# 2.2. Линейные классификаторы и их обобщения

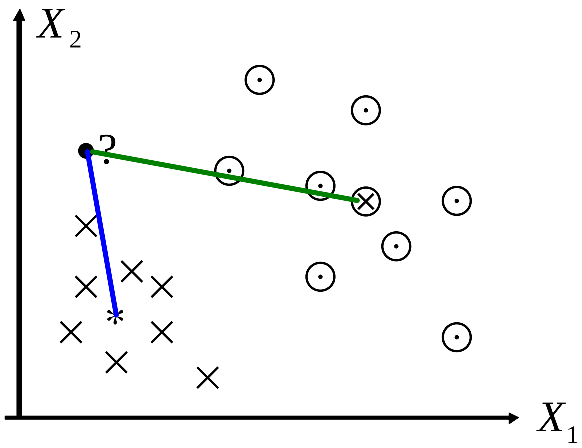
- Не используют вероятностные модели;
- Геометрический подход

## Геометрический подход к задаче классификации

Аналогия: данные = точки в метрическом пространстве.

Гипотеза о компактности: точки одного и того же класса должны быть близки друг к другу в некотором пространстве.

Примеры классификаторов:
a) kNN;
b) основанные на расстоянии: объект относится к классу, расстояние до центра тяжести которого минимально.

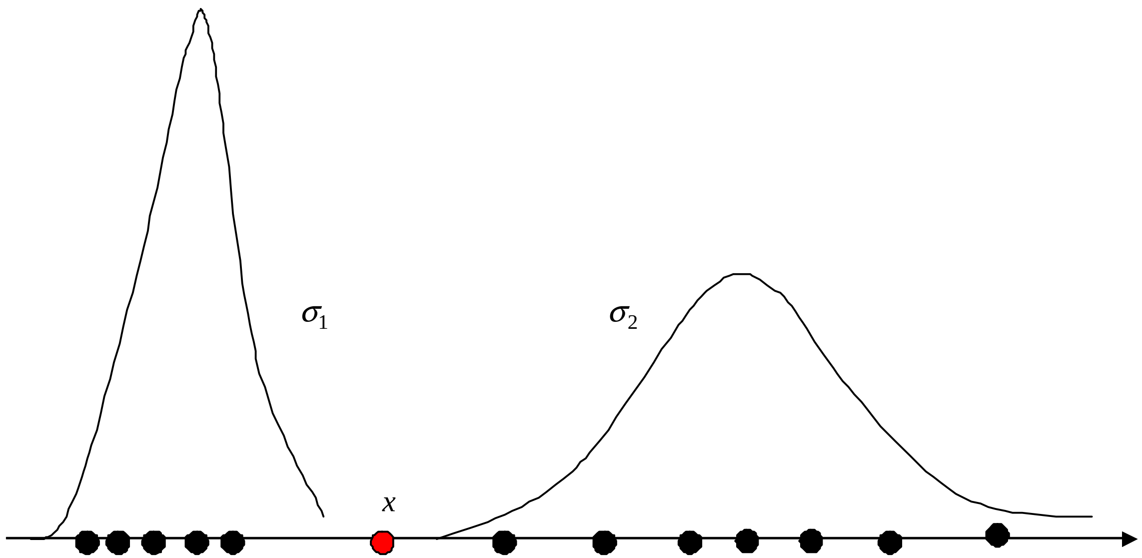


Расстояние Махалонобиса учитывает дисперсию:

$$d_M(a,b) = (a-b)^T \Sigma^{-1} (a-b),$$

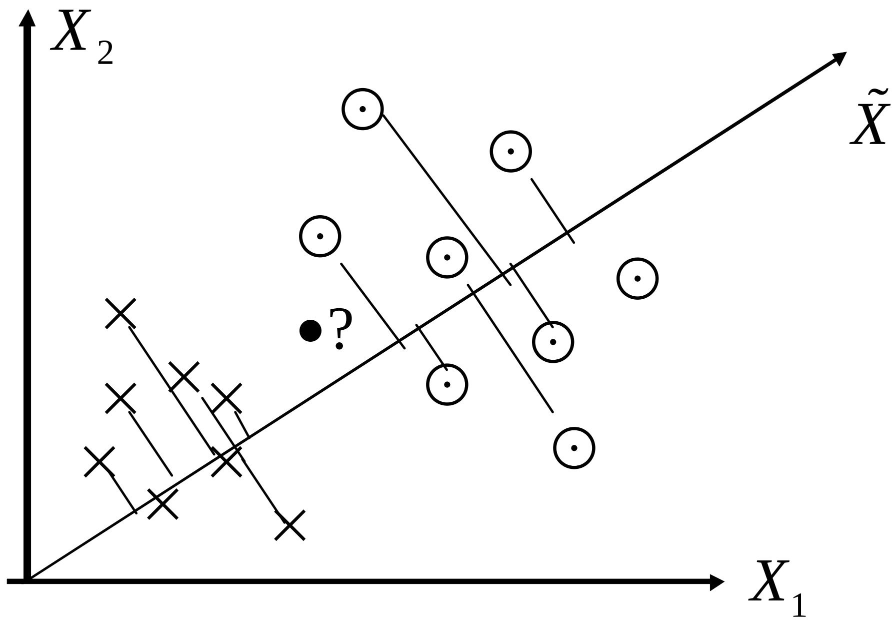
где $\Sigma$ это ковариационная матрица.

$n$=1: $d_M(a,b) = \dfrac{(a-b)^2}{\sigma^2}$ ($\sigma^2$ - дисперсия)




$\sigma_1 < \sigma_2 \Rightarrow x$ ближе ко второму классу

c) точки проецируются на пространство меньшей размерности таким образом, чтобы увеличить различия между классами;



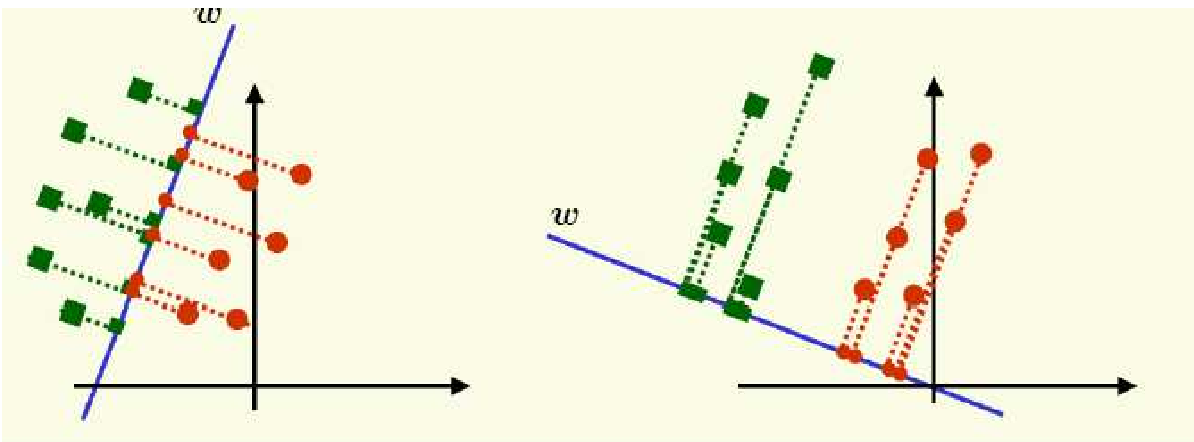
# Линейный дискриминант Фишера

Пусть число классов $K = 2$, и дана выборка

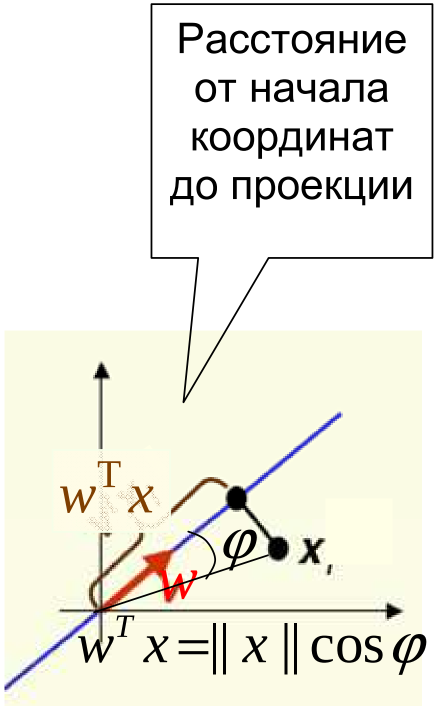
$$s = \left\{ (x^{(1)}, y^{(1)}), \ldots, (x^{(i)}, y^{(i)}), \ldots, (x^{(N)}, y^{(N)}) \right\},$$

