Метод опорных векторов

Support Vector Machine (SVM)

История

- Первые идеи метода были предложены еще в 1950-е годы.
- Метод был создан на основе статистической теории обучения
- Метод стал известен и популярен после замечательной статьи (Вапник и др.) в 1992 г.
- В настоящее время метод успешно используется во многих областях.
- Метод также был модифицирован для задач регрессии.

Метод

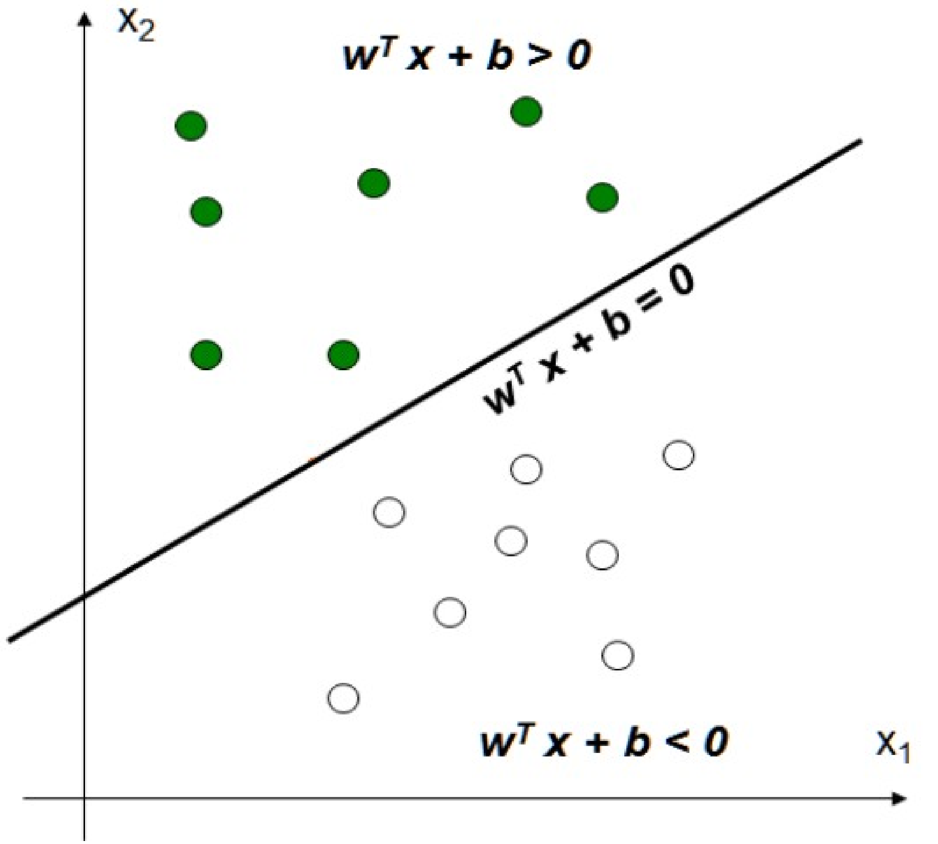
- Метод опорных векторов решает задачу классификации.
- Каждый элемент данных - точка в m–мерном пространстве $\mathbb{R}_m$.
- Формально: есть точки x_i, $i = 1, ..., m$, у точек есть метки $y_i \in \{-1, +1\}$.

Можно ли разделить данные гиперплоскостью и какая она?

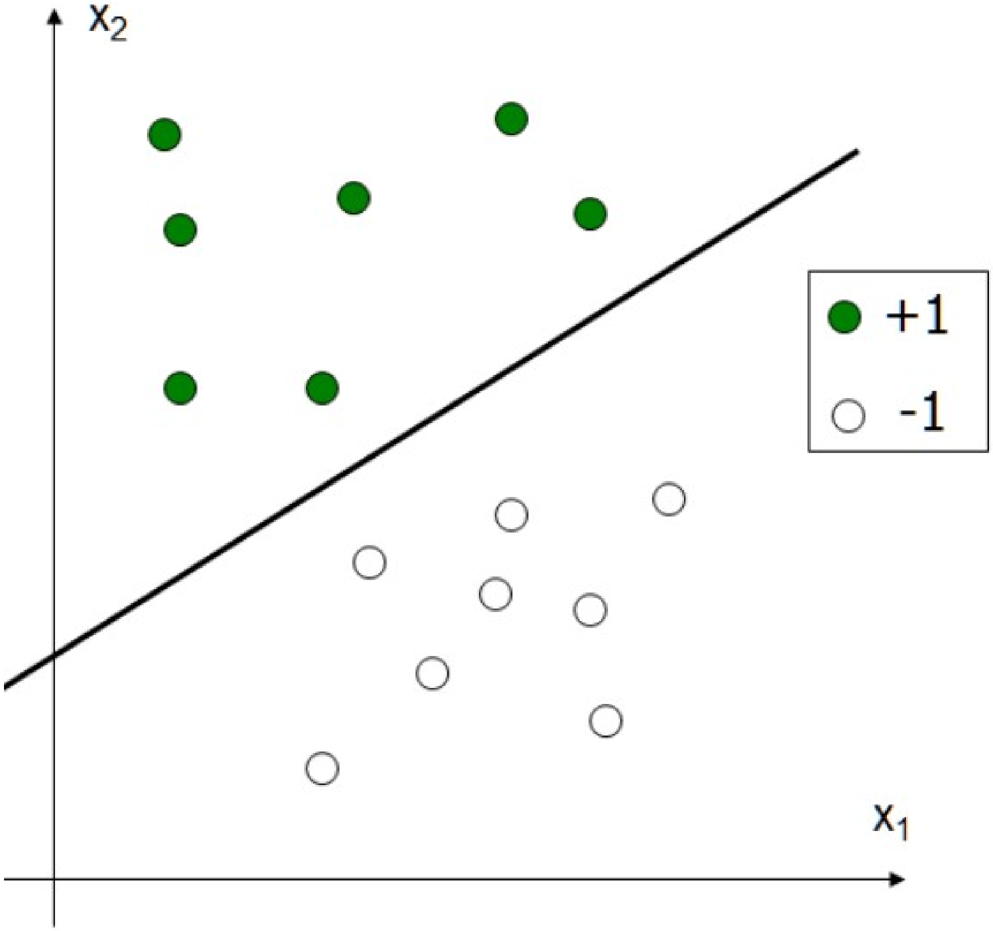
Классификация данных

$g(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b$ - линейная разделяющая функция (гиперплоскость)

