Линейные модели классификации.

Елена Кантонистова

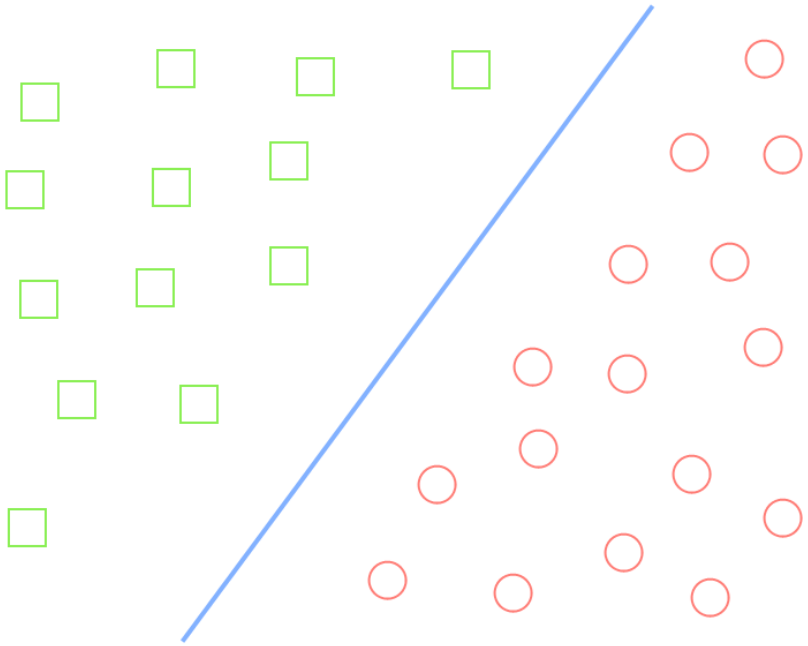
ПЛАН

- Метод опорных векторов
- Калибровка вероятностей

МЕТОД ОПОРНЫХ ВЕКТОРОВ

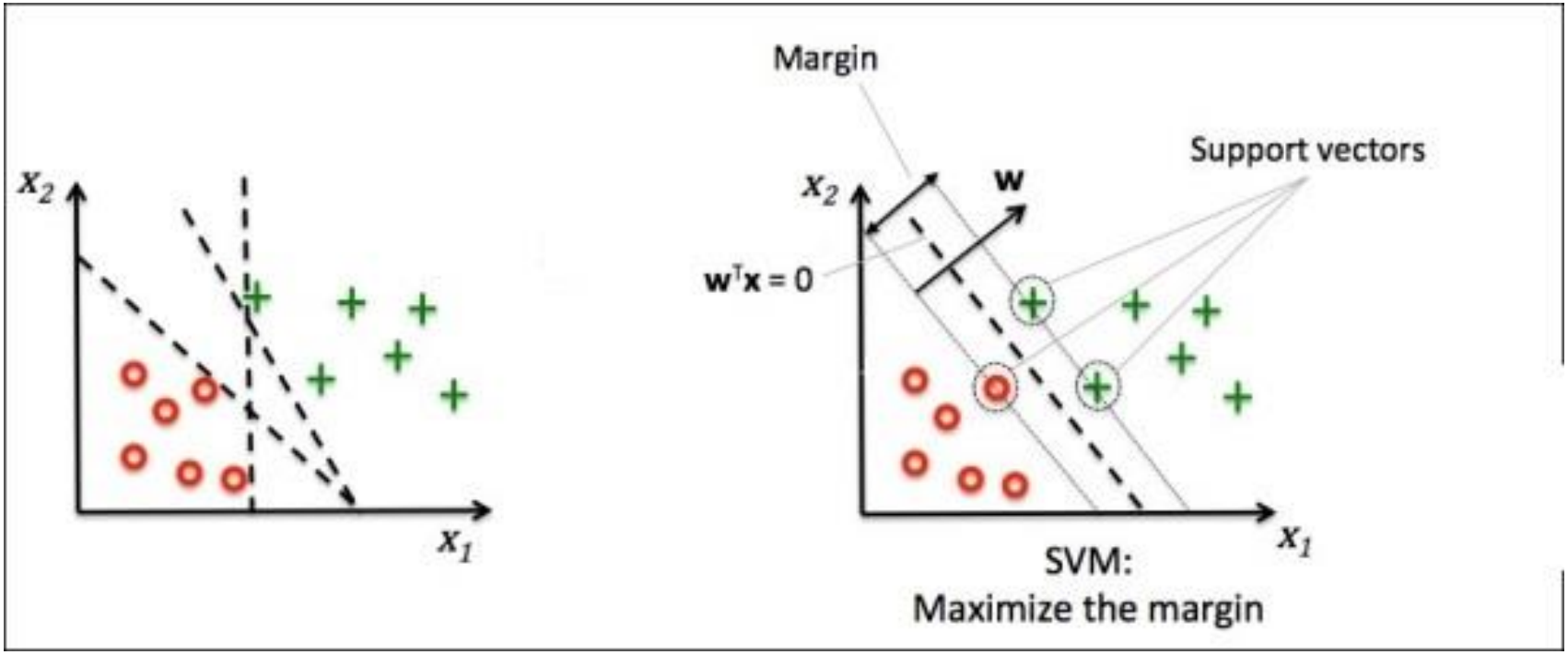
ЛИНЕЙНО РАЗДЕЛИМАЯ ВЫБОРКА

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

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

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

МЕТОД ОПОРНЫХ ВЕКТОРОВ: РАЗДЕЛИМЫЙ СЛУЧАЙ

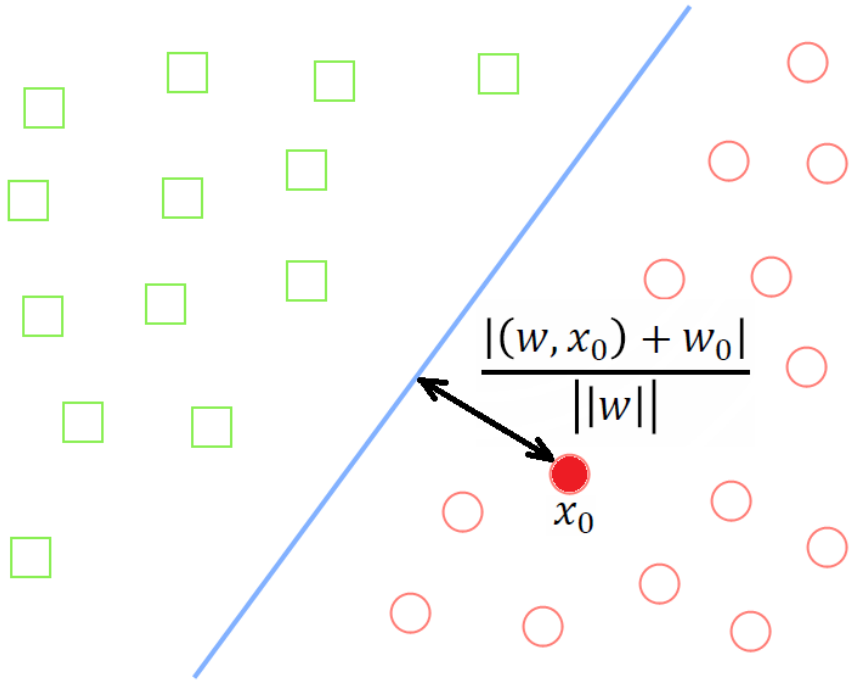
- $a(x) = sign\big((w, x) + w_0\big)$
- Нормируем параметры w и w_0 так, что