$$x^{(i)} = \left( x_1^{(i)}, \ldots, x_j^{(i)}, \ldots, x_n^{(i)} \right)^T, \; y^{(i)} \in \{-1, 1\}.$$

Требуется найти такую прямую линию, чтобы проекции точек $\{x^{(1)}, \ldots, x^{(N)}\}$ были разделены как можно лучше.



Пусть $l$ - прямая, проходящая через начало координат, $w = (w_1, ..., w_n)^T$ - направляющий вектор, $\| w \| = 1$.




Координаты проекций точек: $\tilde{x}^{(i)} =< w, x^{(i)} >= w^T x$ . Центры проекций каждого класса:

$$\tilde{m}^{(k)} = \frac{1}{N(k)} \sum_{i \in C_k} \tilde{x}^{(i)}, \ k = 1, 2.$$

$$\tilde{m}^{(k)} = \frac{1}{N(k)} \sum_{i \in C_k} w^T x^{(i)} = w^T m^{(k)},$$

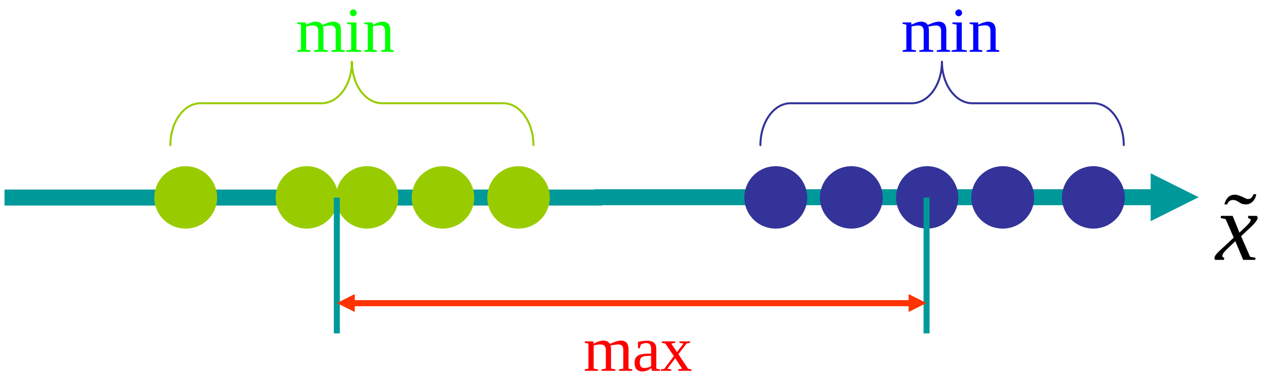
где $m^{(k)}$ это центроид $k$-го класса.

Разброс проекций каждого класса:

$$\tilde{s}_k^2 = \sum_{i \in C_k} \left( \tilde{x}^{(i)} - \tilde{m}^{(k)} \right)^2 .$$

Критерий качества:

$$J(w) = \frac{\left( \tilde{m}^{(1)} - \tilde{m}^{(2)} \right)^2}{\tilde{s}_1^2 + \tilde{s}_2^2} \to \max .$$



Матрицы разброса:

$$S_k = \sum_{i \in C_k} (x^{(i)} - m^{(k)})(x^{(i)} - m^{(k)})^T, \; k = 1, 2,$$

$S_W = S_1 + S_2$ - общая матрица разброса внутри классов.

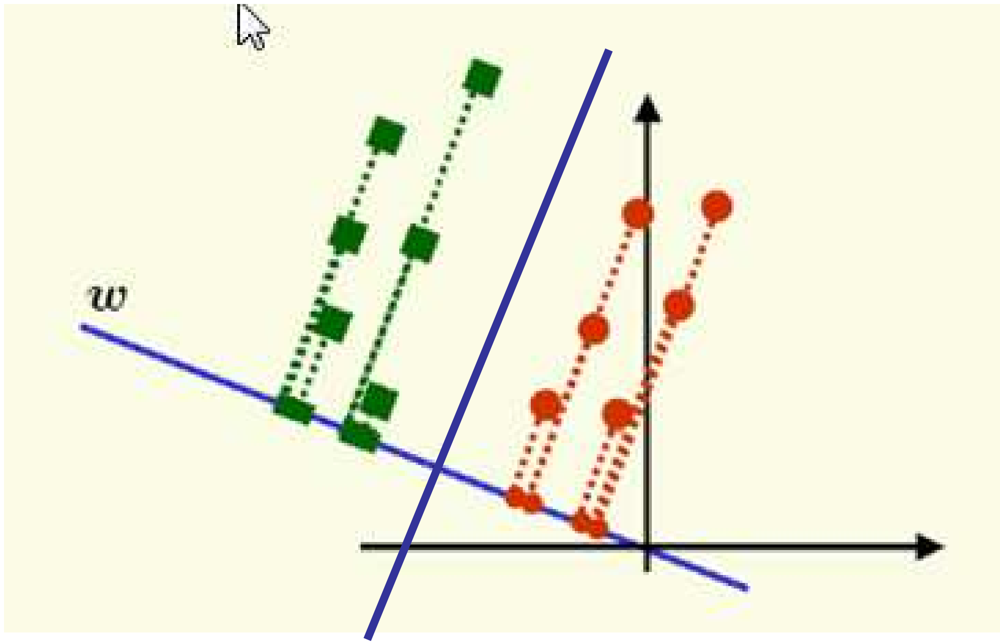
Пусть $S_B = (m_1 - m_2)(m_1 - m_2)^T$ матрица разброса между классами.