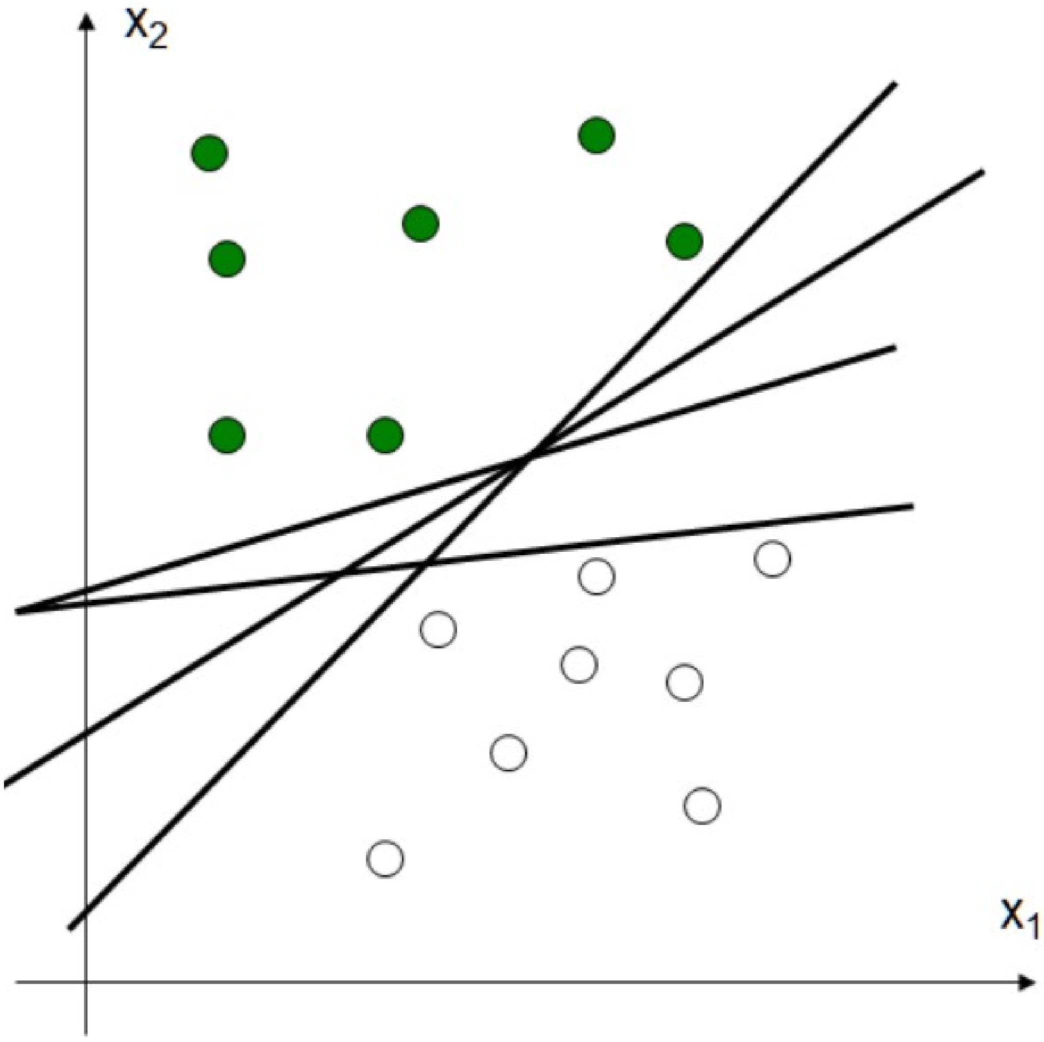
$g(x_1, ..., x_m) = \sum_{i=1}^{m} w_i x_i + b$



