Метод опорных векторов (Support Vector Machine, SVM)

Александр Безносиков

ФПМИ МФТИ

8 ноября 2025

Постановка задачи

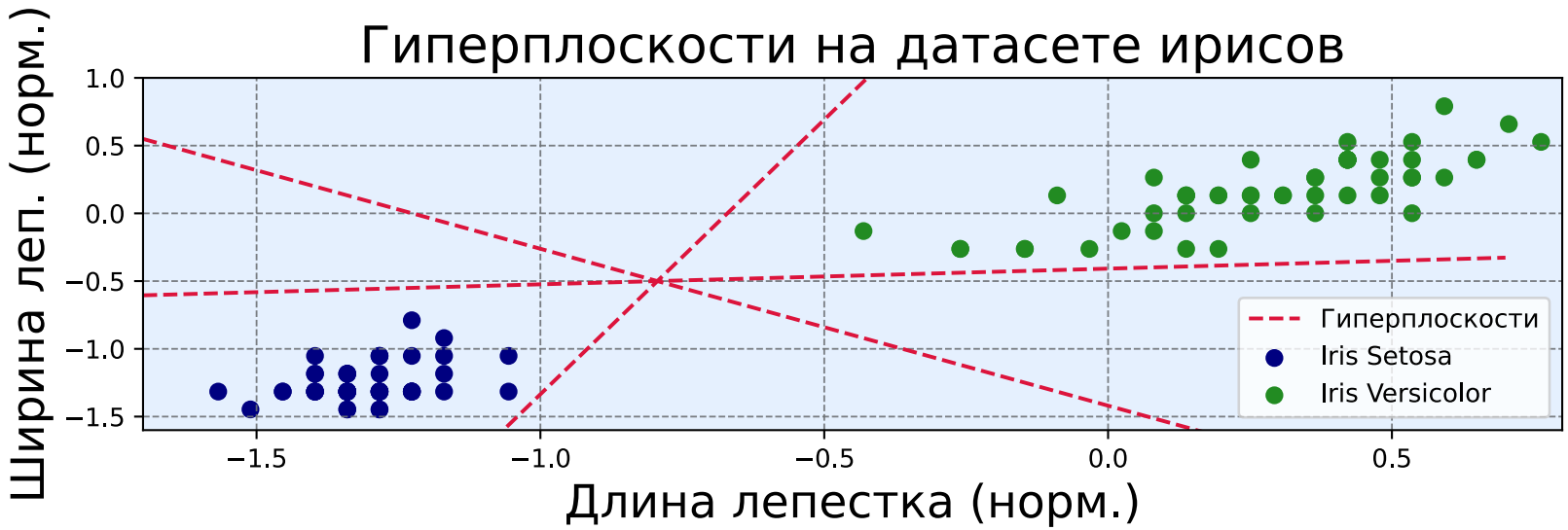
Дано: Выборка $\{(x^i, y^i)\}_{1 \le i \le n}$, где $x^i \in \mathbb{R}^d$, $y^i \in \{-1, 1\}$.
Идея: Хотим построить разделяющую гиперплоскость, задаваемую уравнением $\langle w, x \rangle + b = 0$.

Постановка задачи

Дано: Выборка $\{(x^i, y^i)\}_{1 \le i \le n}$, где $x^i \in \mathbb{R}^d$, $y^i \in \{-1, 1\}$.

