Лекция 5
Линейные модели классификации. Часть 2.

Кантонистова Е.О.

ВШЭ, 2018

ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ

Логистическая регрессия – линейный классификатор, корректно предсказывающий вероятности классов.

ВЕРОЯТНОСТНАЯ ПОСТАНОВКА ЗАДАЧИ

Предположение: В каждой точке x пространства объектов задана вероятность $p(y = +1|x)$

Объекты с одинаковым признаковым описанием могут иметь разные значения целевой переменной.

ВЕРОЯТНОСТНАЯ ПОСТАНОВКА ЗАДАЧИ

Предположение: В каждой точке x пространства объектов задана вероятность $p(y = +1|x)$

Объекты с одинаковым признаковым описанием могут иметь разные значения целевой переменной.

Цель: построить алгоритм $b(x)$, в каждой точке x предсказывающий $p(y = +1|x)$.

ВЕРОЯТНОСТНАЯ ПОСТАНОВКА ЗАДАЧИ

- Пусть объект x встречается в выборке n раз с ответами $\{y_1, \dots, y_n\}$. Хотим, чтобы алгоритм выдавал вероятность классов:

$$b_* = \underset{b \in \mathbb{R}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} L(y_i, b) \approx p(y = +1|x)$$

По ЗБЧ при $n \to \infty$ получаем

$$b_* = \underset{b \in \mathbb{R}}{\operatorname{argmin}} E[L(y, b)|x] = p(y = +1|x)$$

ФУНКЦИИ ПОТЕРЬ

Подходят:

- **Квадратичная**

$$L(y,z) = (y-z)^2$$

- **Логистическая**

$$L(y,z) = [y=+1]\cdot\log\big(b(x,w)\big) + [y=-1]\cdot\log(1-b(x,w))$$

Не подходят:

- **Модуль**

$$L(y,z) = |\mathrm{y}-\mathrm{z}|$$

ПРАВДОПОДОБИЕ И LOG-LOSS

- Вероятности, которые выдает алгоритм $b(x)$, должны согласовываться с выборкой
- Вероятность того, что в выборке встретится объект x с классом y:

$$b(x)^{[y=+1]} \cdot \big(1 - b(x)\big)^{[y=-1]}$$

ПРАВДОПОДОБИЕ И LOG-LOSS

- Вероятности, которые выдает алгоритм $b(x)$, должны согласовываться с выборкой
- Вероятность того, что в выборке встретится объект x с классом y:

$$b(x)^{[y=+1]} \cdot \big(1 - b(x)\big)^{[y=-1]}$$

Правдоподобие выборки:

$$(b, X) = \prod_{i=1}^{l} b(x_i)^{[y_i=+1]} \cdot \big(1 - b(x_i)\big)^{[y_i=-1]}$$

ФУНКЦИЯ ПОТЕРЬ ДЛЯ ОБУЧЕНИЯ

- Можно максимизировать правдоподобие

$$(b, X) = \prod_{i=1}^{l} b(x_i)^{[y_i=+1]} \cdot \big(1 - b(x_i)\big)^{[y_i=-1]} \to \max_b$$

- Или, что эквивалентно (**логарифмическая, log-loss**):

$$-\sum_{i=1}^{l} ([y_i = +1] \log b(x_i) + [y_i = -1] \log(1 - b(x_i))) \to \min_b$$

ФУНКЦИЯ ПОТЕРЬ ДЛЯ ОБУЧЕНИЯ

- Можно максимизировать правдоподобие

$$(b, X) = \prod_{i=1}^{l} b(x_i)^{[y_i=+1]} \cdot \big(1 - b(x_i)\big)^{[y_i=-1]} \to \max_b$$

- Или, что эквивалентно (**логарифмическая, log-loss**):

$$-\sum_{i=1}^{l} ([y_i = +1] \log b(x_i) + [y_i = -1] \log(1 - b(x_i))) \to \min_b$$

Утверждение. Логарифмическая функция потерь корректно предсказывает вероятности.

ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ

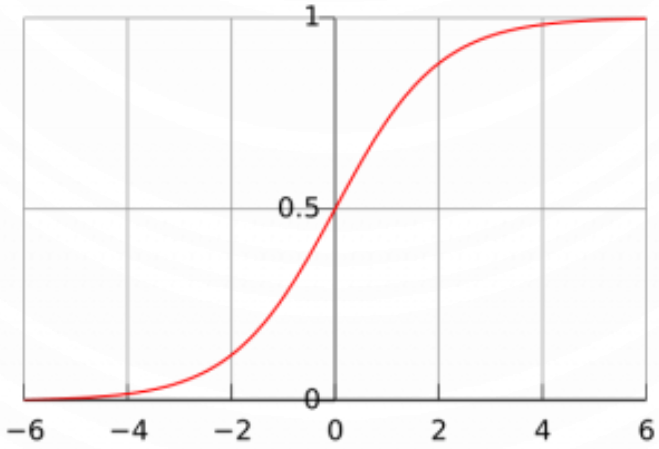
- Хотим, чтобы алгоритм $b(x)$ возвращал числа из отрезка $[0, 1]$.

ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ

- Хотим, чтобы алгоритм $b(x)$ возвращал числа из отрезка $[0, 1]$.
- Можно взять $b(x) = \sigma(w^T x)$, где σ – любая монотонно неубывающая функция с областью значений $[0, 1]$.

ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ

- Хотим, чтобы алгоритм $b(x)$ возвращал числа из отрезка $[0, 1]$.
- Можно взять $b(x) = \sigma(w^T x)$, где σ – любая монотонно неубывающая функция с областью значений $[0, 1]$.
- Возьмем ***сигмоиду***: $\sigma(z) = \frac{1}{1+e^{-z}}$



ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ

- $p(y = +1|x) = \frac{1}{1+e^{-w^T x}}$, следовательно,
- $(w, x) = w^T x = \log \frac{p(y=+1|x)}{p(y=-1|x)}$ - логарифм отношения вероятностей классов.

Утверждение. Логарифмическая функция потерь может быть записана в виде

$$\boldsymbol{L(b, X) = \sum_{i=1}^{l} \log(1 + e^{-y_i(w,x)})}$$

МЕТОД ОПОРНЫХ ВЕКТОРОВ

ЛИНЕЙНЫЙ КЛАССИФИКАТОР

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

Ошибка линейного классификатора:

- $Q(a, X) = \frac{1}{l}\sum_{i=1}^{l}[a(x_i) \neq y_i] =$

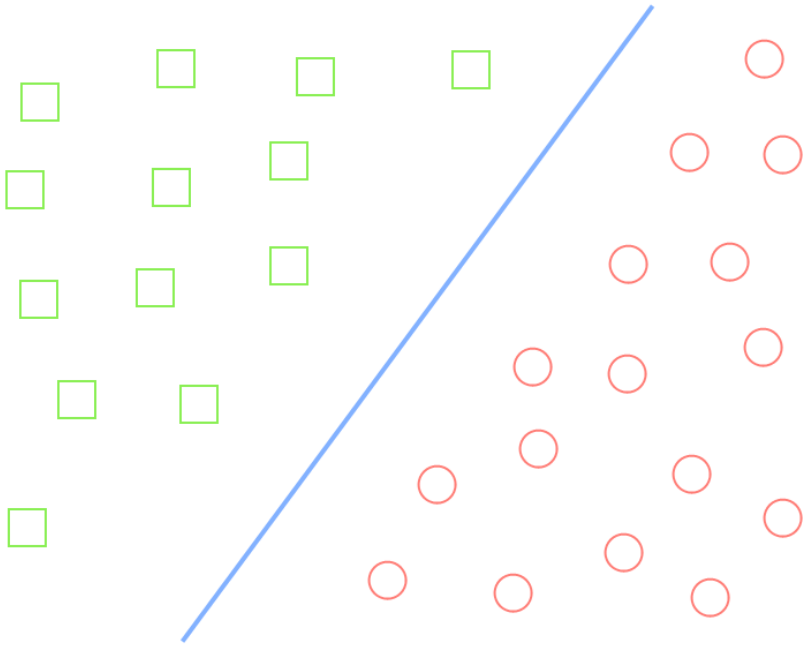
$$= \frac{1}{l}\sum_{i=1}^{l}\big[sign\big((w, x_i) + w_0\big) \neq y_i\big] =$$

$$= \frac{1}{l}\sum_{i=1}^{l}[\boldsymbol{y_i \cdot ((w, x_i) + w_0)} < 0] \to \min_{w, w_0}$$