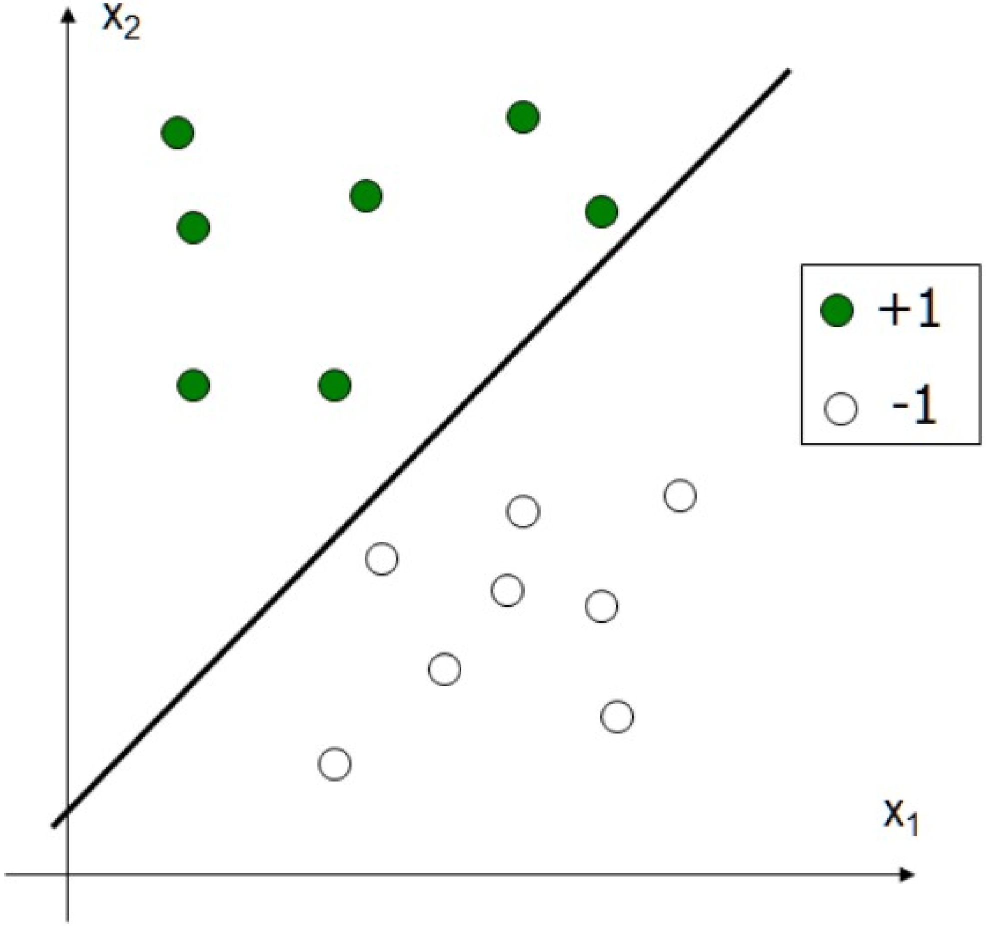
Первый вариант



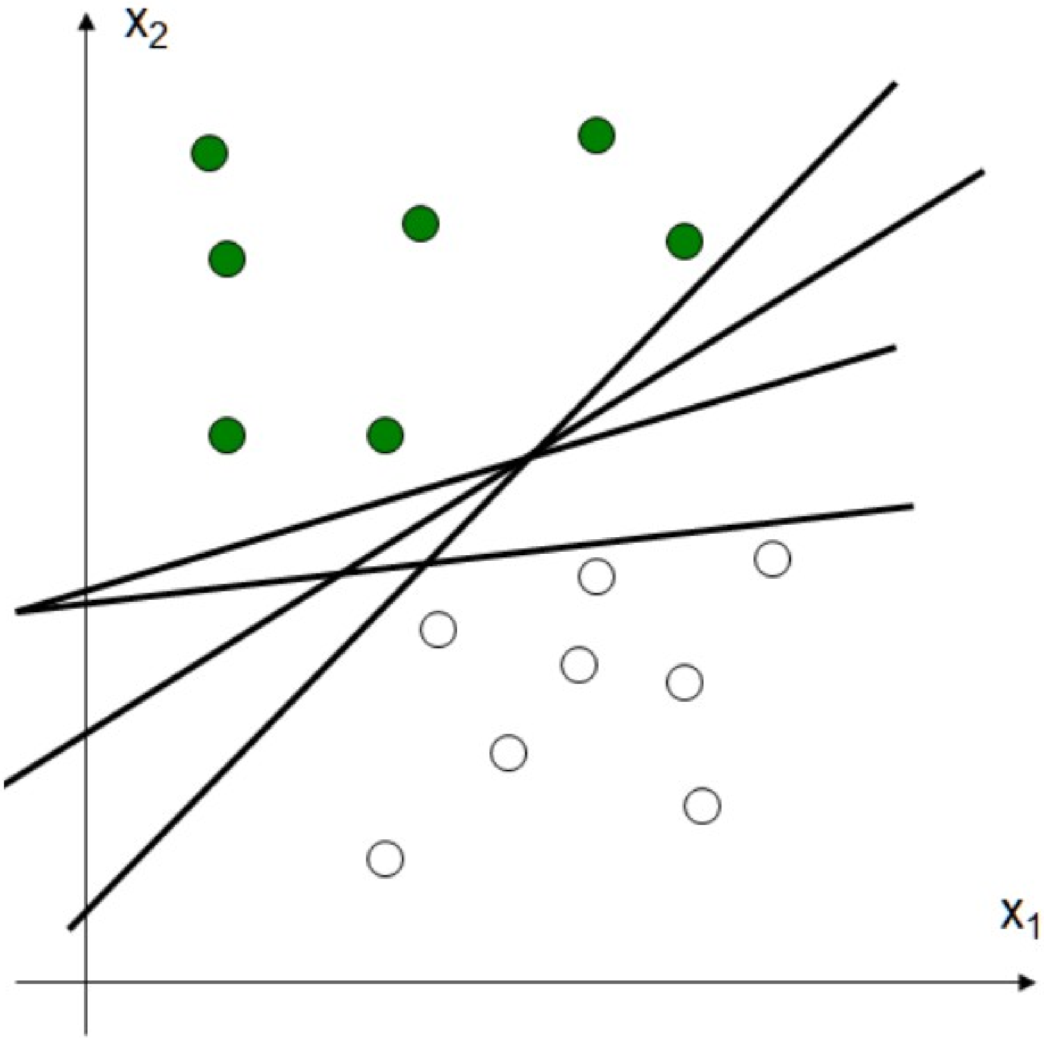
Еще несколько вариантов



Второй вариант

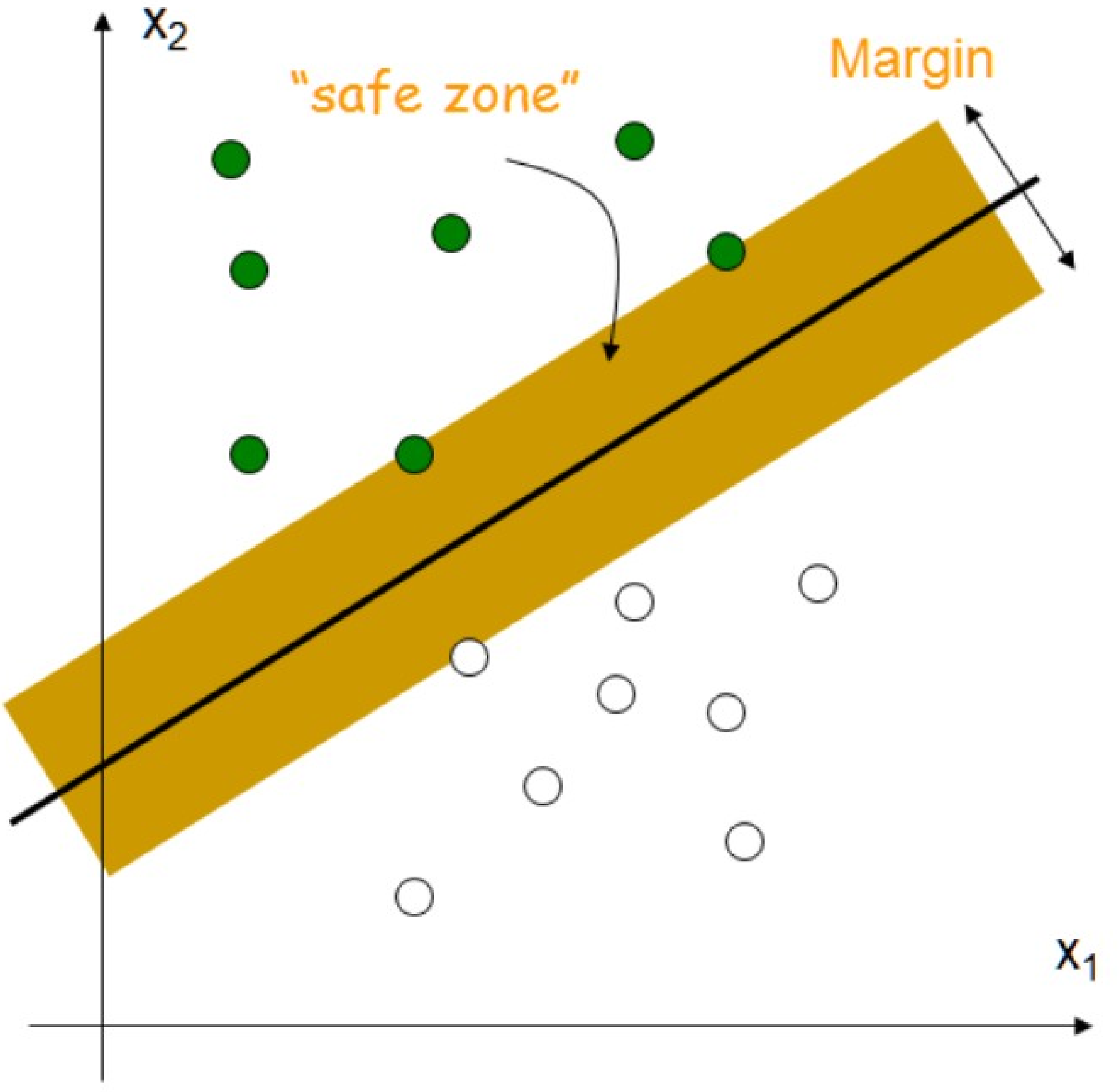


Еще несколько вариантов

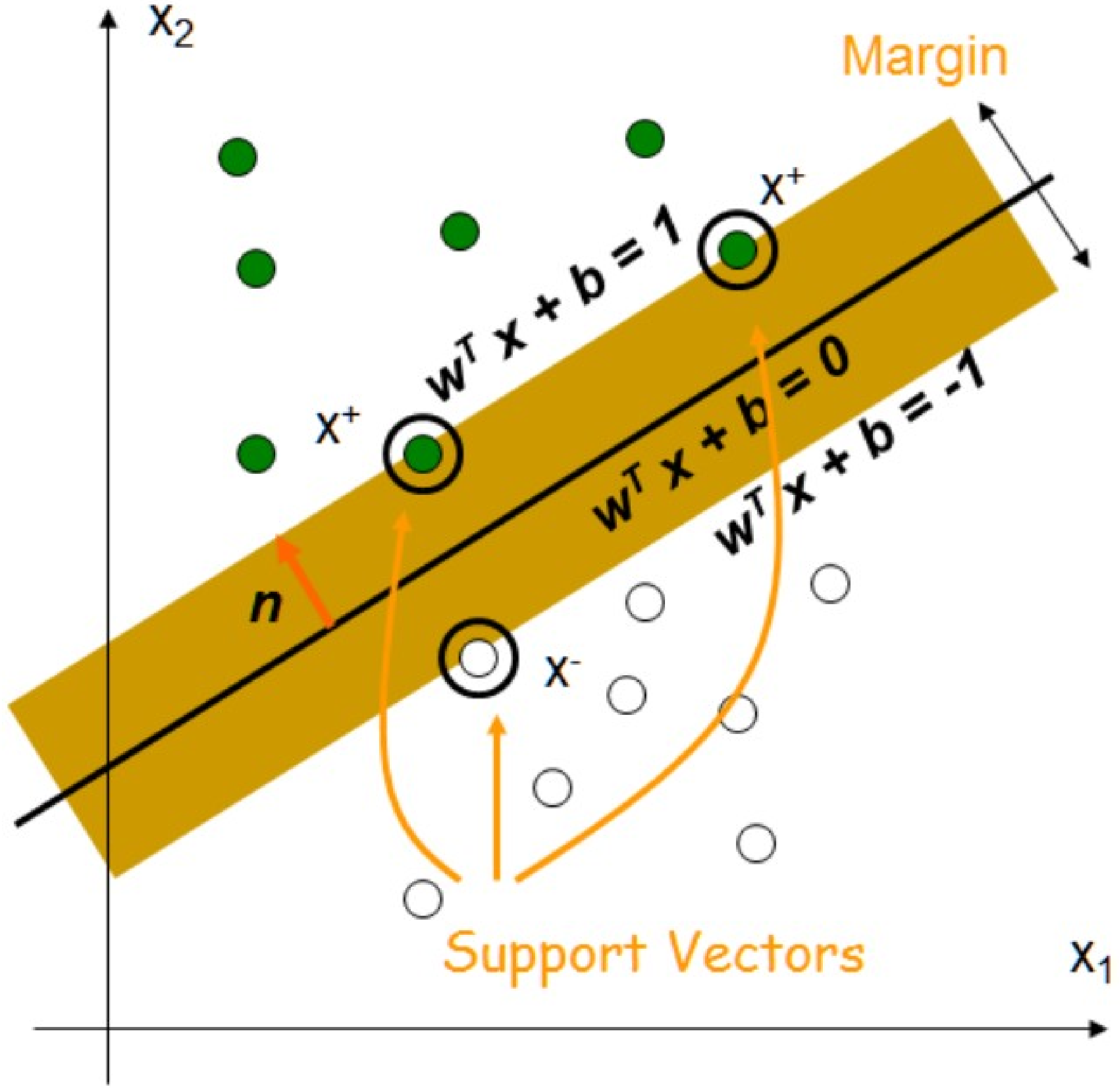


Оптимальная разделяющая гиперплоскость — это гиперплоскость, максимизирующая ширину **разделяющей полосы** и лежащая в середине этой полосы.
Иными словами, оптимальная разделяющая гиперплоскость максимизирует **зазор** (margin) между плоскостью и данными из обучающей выборки.
Если классы линейно разделимы и каждый содержит не менее одного элемента, то оптимальная разделяющая гиперплоскость единственна.

Разделяющая полоса



Поиск лучшей полосы



Как определить ширину полосы?

Без доказательства:

$$M = \frac{2}{\|\mathbf{w}\|} = \frac{2}{w_1^2 + w_2^2 + \ldots + w_m^2}$$

Формулировка задачи оптимизации:

$$\frac{2}{\|\mathbf{w}\|} \to \max_{\mathbf{w}}$$

при условии:

$$\begin{aligned} &\text{для } y_i = +1 : \quad \mathbf{w}^T\mathbf{x} + b \geq 1 \\ &\text{для } y_i = -1 : \quad \mathbf{w}^T\mathbf{x} + b \leq -1 \end{aligned}$$

Задача оптимизации

$$\frac{\|\mathbf{w}\|}{2} = \frac{1}{2}(w_1^2 + w_2^2 + \ldots + w_m^2) \rightarrow \min_{\mathbf{w}}$$

при условии:

$$y_i\left(\mathbf{w}^T\mathbf{x} + b\right) \geq 1$$

Это задача квадратической оптимизации с линейными ограничениями!

Задача оптимизации

Функция Лагранжа:

$$\mathcal{L}(\mathbf{x}, \mathbf{w}, b, \alpha) = \frac{1}{2}\sum_{i=1}^{m} w_i^2 - \sum_{i=1}^{n} \alpha_i \left(y_i \left(\mathbf{w}^T \mathbf{x} + b\right) - 1\right)$$

α_i, $i = 1, ..., n$ - множители Лагранжа.
Необходимые условия седловой точки функции Лагранжа

$$\frac{\partial \mathcal{L}}{\partial b} = -\sum_{i=1}^{n} \alpha_i y_i = 0, \;\; \frac{\partial \mathcal{L}}{\partial w_k} = w_k - \sum_{i=1}^{n} \alpha_i y_i x_i^{(k)} = 0, \;\; k = 1, ..., m$$

Задача оптимизации

Подставляя условия седловой точки в функцию Лагранжа, получаем двойственную задачу