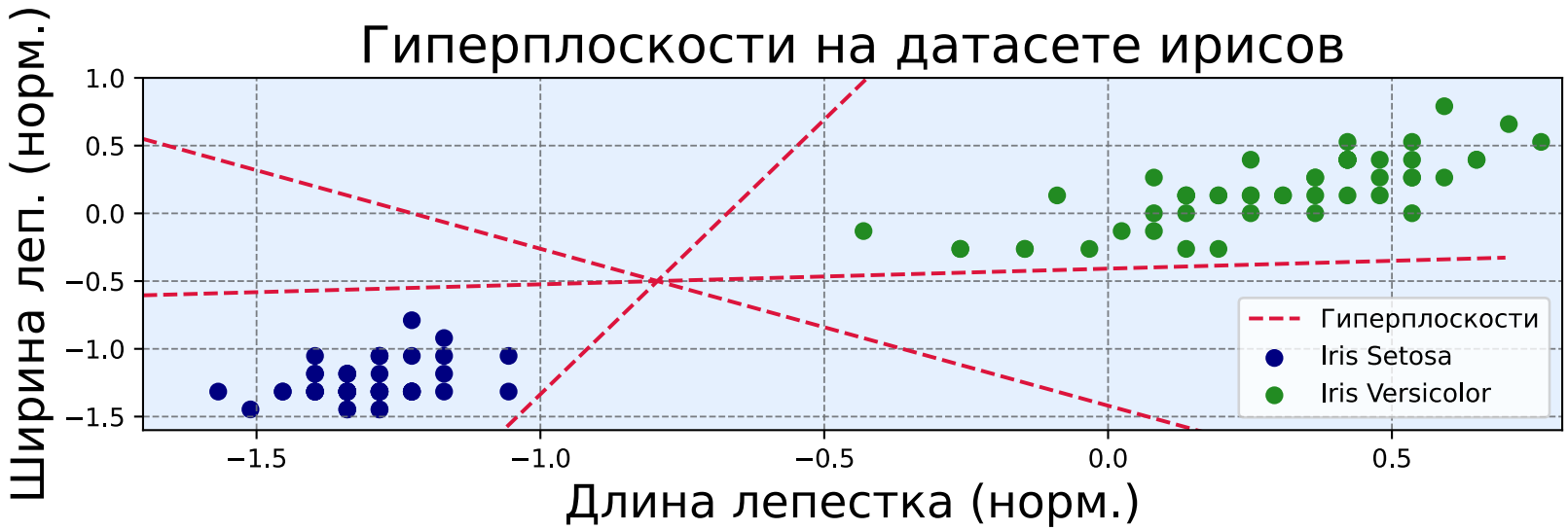
Идея: Хотим построить разделяющую гиперплоскость, задаваемую уравнением $\langle w, x \rangle + b = 0$.

Вопрос: Из каких соображений мы можем это сделать? Какая из плоскостей на картинке кажется наилучшей?



Постановка задачи

Дано: Выборка $\{(x^i, y^i)\}_{1 \le i \le n}$, где $x^i \in \mathbb{R}^d$, $y^i \in \{-1, 1\}$.
Идея: Хотим построить разделяющую гиперплоскость, задаваемую уравнением $\langle w, x \rangle + b = 0$.
Вопрос: Из каких соображений мы можем это сделать? Какая из плоскостей на картинке кажется наилучшей?



Хотим максимизировать расстояния до ближайшего положительного ($y = 1$) и ближайшего отрицательного ($y = -1$) примеров.

Постановка задачи

Формула для нахождения расстояния от точки X до гиперплоскости α:

$$\rho(X, \alpha) = \frac{|\langle w, X \rangle + b|}{\|w\|_2},$$

Постановка задачи

Формула для нахождения расстояния от точки X до гиперплоскости α:

$$\rho(X, \alpha) = \frac{|\langle w, X \rangle + b|}{\|w\|_2},$$

Задача оптимизации

$$\begin{cases} \min\limits_{1 \le i \le n} |\langle w, x^i \rangle + b| \to \max\limits_{w,b} \\ y^i \cdot (\langle w, x^i \rangle + b) > 0, \quad 1 \le i \le n \\ \|w\|_2 = 1 \end{cases}$$

Постановка задачи

Формула для нахождения расстояния от точки X до гиперплоскости α:

$$\rho(X, \alpha) = \frac{|\langle w, X \rangle + b|}{\|w\|_2},$$

Задача оптимизации

$$\begin{cases} \min\limits_{1 \le i \le n} |\langle w, x^i \rangle + b| \to \max\limits_{w,b} & \text{- максимизация расстояний до ближайших объектов} \\ y^i \cdot (\langle w, x^i \rangle + b) > 0, \quad 1 \le i \le n & \text{- объект попал в нужное полупр-во} \\ \|w\|_2 = 1 & \text{- нормировка, т.к. в } \rho \text{ есть } \|w\|_2^{-1} \end{cases}$$

