Занятие 5
Линейные модели классификации.

Елена Кантонистова

ВШЭ, 2023

ПЛАН ЗАНЯТИЯ

1. Тест
2. Продвинутые метрики классификации
3. Логистическая регрессия и метод опорных векторов
4. Практика
5. Борьба с переобучением: регуляризация
6. Продолжение практики

ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ

Хотим предсказывать не только классы, но и ***вероятности классов***.

- Линейная регрессия:

$$a(x, w) = w_0 + w_1 x_1 + w_2 x_2 + \ \dots$$

- Линейный классификатор (любой):

$$a(x, w) = sign(w_0 + w_1 x_1 + w_2 x_2 + \ \dots)$$

ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ

Хотим предсказывать не только классы, но и ***вероятности классов***.

- Линейная регрессия:

$$a(x, w) = w_0 + w_1 x_1 + w_2 x_2 + \ \dots$$

- Линейный классификатор (любой):

$$a(x, w) = sign(w_0 + w_1 x_1 + w_2 x_2 + \ \dots)$$

- Логистическая регрессия:

$$\boldsymbol{a(x, w) = \sigma(w_0 + w_1 x_1 + w_2 x_2 + \ \dots) = \sigma(w, x)},$$

где $\sigma(z) = \frac{1}{1+e^{-z}}$ - сигмоида (логистическая функция)

ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ

Хотим предсказывать не классы, а вероятности классов.

- Логистическая регрессия: $\boldsymbol{a(x,w) = \sigma(w,x)}$,

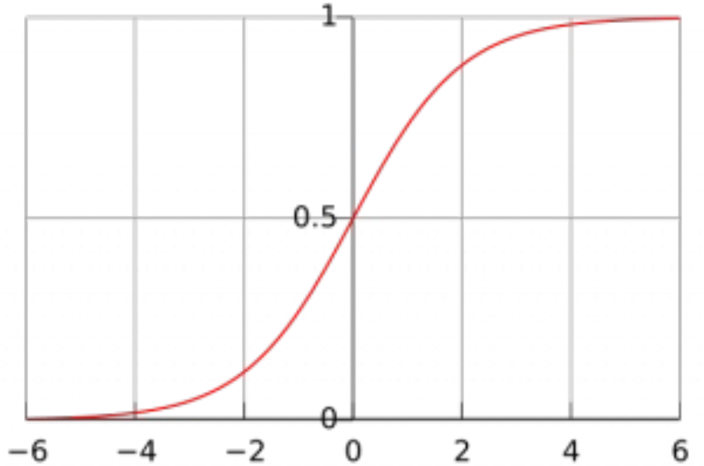
где $\sigma(z) = \frac{1}{1+e^{-z}}$ - сигмоида (логистическая функция),

$\sigma(z) \in (0;1)$.

Логистическая регрессия: $\boldsymbol{a(x,w) = \frac{1}{1+e^{-(w,x)}}}$

РАЗДЕЛЯЮЩАЯ ГРАНИЦА

Предсказываем $y = +1$, если $a(x, w) \geq 0.5$.



$a(x, w) = \sigma(w, x) \geq 0.5$, если $(w, x) \geq 0$.

Получаем, что

- $y = +1$ при $(w, x) \geq 0$
- $y = -1$ при $(w, x) < 0$,

т.е. $(w, x) = w_1 x_1 + w_2 x_2 + \cdots = 0$ – разделяющая гиперплоскость.

ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ

Логистическая регрессия - это линейный классификатор!

ФУНКЦИЯ ПОТЕРЬ ЛОГИСТИЧЕСКОЙ РЕГРЕССИИ

Если взять квадратичную функцию потерь
$L(a, y) = (a - y)^2$,

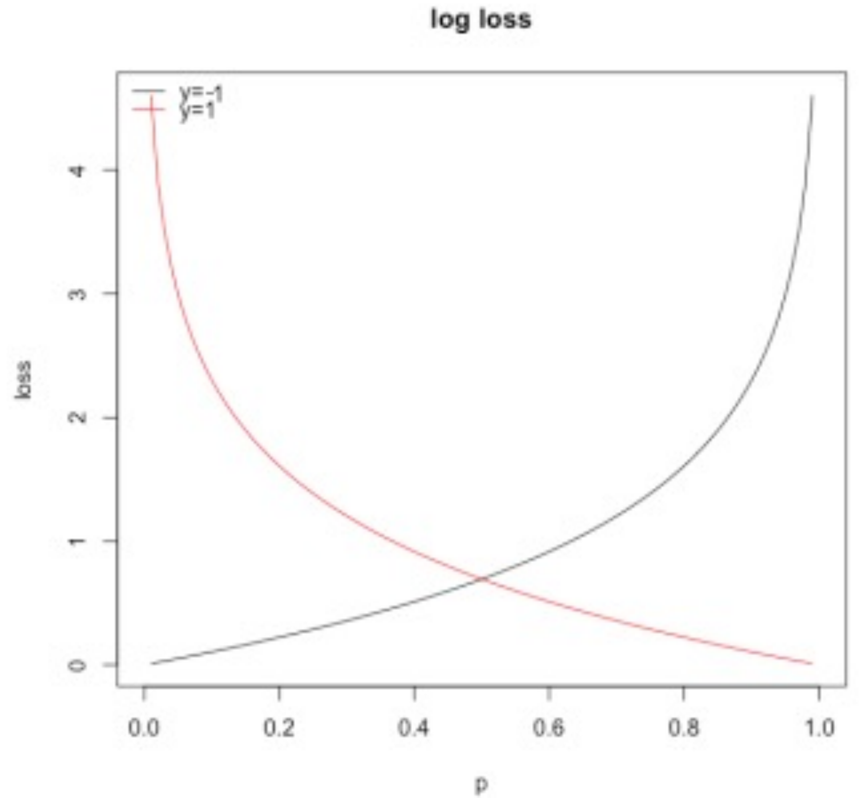
то возникнут проблемы:

- $Q(a, X) = \frac{1}{l} \sum_{i=1}^{l} \left(\frac{1}{1+e^{-w^T x}} - y \right)^2$ - не выпуклая функция (можем не попасть в глобальный минимум при оптимизации)
- На совсем неправильном предсказании маленький штраф (пусть предсказали вероятность 0% на объекте класса $y = +1$, тогда штраф всего $(1 - 0)^2 = 1$)