$M_i = y_i((w, x_i) + w_0)$ – отступ на объекте

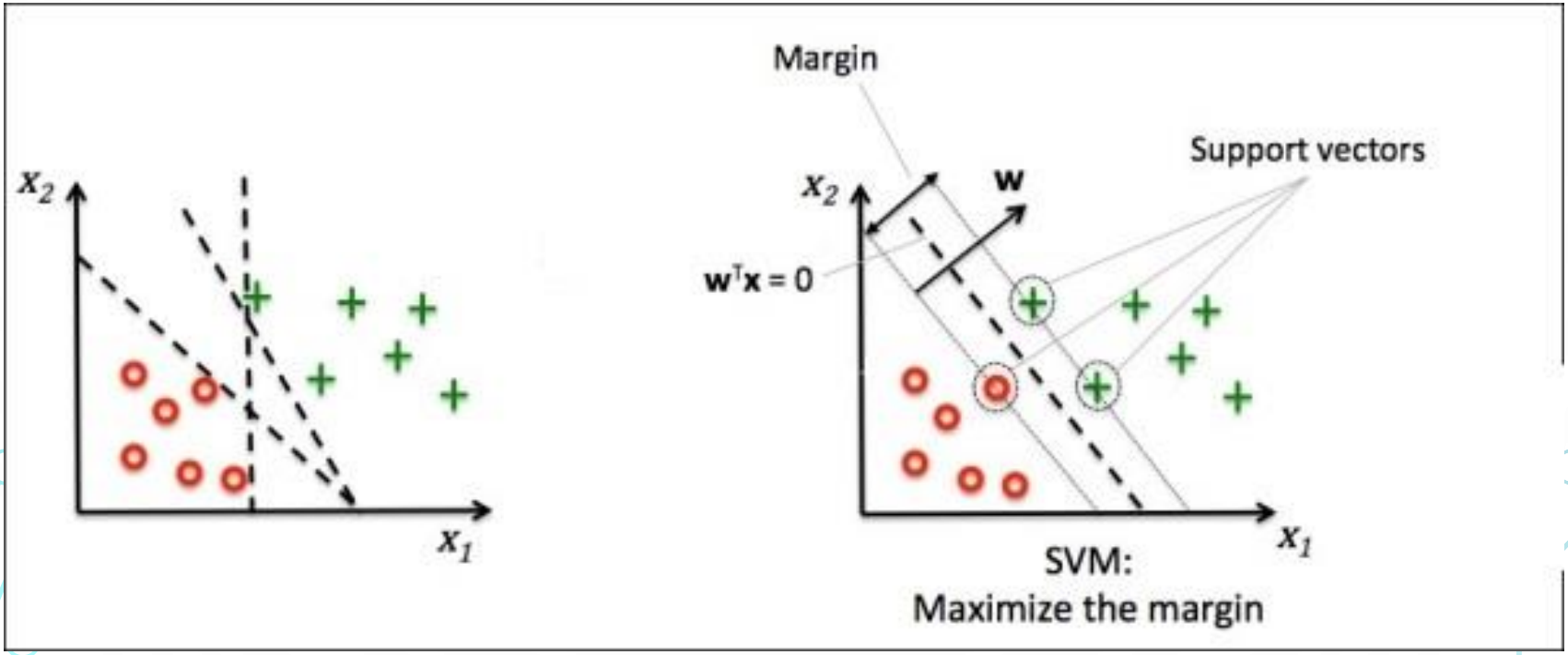
ЛИНЕЙНО РАЗДЕЛИМАЯ ВЫБОРКА

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

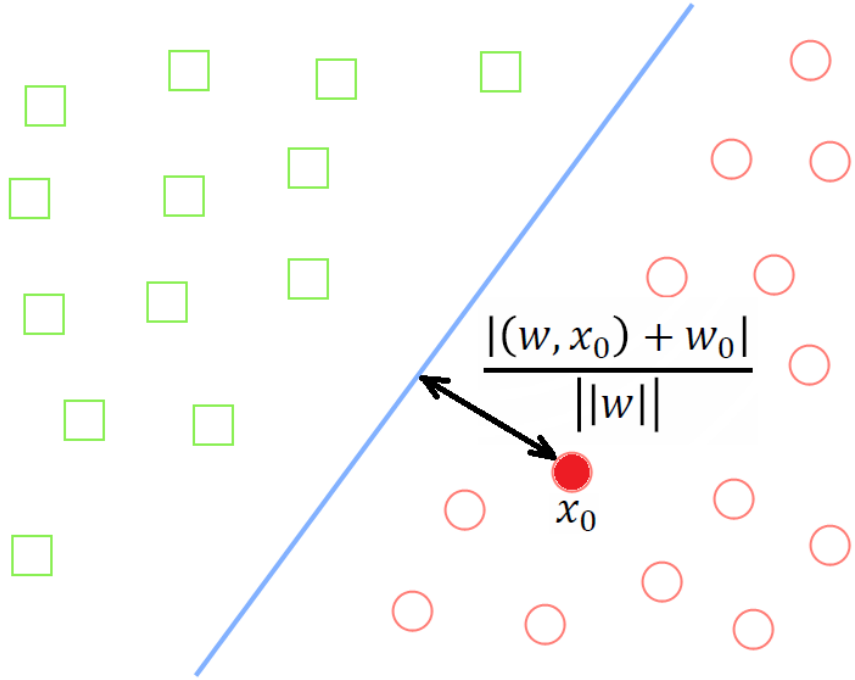
- $a(x) = sign\big((w, x) + w_0\big)$
- Нормируем параметры w и w_0 так, что

$$\min_{x \in X} |(w, x) + w_0| = 1$$

Расстояние от точки x_0 до разделяющей гиперплоскости, задаваемой

классификатором:

$$\rho(x_0, a) = \frac{|(w, x_0) + w_0|}{||w||}$$



МЕТОД ОПОРНЫХ ВЕКТОРОВ: РАЗДЕЛИМЫЙ СЛУЧАЙ

- Нормируем параметры w и w_0 так, что

$$\min_{x \in X} |(w, x) + w_0| = 1$$

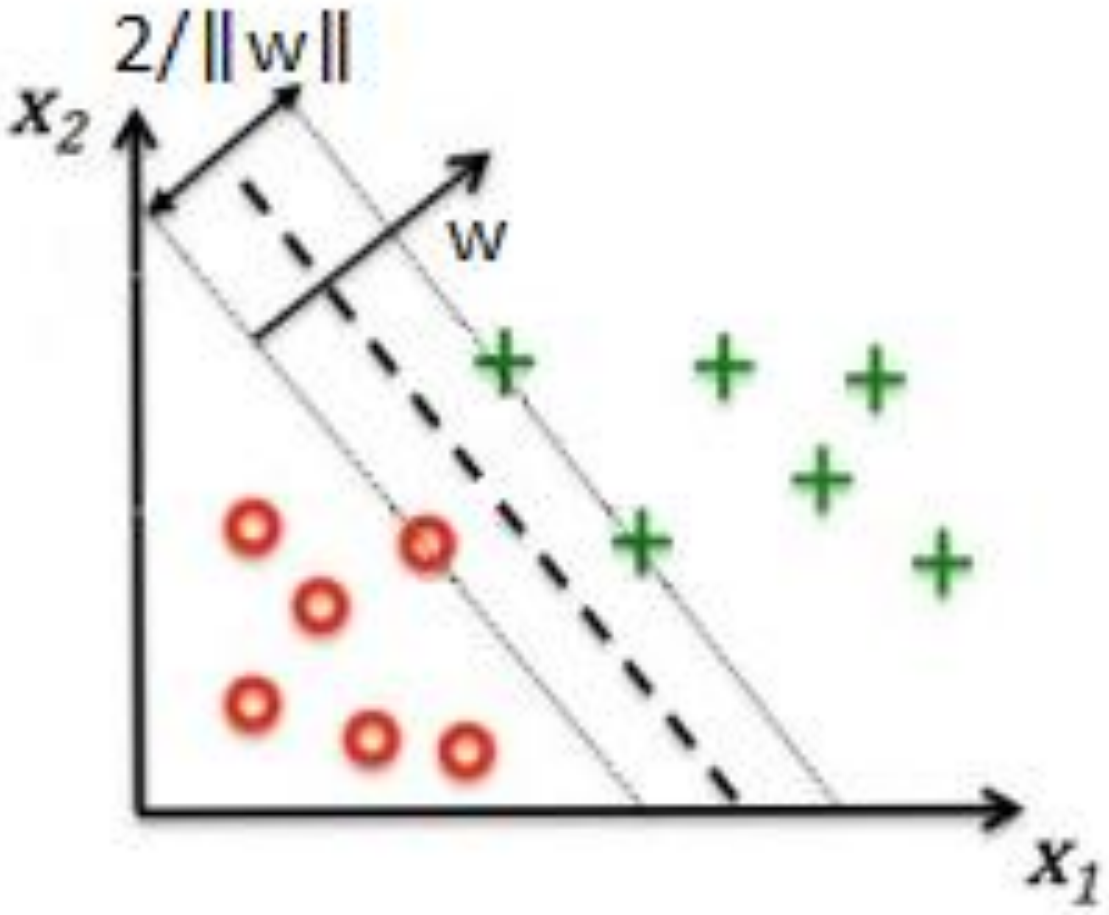
Тогда расстояние от точки x_0 до разделяющей гиперплоскости, задаваемой классификатором:

$$\rho(x_0, a) = \frac{|(w, x_0) + w_0|}{||w||}$$

- Расстояние до ближайшего объекта $x \in X$:

$$\min_{x \in X} \frac{|(w, x) + w_0|}{||w||} = \frac{1}{||w||} \min_{x \in X} |(w, x) + w_0| = \frac{\mathbf{1}}{||\boldsymbol{w}||}$$

РАЗДЕЛЯЮЩАЯ ПОЛОСА



ОПТИМИЗАЦИОННАЯ ЗАДАЧА SVM ДЛЯ РАЗДЕЛИМОЙ ВЫБОРКИ

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

Утверждение. Данная оптимизационная задача имеет единственное решение.

ЛИНЕЙНО НЕРАЗДЕЛИМАЯ ВЫБОРКА

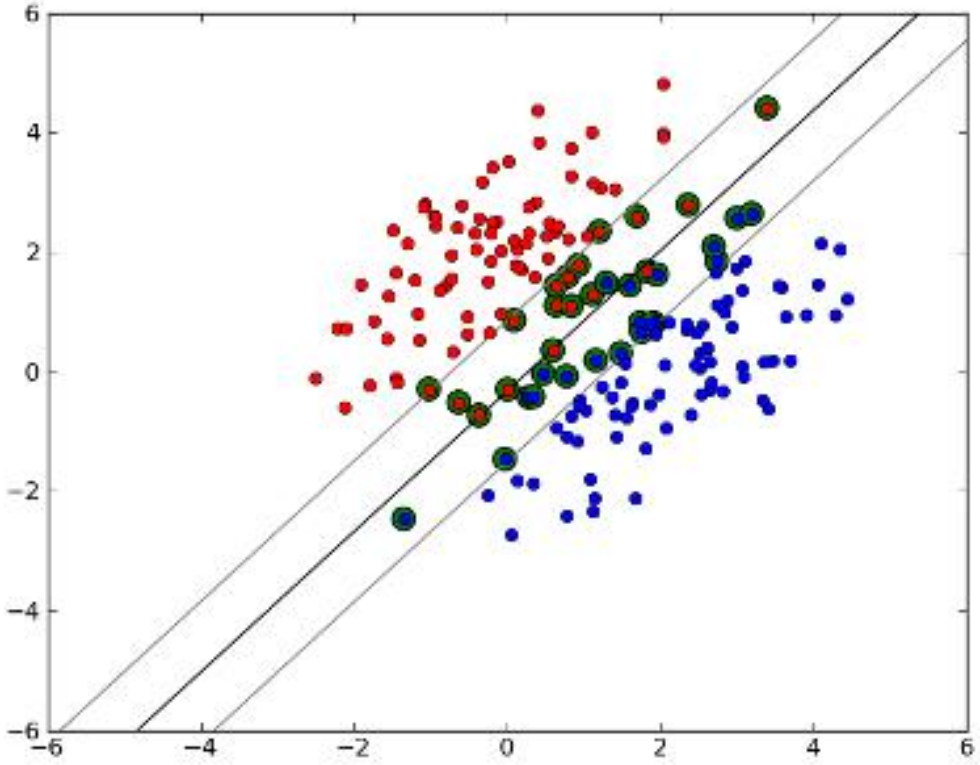
- Существует хотя бы один объект $x \in X$, что

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

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

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

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

Смягчим ограничения, введя штрафы $\xi_i \geq 0$:

$$y_i\big((w, x_i) + w_0\big) \geq 1 - \xi_i, i = 1, \dots, l$$

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:

$$\begin{cases} \frac{1}{2}||w||^2 + C\sum_{i=1}^{l} \xi_i \to \min\limits_{w, w_0, \xi_i} \\ y_i\big((w, x_i) + w_0\big) \geq 1 - \xi_i, i = 1, \dots, l \\ \xi_i \geq 0, i = 1, \dots, l \end{cases}$$

МЕТОД ОПОРНЫХ ВЕКТОРОВ: НЕРАЗДЕЛИМЫЙ СЛУЧАЙ

Утверждение. Задача

$$\begin{cases} \frac{1}{2}||w||^2 + C\sum_{i=1}^{l}\xi_i \to \min\limits_{w,w_0,\xi_i} \\ y_i\big((w,x_i)+w_0\big) \geq 1-\xi_i, i=1,\dots,l \\ \xi_i \geq 0, i=1,\dots,l \end{cases}$$

Является выпуклой и имеет единственное решение.