Критерий качества: $J(w) = \dfrac{w^T S_B w}{w^T S_W w} \to \max$

имеет решение:

$$\boxed{\tilde{w} = S_W^{-1} \, (m^{(1)} - m^{(2)})} \; .$$

Направляющий вектор: $w = \tilde{w} / \|\tilde{w}\|$ :



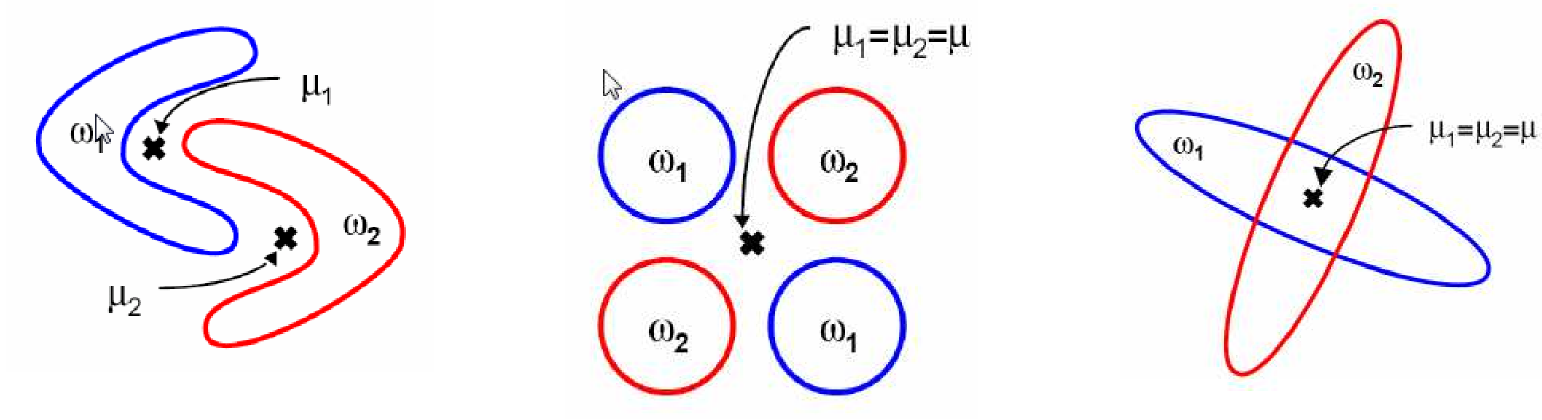

Решающая функция:

$$f(x) = sign\left(w^T x - w_0\right),$$

где $w_0 = w^T m, \quad m = \frac{1}{N}\left(N(1)m^{(1)} + N(2)m^{(2)}\right).$

# Недостатки

1. Неявно предполагается нормальное распределение, отклонение от которого приводит к невозможности классификации:




**2. Требуется обращение матрицы**

**3. Переобучение в пространстве большой размерности**

# Метод опорных векторов (SVM)

# Метод опорных векторов (SVM)

Пусть $K = 2$, дана выборка:

$$s = \left\{ (x^{(1)}, y^{(1)}), \ldots, (x^{(i)}, y^{(i)}), \ldots, (x^{(N)}, y^{(N)}) \right\},$$

$$x^{(i)} = \left( x_1^{(i)}, \ldots, x_j^{(i)}, \ldots, x_n^{(i)} \right).$$

$$Y(a) = \begin{cases} +1, & a \in class\ 1; \\ -1, & a \in class\ 2. \end{cases}$$

Линейный классификатор:

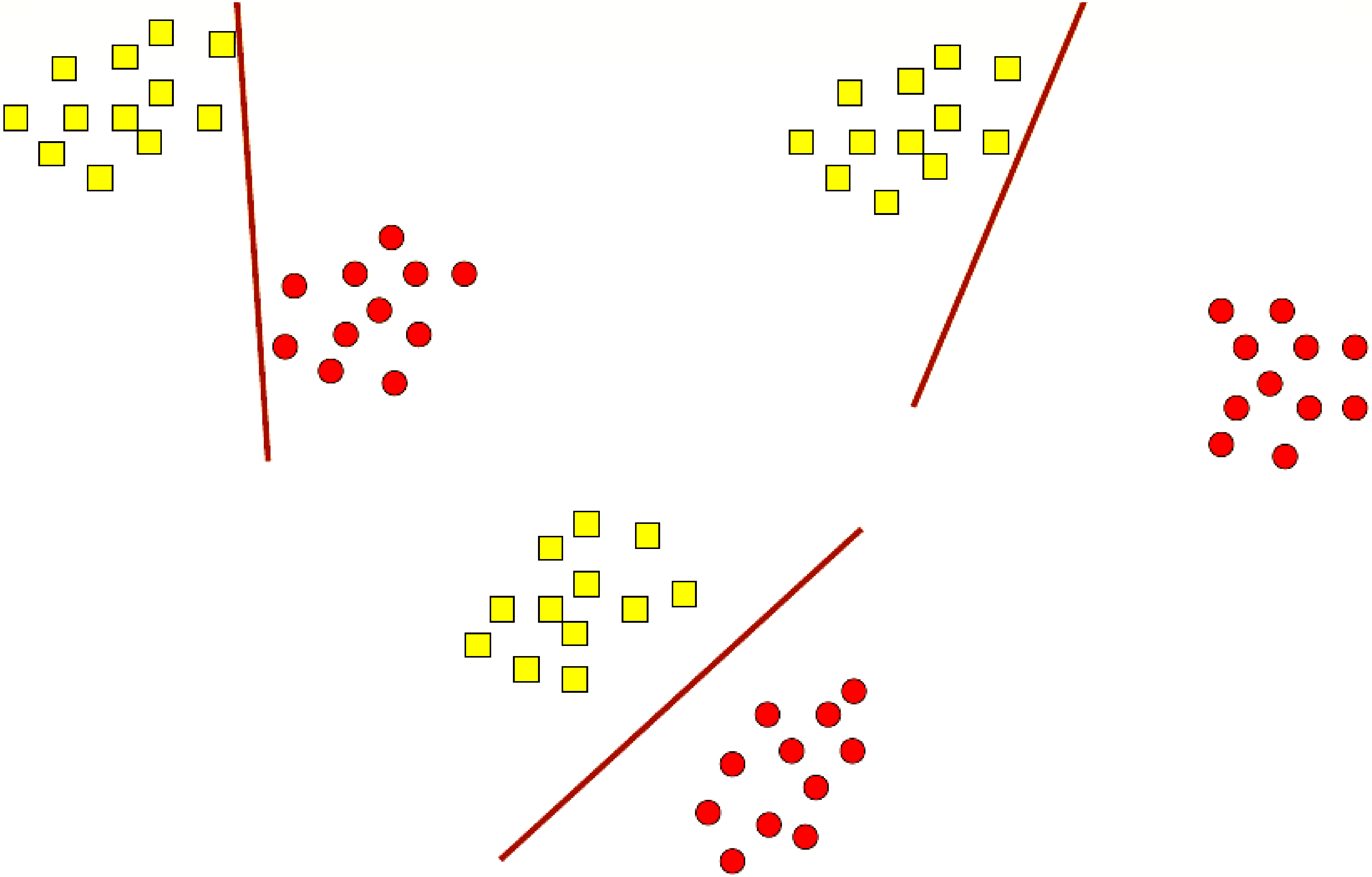
$$f(x) = sign\left( w^T x + b \right),$$

где $w = \left( w_1, \ldots, w_n \right)$ - вектор весов, $b$ - «смещение» или «порог».

Необходимо найти $w, b$ чтобы критерий качества был оптимальным.

Уравнение $w^T x + b = 0$ определяет гиперплоскость в $\mathbf{R}^n$.
Расстояние (отступ) от $z \in \mathbf{R}^n$ до разделяющей гиперплоскости пропорционально $| w^T z + b |$.