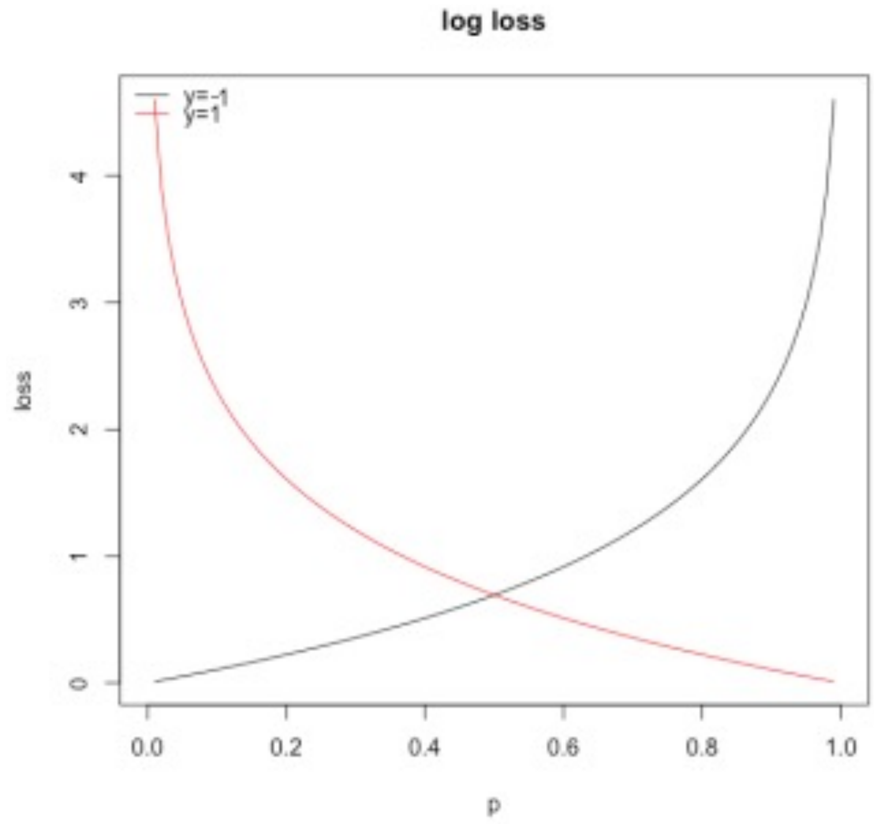
ФУНКЦИЯ ПОТЕРЬ ЛОГИСТИЧЕСКОЙ РЕГРЕССИИ

Возьмем логистическую функцию потерь (***log-loss***):

$$Q(w) = -\sum_{i=1}^{l}\left([y_i = +1] \cdot \log\left(a(x_i, w)\right) + [y_i = -1] \cdot \log(1 - a(x_i, w))\right)$$



ЛОГИСТИЧЕСКАЯ ФУНКЦИЯ ПОТЕРЬ

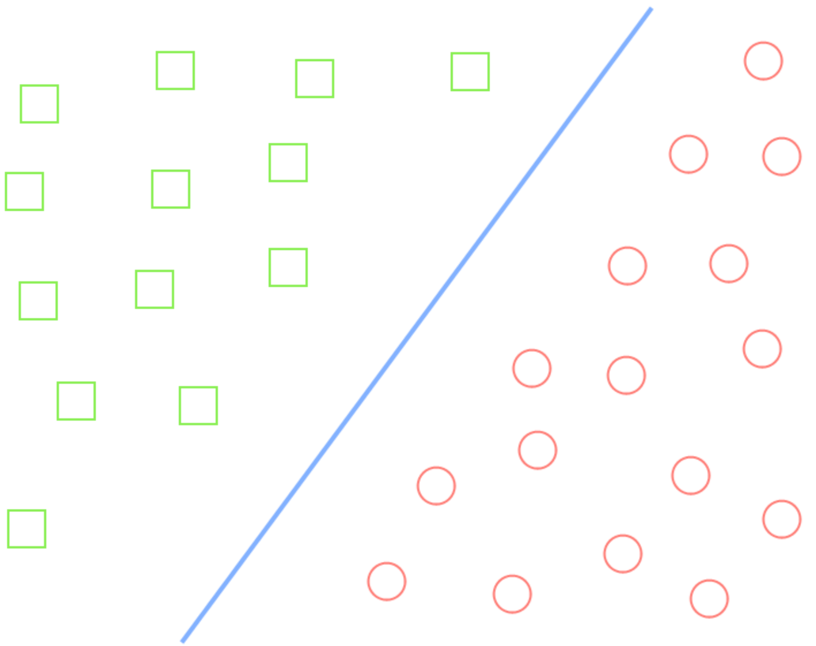


- если $a(x, w) = 1$ и $y = +1$, то штраф $L(a, y) = 0$
- если $a(x, w) \to 0$, а $y = +1$, то штраф $L(a, y) \to +\infty$

МЕТОД ОПОРНЫХ ВЕКТОРОВ

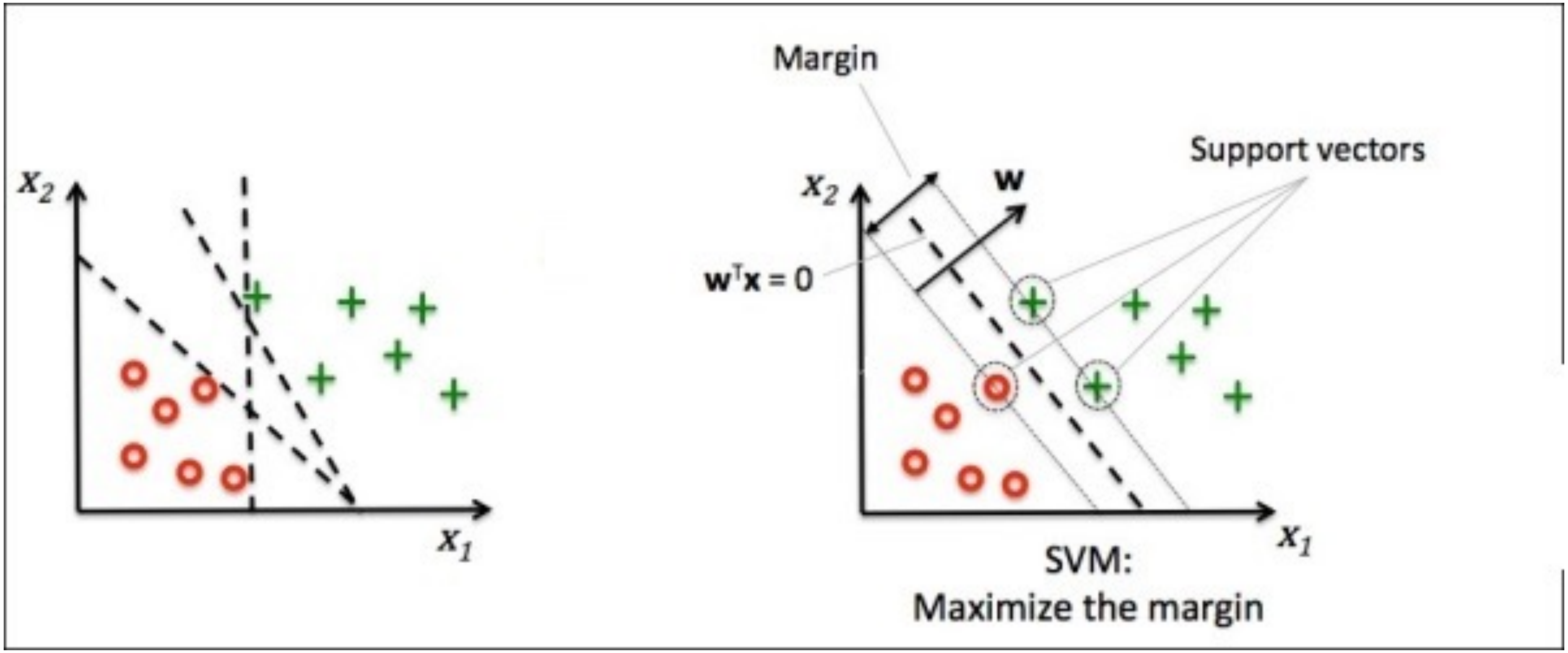
ЛИНЕЙНО РАЗДЕЛИМАЯ ВЫБОРКА

Выборка ***линейно разделима***, если существует такой вектор параметров w^*, что соответствующий классификатор $a(x)$ не допускает ошибок на этой выборке.