Эквивалентная формулировка

Эквивалентная формулировка задачи:

$$\begin{cases} \|\hat{w}\|^2 \to \min\limits_{\hat{w},\hat{b}} \\ y^i \cdot (\langle \hat{w}, x^i \rangle + \hat{b}) \geq 1, \quad 1 \leq i \leq n \end{cases}$$

Эквивалентная формулировка

Эквивалентная формулировка задачи:

$$\begin{cases} \|\hat{w}\|^2 \to \min\limits_{\hat{w},\hat{b}} \\ y^i \cdot (\langle \hat{w}, x^i \rangle + \hat{b}) \geq 1, \quad 1 \leq i \leq n \end{cases}$$

В самом деле, обозначим $C = \min\limits_{1 \leq i \leq n} |\langle w, x^i \rangle + b|$ и сделаем замену $\hat{w} = \frac{w}{C}, \hat{b} = \frac{b}{C}$. Тогда для любого $i, 1 \leq i \leq n$, будет выполнено

$$y^i \cdot (\langle \hat{w}, x^i \rangle + \hat{b}) = \frac{y^i}{C} \cdot (\langle w, x^i \rangle + b) > 1$$

В то же время задача максимизации C эквивалентна минимизации $\|\hat{w}\| = \frac{1}{C}$.

Эквивалентная формулировка

Эквивалентная формулировка задачи:

$$\begin{cases} \|\hat{w}\|^2 \to \min\limits_{\hat{w},\hat{b}} \\ y^i \cdot (\langle \hat{w}, x^i \rangle + \hat{b}) \geq 1, \quad 1 \leq i \leq n \end{cases}$$

В самом деле, обозначим $C = \min\limits_{1 \leq i \leq n} |\langle w, x^i \rangle + b|$ и сделаем замену $\hat{w} = \frac{w}{C}, \hat{b} = \frac{b}{C}$. Тогда для любого $i, 1 \leq i \leq n$, будет выполнено

$$y^i \cdot (\langle \hat{w}, x^i \rangle + \hat{b}) = \frac{y^i}{C} \cdot (\langle w, x^i \rangle + b) > 1$$

В то же время задача максимизации C эквивалентна минимизации $\|\hat{w}\| = \frac{1}{C}$.

Вопрос: Как делать предсказания с помощью полученной модели?

Эквивалентная формулировка

Эквивалентная формулировка задачи:

$$\begin{cases} \|\hat{w}\|^2 \to \min\limits_{\hat{w},\hat{b}} \\ y^i \cdot (\langle \hat{w}, x^i \rangle + \hat{b}) \geq 1, \quad 1 \leq i \leq n \end{cases}$$

В самом деле, обозначим $C = \min\limits_{1 \leq i \leq n} |\langle w, x^i \rangle + b|$ и сделаем замену $\hat{w} = \frac{w}{C}, \hat{b} = \frac{b}{C}$. Тогда для любого $i, 1 \leq i \leq n$, будет выполнено

$$y^i \cdot (\langle \hat{w}, x^i \rangle + \hat{b}) = \frac{y^i}{C} \cdot (\langle w, x^i \rangle + b) > 1$$

В то же время задача максимизации C эквивалентна минимизации $\|\hat{w}\| = \frac{1}{C}$.

Вопрос: Как делать предсказания с помощью полученной модели?

Решающее правило: $\text{sign}(\langle \hat{w}, x \rangle + \hat{b})$

Условия

Для простоты $\hat{w}, \hat{b} \to w, b$

Задача в новых обозначениях:

$$\begin{cases} \|w\|^2 \to \min\limits_{w,b} \\ y^i \cdot (\langle w, x^i \rangle + b) > 1, \quad 1 \leq i \leq n \end{cases}$$

Вопрос: Как решать эту задачу с ограничениями?