# Какая гиперплоскость лучше?



**Определение.** Гиперплоскость с $\Delta$- полосой называется гиперплоскость $w^T x + b = 0$ , $\| w \| = 1$, которая классифицирует $x$ следующим образом:

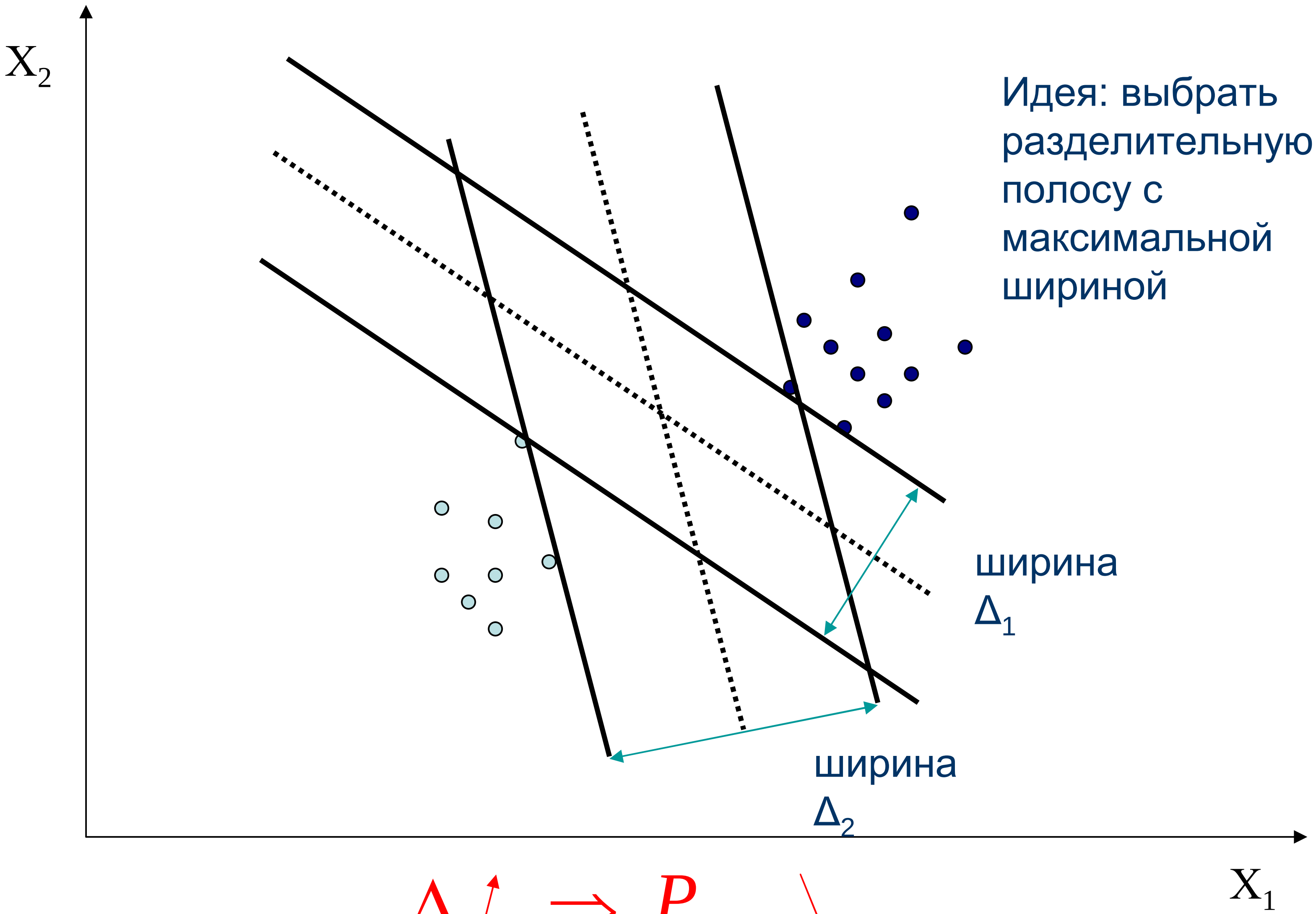
$$y = \begin{cases} 1, & w^T x + b \geq \Delta, \\ -1, & w^T x + b \leq -\Delta. \end{cases}$$

**Теорема (V.N.Vapnik).**
Предположим, что все $x$ принадлежат сфере радиуса $R$. Пусть построена гиперплоскость с $\Delta$- полосой, так что классифицирует выборку объемом $N$ с числом ошибок $m$. Тогда с вероятностью $1-\eta$ можно утверждать, что вероятность ошибочной классификации не превосходит

$$P_{error} < \frac{m}{N} + \frac{\varepsilon}{2}\left(1 + \sqrt{1 + \frac{4m}{N\varepsilon}}\right),$$

где $\varepsilon = 4\dfrac{h\left(\ln\dfrac{2N}{h} + 1\right) - \ln\dfrac{\eta}{4}}{N}$, $h \leq \min\left(\dfrac{R^2}{\Delta^2}, n\right) + 1.$

# Максимизация полосы



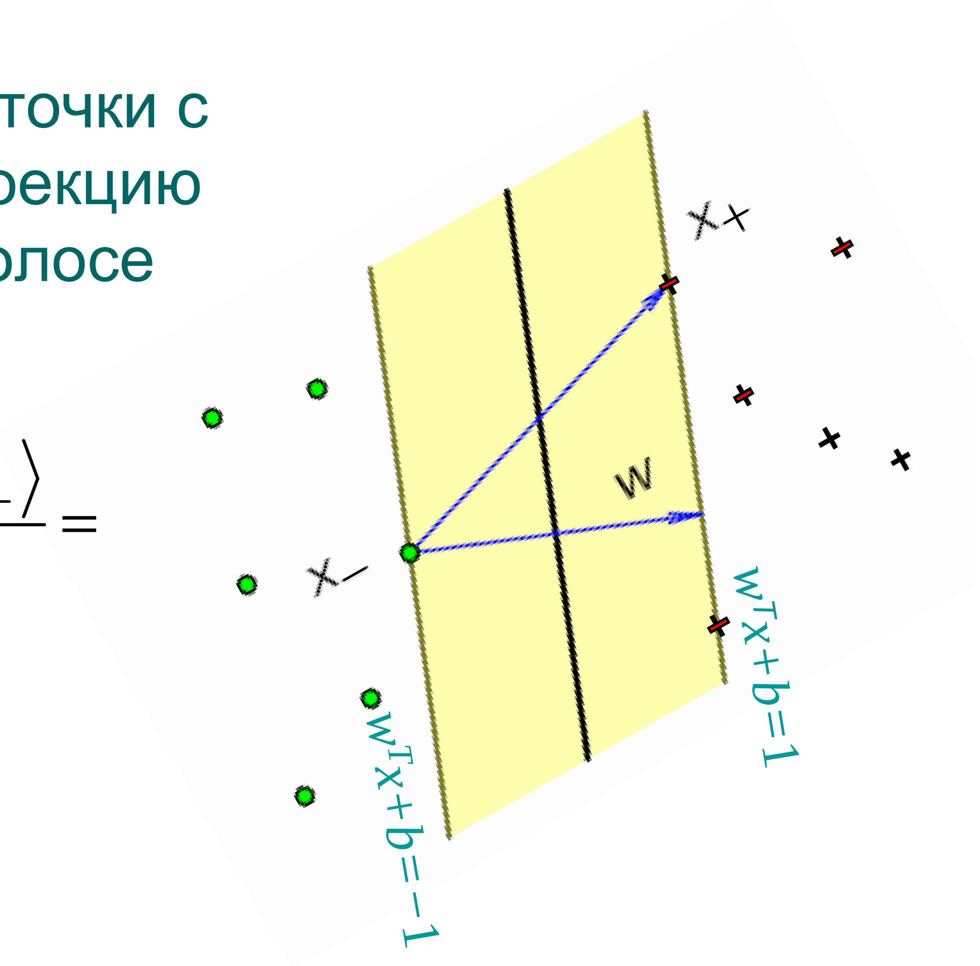

Идея: выбрать разделительную полосу с максимальной шириной

$$\Delta \uparrow \;\Rightarrow\; P_{error} \downarrow$$

Предположим, что классы линейно разделимы:
$\exists w, b : (w^T x^{(i)} + b) \cdot y^{(i)} > 0$ for all $i = 1, \ldots, N$.

Решающая функция не изменится, если умножить $w, b$ на константу. Сделаем нормировку: $\min\limits_i (w^T x^{(i)} + b) \cdot y^{(i)} = 1$.