МЕТОД ОПОРНЫХ ВЕКТОРОВ: РАЗДЕЛИМЫЙ СЛУЧАЙ

Цель метода опорных векторов (Support Vector Machine) – максимизировать ширину разделяющей полосы.

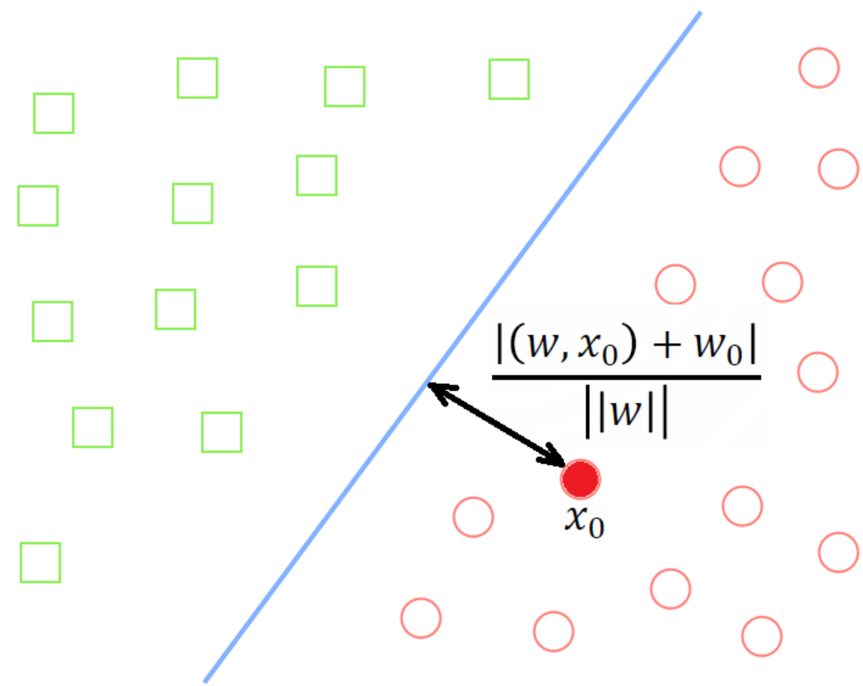


МЕТОД ОПОРНЫХ ВЕКТОРОВ: РАЗДЕЛИМЫЙ СЛУЧАЙ

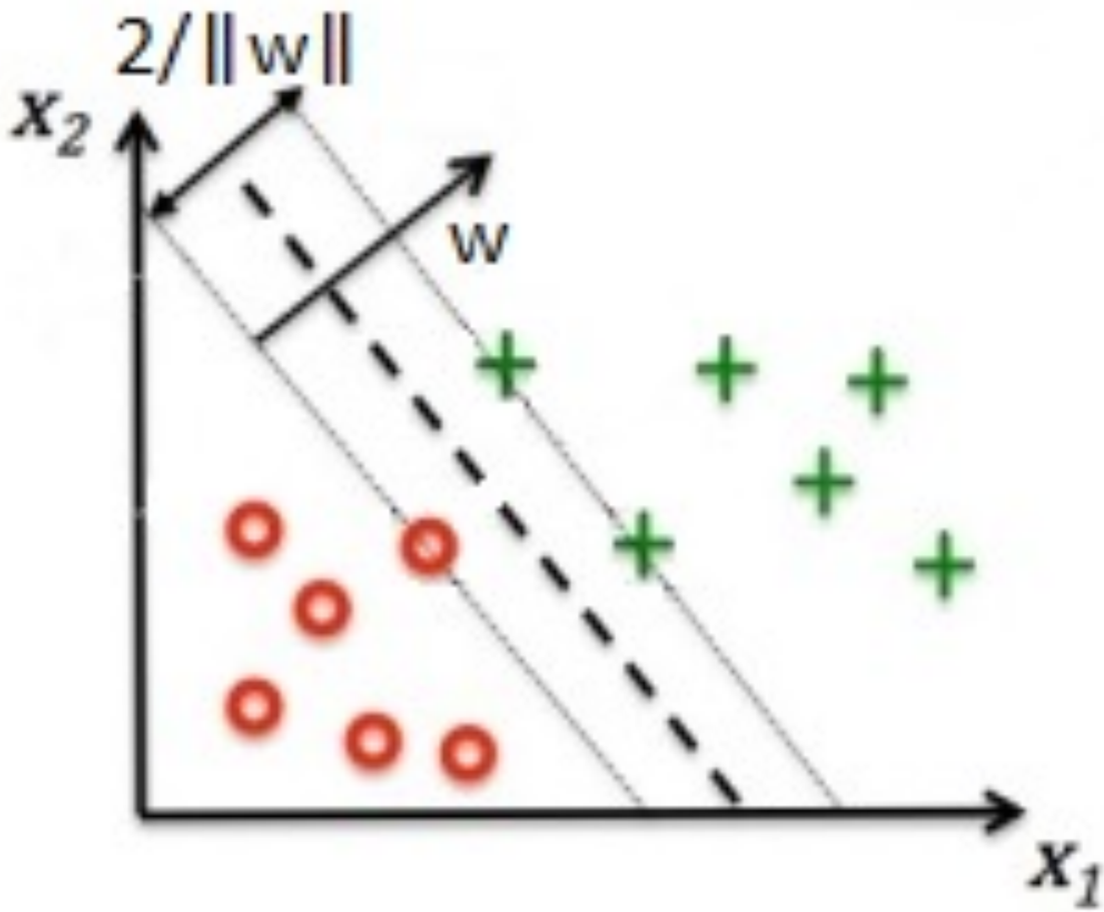
- $a(x) = sign\big((w, x) + w_0\big)$

Расстояние от разделяющей гиперплоскости, задаваемой классификатором до ближайшей точки выборки:

$$\rho = \frac{1}{||w||}$$



РАЗДЕЛЯЮЩАЯ ПОЛОСА



ОПТИМИЗАЦИОННАЯ ЗАДАЧА SVM ДЛЯ РАЗДЕЛИМОЙ ВЫБОРКИ

$$\begin{cases} \frac{1}{2}\left|\left|w\right|\right|^2 \to \min\limits_{w} \\ y_i\big((w, x_i) + w_0\big) \geq 1 \ (*), i = 1, \dots, l \end{cases}$$

(*) **–** эти неравенства означают, что все объекты попадают вне разделяющей полосы.