Функция Лагранжа

Задача. Пусть стоит задача с ограничениями типа равенств и неравенств вида $\min\limits_{h_i(x)\leqslant 0, i=1,\cdots,p; x\in Q} f(x)$, где Q - выпуклое.

Функция Лагранжа в этой задаче: $L(x,\mu) = \mu_0 f(x) + \sum_{i=1}^{p} \mu_i h_i(x)$. Тогда точка, в которой достигается минимум функции Лагранжа, является решением задачи выше:

$$x^* = \operatorname*{argmin}_{x\in Q} L(x,\mu) \Rightarrow \min_{h_i(x)\leqslant 0, i=1,\cdots,p; x\in Q} f(x) = f(x^*).$$

Функция Лагранжа для нашей задачи с предыдущего слайда:

$$L(w,b,\mu) = \|w\|^2 + \sum_{i=1}^{n} \mu_i \left(1 - y^i \cdot (\langle w, x^i\rangle + b)\right)$$

Таким образом, можно оптимизировать

$$L : \min_{w,b} \max_{\mu_i \geqslant 0} L$$

Теорема Каруша-Куна-Таккера

Теорема Каруша-Куна-Таккера

Если x^* — решение задачи выше, то $\exists \mu = (v_0, v_1, \ldots, v_p) \neq 0$ - вектор множителей Лагранжа, такой, что для $L(x, \mu)$ выполняются:

1. принцип минимума для функции Лагранжа $\min\limits_{x \in Q} L(x, \mu) = L(x^*, \mu)$;
2. условия дополняющей нежёсткости $\mu_i h_i(x^*) = 0$, $i = 1, \ldots, p$;
3. условия неотрицательности (двойственные ограничения) $\mu_i \geq 0$, $i = 0, \ldots, p$

Причём, если $\mu_0 \neq 0$, то условия 1-3) достаточны для того, чтобы точка x^* была решением задачи.

Вопрос: Для чего нужно каждое из условий?

Теорема Каруша-Куна-Таккера

Теорема Каруша-Куна-Таккера

Если x^* — решение задачи выше, то $\exists \mu = (v_0, v_1, \ldots, v_p) \neq 0$ - вектор множителей Лагранжа, такой, что для $L(x, \mu)$ выполняются:

1. принцип минимума для функции Лагранжа $\min\limits_{x \in Q} L(x, \mu) = L(x^*, \mu)$;
2. условия дополняющей нежёсткости $\mu_i h_i(x^*) = 0$, $i = 1, \ldots, p$;
3. условия неотрицательности (двойственные ограничения) $\mu_i \geq 0$, $i = 0, \ldots, p$

Причём, если $\mu_0 \neq 0$, то условия 1-3) достаточны для того, чтобы точка x^* была решением задачи.

Вопрос: Для чего нужно каждое из условий?
Принцип минимума означает, что это действительно минимум (можно проверять по производным); дополняющая нежёсткость "включает"и "выключает"ограничения - если $\mu_i = 0$, то h_i не действует; если $h_i = 0$, то мы "упираемся в границу"множества и "улучшить"функцию не сможем.

Теорема Каруша-Куна-Таккера

Функция Лагранжа для нашей задачи:

$L(w, b, \mu) = \|w\|^2 + \sum_{i=1}^{n} \mu_i \left(1 - y^i \cdot (\langle w, x^i \rangle + b)\right)$

Задача поиска минимума функции Лагранжа L : $\min\limits_{w,b} \max\limits_{\mu_i \geqslant 0} L$