СВЕДЕНИЕ К БЕЗУСЛОВНОЙ ЗАДАЧЕ

$$\begin{cases} \frac{1}{2}\left|\left|w\right|\right|^2 + C\sum_{i=1}^{l}\xi_i \to \min\limits_{w,w_0,\xi_i} (1) \\ y_i\big((w,x_i)+w_0\big) \geq 1-\xi_i, i=1,\dots,l\ (2) \\ \xi_i \geq 0, i=1,\dots,l\ (3) \end{cases}$$

- Перепишем (2) и (3):

$$\begin{cases} \xi_i \geq 1 - y_i\big((w,x_i)+w_0\big) = 1 - M_i \\ \xi_i \geq 0 \end{cases}$$

СВЕДЕНИЕ К БЕЗУСЛОВНОЙ ЗАДАЧЕ

$$\begin{cases} \frac{1}{2}\left|\left|w\right|\right|^2 + C\sum_{i=1}^{l}\xi_i \to \min\limits_{w,w_0,\xi_i} (1) \\ y_i\big((w,x_i)+w_0\big) \geq 1-\xi_i, i=1,\dots,l\ (2) \\ \xi_i \geq 0, i=1,\dots,l\ (3) \end{cases}$$

- Перепишем (2) и (3):

$$\begin{cases} \xi_i \geq 1 - y_i\big((w,x_i)+w_0\big) \\ \xi_i \geq 0 \end{cases} \Rightarrow \xi_i = \max(0, 1 - y_i\big((w,x_i)+w_0\big))$$

СВЕДЕНИЕ К БЕЗУСЛОВНОЙ ЗАДАЧЕ

$$\begin{cases} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\xi_i \to \min\limits_{w,w_0,\xi_i} (1) \\ y_i\big((w,x_i)+w_0\big) \geq 1-\xi_i, i=1,\dots,l\ (2) \\ \xi_i \geq 0, i=1,\dots,l\ (3) \end{cases}$$

- Перепишем (2) и (3):

$$\begin{cases} \xi_i \geq 1 - y_i\big((w,x_i)+w_0\big) \\ \xi_i \geq 0 \end{cases} \Rightarrow \xi_i = \max(0, 1 - y_i\big((w,x_i)+w_0\big))$$

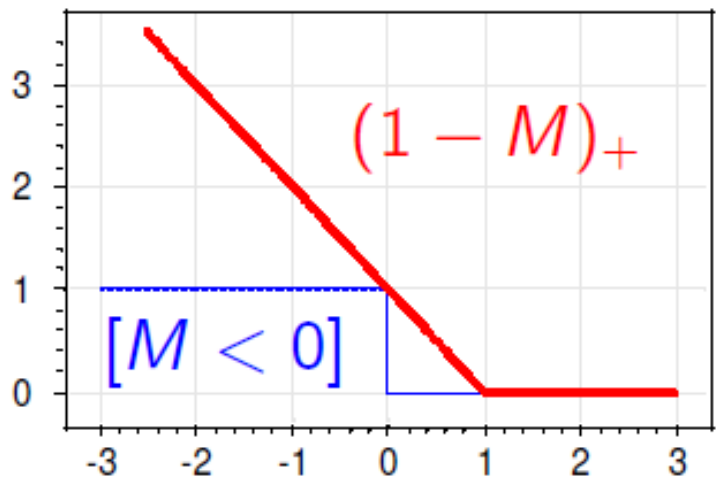
Получаем безусловную задачу оптимизации:

$$\boldsymbol{\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\max\Big(0, 1 - y_i\big((w,x_i)+w_0\big)\Big) \to \min\limits_{w,w_0}}$$

МЕТОД ОПОРНЫХ ВЕКТОРОВ: ЗАДАЧА ОПТИМИЗАЦИИ

- На задачу оптимизации SVM можно смотреть, как на *оптимизацию функции потерь* $L(M) = max(0,1 - M) = (1 - M)_+$ *с регуляризацией*:

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

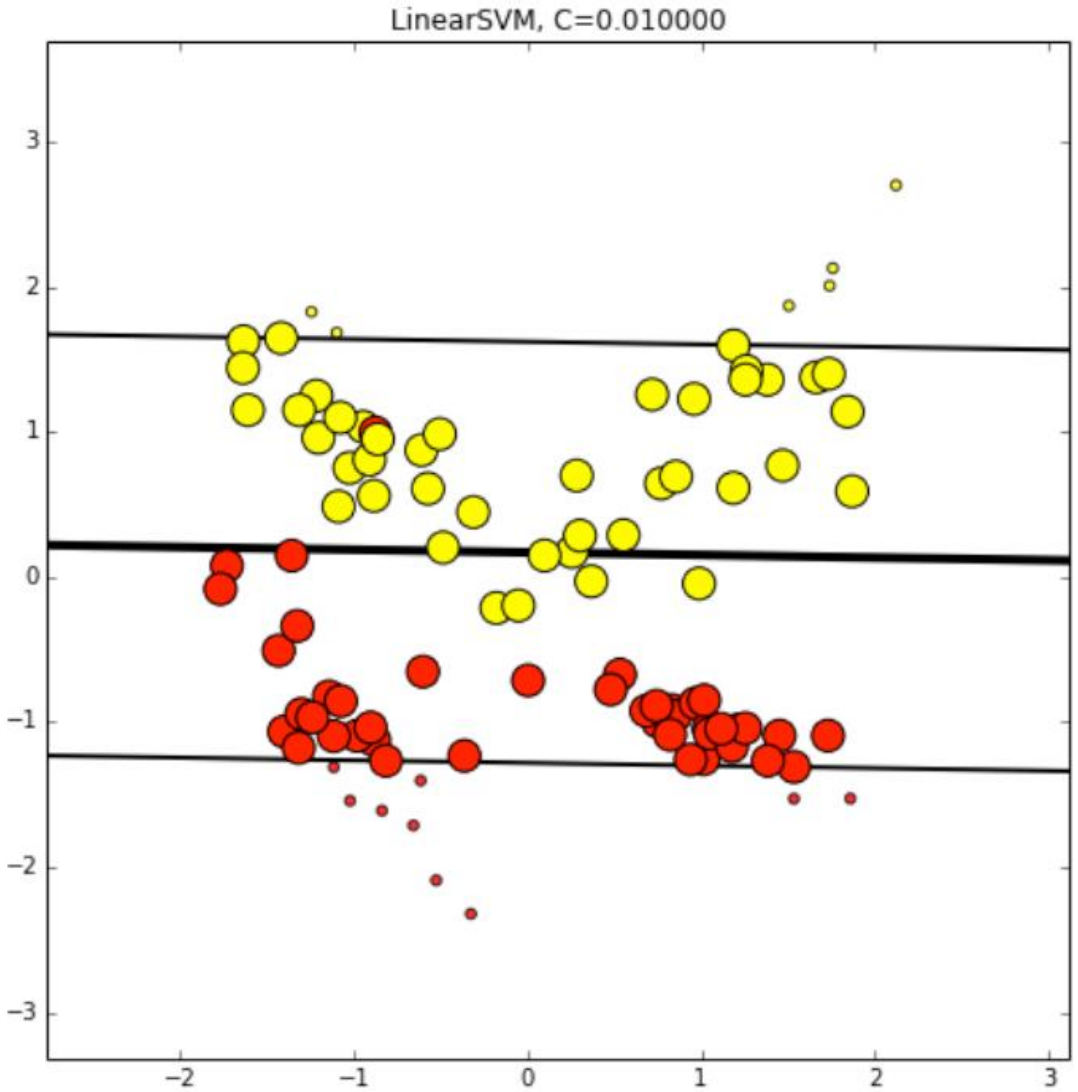


ЗНАЧЕНИЕ КОНСТАНТЫ C

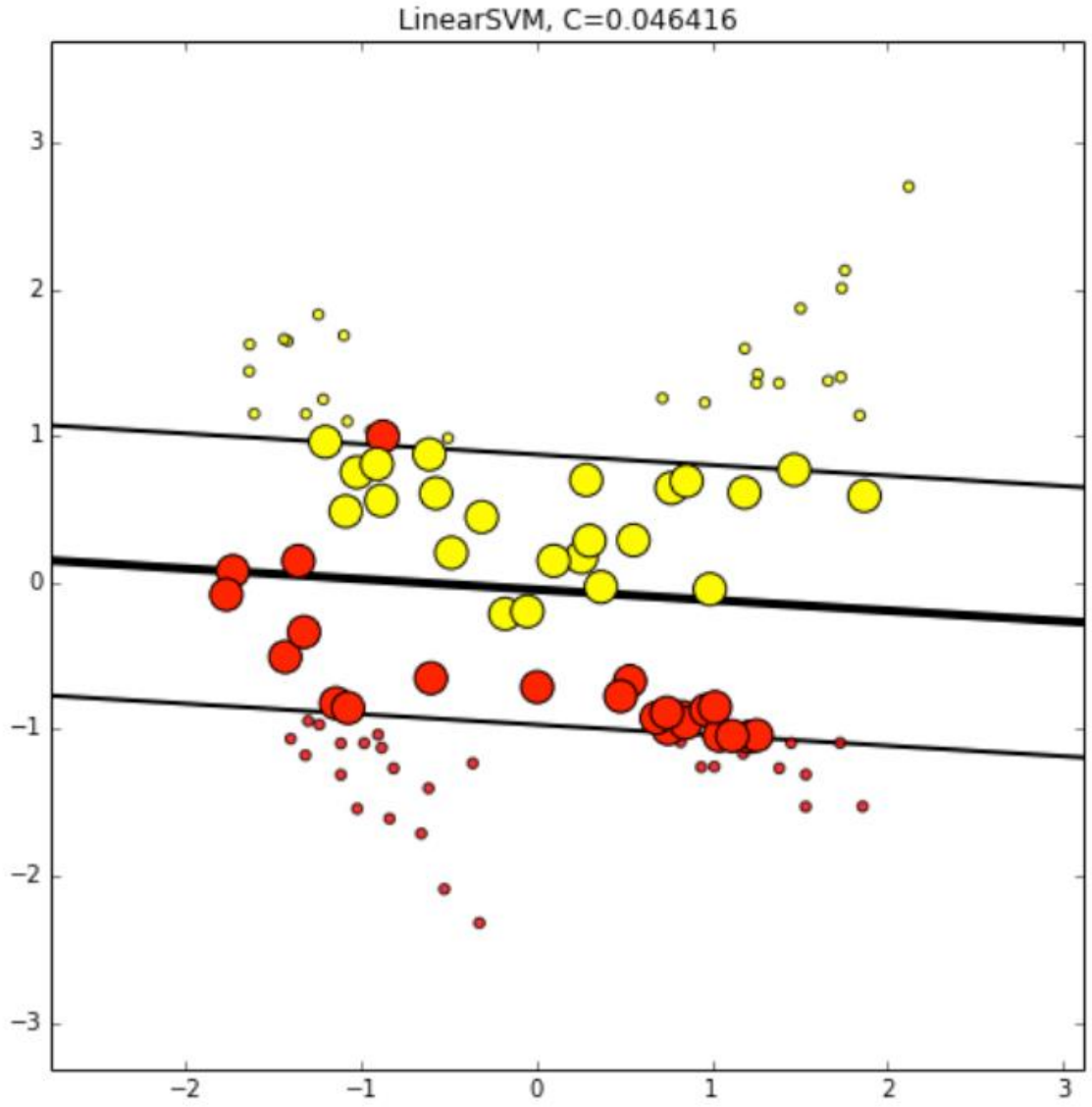
$$\begin{cases} \frac{1}{2}||w||^2 + \boldsymbol{C}\sum_{i=1}^{l}\xi_i \to \min\limits_{w,w_0,\xi_i} (1) \\ y_i\big((w,x_i)+w_0\big) \geq 1-\xi_i, i=1,\dots,l\ (2) \\ \xi_i \geq 0, i=1,\dots,l\ (3) \end{cases}$$