Утверждение. Данная оптимизационная задача имеет единственное решение.

ЛИНЕЙНО НЕРАЗДЕЛИМАЯ ВЫБОРКА

- Существует хотя бы один объект $x \in X$, что

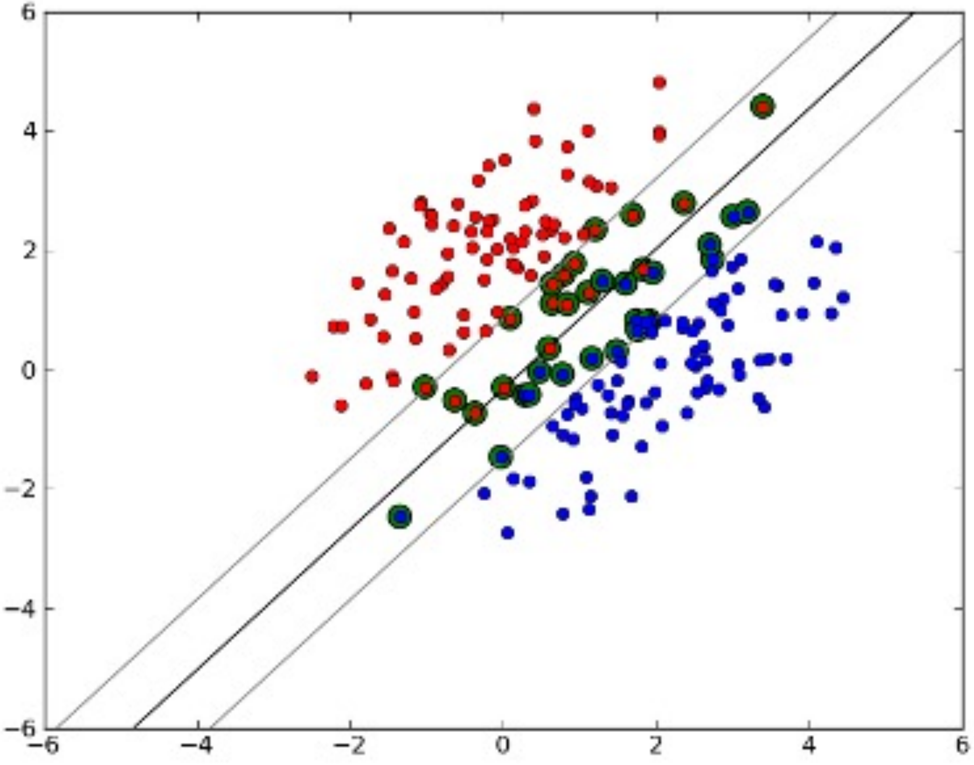
$$y_i\big((w, x_i) + w_0\big) < 1$$

То есть существует хотя бы один объект, попадающий внутрь разделяющей полосы.

ЛИНЕЙНО НЕРАЗДЕЛИМАЯ ВЫБОРКА

- Существует хотя бы один объект $x \in X$, что

$$y_i\big((w, x_i) + w_0\big) < 1$$



ЛИНЕЙНО НЕРАЗДЕЛИМАЯ ВЫБОРКА

- Будем штрафовать объекты, попадающие внутрь разделяющей полосы!
- ξ_i - штраф на i-м объекте (равен нулю на объектах, попадающих в свой класс вне разделяющей полосы)

МЕТОД ОПОРНЫХ ВЕКТОРОВ: НЕРАЗДЕЛИМЫЙ СЛУЧАЙ

Хотим:

- Минимизировать штрафы $\sum_{i=1}^{l} \xi_i$
- Максимизировать отступ $\frac{1}{||w||}$

МЕТОД ОПОРНЫХ ВЕКТОРОВ: НЕРАЗДЕЛИМЫЙ СЛУЧАЙ

Хотим:

- Минимизировать штрафы $\sum_{i=1}^{l} \xi_i$
- Максимизировать отступ $\frac{1}{||w||}$

Задача оптимизации:

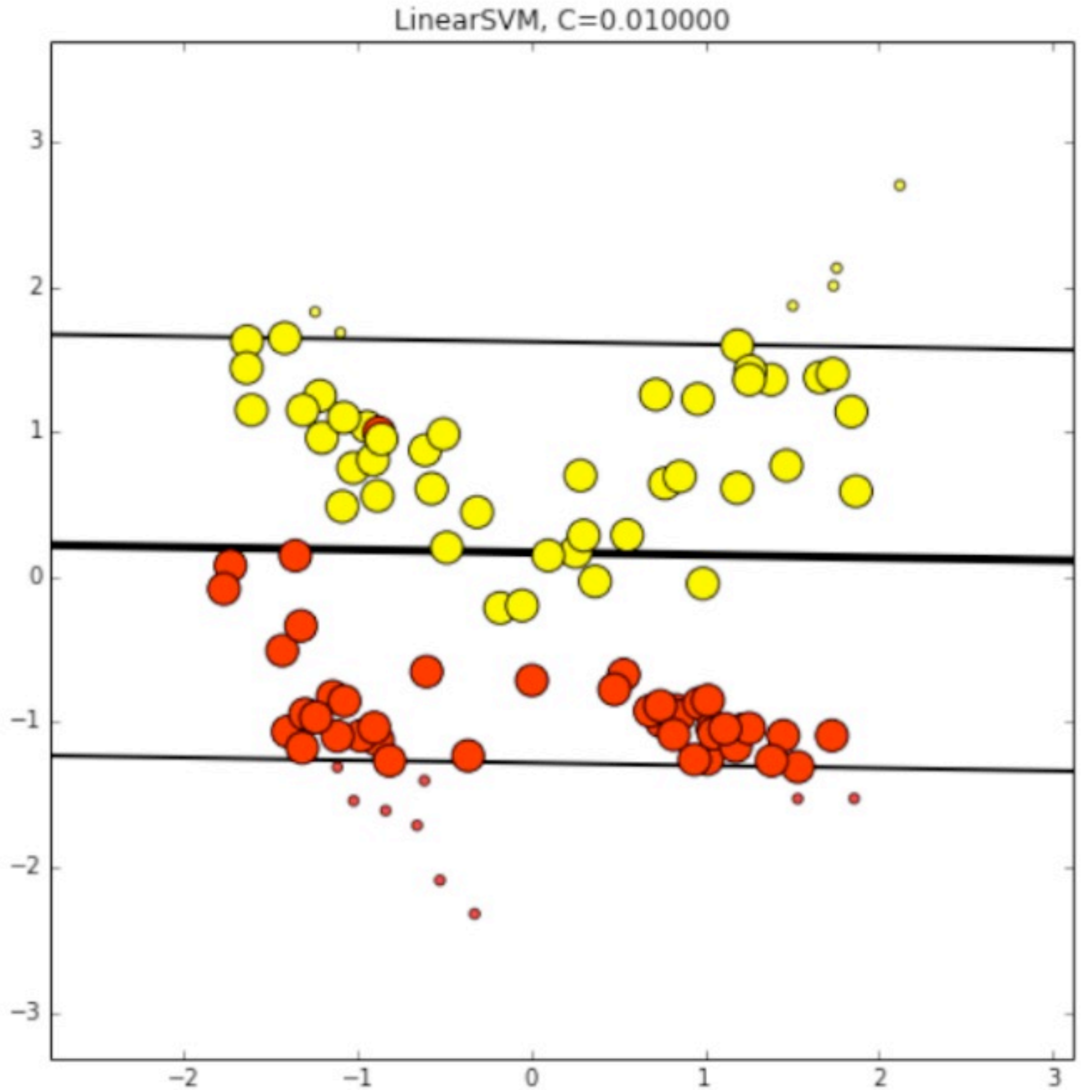
$$\frac{1}{2}\left||w|\right|^2 + C\sum_{i=1}^{l} \xi_i \to \min_{w, w_0, \xi_i}$$