Ширина полосы: возьмем 2 точки с границ полосы и найдем проекцию на линию, ортогональную полосе




$$\left\langle (x_+ - x_-), \frac{w}{\|w\|} \right\rangle = \frac{\langle w, x_+ \rangle - \langle w, x_- \rangle}{\|w\|} =$$

$$= \frac{1 - b - (-1 - b)}{\|w\|} = \frac{2}{\|w\|}$$

# Задача оптимизации:

$$\max \frac{2}{\|w\|}:$$
$$(w^T x^{(i)} + b) y^{(i)} \geq 1,\ i = 1, \ldots, N$$

=>

$$\frac{1}{2}\|w\|^2 \quad \rightarrow \min:$$
$$(w^T x^{(i)} + b) y^{(i)} \geq 1,\ i = 1, \ldots, N$$

Задача нелинейной оптимизации:

найти $\min\limits_x f(x)$ при условии $g_i(x) \leq 0,\ i = 1, \ldots, m$.

Условия Круша-Куна-Такера (необходимые):
если $x^*$ решение, то существует ненулевой вектор $\alpha \in \mathbf{R}^m$ такой что функция Лагранжа

$$L(x) = f(x) + \sum_{i=1}^{m} \alpha_i g_i(x),$$

удовлетворяет условиям:
- стационарности: $\min\limits_x L(x) = L(x^*)$;
- дополняющей нежесткости: $\alpha_i g_i(x^*) = 0,\ i = 1, \ldots, m$;
- неотрицательности: $\alpha_i \geq 0,\ i = 1, \ldots, m$.

Функция Лагранжа:

$$L(w,b,\alpha) = \frac{1}{2}\|w\|^2 - \sum_i \alpha^{(i)}\left((w^T x^{(i)} + b)y^{(i)} - 1\right).$$

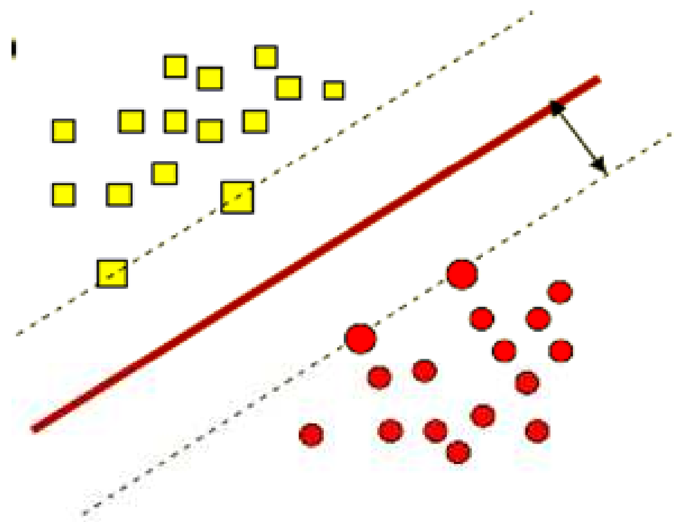
в точке минимума производные равны нулю:

$$\frac{\partial L}{\partial w} = w - \sum_i \alpha^{(i)} y^{(i)} x^{(i)} = 0 \quad \Rightarrow \quad w = \sum_i \alpha^{(i)} y^{(i)} x^{(i)} \ ,$$

$$\frac{\partial L}{\partial b} = -\sum_i \alpha^{(i)} y^{(i)} = 0 \quad \Rightarrow \quad \sum_i \alpha^{(i)} y^{(i)} = 0 \ .$$

$$\forall i: \ \alpha^{(i)}\left((w^T x^{(i)} + b)y^{(i)} - 1\right) = 0 \ \Rightarrow \ \alpha^{(i)} = 0 \text{ или } \ (w^T x^{(i)} + b)y^{(i)} = 1.$$

Ненулевые $\alpha^{(i)}$ будут иметь только те точки, которые лежат на границе полосы. Эти точки называются опорными векторами (support vectors).



Решение исходной задачи через множители Лагранжа:

$$w = \sum_{i:\text{sup vect.}} \alpha^{(i)} y^{(i)} x^{(i)} \ ,$$

$$b = 1/y^{(i)} - w^T x^{(i)} = y^{(i)} - w^T x^{(i)}$$ для любого $i : \alpha^{(i)} > 0$,

разделяющая функция:

$$l(x) = \langle w, x \rangle + b = \sum_i \alpha^{(i)} y^{(i)} \left\langle x, x^{(i)} \right\rangle + b.$$

Все точки кроме опорных векторов не влияют на решение.

Необходимое условие минимума: $(x^*, \alpha^*)$ это седловая точка функции Лагранжа

$$L(w,b,\alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1} \alpha^{(i)}\left((w^T x^{(i)} + b) y^{(i)} - 1\right) \to \min_{w,b} \max_{\alpha}$$

при условии: $\alpha^{(i)} \geq 0, \ \sum_i \alpha^{(i)} y^{(i)} = 0.$

$$L(w,b,\alpha) = \frac{1}{2} < w,w > - \sum \alpha^{(i)} y^{(i)} w^T x^{(i)} - b \underbrace{\sum \alpha^{(i)} y^{(i)}}_{=0} + \sum \alpha^{(i)} =$$

$= \frac{1}{2} < w,w > - \underbrace{\sum \alpha^{(i)} y^{(i)} \left(\sum \alpha^{(i)} y^{(i)} x^{(i)}\right) \cdot x^{(i)}}_{<w,w>} + \sum \alpha^{(i)}$. Так как $w = \sum \alpha^{(i)} y^{(i)} x^{(i)}$,

мы получаем задачу квадратичного программирования:

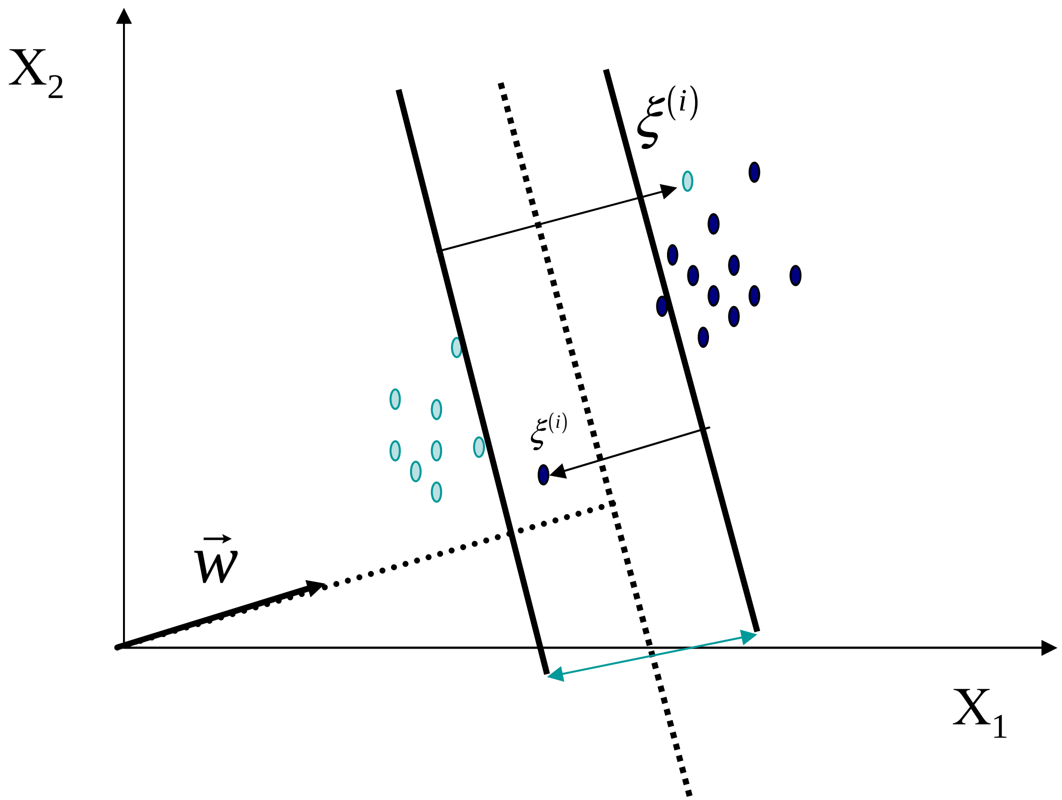
$$\boxed{W(\alpha) = \sum_i \alpha^{(i)} - \frac{1}{2} \sum_{i,j} y^{(i)} \alpha^{(i)} y^{(j)} \alpha^{(j)} \left\langle x^{(i)}, x^{(j)} \right\rangle \to \max}$$

