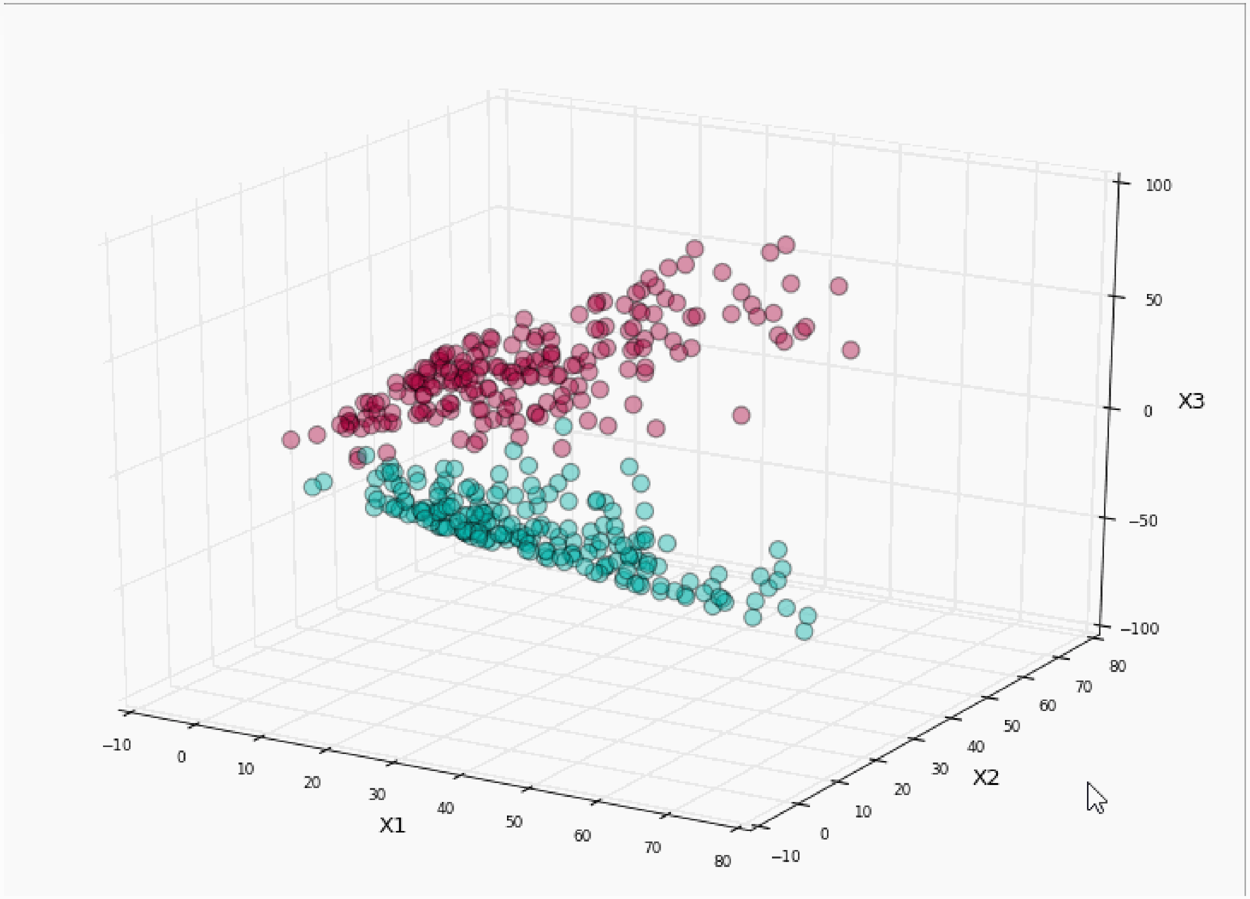


Ядерный трюк

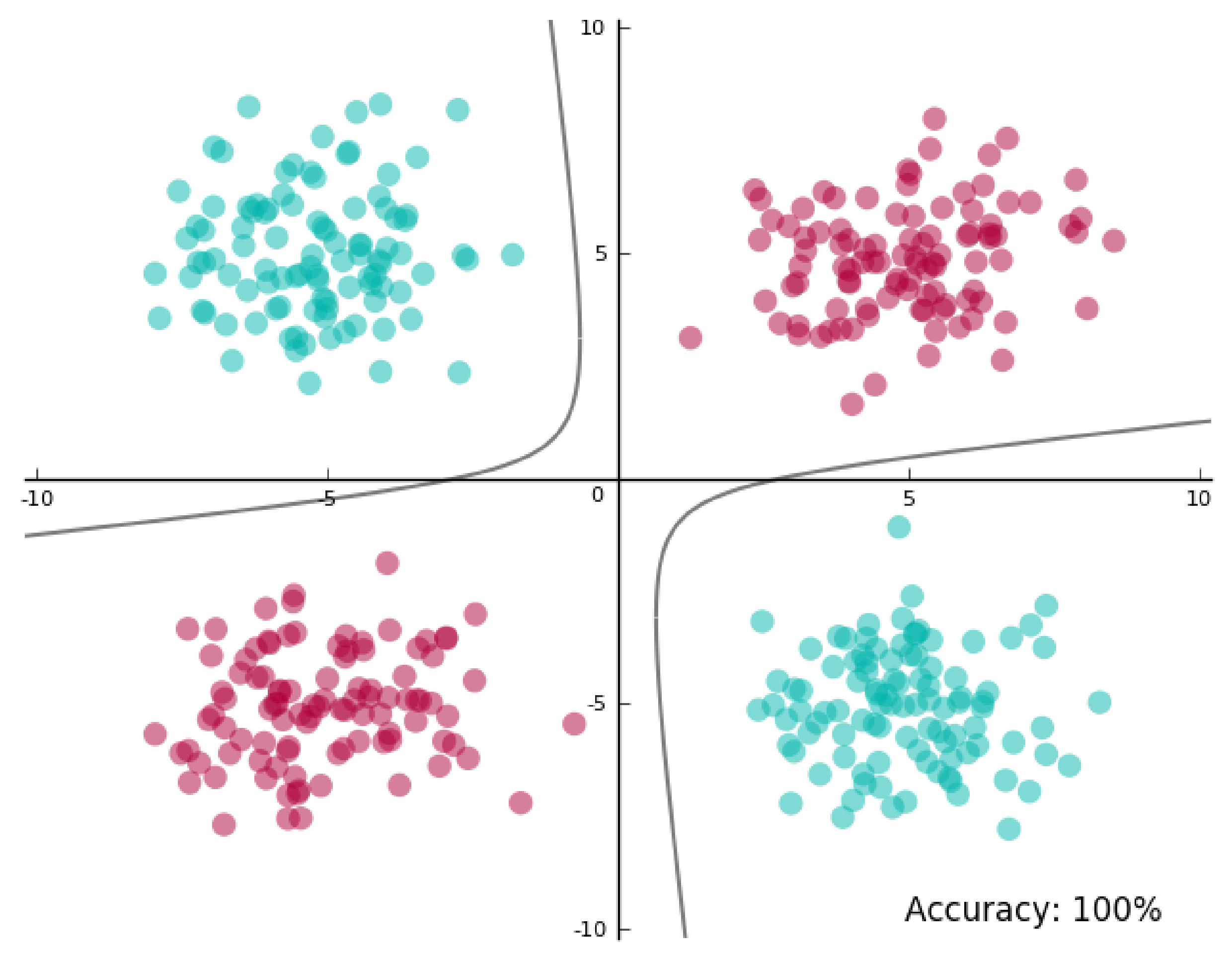
Основная идея: найти отображение в многомерное пространство, такое, что точки в новом пространстве будут линейно разделимы.

Суть: пусть разделяющая поверхность хорошо аппроксимируется суммой функций, зависящих от $x_1, \ldots, x_n$: $c_1 x_1 + \cdots + c_n x_n + f_1(x_1, \ldots, x_n) + \cdots + f_k(x_1, \ldots, x_n)$
Если мы добавим признаки $f_1(x_1, \ldots, x_n), \ldots, f_k(x_1, \ldots, x_n)$, тогда у нас будет новое пространство над переменными $x_1, \ldots, x_n, x_{n+1}, \ldots, x_{n+k}$, точки которых будут линейно разделимы.