$$\max\left(\sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j\right)$$

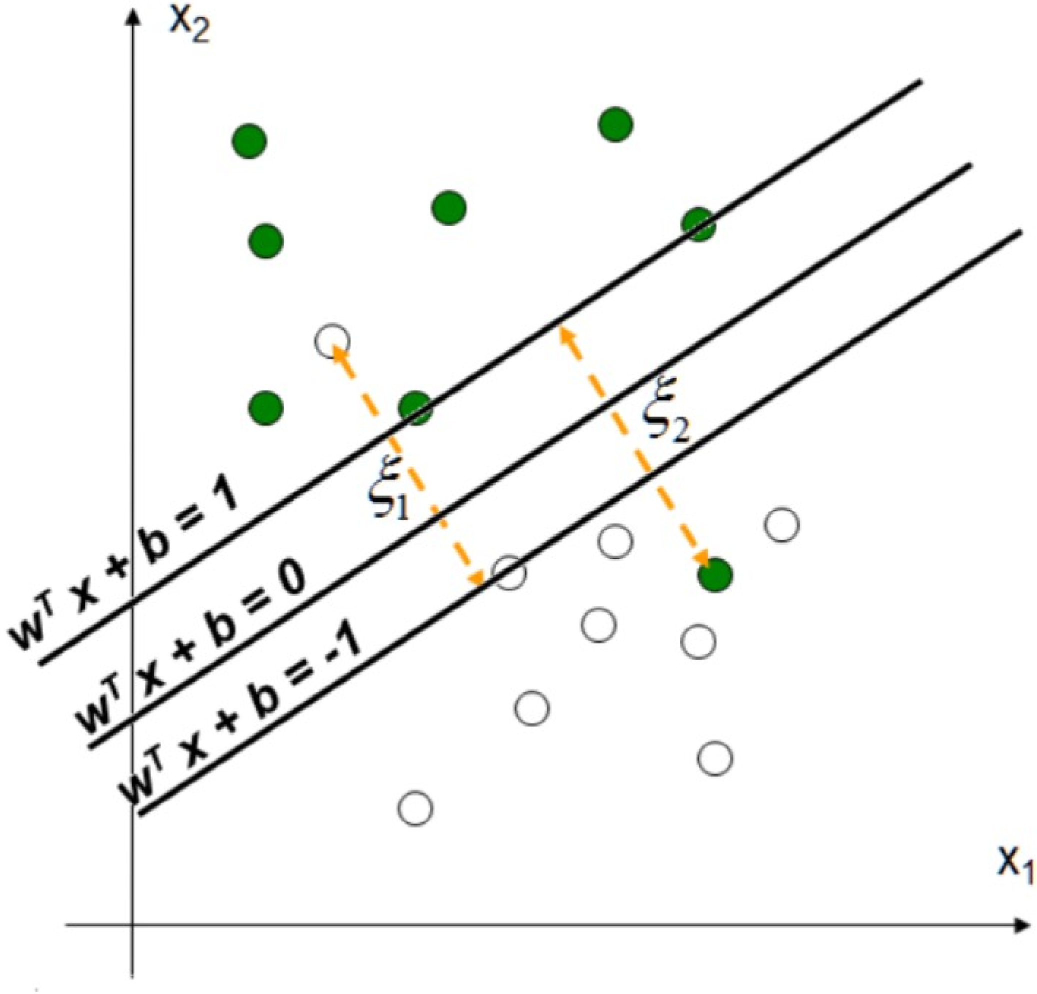
при ограничениях

$$\alpha_i \geq 0, i = 1, \ldots, n, \ \sum_{i=1}^{n}\alpha_i y_i = 0.$$

Разделяющая функция:

$$g(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b = \sum_{i=1}^{n}\alpha_i\mathbf{x}_i^T\mathbf{x} + b$$

Данные линейно не разделимы



Вспомогательные переменные ξ_i (неотрицательные ошибки) могут быть добавлены.

Переформулирование задачи оптимизации

$$\frac{\|\mathbf{w}\|}{2} = \frac{1}{2}(w_1^2 + w_2^2 + \ldots + w_m^2) \to \min_{\mathbf{w}}$$

при условии: $y_i\left(\mathbf{w}^T\mathbf{x} + b\right) \geq 1,\ i = 1, \ldots, n.$

$$\Downarrow$$

$$\frac{\|\mathbf{w}\|}{2} + C\sum_{i=1}^{n}\xi_i$$

при условии:

$$y_i\left(\mathbf{w}^T\mathbf{x} + b\right) \geq 1 - \xi_i, \quad \xi_i \geq 0,\ i = 1, \ldots, n.$$

C - штрафной параметр

Переформулирование задачи оптимизации

$$\max_{\alpha}\left(\sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j\right)$$

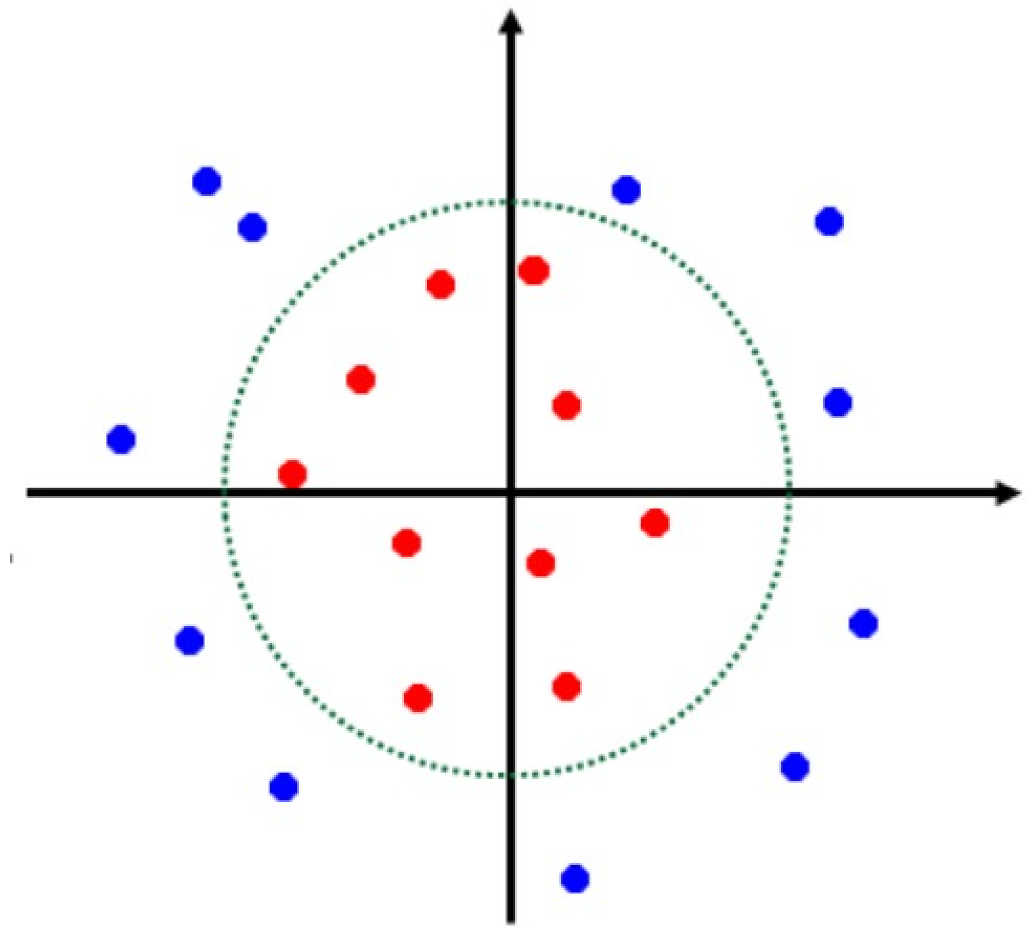
при ограничениях

$$0 \le \alpha_i \le C, i = 1, ..., n, \ \sum_{i=1}^{n}\alpha_i y_i = 0.$$

Разделяющая функция:

$$g(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b = \sum_{i=1}^{n}\alpha_i\mathbf{x}_i^T\mathbf{x} + b$$