ЗНАЧЕНИЕ КОНСТАНТЫ C

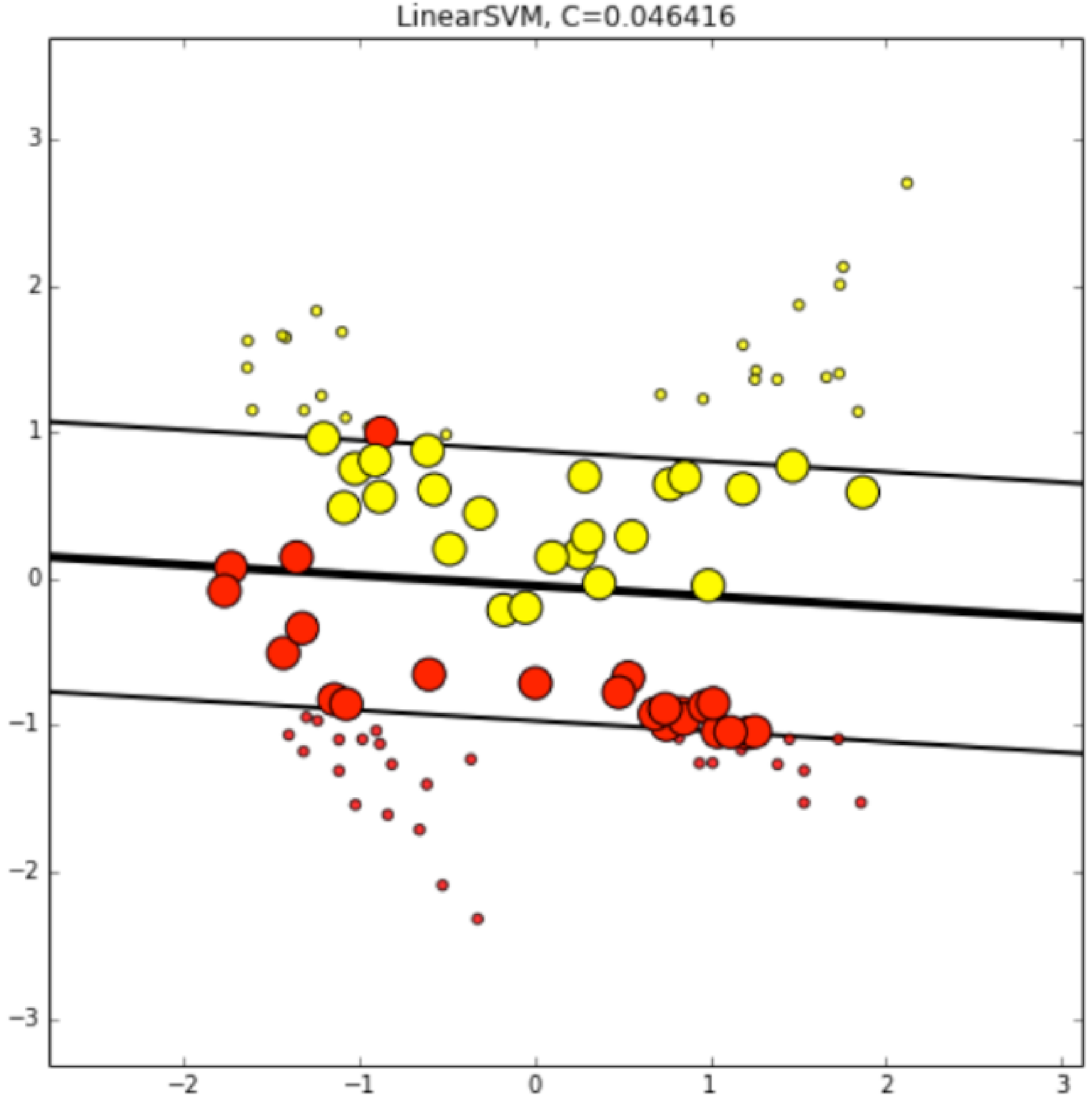
$$\frac{1}{2}\left|\left|w\right|\right|^{2} + C\sum_{i=1}^{l}\xi_i \to \min_{w,w_0,\xi_i}$$

Положительная константа C является гиперпараметром метода и позволяет находить компромисс между максимизацией разделяющей полосы и минимизацией суммарной ошибки.