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

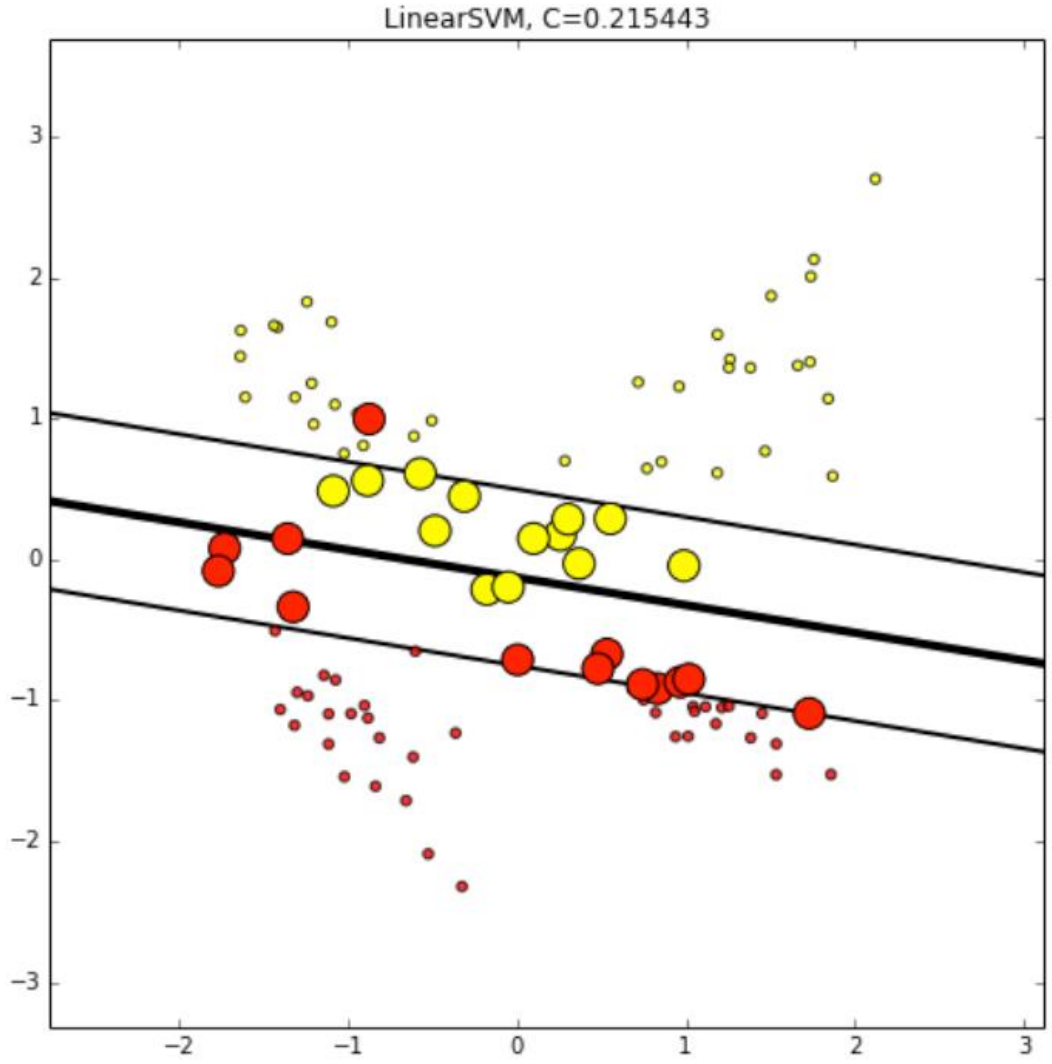
ЗНАЧЕНИЕ КОНСТАНТЫ C



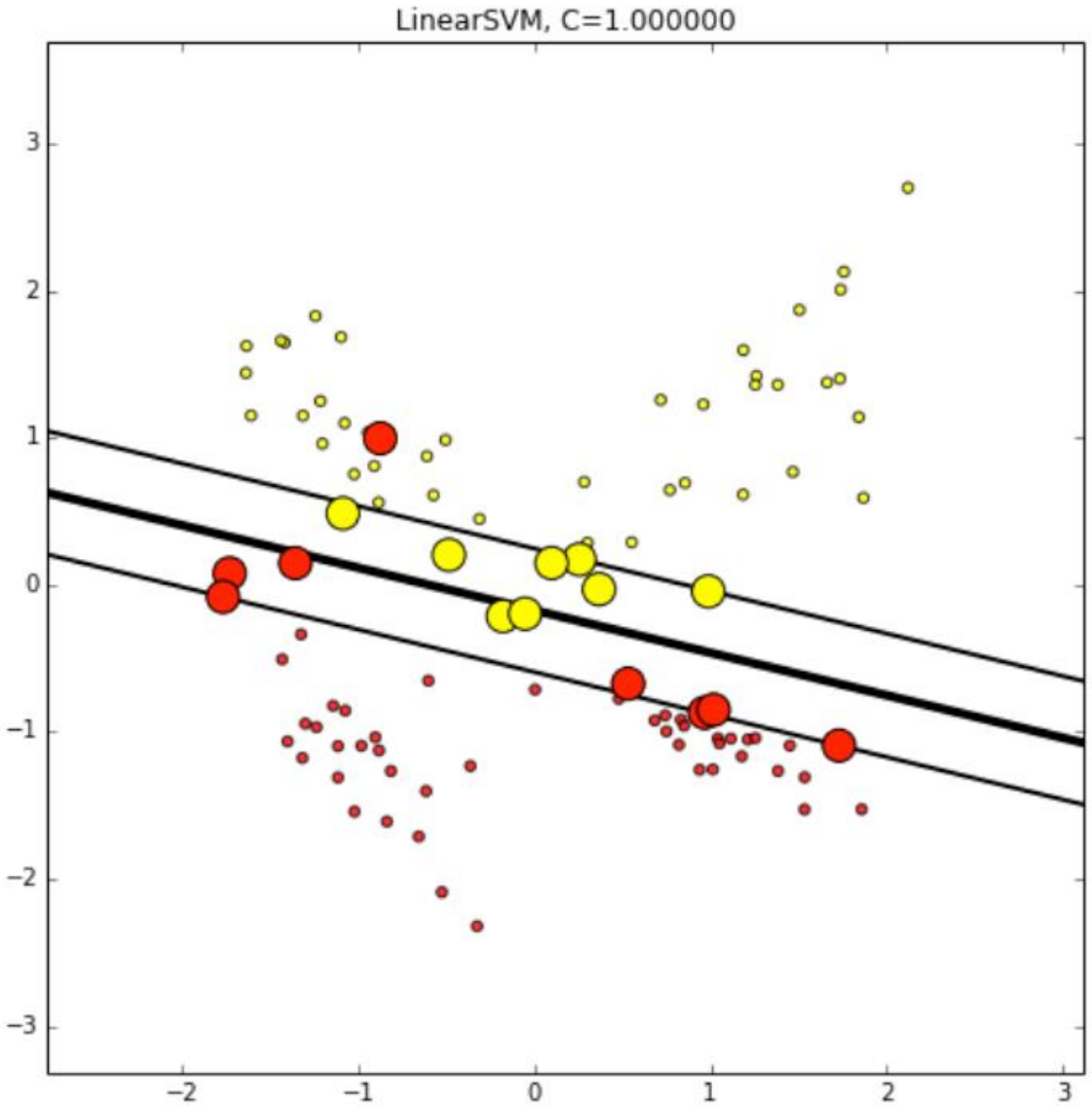
ЗНАЧЕНИЕ КОНСТАНТЫ C



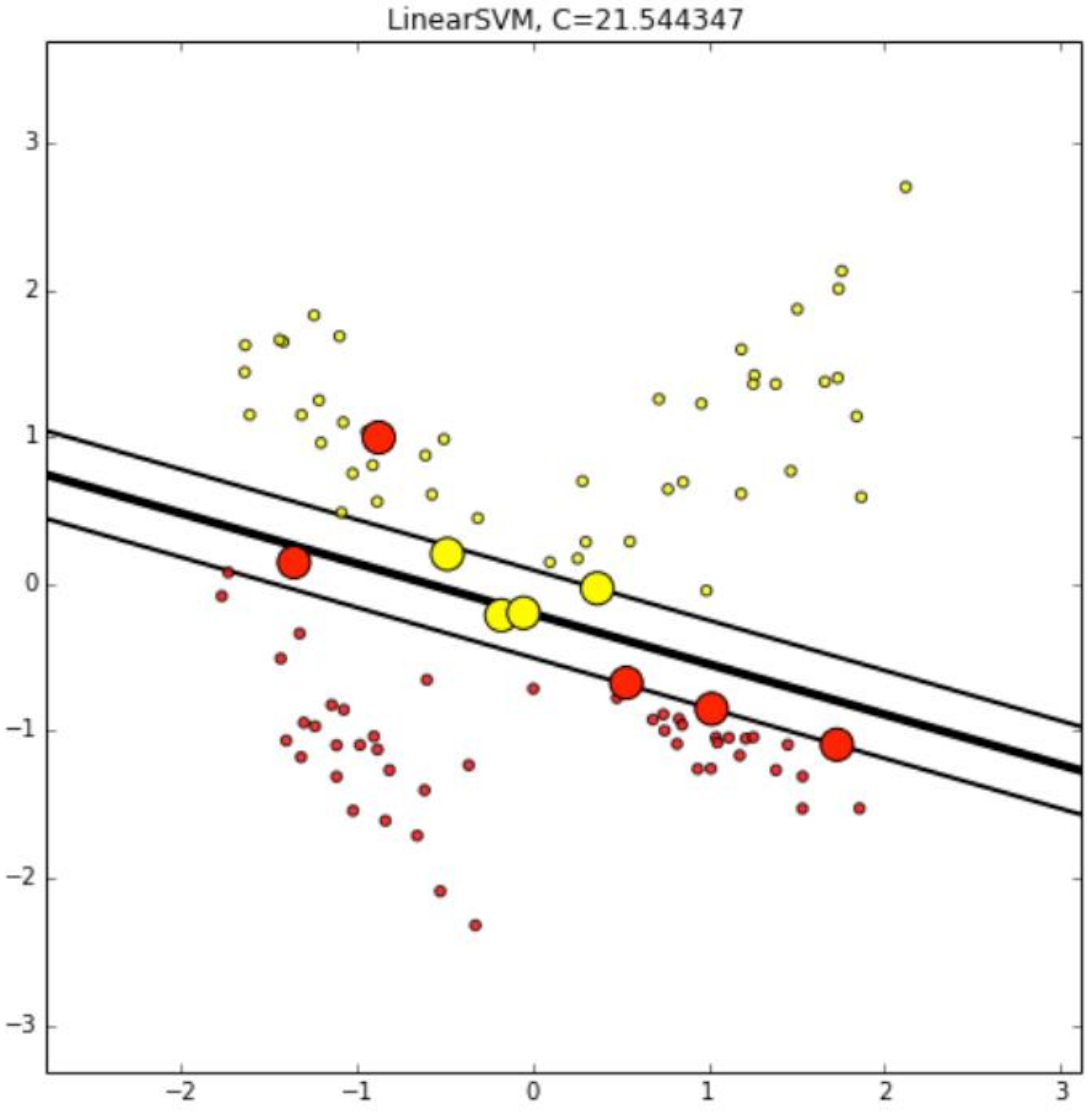
ЗНАЧЕНИЕ КОНСТАНТЫ C



ЗНАЧЕНИЕ КОНСТАНТЫ C



ЗНАЧЕНИЕ КОНСТАНТЫ C



УСЛОВИЯ КАРУША-КУНА-ТАККЕРА (ККТ)

Задача математического программирования:

$$(*)\begin{cases} f(x) \to \min\limits_{x} \\ g_i(x) \le 0, i = 1, \dots, m \\ h_j(x) = 0, j = 1, \dots, k \end{cases}$$