Сложный случай

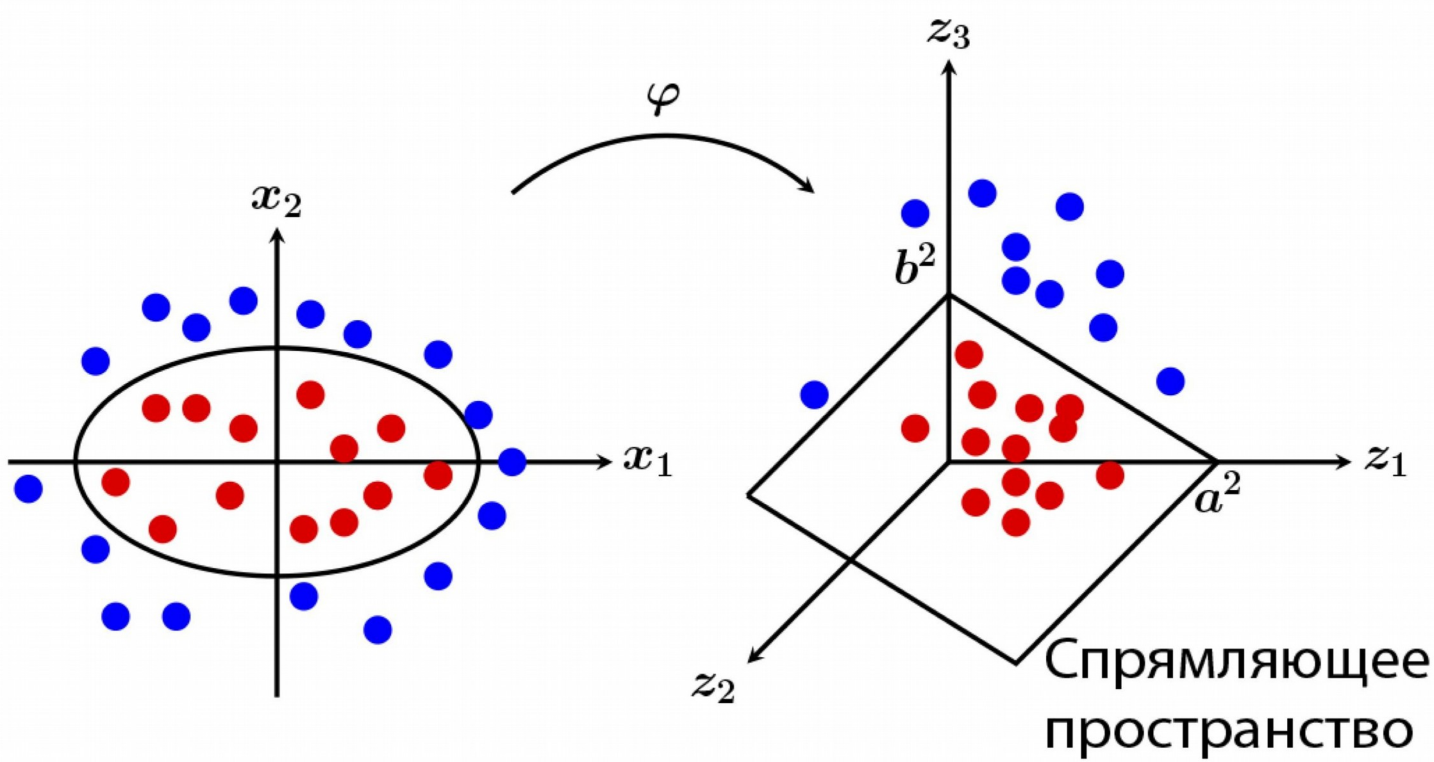


Как построить разделительную плоскость?

Ядра и спрямляющие пространства

Необходимо перейти от исходного пространства X в другое большей размерности, называемое **спрямляющим пространством**, с помощью некоторого отображения Φ.

Ядра и спрямляющие пространства



$$\varphi : (x_1, x_2) \to (x_1^2, \sqrt{2}x_1x_2, x_2^2)$$

$$\left(\frac{x_1}{a}\right)^2 + \left(\frac{x_2}{b}\right)^2 = 1 \;\to\; \frac{z_1}{a^2} + \frac{z_3}{b^2} = 1$$

Ядра и спрямляющие пространства

- Чтобы в двумерном пространстве с координатами $(x_1,\ x_2)$ решить задачу классификации квадратичной функцией, надо перейти в пятимерное пространство:
$$(x_1, \quad x_2, \quad x_1 \cdot x_2, \quad x_1^2, \quad x_2^2)$$
- Если решить задачу линейного разделения в этом новом пространстве, тем самым решится задача квадратичного разделения в исходном.
- Отображение
$$\varphi : \mathbb{R}^2 \to \mathbb{R}^5 : \varphi(x_1,\ x_2) = (x_1, \quad x_2, \quad x_1 \cdot x_2, \quad x_1^2, \quad x_2^2)$$
- Вектор в $\mathbb{R}^5$ теперь соответствует квадратичной кривой общего положения в $\mathbb{R}^2$, а функция классификации выглядит как
$$f(\mathbf{x}) = \mathrm{sign}\left(\varphi(\mathbf{w})\varphi(\mathbf{x}) - b\right)$$

Ядра и спрямляющие пространства

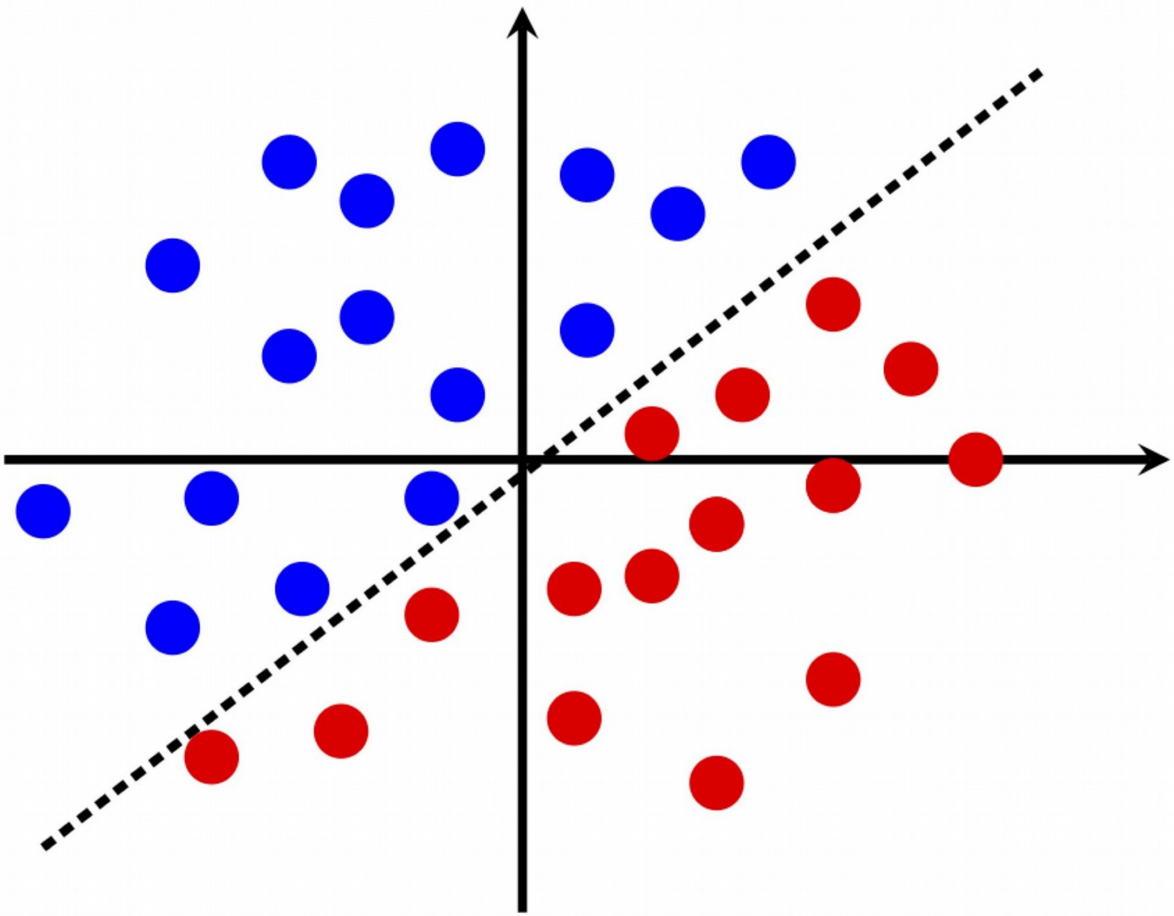
Разделяющая плоскость в спремляющем пространстве:

$$g(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i \mathbf{x}_i^T \mathbf{x} + b \rightarrow g(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i \varphi(\mathbf{x}_i)^T \varphi(\mathbf{x}) + b$$

Произведение $\varphi(\mathbf{x}_i)^T \varphi(\mathbf{x})$ играет необыкновенную роль.
Ядро

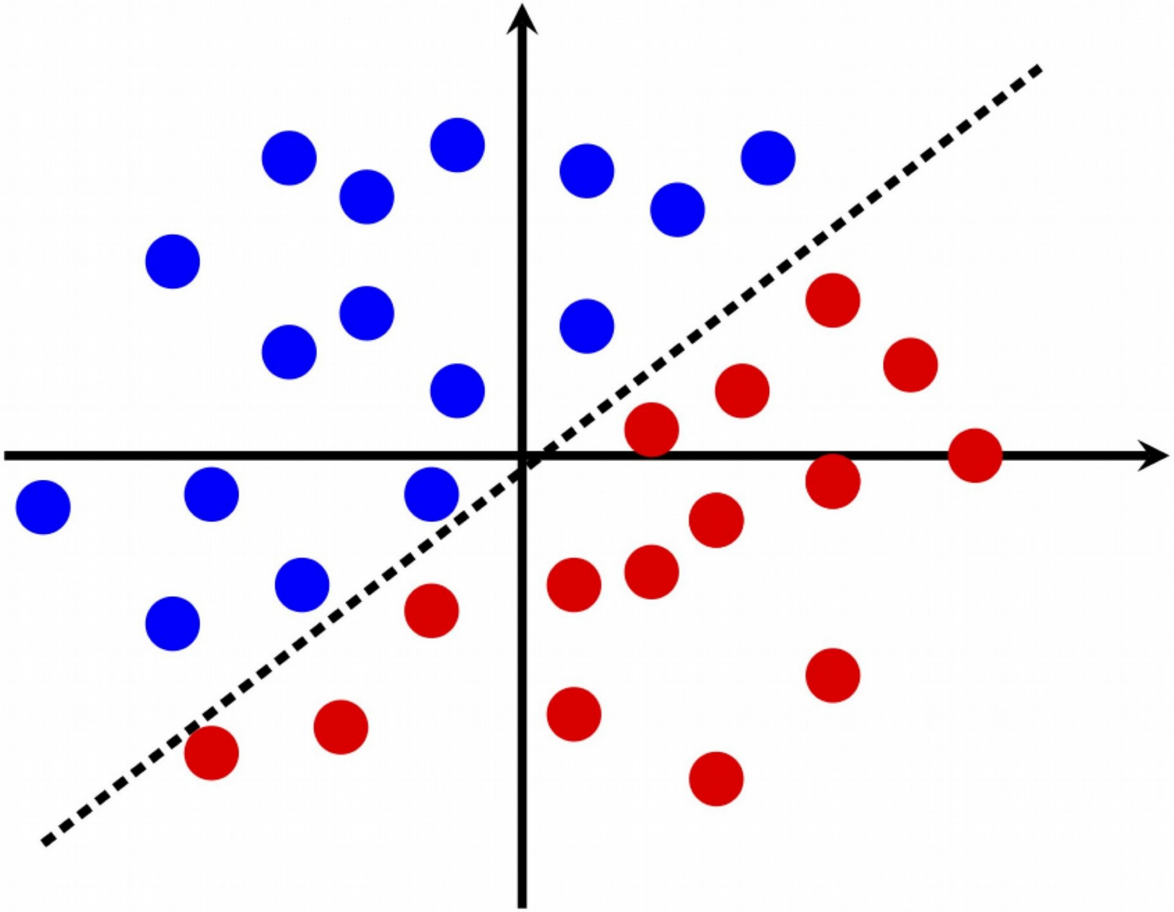
$$K(\mathbf{x}_i, \mathbf{x}_j) = \varphi(\mathbf{x}_i)\varphi(\mathbf{x}_j).$$