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

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

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

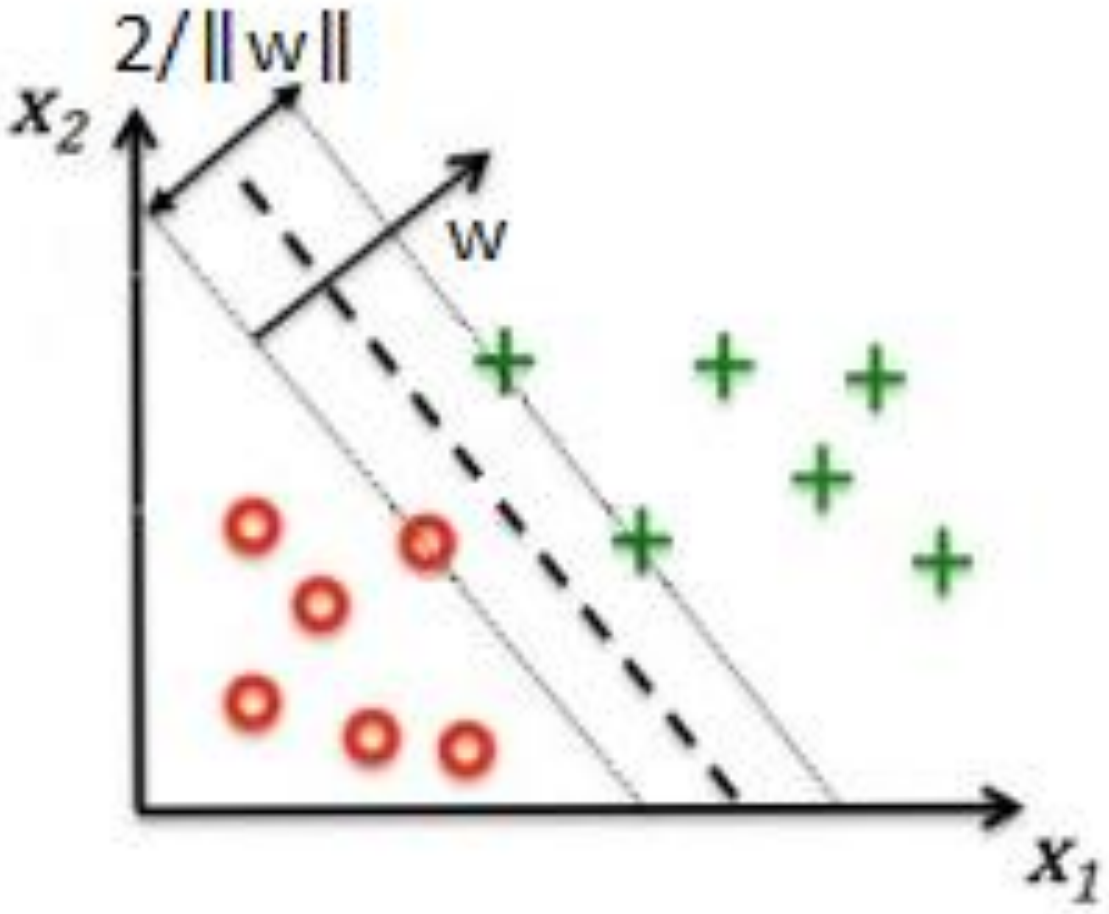
Тогда расстояние от точки x_0 до разделяющей гиперплоскости, задаваемой классификатором:

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

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

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

РАЗДЕЛЯЮЩАЯ ПОЛОСА



ОПТИМИЗАЦИОННАЯ ЗАДАЧА SVM ДЛЯ РАЗДЕЛИМОЙ ВЫБОРКИ

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

Утверждение. Данная оптимизационная задача имеет единственное решение.

ЛИНЕЙНО НЕРАЗДЕЛИМАЯ ВЫБОРКА

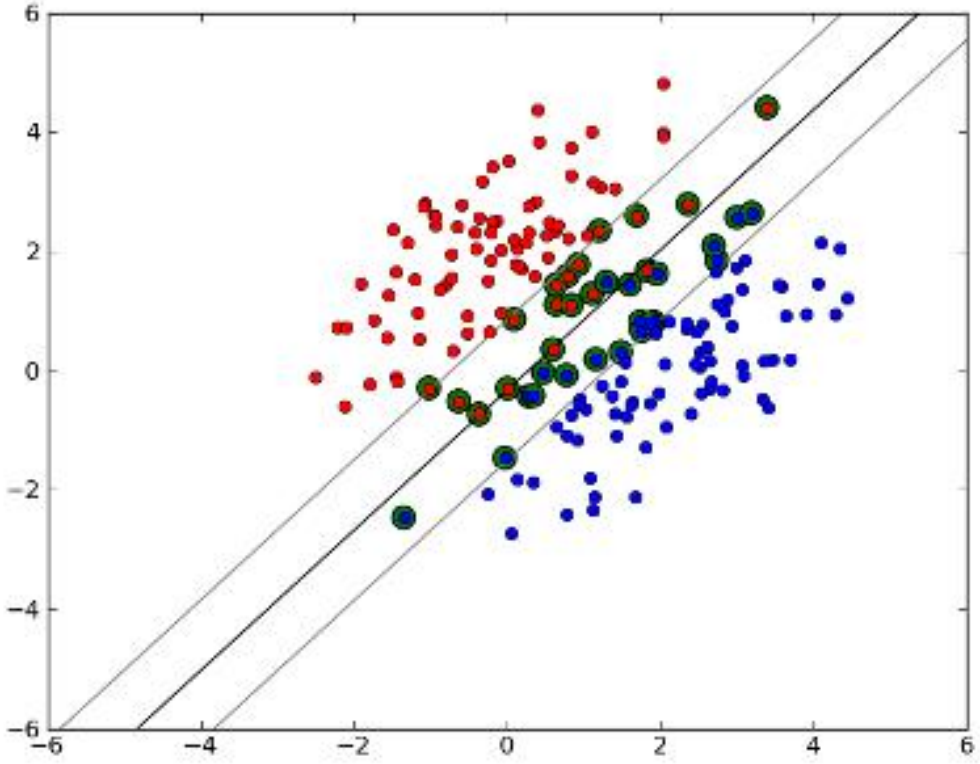
- Существует хотя бы один объект $x \in X$, что

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

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:

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

МЕТОД ОПОРНЫХ ВЕКТОРОВ: НЕРАЗДЕЛИМЫЙ СЛУЧАЙ

Утверждение. Задача

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

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

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

- Перепишем (2) и (3):

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

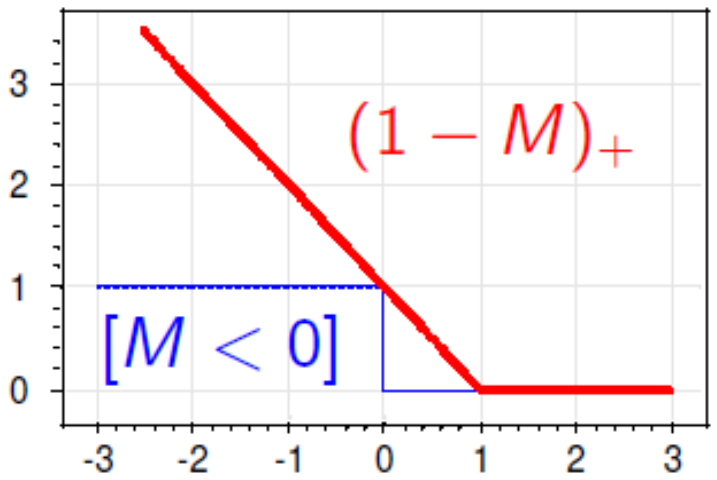
Получаем безусловную задачу оптимизации:

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

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*:

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

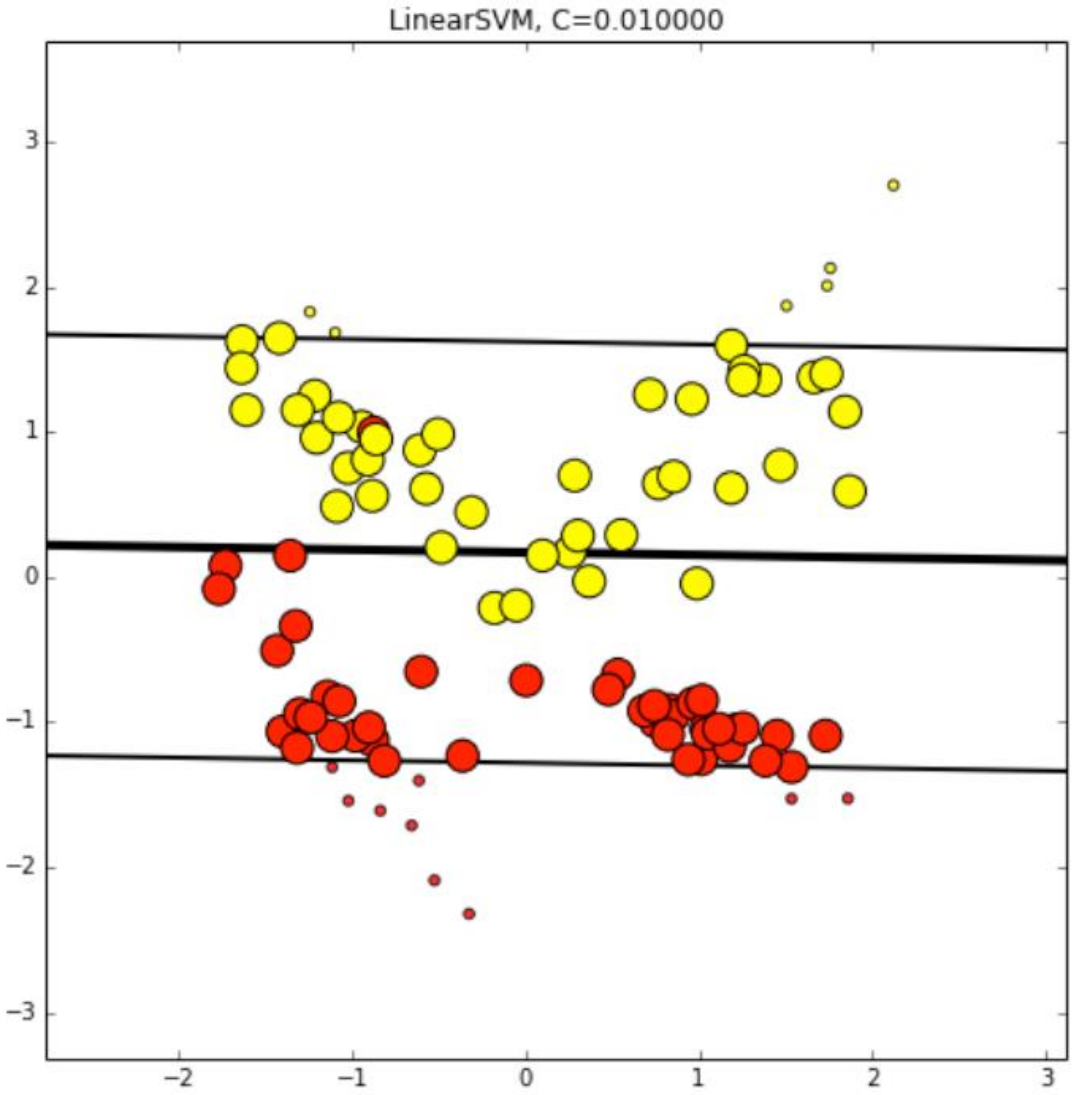


ЗНАЧЕНИЕ КОНСТАНТЫ C

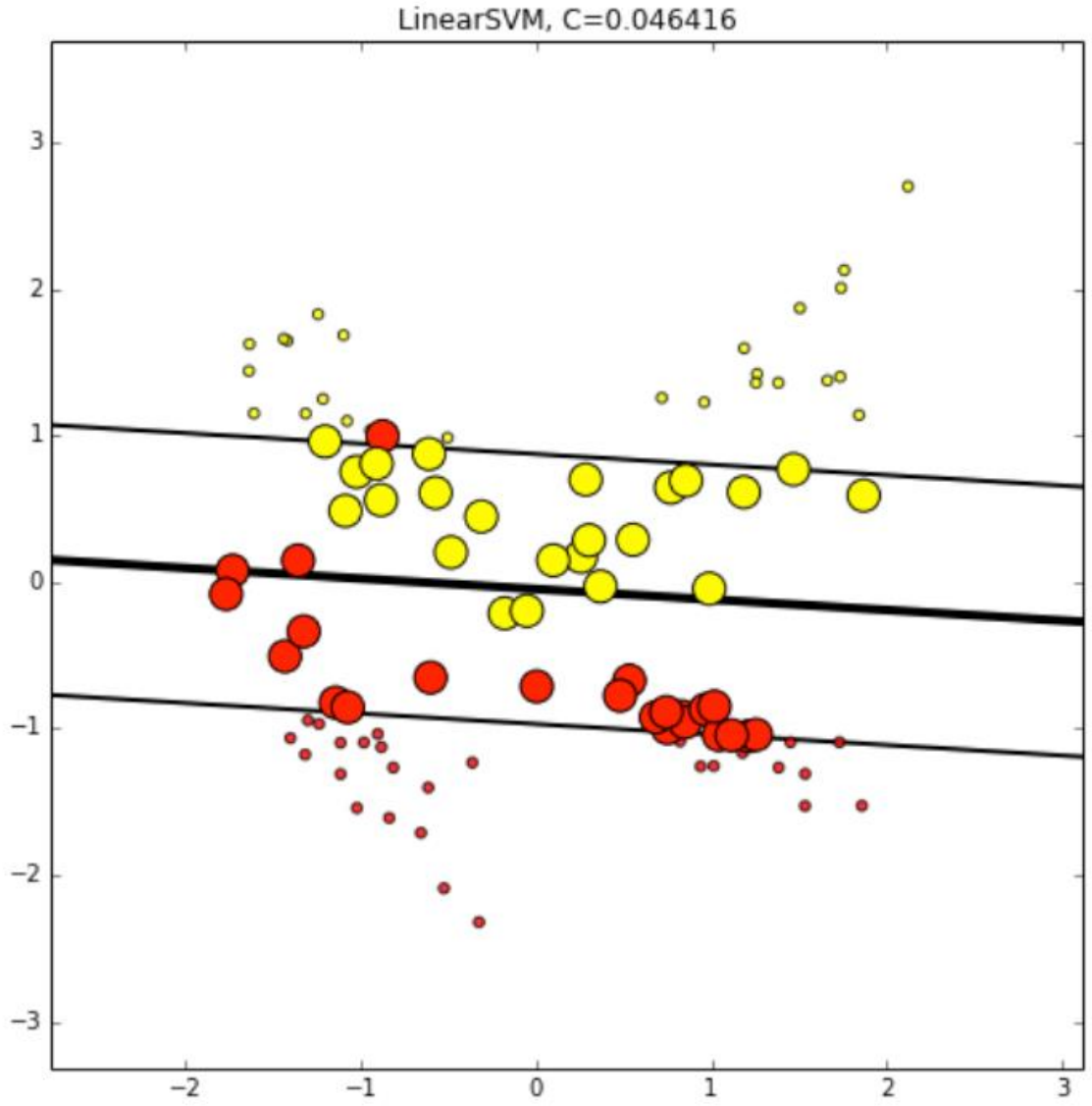
$$\begin{cases} \frac{1}{2}\left|\left|w\right|\right|^2 + \boldsymbol{C}\sum_{i=1}^{l}\xi_i \to \min\limits_{w,w_0,\xi_i} (1) \\ y_i\left((w, x_i) + w_0\right) \geq 1 - \xi_i, i = 1, \dots, l\ (2) \\ \xi_i \geq 0, i = 1, \dots, l\ (3) \end{cases}$$