Условия Каруша-Куна-Таккера для нашей задачи

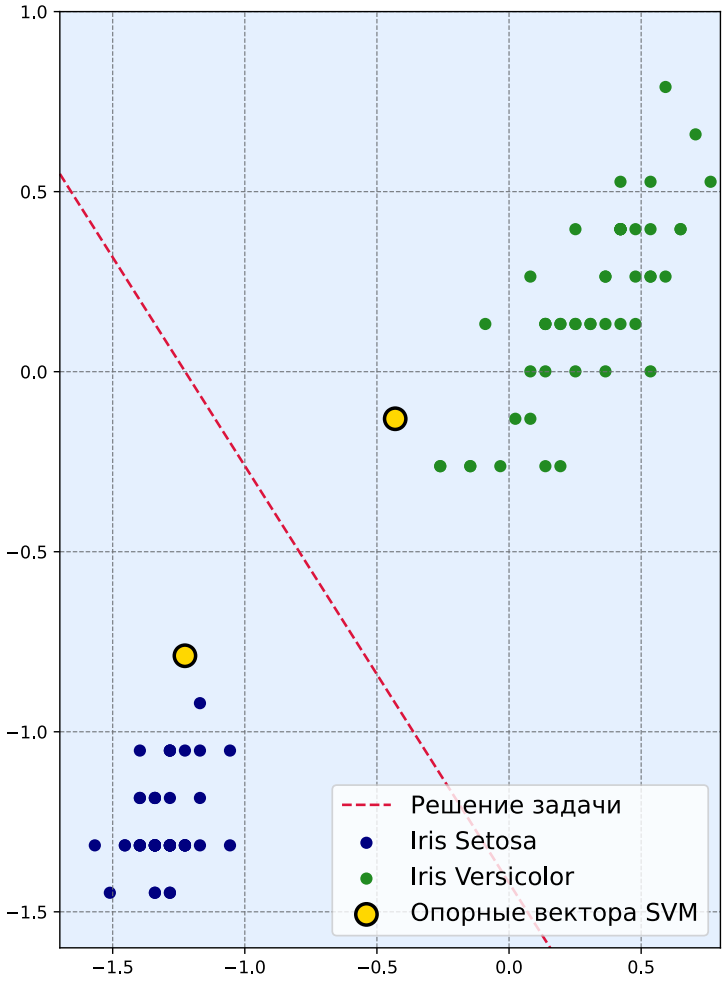
Если x — точка локального минимума, то существуют множители $\{\mu_i\}_{i=1,\ldots,m}$:

$$\begin{cases} \frac{\partial L}{\partial w} = 2w - \sum_{i=1}^{n} \mu_i y^i x^i = 0, \quad \frac{\partial L}{\partial b} = \sum_{i=1}^{n} \mu_i y^i = 0; \\ g_i(w) \leq 0; \\ \mu_i \geq 0; \quad \text{(двойственные ограничения)} \\ \mu_i g_i(w) = 0; \quad \text{(условие дополняющей нежёсткости)} \end{cases}$$

Выводы из условий KKT

Выводы:

- $w = \frac{1}{2}\sum_{i=1}^{n} \mu_i y^i x^i$, т.е. w - комбинация x и y при $\mu \neq 0$;
- $\mu_j \neq 0$ только в случае $1 = |\langle w; x^j \rangle + b|$. Это означает, что на этом x^j достигается минимум. Такие x^j и называются **опорными векторами**.



Случай неразделимой выборки

Вопрос. Что делать, если не существует гиперплоскости, разделяющей выборку на два класса? Это называется **неразделимой выборкой**.

Случай неразделимой выборки

Вопрос. Что делать, если не существует гиперплоскости, разделяющей выборку на два класса? Это называется **неразделимой выборкой**.
В случае неразделимой выборки мы немного модифицируем задачу и добавляем **мягкие отступы** ξ_i:

$$\begin{cases} \lambda\|w\|^2 + \frac{1}{n}\sum_{i=1}^{n}\xi_i \to \min\limits_{w,b,\xi_i} \\ y^i \cdot (\langle w, x^i\rangle + b) \geq 1 - \xi_i, & 1 \leq i \leq n, \\ \xi_i \geq 0, & 1 \leq i \leq n, \end{cases}$$

Так ξ_i разрешают чуть-чуть «нарушить» условия для точек из выборки, а λ^{-1} управляет штрафом за степень нарушения ξ_i.
Эквивалентная формулировка:

$$\min_{w,b} \|w\|_2^2 + C\sum_{i=1}^{n} \max\left\{0, (1 - y^i \cdot (\langle w, x_i\rangle + b))\right\},$$

где $C = \frac{1}{n\lambda}$.