при условии $\alpha^{(i)} \geq 0, \ \sum_i \alpha^{(i)} y^{(i)} = 0.$

a) целевая функция зависит от скалярного произведения точек

b) целевая функция выпуклая => имеется единственное решение.

# Линейно неразделимые классы




Фиктивные переменные $\xi^{(i)}$ - показывают насколько далеко нарушители находятся от границы.

Ограничения:

$$y^{(i)}\,\tilde{y}^{(i)} \geq 1-\xi^{(i)},$$

$$\forall x^{(i)}, \quad \xi^{(i)} \geq 0,$$

$$\tilde{y}^{(i)} = w^T x^{(i)} + b,$$

- расстояние от гиперплоскости (со знаком).

Требуется минимизировать $\frac{1}{2}\|w\|^2$ и $\sum_i \xi^{(i)}$.

Компромиссный вариант: $\frac{1}{2}\|w\|^2 + C\sum_i \xi^{(i)} \rightarrow \min$, $C$ это коэффициент штрафа.

# Задача Soft margin SVM:

$$\frac{1}{2}\|w\|^2 + C\sum_i \xi^{(i)} \to \min:$$

$$y^{(i)}\,\tilde{y}^{(i)} \geq 1 - \xi^{(i)},$$

$$\xi^{(i)} \geq 0, \quad i = 1, \ldots, N.$$

- необходимо максимизировать ширину разделительной полосы, исключая большие ошибки;
- При $C \to \infty$ задача сводится к предыдущей (Hard margin SVM);
- Требуется минимизировать не число неправильно классифицированных объектов, а суммарные расстояния до разделяющей гиперплоскости.

Задача квадратичного программирования:

$$\boxed{W(\alpha) = \sum_i \alpha^{(i)} - \frac{1}{2} \sum_{i,j} y^{(i)} \alpha^{(i)} y^{(j)} \alpha^{(j)} \left\langle x^{(i)}, x^{(j)} \right\rangle \to \max}$$

при условии:

$$0 \le \alpha^{(i)} \le C, \ \sum_i \alpha^{(i)} y^{(i)} = 0, \ \ i = 1, \ldots, N.$$

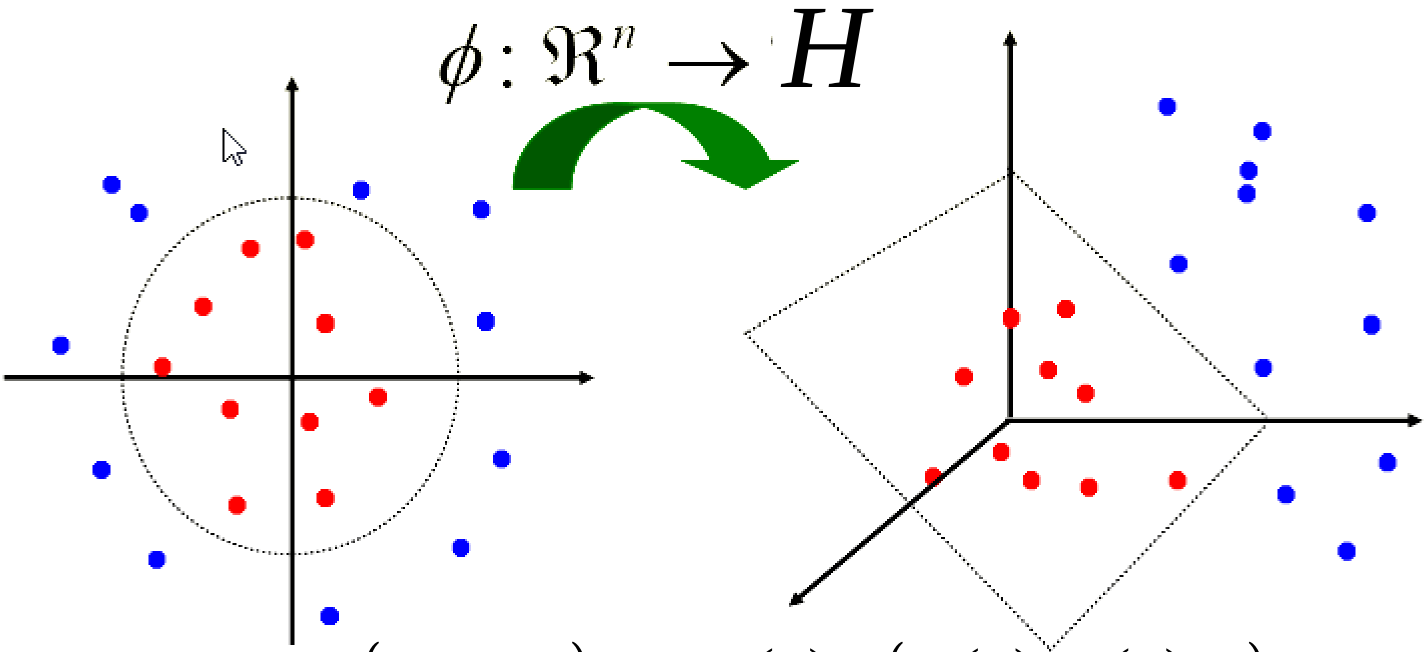
Оптимальная гиперплоскость:

$w = \sum_{i:\alpha^{(i)}>0} \alpha^{(i)} y^{(i)} x^{(i)}$ , $b = y^{(i)} - w \cdot x^{(i)}$ для любого $i : \alpha^{(i)} > 0$.

Решение зависит от опорных векторов и точек-нарушителей.

# Использование ядра для случая линейно неразделимых классов

Теорема (Cover). Нелинейное преобразование сложной задачи классификации образов в пространство более высокой размерности повышает возможность линейной разделимости классов.