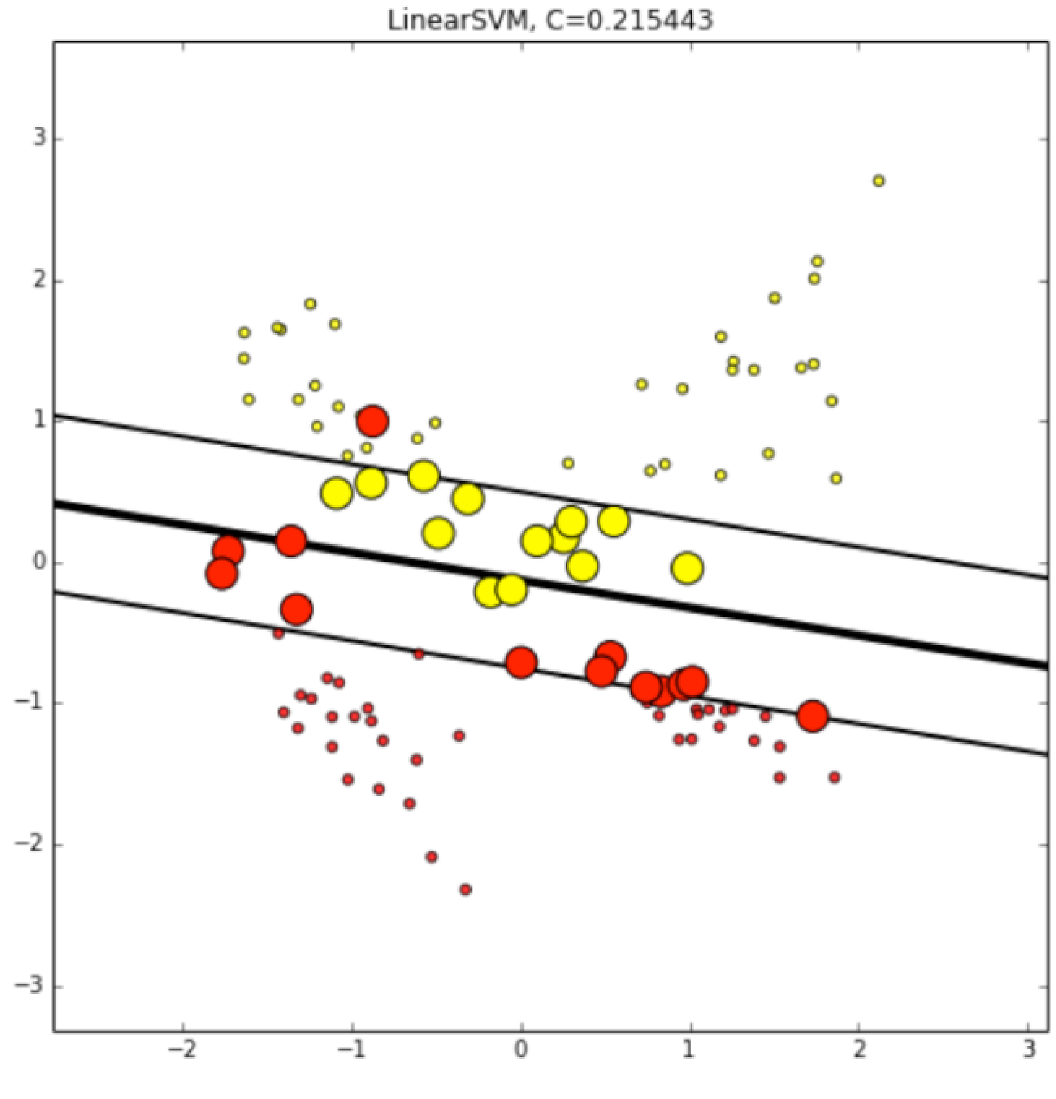
ЗНАЧЕНИЕ КОНСТАНТЫ C



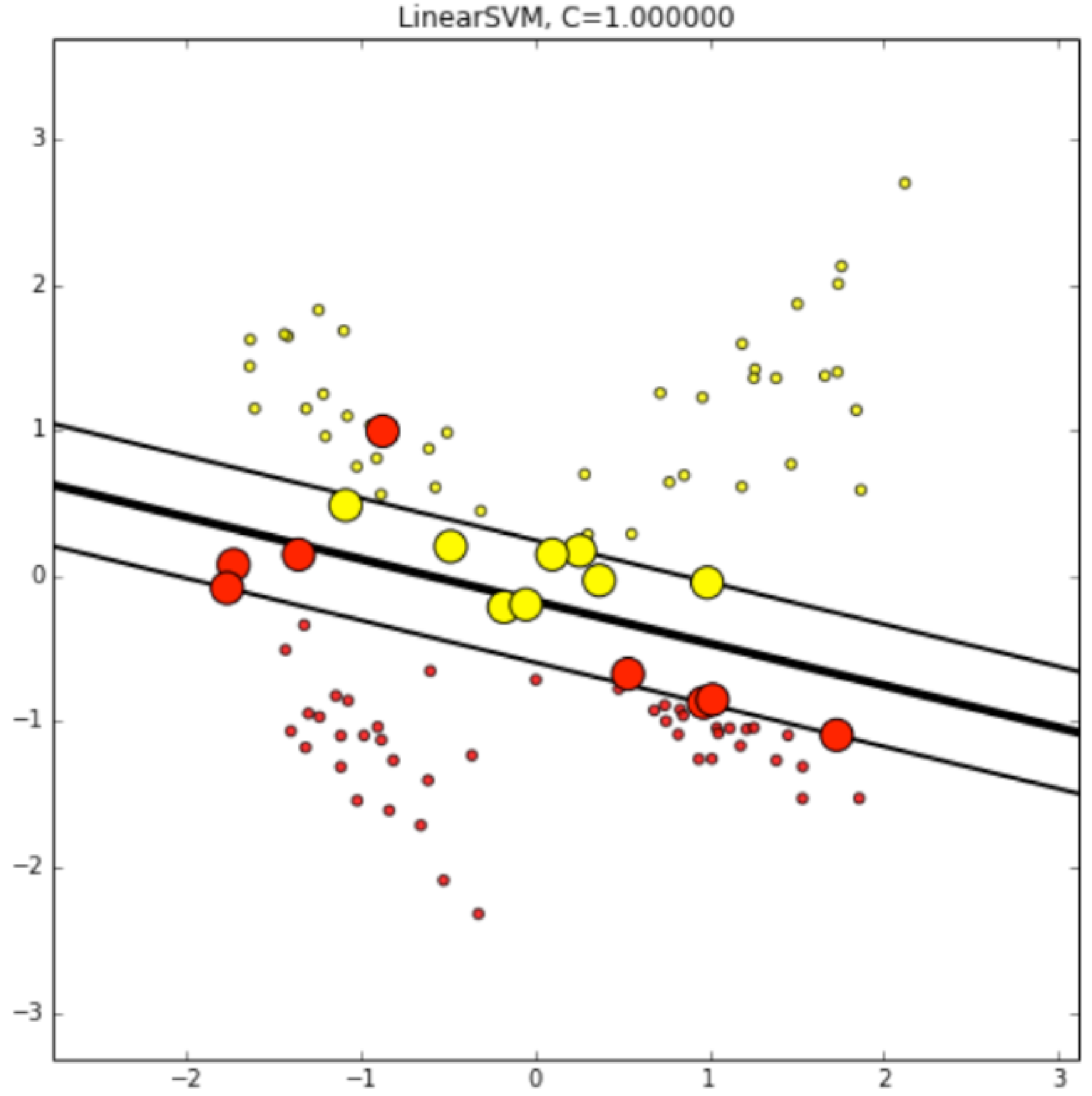
ЗНАЧЕНИЕ КОНСТАНТЫ C



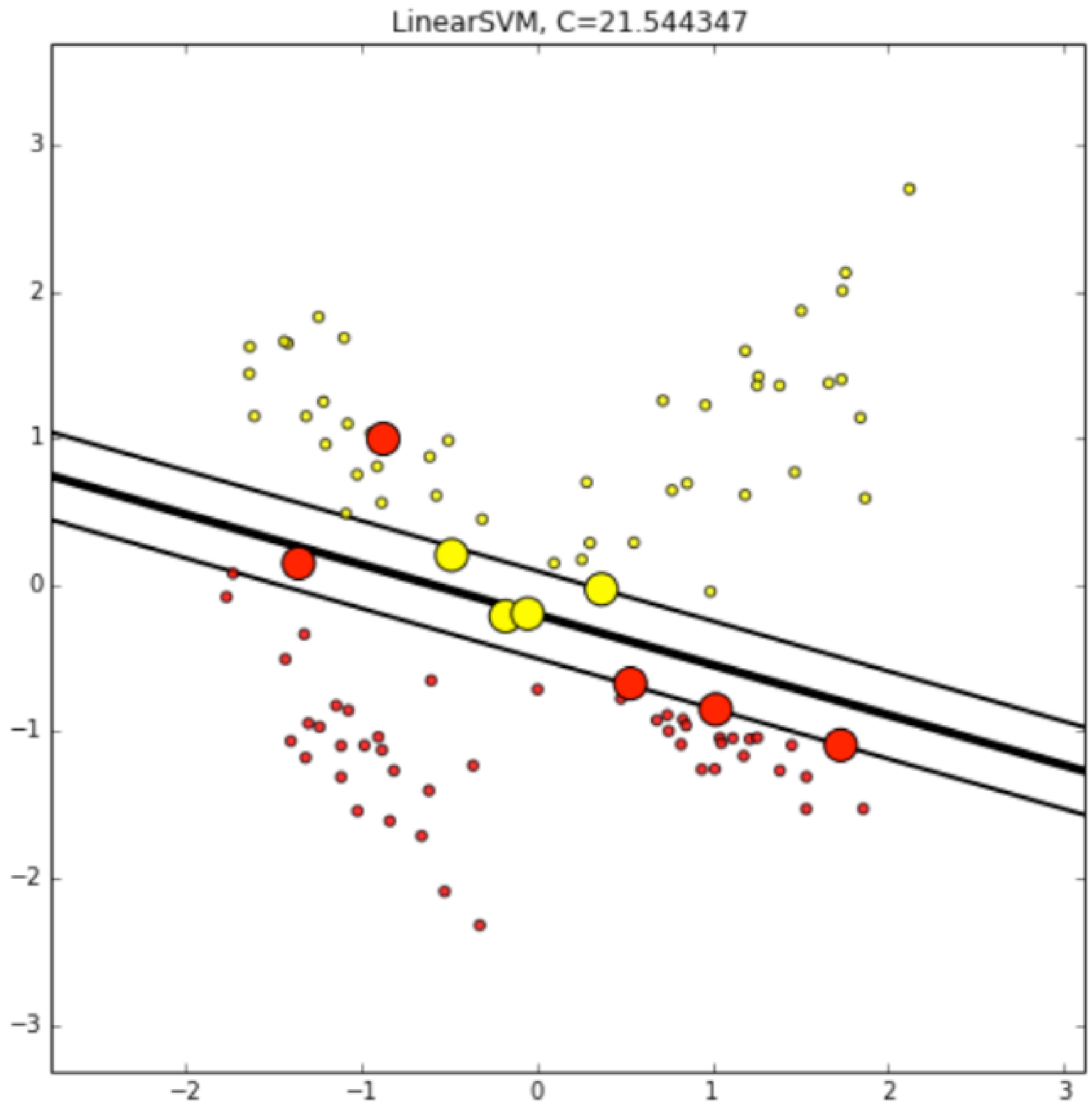
ЗНАЧЕНИЕ КОНСТАНТЫ C



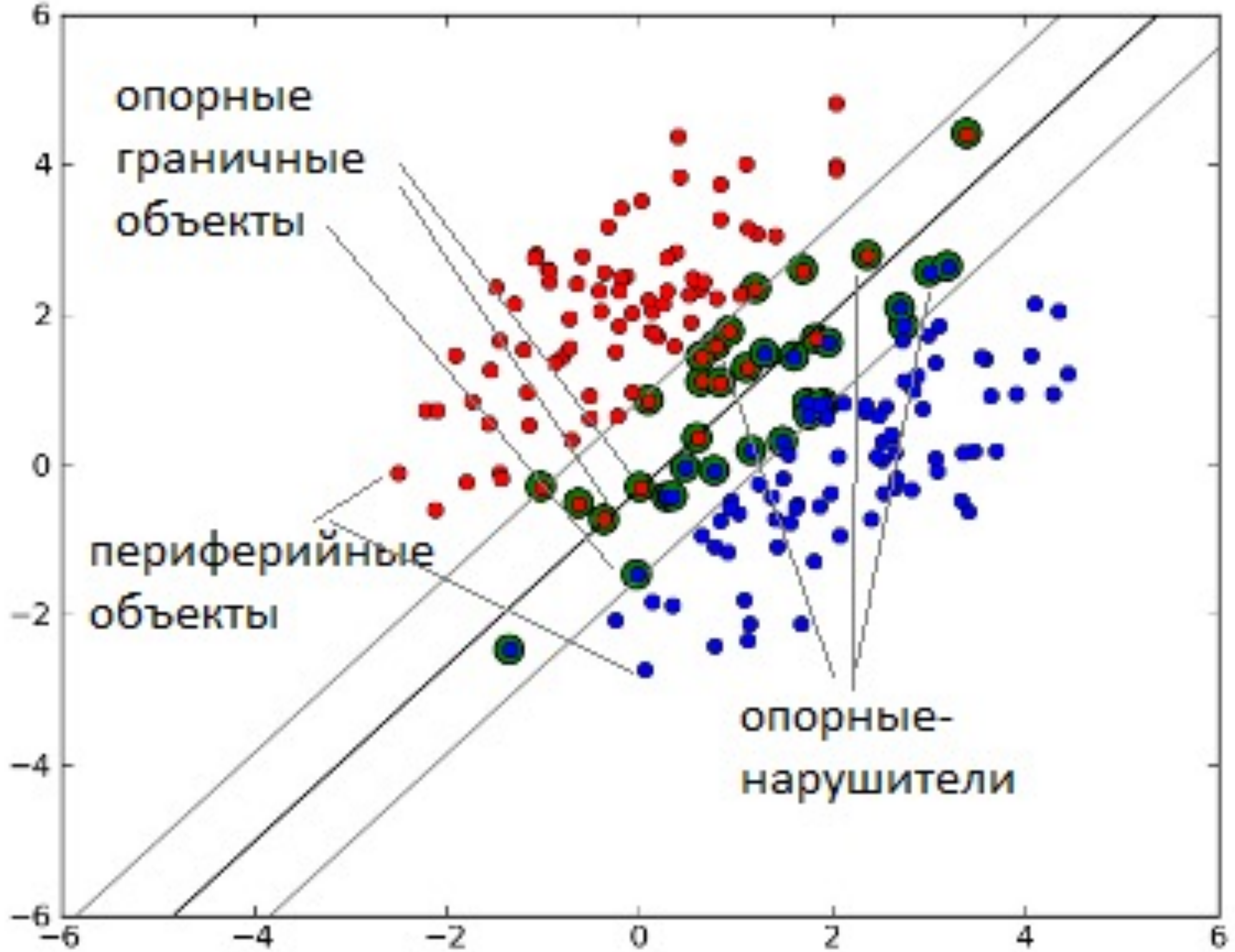
ЗНАЧЕНИЕ КОНСТАНТЫ C



ЗНАЧЕНИЕ КОНСТАНТЫ C



ТИПЫ ОБЪЕКТОВ В SVM



ПЕРЕОБУЧЕНИЕ И РЕГУЛЯРИЗАЦИЯ

МЕТОД БОРЬБЫ С ПЕРЕОБУЧЕНИЕМ: РЕГУЛЯРИЗАЦИЯ

Большие значения параметров (весов) модели w – признак переобучения.