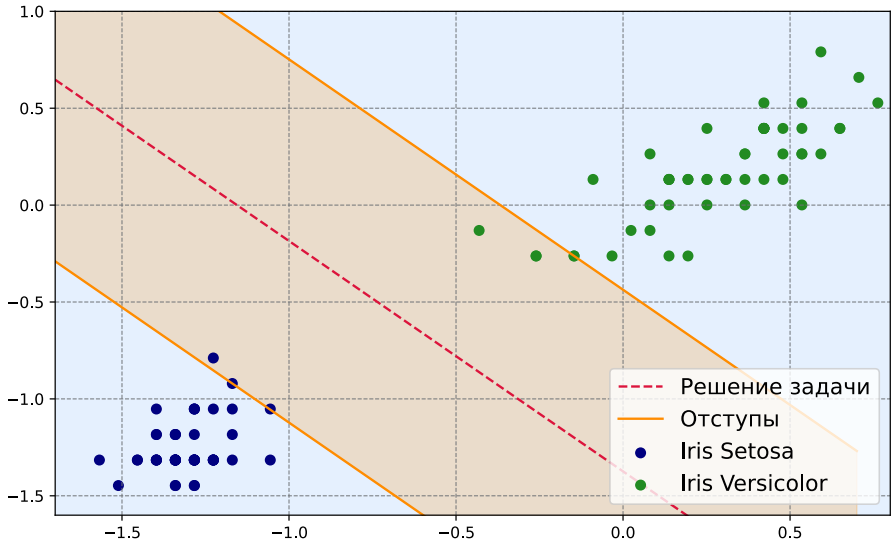
Случай неразделимой выборки

Система условий KKT

$$\begin{cases} w = \sum\limits_{i=1}^{\ell} \mu_i y^i x^i; & \sum\limits_{i=1}^{\ell} \mu_i y^i = 0; \\ \xi_i \geq 0, & \mu_i \geq 0, \quad \eta_i \geq 0, \\ \mu_i = 0 \quad \text{либо} \quad M_i(w, w_0) = 1 - \xi_i; & \\ \eta_i = 0 \quad \text{либо} \quad \xi_i = 0; & \end{cases}$$

Выводы из условий KKT:

1) $y^i \cdot (\langle w, x^i \rangle + b) > 1$
вектор периферийный
2) $y^i \cdot (\langle w, x^i \rangle + b) = 1$
опорный (опорный-граничный)
3) $y^i \cdot (\langle w, x^i \rangle + b) < 1$
нарушитель (опорный-нарушитель)



Аппроксимация и регуляризация эмпирического риска

Напоминание: Наша задача имеет следующий вид:

$$\min_{w,b} \|w\|_2^2 + C \sum_{i=1}^{n} \max\left\{0, (1 - y^i \cdot (\langle w, x^i \rangle + b))\right\}.$$

Замечание: Это можно рассматривать как минимизацию эмпирического риска с
$L(w,b) = \min_{w,b} \|w\|_2^2 + \frac{1}{n\lambda} \sum_{i=1}^{n} \max\left\{0, (1 - y^i \cdot (\langle w, x^i \rangle + b))\right\}$.

Вопрос: В чём проблема такого подхода?

Аппроксимация и регуляризация эмпирического риска

Напоминание: Наша задача имеет следующий вид:

$$\min_{w,b} \|w\|_2^2 + C \sum_{i=1}^{n} \max\left\{0, (1 - y^i \cdot (\langle w, x^i \rangle + b))\right\}.$$

Замечание: Это можно рассматривать как минимизацию эмпирического риска с
$L(w,b) = \min_{w,b} \|w\|_2^2 + \frac{1}{n\lambda} \sum_{i=1}^{n} \max\left\{0, (1 - y^i \cdot (\langle w, x^i \rangle + b))\right\}$.

Вопрос: В чём проблема такого подхода?

С такой «ступенькой» трудно работать: её производная почти везде равна нулю.