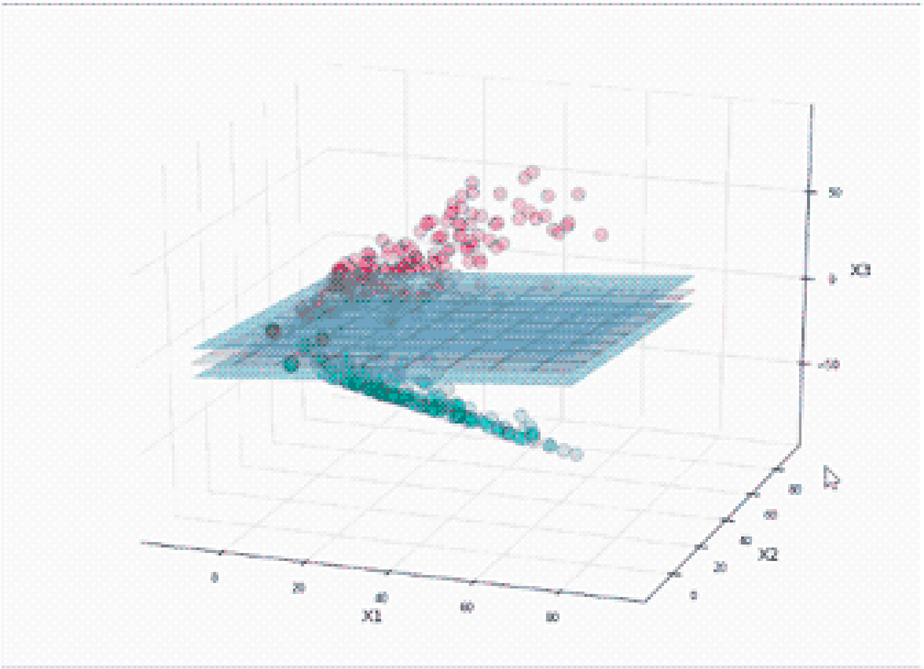
Разделимость в пространстве большей размерности

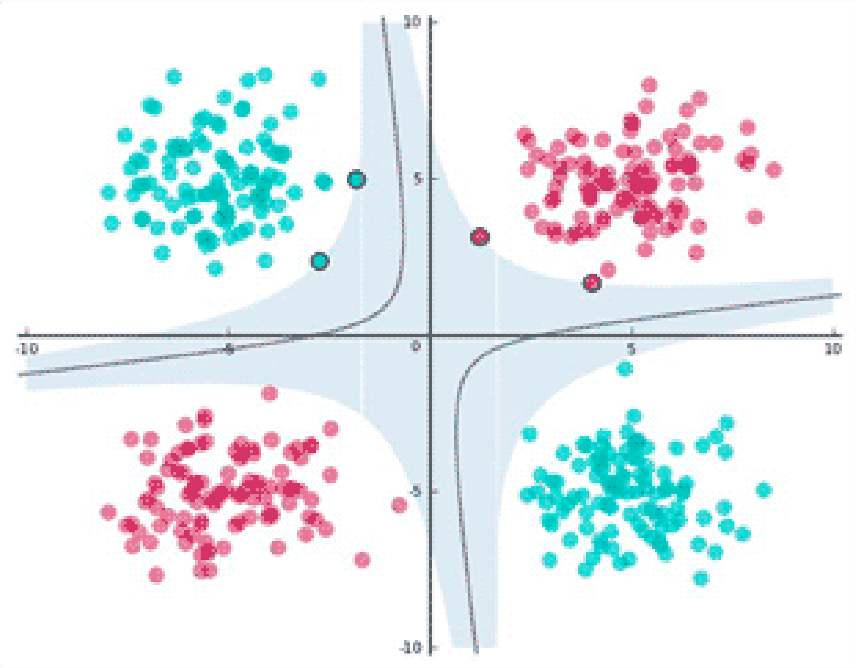


Как это выглядит в оригинальном пространстве



Результирующая разделяющая поверхность





Почему ядра?