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

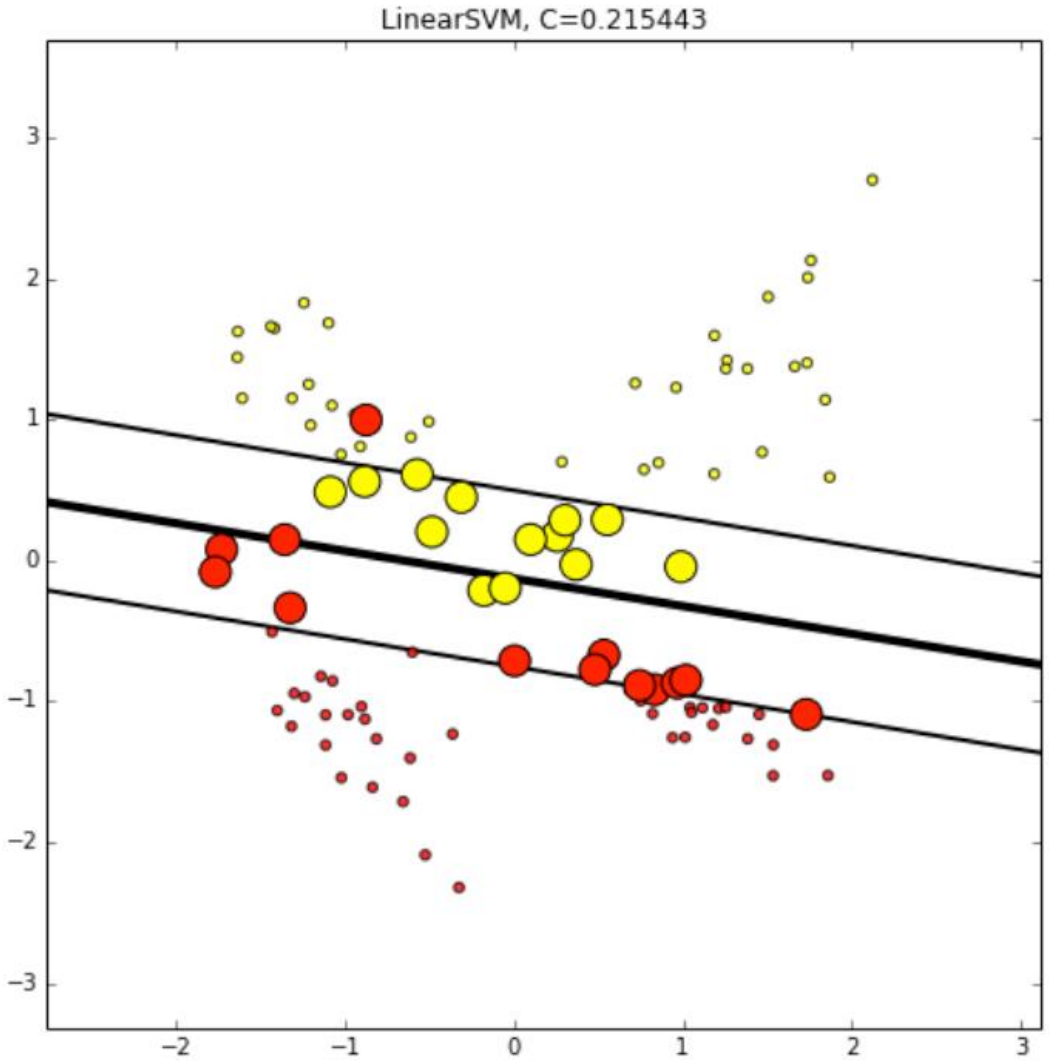
ЗНАЧЕНИЕ КОНСТАНТЫ C



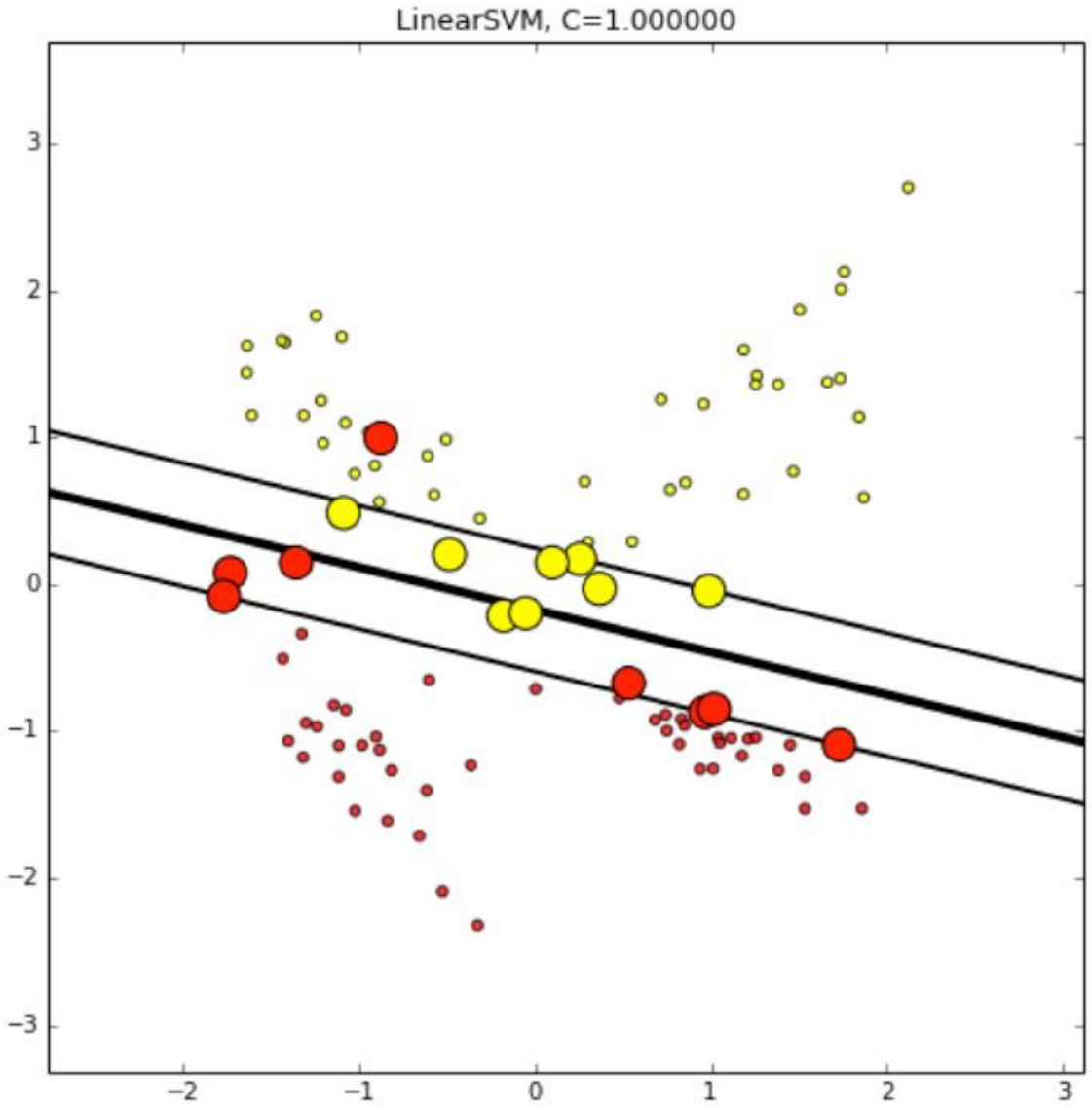
ЗНАЧЕНИЕ КОНСТАНТЫ C



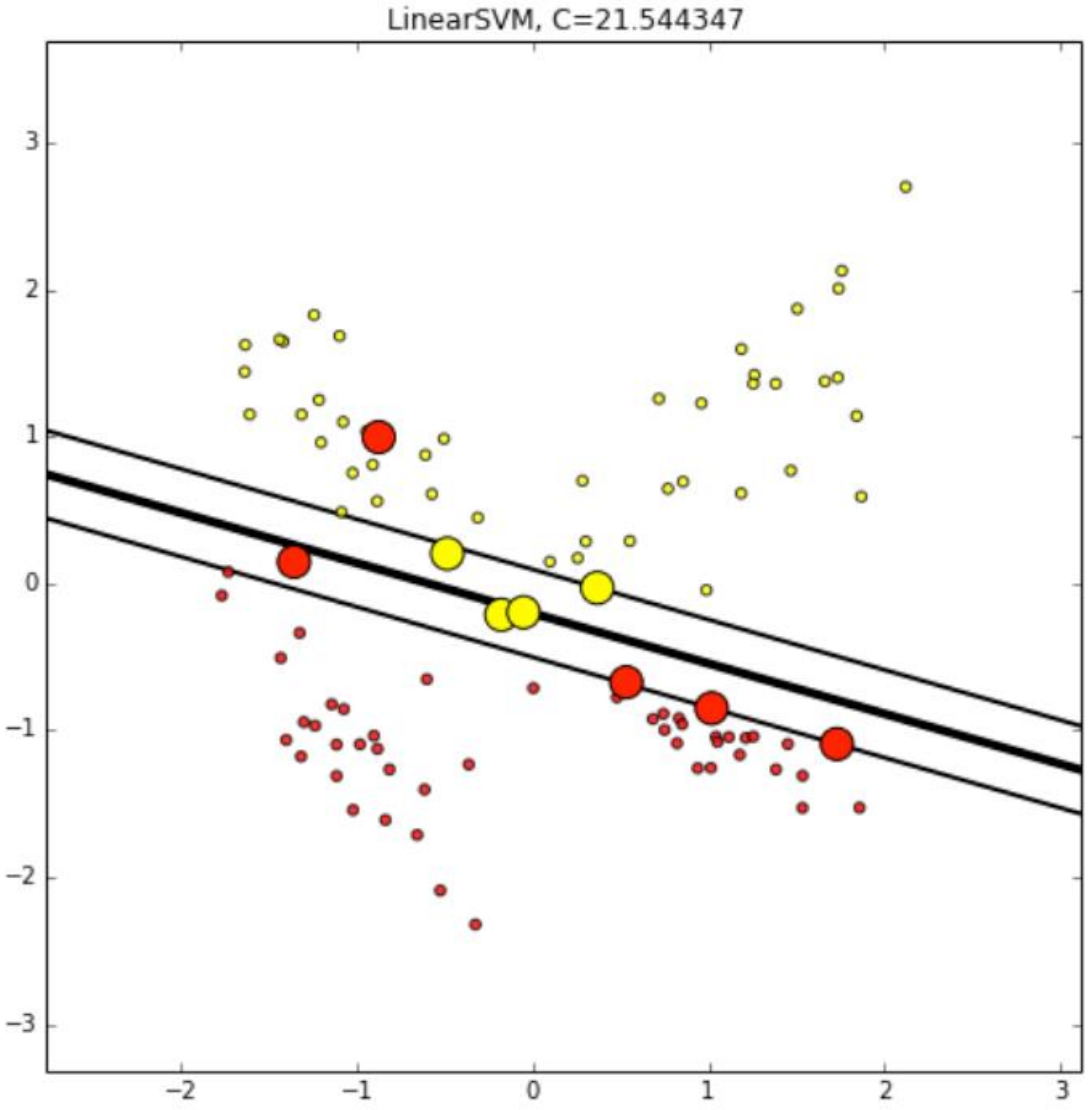
ЗНАЧЕНИЕ КОНСТАНТЫ C