P.S. Если в данных есть линейно-зависимые признаки, они тоже приводят к переобучению (и к большим весам).

МЕТОД БОРЬБЫ С ПЕРЕОБУЧЕНИЕМ: РЕГУЛЯРИЗАЦИЯ

Большие значения параметров (весов) модели w – признак переобучения.

Решение проблемы **–** ***регуляризация.***

Будем минимизировать регуляризованный функционал ошибки:

$$\boldsymbol{Q_{alpha}(w) = Q(w) + \alpha \cdot R(w) \to \min_{w}} \quad ,$$

где $R(w)$ - регуляризатор.

РЕГУЛЯРИЗАЦИЯ

- **Регуляризация штрафует за слишком большие веса.**

Наиболее используемые регуляризаторы:

- L_2-регуляризатор: $R(w) = \left||w|\right|_2 = \sum_{i=1}^{d} w_i^2$
- L_1-регуляризатор: $R(w) = \left||w|\right|_1 = \sum_{i=1}^{d} |w_i|$

РЕГУЛЯРИЗАЦИЯ

- **Регуляризация штрафует за слишком большие веса.**

Наиболее используемые регуляризаторы:

- L_2-регуляризатор: $R(w) = ||w||_2 = \sum_{i=1}^{d} w_i^2$
- L_1-регуляризатор: $R(w) = ||w||_1 = \sum_{i=1}^{d} |w_i|$

Пример регуляризованного функционала:

$$Q(a(w), X) = \frac{1}{l}\sum_{i=1}^{l}((w, x_i) - y_i)^2 + \alpha \sum_{i=1}^{d} w_i^2 ,$$

где α – коэффициент регуляризации.

ПОЛЕЗНОЕ СВОЙСТВО L1-РЕГУЛЯРИЗАЦИИ

Все ли признаки в задаче нужны?

- Некоторые признаки могут не иметь отношения к задаче, т.е. они не нужны.
- Если есть ограничения на скорость получения предсказаний, то чем меньше признаков, тем быстрее
- Если признаков больше, чем объектов, то решение задачи будет неоднозначным.

Поэтому в таких случаях надо делать отбор признаков, то есть убирать некоторые признаки.

L_1-РЕГУЛЯРИЗАЦИЯ

Утверждение. В результате обучения модели с L_1-регуляризатором происходит зануление некоторых весов, т.е. отбор признаков.

Можно показать, что задачи

$$(1) \quad Q(w) + \alpha \left|\left|w\right|\right|_1 \to \min_w$$

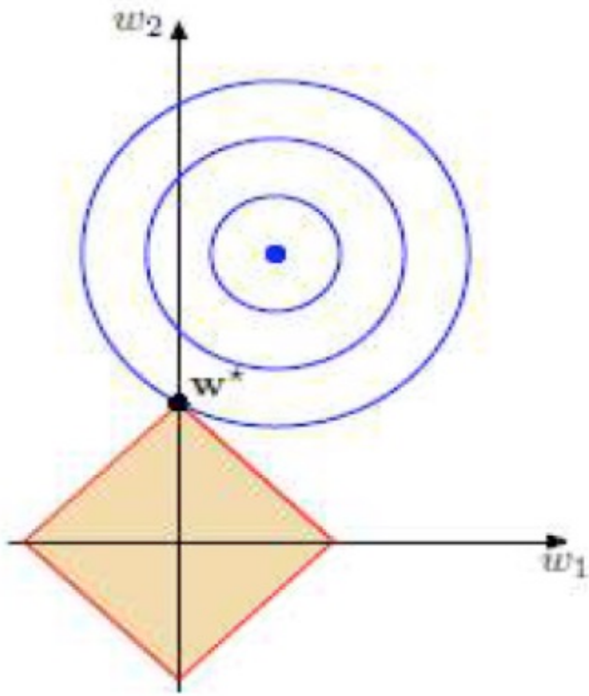
и

$$(2) \quad \begin{cases} Q(w) \to \min\limits_w \\ \left|\left|w\right|\right|_1 \le C \end{cases}$$

эквивалентны.

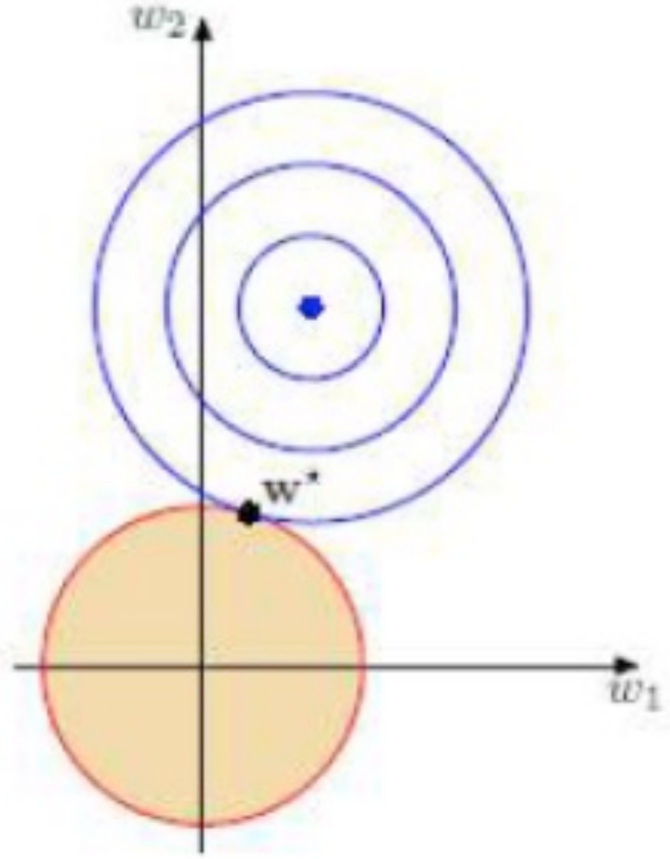
ОТБОР ПРИЗНАКОВ ПО L1-РЕГУЛЯРИЗАЦИИ

Нарисуем линии уровня $Q(w)$ и область $||w||_1 \leq C$:



Если признак незначимый, то соответствующий вес близок к 0. Отсюда получим, что в большинстве случаев решение нашей задачи попадает в вершину ромба, т.е. обнуляет незначимый признак.

L2-РЕГУЛЯРИЗАЦИЯ НЕ ОБНУЛЯЕТ ПРИЗНАКИ



РАЗРЕЖЕННЫЕ МОДЕЛИ

Модели, в которых часть весов равна 0, называются ***разреженными моделями***.

- L1-регуляризация зануляет часть весов, то есть делает модель разреженной.