Необходимые условия. Если x – точка локального минимума, то существуют множители $\mu_i, i = 1, \dots, m, \lambda_j, j = 1, \dots, k$:

$$\begin{cases} \dfrac{\partial \mathcal{L}}{\partial x} = 0, \qquad \mathcal{L}(x;\mu;\lambda) = f(x) + \sum\limits_{i=1}^{m} \mu_i g_i(x) + \sum\limits_{j=1}^{k} \lambda_j h_j(x) \\ g_i(x) \le 0; h_j(x) = 0 \text{ (исходные ограничения)} \\ \mu_i \ge 0 \text{ (двойственные ограничения)} \\ \mu_i g_i(x) = 0 \text{ (условие дополняющей нежесткости)} \end{cases}$$

ПРИМЕНЕНИЕ УСЛОВИЙ ККТ К ЗАДАЧЕ SVM

Функция Лагранжа: $\mathcal{L}(w; w_0; \xi; \lambda; \eta) =$

$= \frac{1}{2}||w||^2 - \sum_{i=1}^{l} \lambda_i (M_i(w, w_0) - 1) - \sum_{i=1}^{l} \xi_i (\lambda_i + \eta_i - C)$

λ_i - переменные, двойственные к ограничениям $M_i \geq 1 - \xi_i$

η_i - переменные, двойственные к ограничениям $\xi_i \geq 0$

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial w} = 0, \frac{\partial \mathcal{L}}{\partial w_0} = 0, \frac{\partial \mathcal{L}}{\partial \xi} = 0 \\ \xi_i \geq 0, \lambda_i \geq 0, \eta_i \geq 0 \\ \lambda_i = 0 \text{ или } M_i(w, w_0) = 1 - \xi_i \\ \eta_i = 0 \text{ или } \xi_i = 0, \end{cases}$$

$i = 1, \dots, l.$

НЕОБХОДИМЫЕ УСЛОВИЯ СЕДЛОВОЙ ТОЧКИ

Функция Лагранжа: $\mathcal{L}(w; w_0; \xi; \lambda; \eta) =$

$$= \frac{1}{2}\left|\left|w\right|\right|^2 - \sum_{i=1}^{l} \lambda_i (M_i(w, w_0) - 1) - \sum_{i=1}^{l} \xi_i (\lambda_i + \eta_i - C)$$

Необходимые условия седловой точки функции Лагранжа:

$$\frac{\partial \mathcal{L}}{\partial w} = w - \sum_{i=1}^{l} \lambda_i y_i x_i = 0 \;\; \Rightarrow w = \sum_{i=1}^{l} \lambda_i y_i x_i \,,$$

$$\frac{\partial \mathcal{L}}{\partial w_0} = -\sum_{i=1}^{l} \lambda_i y_i = 0 \;\Rightarrow \sum_{i=1}^{l} \lambda_i y_i = 0 \,,$$

$$\frac{\partial \mathcal{L}}{\partial \xi_i} = -\lambda_i - \eta_i + C = 0 \Rightarrow \eta_i + \lambda_i = C, i = 1, \ldots l.$$

ОПОРНЫЕ ВЕКТОРЫ

- $\lambda_i = 0; \eta_i = C; \xi_i = 0; M_i \geq 1$

- периферийные (неинформативные) объекты

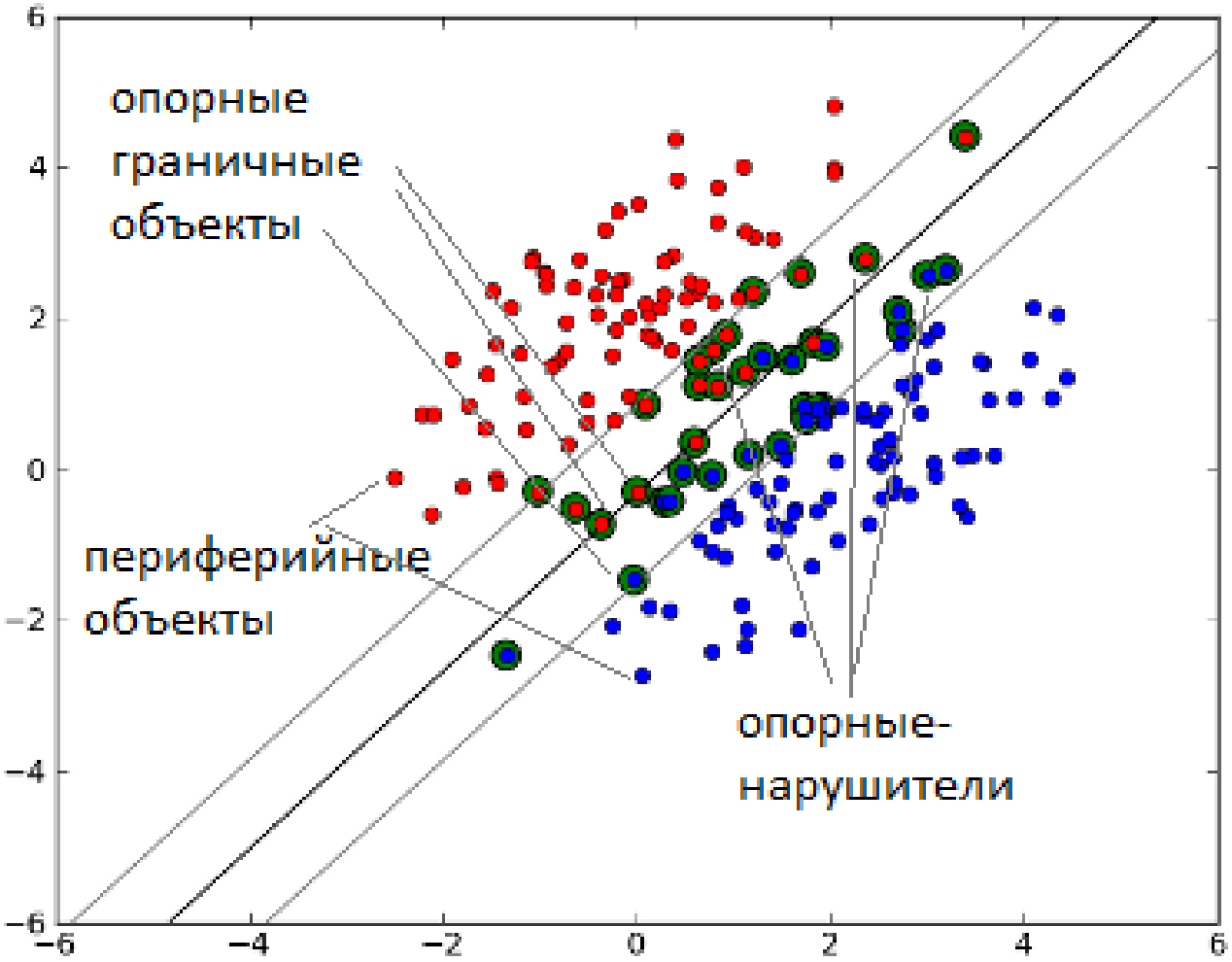
- $0 < \lambda_i < C; 0 < \eta_i < C; \xi_i = 0; M_i = 1$

- опорные граничные объекты

- $\lambda_i = C; \eta_i = 0; \xi_i > 0; M_i < 1$

- опорные-нарушители

ТИПЫ ОБЪЕКТОВ В SVM



ДВОЙСТВЕННАЯ ЗАДАЧА

- $\mathcal{L}(x, \lambda, \mu)$ – лагранжиан

$$\boldsymbol{g(\lambda, \mu) = \min_{x} \mathcal{L}(x, \lambda, \mu)} - \text{двойственная функция}$$

Двойственная задача к задаче $(*)$:

$$\boldsymbol{g(\lambda, \mu) \to \max_{\lambda, \mu}}$$

$$\boldsymbol{\lambda_i \geq 0, i = 1, \dots, m}$$

ДВОЙСТВЕННАЯ ЗАДАЧА

- $\mathcal{L}(x, \lambda, \mu)$ – лагранжиан

$$\boldsymbol{g(\lambda, \mu) = \min_{x} \mathcal{L}(x, \lambda, \mu)} - \text{двойственная функция}$$

Двойственная задача к задаче $(*)$:

$$\boldsymbol{g(\lambda, \mu) \to \max_{\lambda, \mu}}$$

$$\boldsymbol{\lambda_i \geq 0, i = 1, \dots, m}$$

- Пусть (λ_*, μ_*) – решение двойственной задачи.