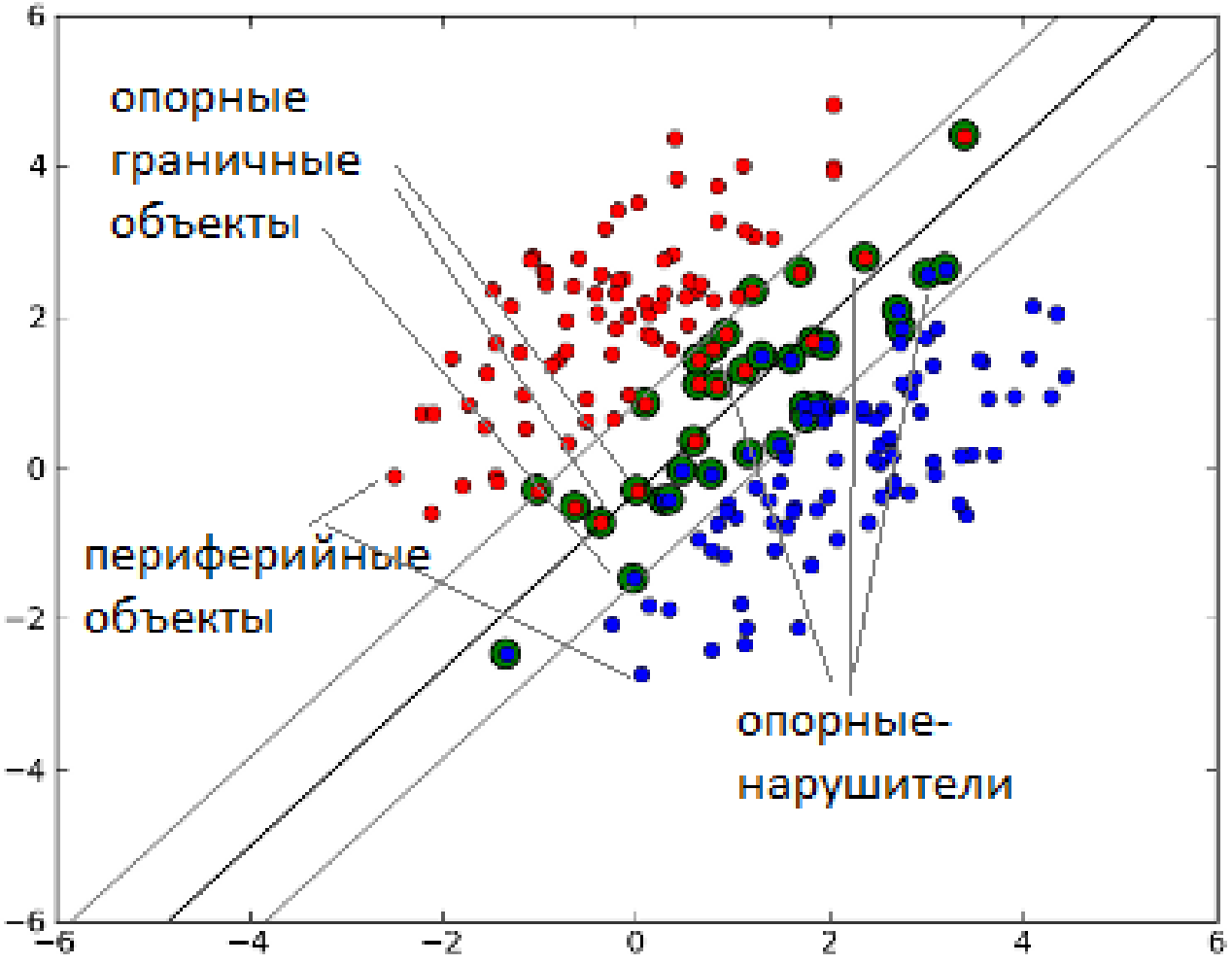
ЗНАЧЕНИЕ КОНСТАНТЫ C



ЗНАЧЕНИЕ КОНСТАНТЫ C



ТИПЫ ОБЪЕКТОВ В SVM

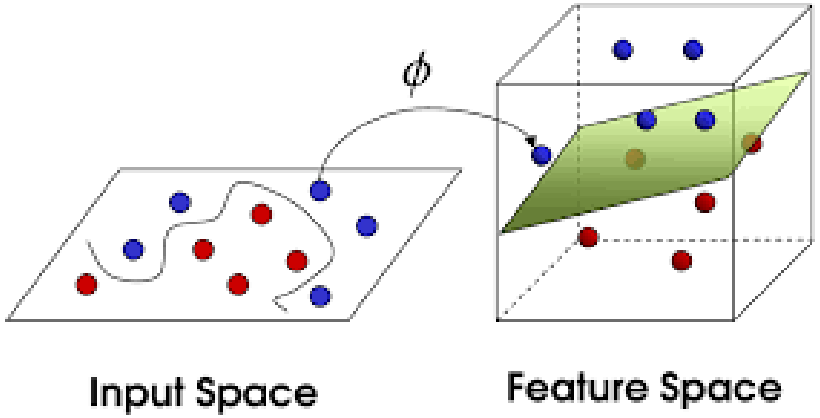


ЯДРОВОЙ МЕТОД ГЛАВНЫХ КОМПОНЕНТ

Пусть исходная выборка (с признаками $x_1, x_2, \dots, x_n$) *линейно не разделима*.