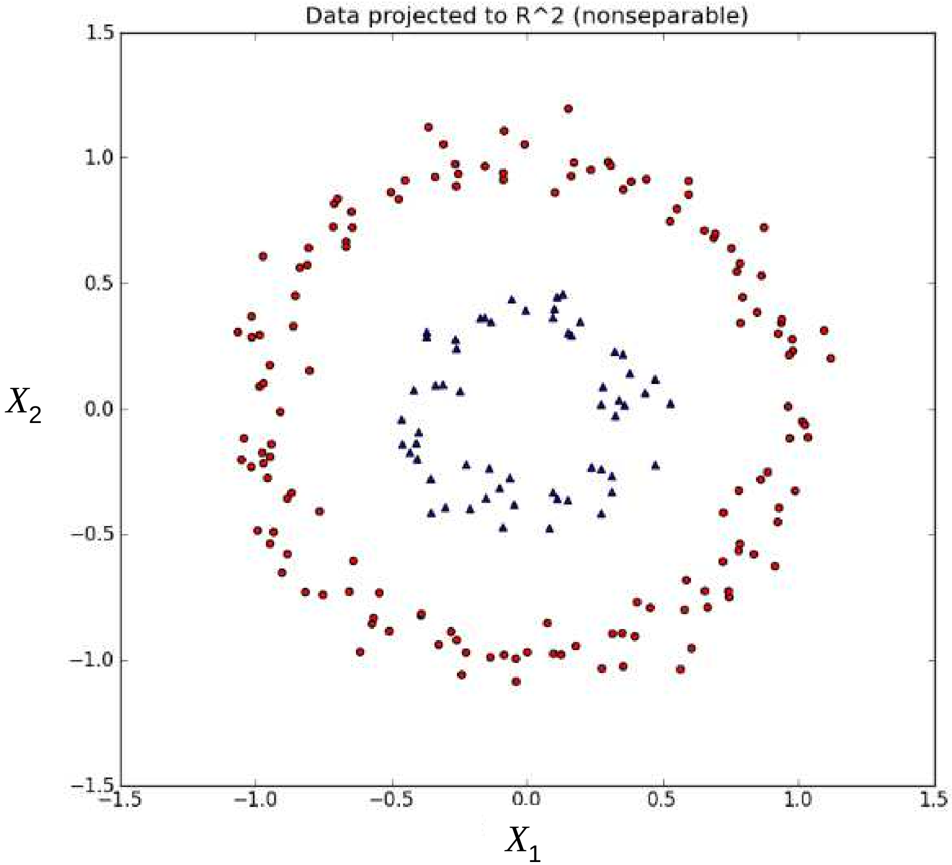
$$x = (x_1, \ldots, x_n) \Rightarrow \varphi(x) = (\varphi_1(x), \varphi_2(x), \ldots);$$

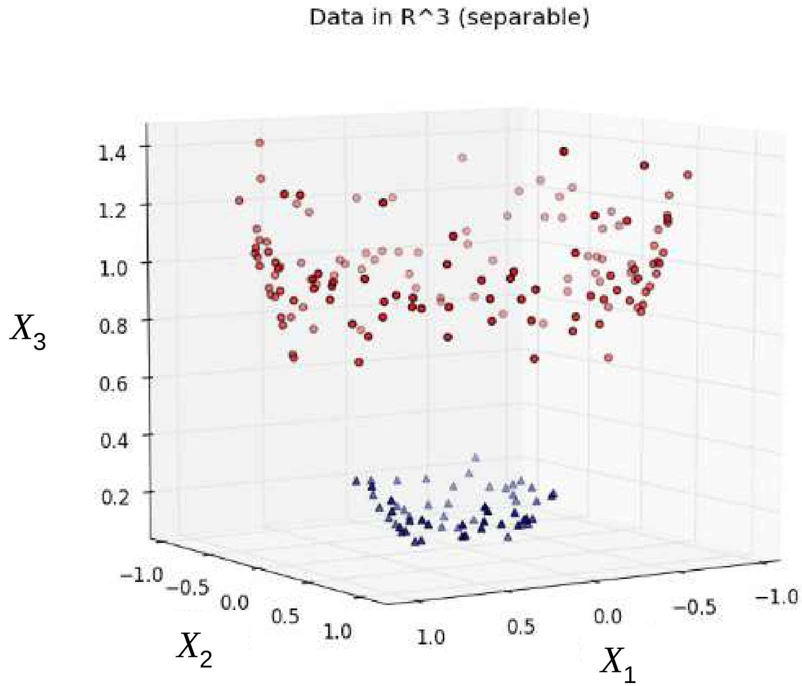
$$\langle x, x' \rangle \Rightarrow \langle \varphi(x), \varphi(x') \rangle = K(x, x') = \sum_{k=1}^{\infty} a_k \varphi_k(x) \varphi_k(x'),$$

where $K(x, x')$ is kernel

# Пример

$$(x_1, x_2) \Rightarrow (x_1, x_2, x_1^2 + x_2^2)$$



# Kernel Trick

Используем вместо скалярного произведения некоторое ядро:

$$\boxed{W(\alpha) = \sum_i \alpha^{(i)} - \frac{1}{2}\sum_{i,j} y^{(i)}\alpha^{(i)} y^{(j)}\alpha^{(j)} K\left(x^{(i)}, x^{(j)}\right) \to \max}$$

при условии:

$$0 \le \alpha^{(i)} \le C, \ \sum_i \alpha^{(i)} y^{(i)} = 0.$$

Получается, что мы неявным образом преобразуем пространство переменных, и в нем ищем линейное разделение. При этом никаких явных преобразований не требуется

На вероятность ошибки влияет в большей степени не размерность нового пространства (она может быть даже бесконечна), а ширина разделительной полосы

## Теорема Мерсера

Симметричная функция $K(x, x')$ из $L_2$, является ядром (т.е. определяет скалярное произведение в некотором пространстве), тогда и только тогда, когда

$$\iint K(x, x') g(x) g(x') \, dx dx' \geq 0 \text{ for any } g \in L_2,$$

или:

для любого набора точек $\{x_i\}_{i=1}^N$ in $\mathbf{R}^n$ и вещественных чисел $\{c_i\}_{i=1}^N$, матрица $\mathbf{K} = (K(x^{(i)}, x^{(j)}))_{i,j=1}^N$ является неотрицательно определенной:

$$\sum_{i,j=1}^N c_i c_j K(x^{(i)}, x^{(j)}) \geq 0.$$

## Примеры ядер

1. Квадратичное

$$K(x,x') = <x,x'>^2;$$

2. Полиномиальное ядро

$$K(x,x') = <x,x'>^d;$$

3. Радиальная базисная функция (RBF):