Ядра в методе опорных векторов
Линейное ядро



$$K(\mathbf{w}, x) = \langle \mathbf{w}, x \rangle$$

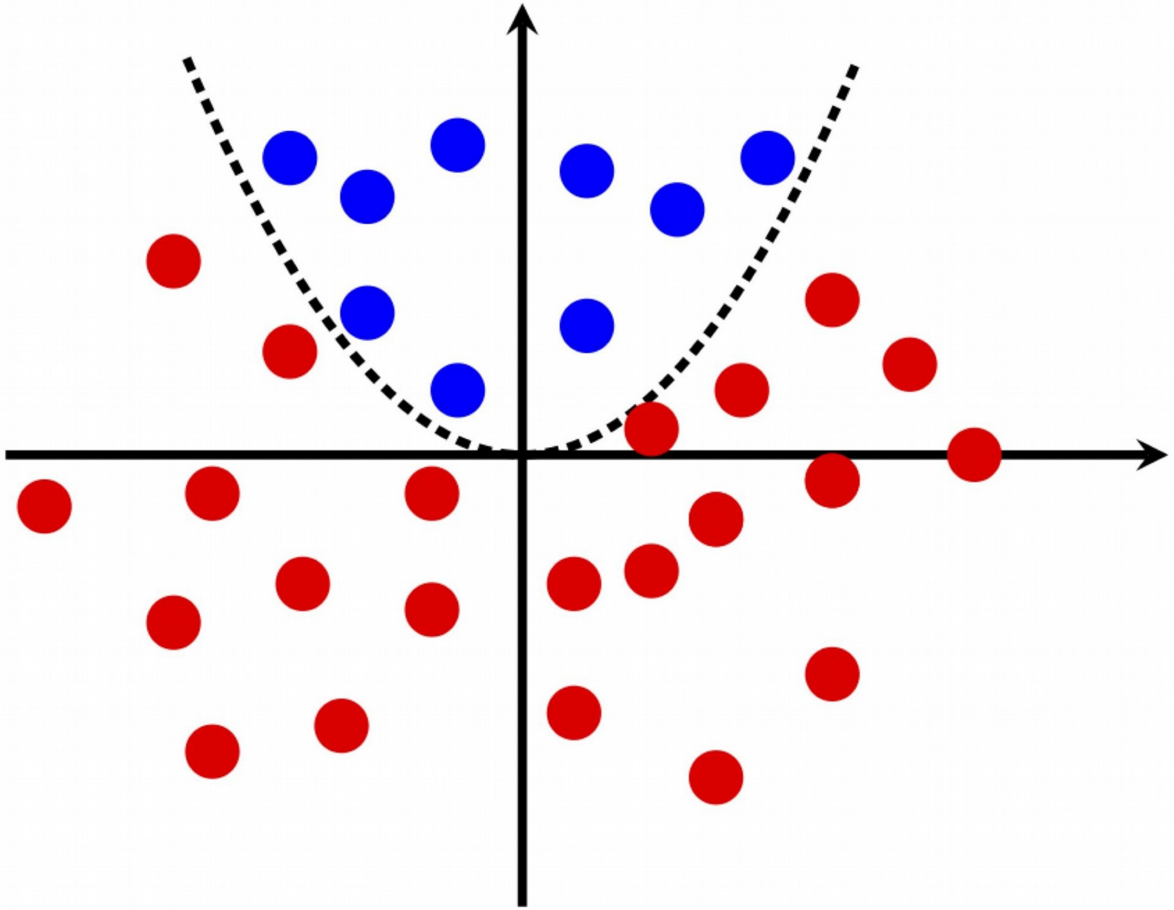
Ядра в методе опорных векторов
Линейное ядро



$$K(\mathbf{w}, x) = \langle \mathbf{w}, x \rangle$$

МФТИ

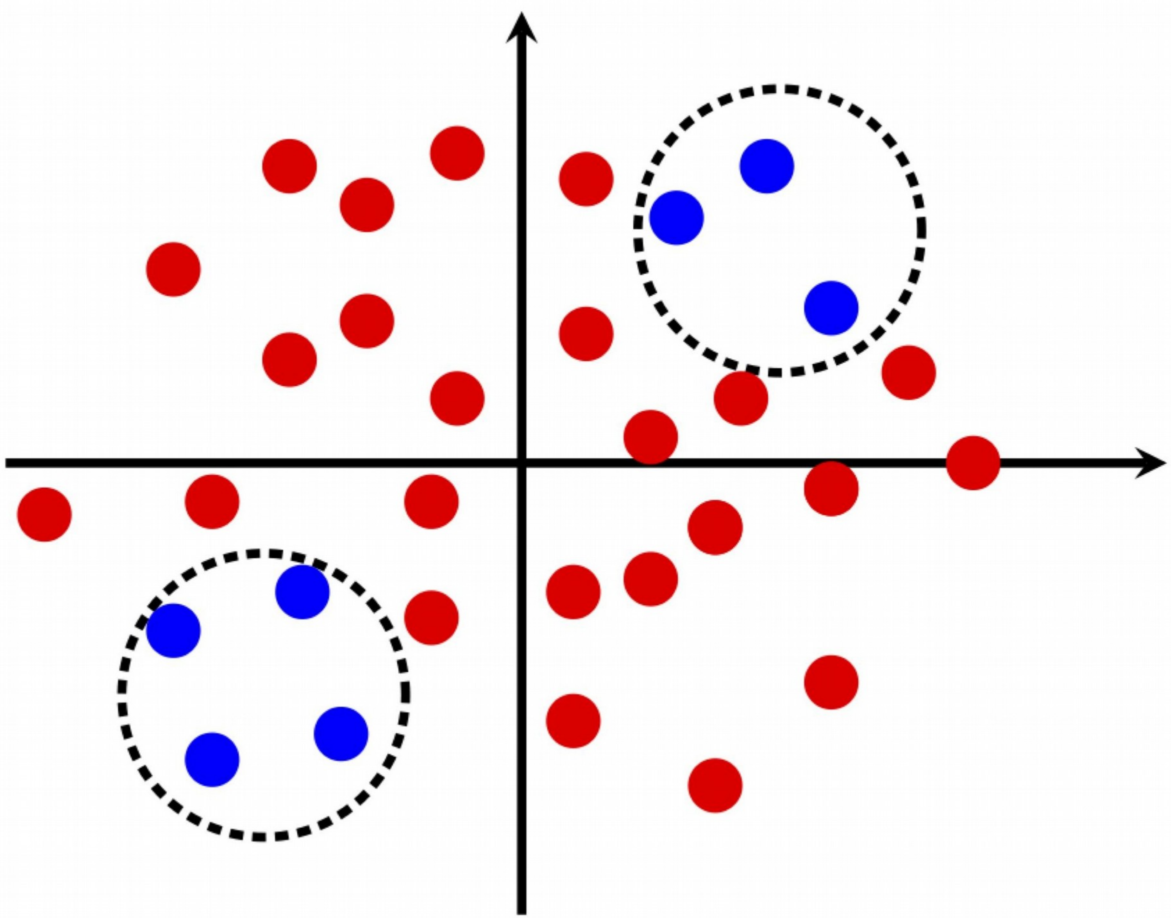
Ядра в методе опорных векторов
Полиномиальное ядро



$$K(\mathrm{w}, x) = (\gamma \langle \mathrm{w}, x \rangle + r)^d$$

Ядра в методе опорных векторов

Радиальное ядро



$$K(\mathrm{w}, x) = e^{-\gamma \| \mathrm{w} - x \|^2}$$

Ядра в методе опорных векторов
Ядра и библиотеки

- LibSVM — можно выбирать ядра
- LibLinear — только линейное ядро
- Scikit-learn — обёртка над LibSVM и LibLinear
- Vowpal Wabbit — только линейное ядро

Scikit-learn code

```
import numpy as np
X = np.array([[-1, -1], [-2, -1], [1, 1], [2, 1]])
y = np.array([1, 1, 2, 2])
```

```
from sklearn.svm import SVC
clf = SVC(kernel='linear')
clf = SVC.fit(X, y)
```

```
prediction = clf.predict([[0,6]])
```

Преимущества и недостатки SVM

- Преимущества SVM:
  - Задача выпуклого квадратичного программирования имеет единственное решение.
  - Позволяет рассматривать различные виды нелинейности, изменяя ядра или их параметры.
- Недостатки SVM:
  - Неустойчивость к шуму.
  - Нет общих подходов к оптимизации ядра под задачу.
  - Приходится подбирать параметр C.
  - Нет отбора признаков.