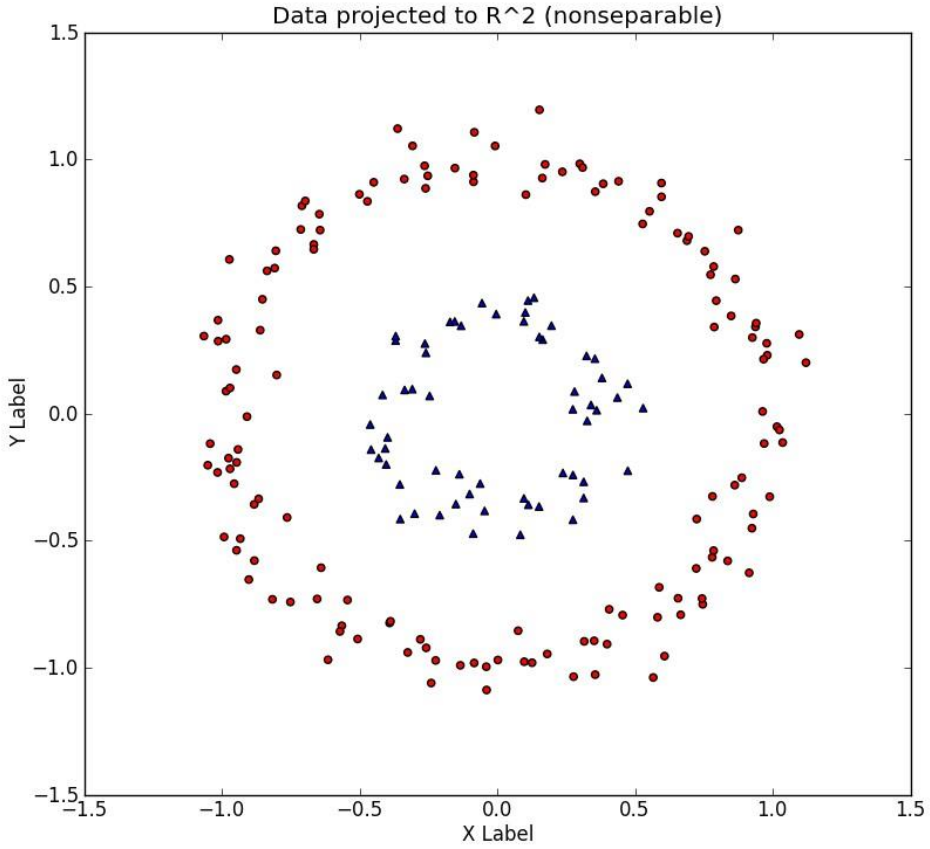
Может существовать такое преобразование координат $(y_1, y_2, \dots, y_N) = f(x_1, x_2, \dots, x_n)$.

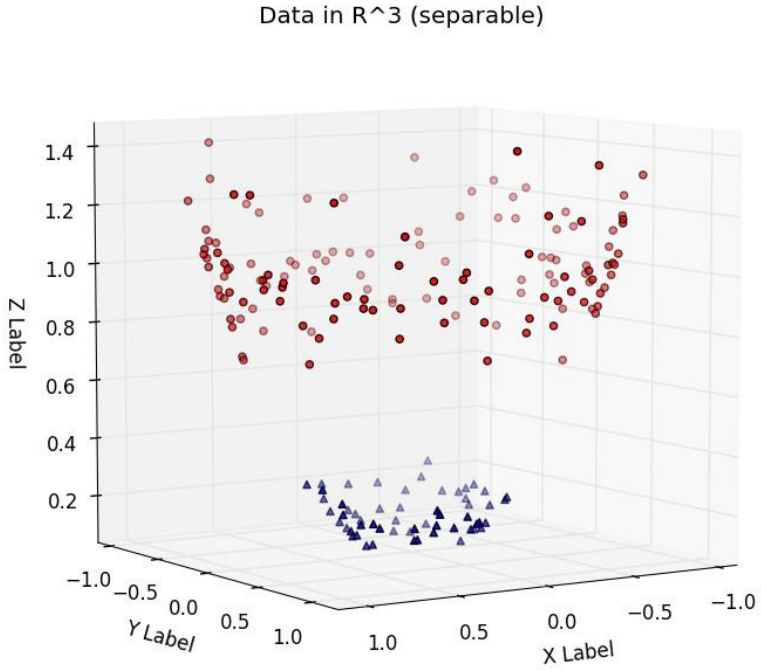
что в пространстве новых координат выборка становится *линейно разделимой*.



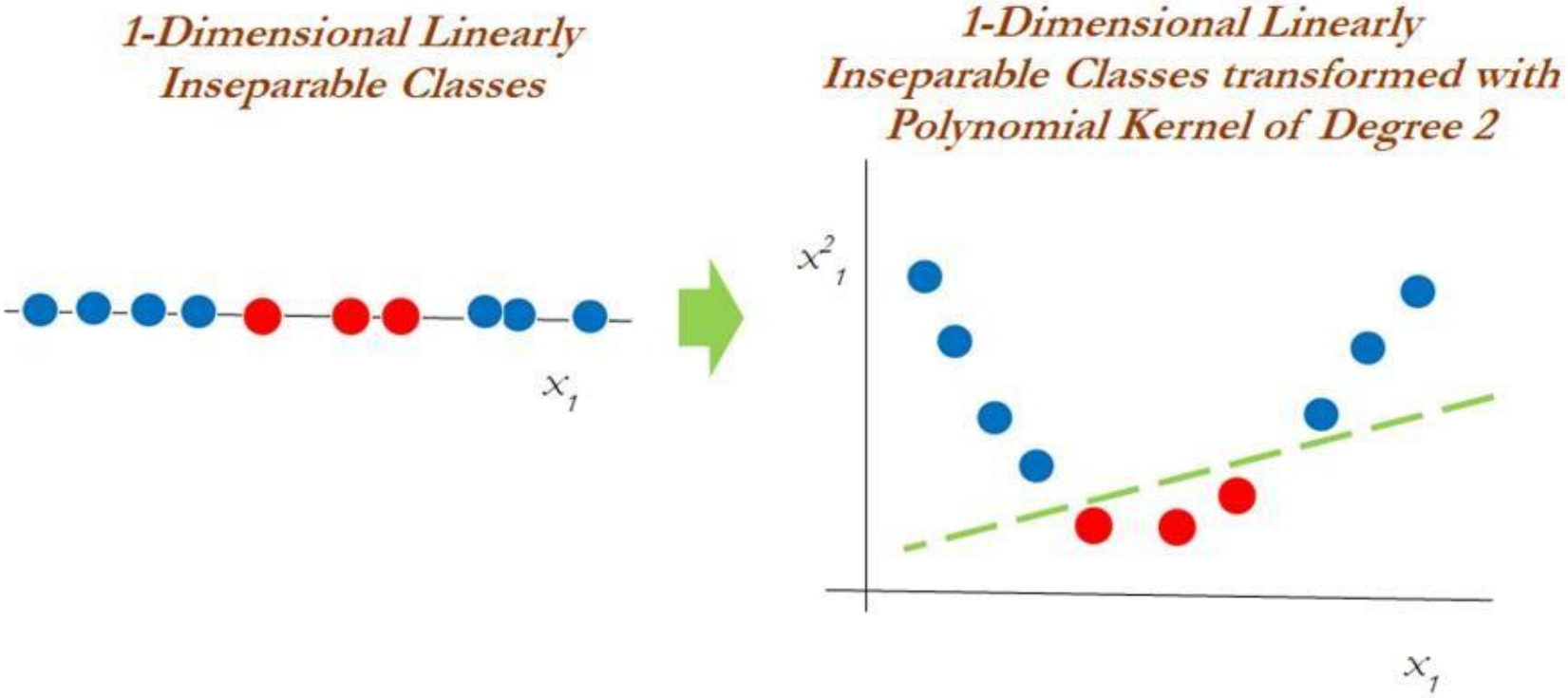
- ***Применение преобразования координат и метода главных компонент называется ядровым методом главных компонент (kernel SVM).***