Мы можем построить дистанционный классификатор для опорных объектов (векторов). Использование функции ядра равносильно использованию определенного отображения.

Основная проблема — найти ядро, которое переводит исходное пространство в линейно разделимое.

План лекции

- Линейно разделимый случай
- Линейно неразделимый случай
- Ядерный трюк
- Выбор и синтез ядер
- Регуляризация для метода опорных векторов

Функции ядра

Функция $K: X \times X \to \mathbb{R}$ — **функция ядра**, если её можно представить как $K(x, x') = \langle \psi(x), \psi(x') \rangle$ с отображением $\psi: X \to H$, где H — пространство со скалярным произведением.

Теорема (Мерсер)

Функция $K(x, x')$ — ядерная функция, если она симметричная, $K(x, x') = K(x', x)$, и неотрицательно определена на $\mathbb{R}$:

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

для любой функции $g: X \to \mathbb{R}$.

Примеры функций ядра

Квадратичное:

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

Многочлен с мономиальной степенью, равной d:

$$K(x, x') = \langle x, x' \rangle^d$$

Многочлен с мономиальной степенью $\leq d$:

$$K(x, x') = (\langle x, x' \rangle + 1)^d$$

Нейронные сети:

$$K(x, x') = \sigma(\langle x, x' \rangle)$$

Радиально-базисный:

$$K(x, x') = \exp(-\beta \|x - x'\|^2)$$

Синтез ядер

- $K(x, x') = \langle x, x' \rangle$ – функция ядра;
- константа $K(x, x') = 1$ – функция ядра;
- $K(x, x') = K_1(x, x')K_2(x, x')$ – функция ядра;
- $\forall \psi : X \to \mathbb{R}$ $K(x, x') = \psi(x)\psi(x')$ – функция ядра;
- $K(x, x') = \alpha_1 K_1(x, x') + \alpha_2 K_2(x, x')$ при $\alpha_1, \alpha_2 > 0$ – функция ядра;
- $\forall \phi : X \to X$ если K_0 – функция ядра, тогда $K(x, x') = K_0(\phi(x), \phi(x'))$ также является функцией ядра;
- если $s: X \times X \to \mathbb{R}$ – симметричная и интегрируемая, тогда

$$K(x, x') = \int_X s(x, z)s(x', z)dz$$ – функция ядра.

Анализ метода опорных векторов

Преимущество:

- Задача выпуклого квадратичного программирования имеет единственное решение
- Любая разделяющая поверхность
- Небольшое количество опорных объектов, используемых для обучения

Недостатки:

- Чувствителен к шуму
- Нет общих правил выбора функций ядра
- Константу C требуется выбирать
- Нет возможности выбора признаков

План лекции

- Линейно разделимый случай
- Линейно неразделимый случай
- Ядерный трюк
- Выбор и синтез ядер
- Регуляризация для метода опорных векторов

Регуляризация (напоминание)

Ключевая гипотеза: w «скачет», что и вызывает переобучение

Основная идея: ограничим норму w.

Добавим штраф регуляризации для нормы весов:

$$\mathcal{L}_\tau(a_w, \mathcal{D}) = \mathcal{L}(a_w, \mathcal{D}) + \frac{\tau}{2}\|w\|^2 \to \min_w.$$

Задача оптимизации для метода опорных векторов:

$$\sum_{i=1}^{|\mathcal{D}|}\big(1 - M_i(w, w_0)\big)_+ + \frac{1}{2C}\|w\|^2 \to \min_{w,w_0}$$

Другие функции штрафов

Вектор релевантности:

$$\frac{1}{2}\sum_{i=1}^{|\mathcal{D}|}\left(\ln\lambda_i + \frac{\alpha_i^2}{\lambda_i}\right)$$

LASSO SVM:

$$\mu\sum_{i=1}^{|\mathcal{D}|}|w_i|$$

Машина опорных признаков (Support feature machine):

$$\sum_{i=1}^{|\mathcal{D}|}R_\mu(w_i),$$

где μ — параметр селективности, $R_\mu = \begin{cases} 2\mu|w_i|, & \text{если } |w_i| < \mu, \\ \mu^2 + w_i^2, & \text{иначе.} \end{cases}$