Утверждение. Если все функции в прямой задаче выпуклые, то оптимальное значение функционала в прямой и двойственной задаче совпадают

$$g(\lambda, \mu) = f(x_*)$$

ДВОЙСТВЕННАЯ ЗАДАЧА

Двойственная задача является выпуклой (даже если прямая задача выпуклой не является).

ДВОЙСТВЕННАЯ ЗАДАЧА SVM

$$\begin{cases} -\mathcal{L}(\lambda) = -\sum_{i=1}^{l} \lambda_i + \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} \lambda_i \lambda_j y_i y_j (x_i, x_j) \to \min\limits_{\lambda} \\ 0 \le \lambda_i \le C, i = 1, \dots, l \\ \sum_{i=1}^{l} \lambda_i y_i = 0 \end{cases}$$

Решение прямой задачи выражается через решение двойственной:

$$\begin{cases} w = \sum_{i=1}^{l} \lambda_i y_i x_i \\ w_0 = (w, x_i) - y_i, \text{для любого } i: \lambda_i > 0, M_i = 1 \end{cases}$$

Линейный классификатор:

$$a(x) = sign(\sum_{i=1}^{l} \lambda_i y_i (x_i, x) - w_0)$$

КАЛИБРОВКА ВЕРОЯТНОСТЕЙ

Калибровка вероятностей - приведение ответов алгоритма к значениям, близким к вероятностям объектов принадлежать конкретному классу.

Зачем это нужно?

- Вероятности гораздо проще интерпретировать
- Вероятности могут дать дополнительную информацию о результатах работы алгоритма

КАЛИБРОВКА ПЛАТТА

- Пусть есть два класса, $Y = \{+1, -1\}$

Задача: для классификатора $a(x)$, предсказывающего значения из отрезка $[0, 1]$, либо предсказывающего класс (+1 или -1), сделать калибровку, чтобы предсказания были вероятностями $p(y = +1|x)$.

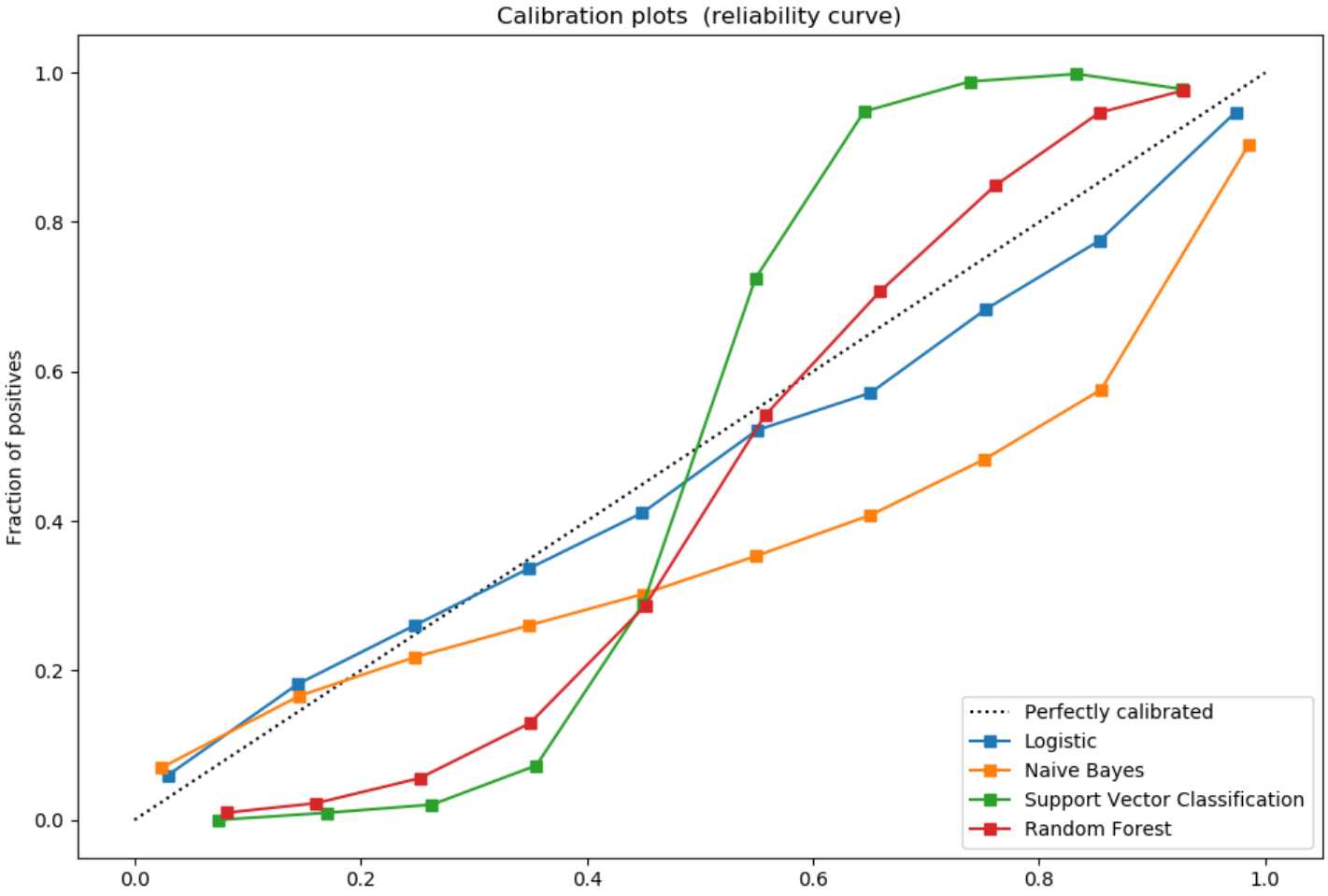
КАЛИБРОВКА ПЛАТТА

- Пусть есть два класса, $Y = \{+1, -1\}$

Задача: для классификатора $a(x)$, предсказывающего значения из отрезка $[0, 1]$, либо предсказывающего класс (+1 или -1), сделать калибровку, чтобы предсказания были вероятностями $p(y = +1|x)$.

Идея: обучаем логистическую регрессию на ответах классификатора $a(x)$.

ПРИМЕР ИЗ SKLEARN



КАЛИБРОВКА ПЛАТТА

- Пусть есть два класса, $Y = \{+1, -1\}$

Задача: для классификатора $a(x)$, предсказывающего значения из отрезка $[0, 1]$, либо предсказывающего класс (+1 или -1), сделать калибровку, чтобы предсказания были вероятностями $p(y = +1|x)$.

Идея: ***обучаем логистическую регрессию на ответах классификатора*** $a(x)$.

- $\pi(x; \alpha; \beta) = \sigma(\alpha \cdot a(x) + \beta) = \frac{1}{1+e^{-(\alpha \cdot a(x)+\beta)}}$
- Находим α и β, минимизируя логистическую функцию потерь:

$$-\sum_{y_i=-1} \log\big(1 - \pi(x; \alpha; \beta)\big) - \sum_{y_i=+1} \log\big(\pi(x; \alpha; \beta)\big) \to \min_{\alpha,\beta}$$