РАДИАЛЬНОЕ ЯДРО



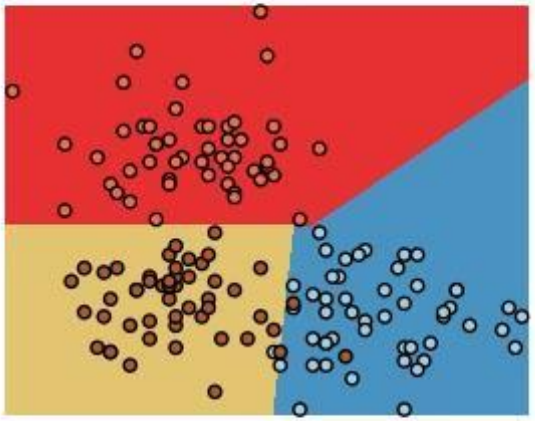


ПОЛИНОМИАЛЬНОЕ ЯДРО

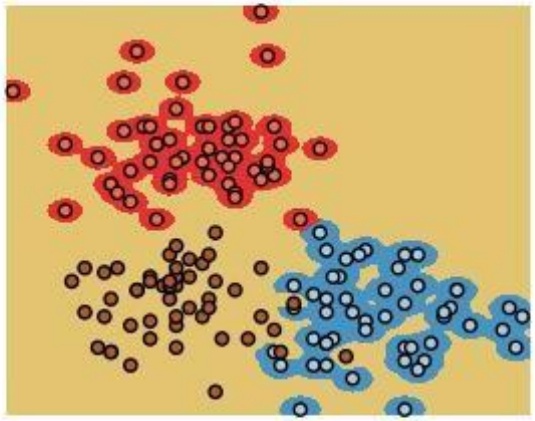


ПРИМЕРЫ

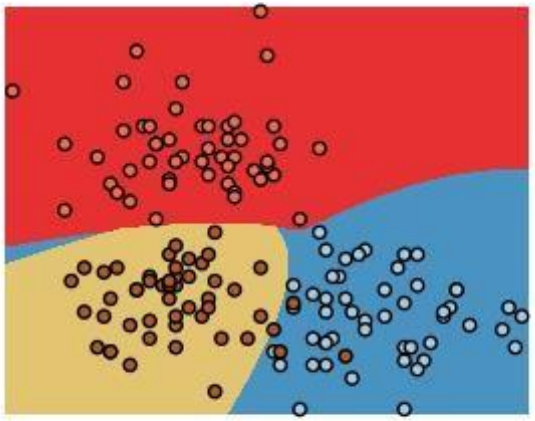
SVC with linear kernel



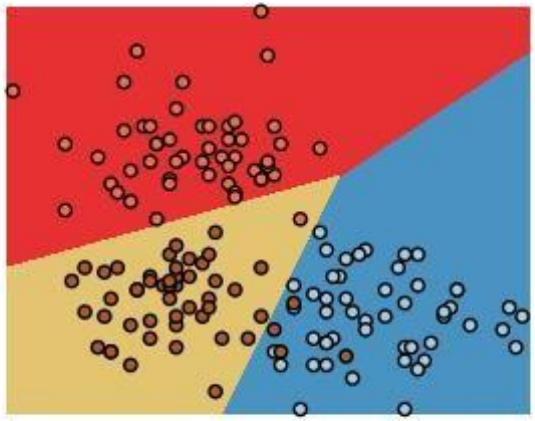
SVC with RBF kernel



SVC with polynomial (degree 3) kernel



LinearSVC (linear kernel)



КАЛИБРОВКА ВЕРОЯТНОСТЕЙ

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 $p(y = +1|x)$.

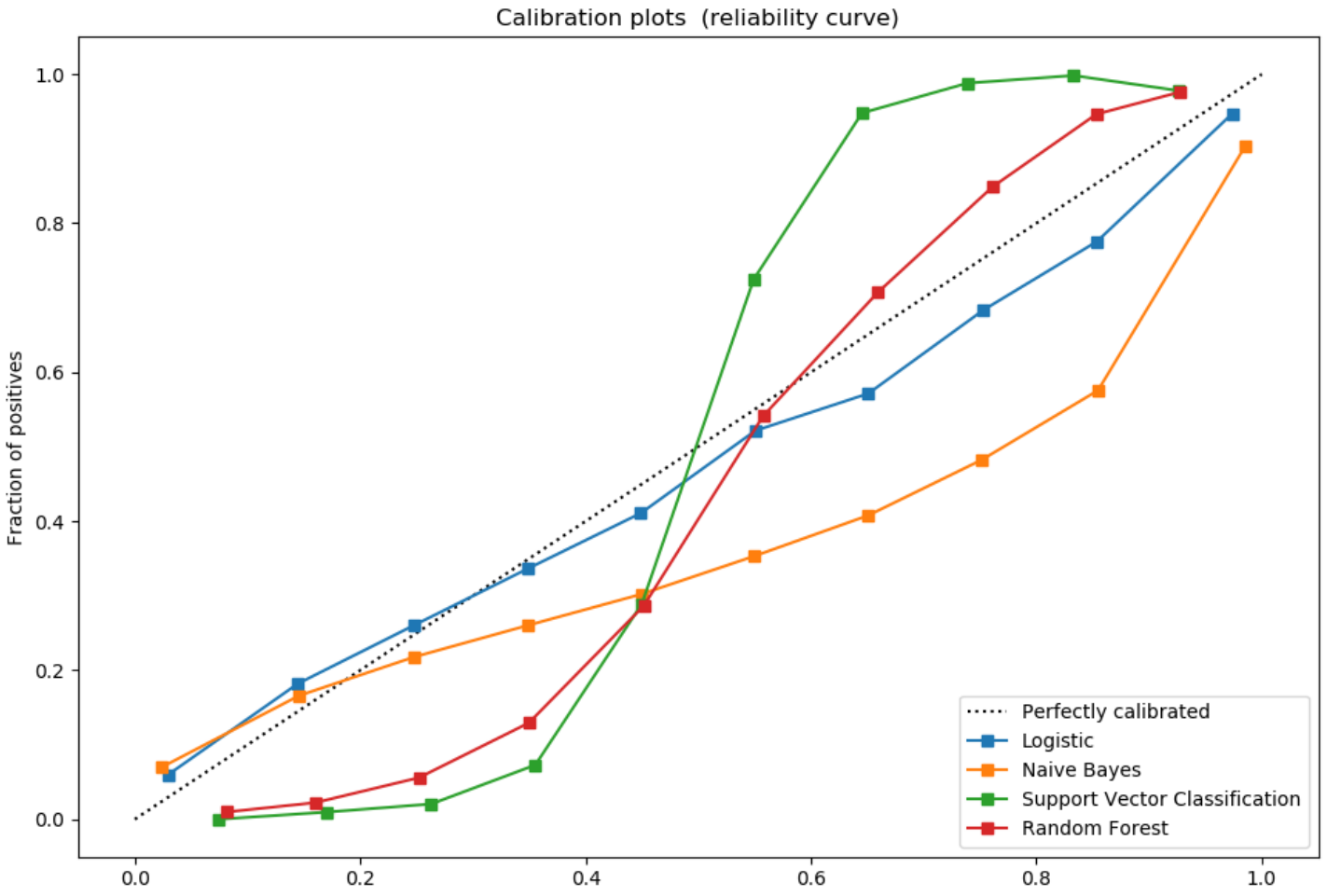
КАЛИБРОВКА ПЛАТТА

- 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*** $a(x)$.