$$K(x,x') = \exp(-\gamma \| x - x' \|^2) \quad \text{или}$$

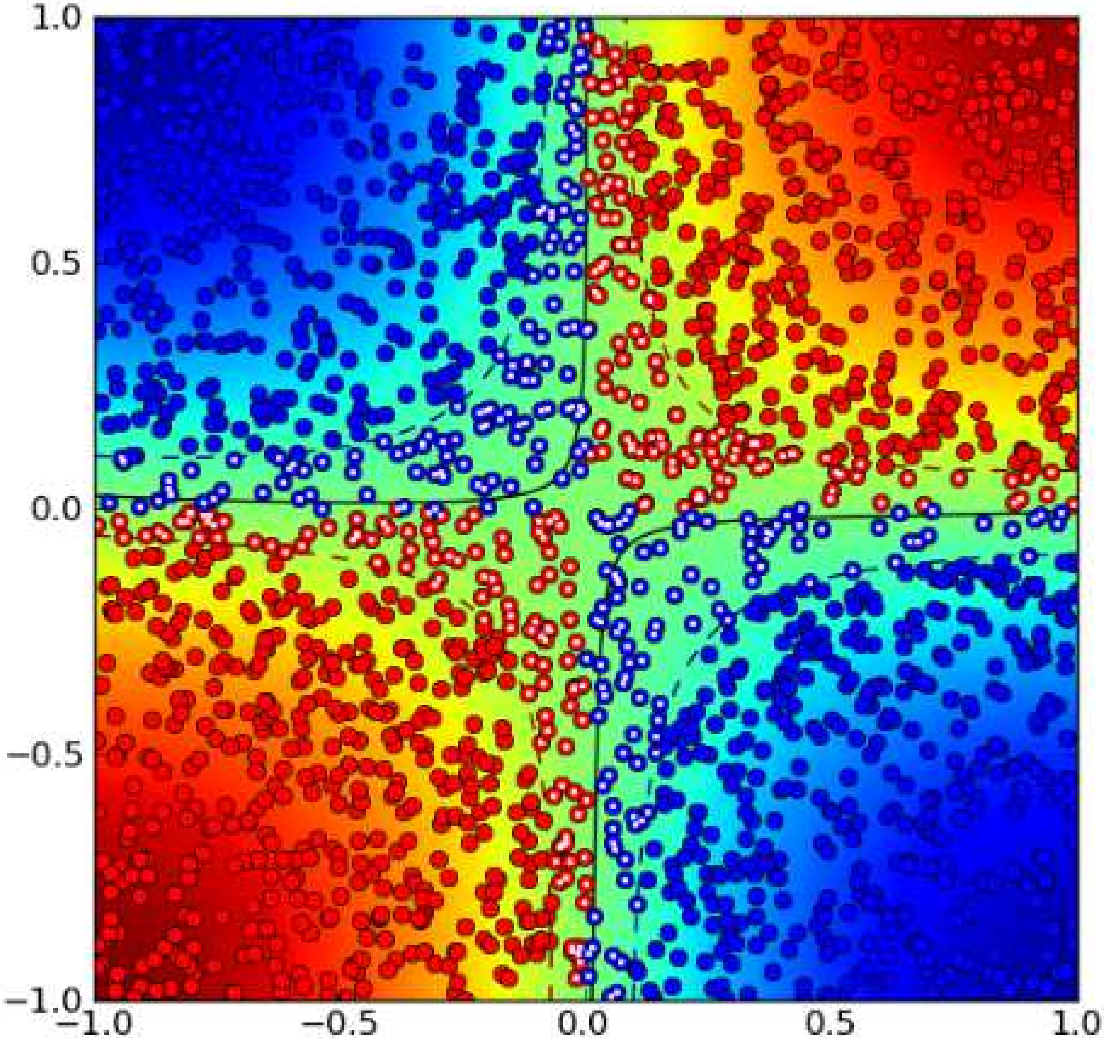
$$K(x,x') = \exp\left( -\frac{\| x - x' \|^2}{2\sigma^2} \right)$$

## Конструирование ядер

Пусть $K(x,x') = \langle x,x' \rangle$ это ядро. Тогда ядром является:

1. константа $K(x,x') = 1$;
2. произведение $K(x,x') = K_1(x,x')K_2(x,x')$;
3. $\forall \varphi : X \to R$ произведение $K(x,x') = \varphi(x)\varphi(x')$;
4. $\alpha_1, \alpha_2 > 0$: $K(x,x') = \alpha_1 K_1(x,x') + \alpha_2 K_2(x,x')$;
5. $\forall \varphi : X \to X$, $K_0$ - ядро: $K(x,x') = K_0(\varphi(x), \varphi(x'))$;
6. Пусть $s : X \times X \to R$ - симметричная интегрируемая функция, тогда $K(x,x') = \int_X s(x,z)s(x',z)dz$;

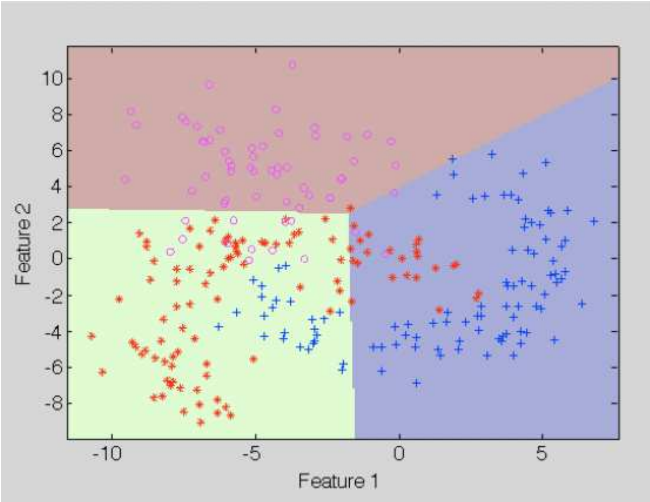
# Пример (RBF kernel)



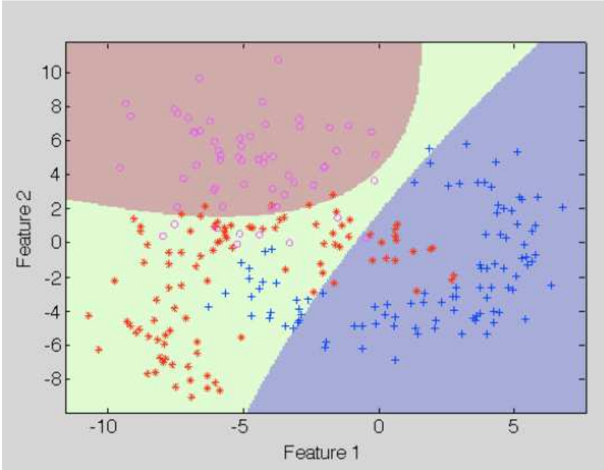
# Особенности метода SVM

- единственность экстремума;
- хорошая обобщающая способность;
- возможность классифицировать данные со сложной структурой;
- анализирует большие объемы данных; в решающем правиле используются только опорные объекты;
- в качестве данных можно использовать объекты любой природы (изображения, тексты и т.д.); возможно безпризнаковое распознавание;
- неустойчивость к шуму, непонятно, как выбирать штраф $C$;
- неясно, как выбрать наилучшее ядро и его параметры в конкретной задаче;
- работает только с количественными переменными;
- предназначен для разделения только двух классов;

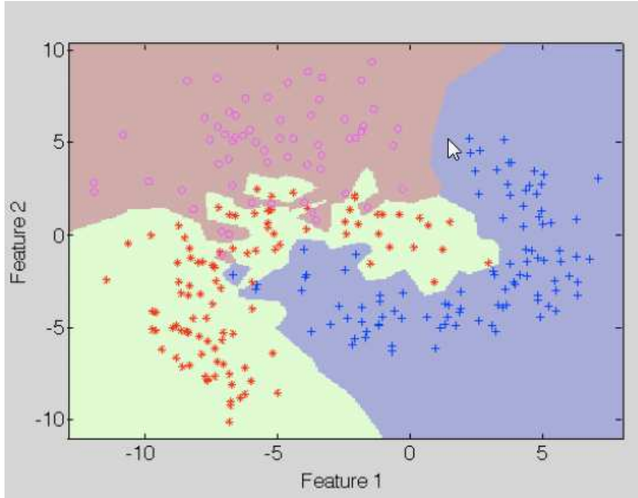
# Примеры разных классификаторов



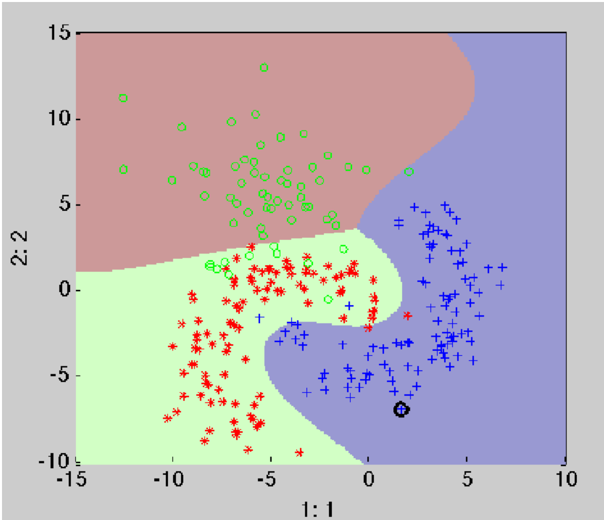

linear discriminant functions




quadratic discriminant functions




nearest neighbor method




multinomial kernel SVM