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*** $\boldsymbol{a(x)}$.

- $\pi(x; \alpha; \beta) = \sigma(\alpha \cdot \boldsymbol{a(x)} + \beta) = \frac{1}{1+e^{-(\alpha \cdot a(x)+\beta)}}$

Изотоническая регрессия для калибровки вероятностей

Калибровка вероятностей — это процесс преобразования выходов модели (обычно вероятностей) таким образом, чтобы они соответствовали истинным вероятностям события. Например, если модель предсказывает вероятность 0.8 для класса 1, мы хотим, чтобы примерно 80% таких объектов действительно принадлежали к классу 1.

Изотоническая регрессия используется для этого, так как она:

1. Обеспечивает монотонную зависимость между входными значениями (сырыми вероятностями модели) и откалиброванными значениями.
2. Минимизирует квадратичную ошибку на обучающих данных.

Математическая постановка

Пусть:

- $\hat{p}_i$ — предсказания модели (сырые вероятности или оценки).
- $y_i \in \{0, 1\}$ — истинные метки классов.

Изотоническая регрессия минимизирует следующую ошибку:

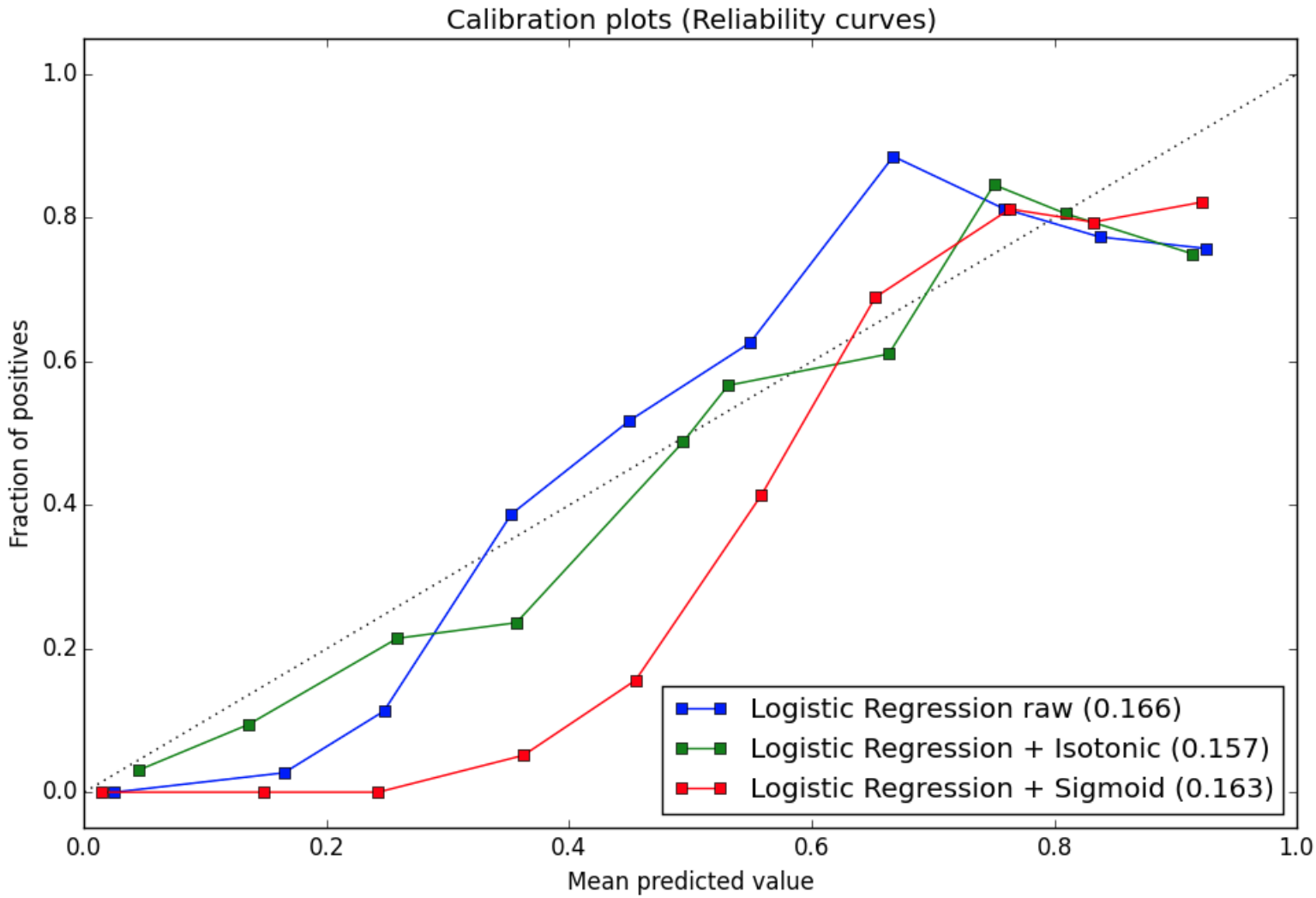
$$\min_f \sum_{i=1}^{n} (y_i - f(\hat{p}_i))^2,$$

где f — монотонно неубывающая функция (изотоническая регрессия), которая калибрует вероятности.

После обучения, для любого нового предсказания $\hat{p}$, откалиброванная вероятность вычисляется как:

$$\hat{p}_{\text{cal}} = f(\hat{p}).$$

РАЗЛИЧНЫЕ КАЛИБРОВКИ



Как это работает?

1. **Сортируем предсказания**.
   Сначала мы располагаем все предсказания модели $\hat{p}$ по возрастанию:

   $$\hat{p}_1 \leq \hat{p}_2 \leq \cdots \leq \hat{p}_n.$$

   К каждому из них привязываем соответствующую истинную метку y (0 или 1).

2. **Группируем соседей, нарушающих порядок монотонности**.
   Например, если модель предсказала вероятность $\hat{p}_i = 0.4$, а из реальных данных видно, что только 30% таких случаев относятся к положительному классу, то это нужно исправить. Изотоническая регрессия объединяет такие точки в группы и вычисляет для них общее значение вероятности, чтобы соблюдалась монотонность.

3. **Вычисляем вероятности в группах**.
   Для каждой группы вычисляем среднее истинных значений y (процент положительных примеров). Это значение станет откалиброванной вероятностью для всех точек в группе.

4. **Создаём преобразующую функцию**.
   В итоге мы получаем кусочно-постоянную монотонную функцию $f(\hat{p})$, которая отображает "сырые" вероятности $\hat{p}$ в откалиброванные.

Предположим, у нас есть следующие данные:

- Сырые вероятности модели: $\hat{p} = [0.1, 0.4, 0.35, 0.8]$;
- Истинные метки классов: $y = [0, 0, 1, 1]$.

1. **Сортируем данные**:
   - Упорядочим $\hat{p}$:
     $[0.1, 0.35, 0.4, 0.8]$,
     и соответствующие метки y:
     $[0, 1, 0, 1]$.
2. **Ищем нарушения монотонности**:
   - Для $\hat{p} = 0.35$ модель ошиблась: вероятность выше, чем у $\hat{p} = 0.4$, но реальная метка говорит обратное.
   - Изотоническая регрессия объединяет эти точки ($\hat{p} = 0.35$ и $\hat{p} = 0.4$).
3. **Вычисляем средние вероятности для групп**:
   - Для объединённой группы ($\hat{p} = 0.35, 0.4$): средняя истинная вероятность $= (1 + 0)/2 = 0.5$.
   - Обновляем предсказания: $\hat{p}_{\text{cal}} = [0.1, 0.5, 0.5, 1.0]$.

Теперь все предсказания $\hat{p}_{\text{cal}}$ соответствуют вероятностной интерпретации:

- $\hat{p}_{\text{cal}} = 0.5$ означает, что 50% таких случаев действительно положительные.
- $\hat{p}_{\text{cal}} = 1.0$ означает, что 100% таких случаев действительно положительные.