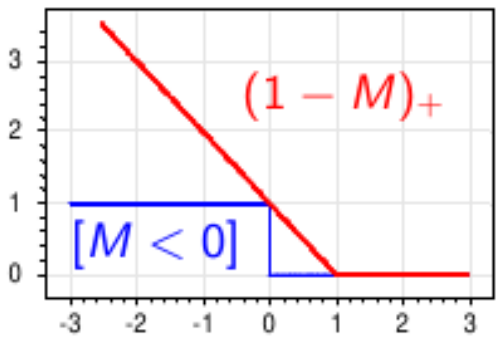
Аппроксимация и регуляризация эмпирического риска

Заменим эмпирический риск оценкой сверху, непрерывной по параметрам:

$$Q(w, w_0) = \sum_{i=1}^{\ell} [M_i(w, w_0) < 0] \leq$$

$$\leq \sum_{i=1}^{\ell} (1 - M_i(w, w_0))_+ + \frac{1}{2C}$$

- **Аппроксимация** штрафует объекты за приближение к границе классов, увеличивая зазор между классами.
- **Регуляризация** штрафует неустойчивые решения в случае мультиколлинеарности.



Задача обучения линейного классификатора

Дано: Обучающая выборка $\{(x^i, y^i)\}_{1 \le i \le n}$,
x^i — объекты, $x^i \in \mathbb{R}^n$; y^i — метки классов, $y^i \in \{-1, +1\}$.

Задача: Найти параметры $w \in \mathbb{R}^n$, $w_0 \in \mathbb{R}$ линейной модели классификации

$$a(x; w, w_0) = \text{sign}\left(\langle x, w\rangle - w_0\right).$$

Критерий — минимизация эмпирического риска:

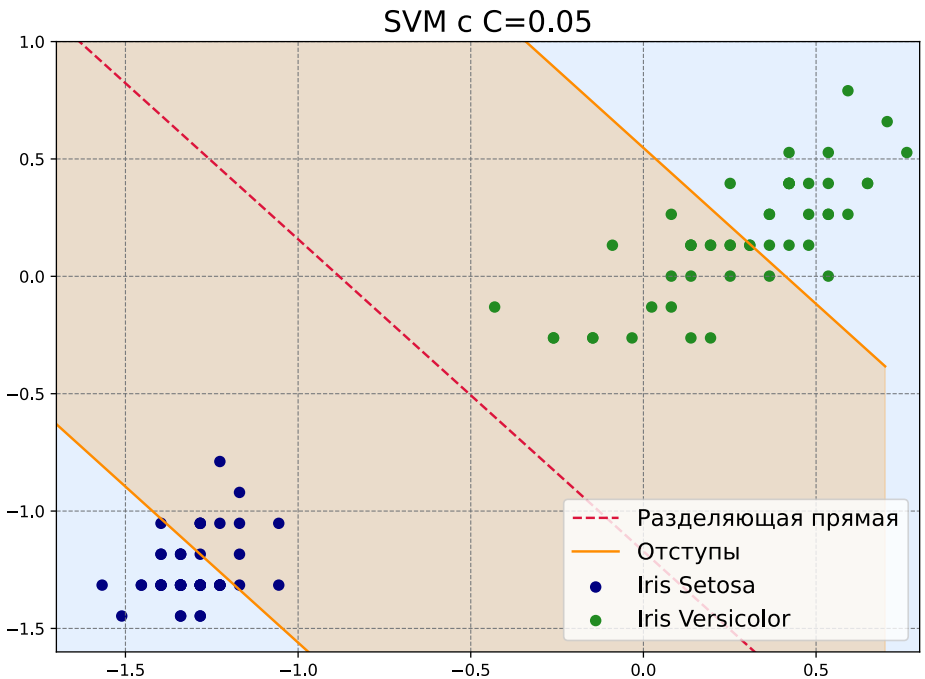
$$\sum_{i=1}^{n} \left[a(x^i; w, w_0) \neq y_i\right] = \sum_{i=1}^{n} \left[M_i(w, w_0) < 0\right] \to \min_{w, w_0},$$

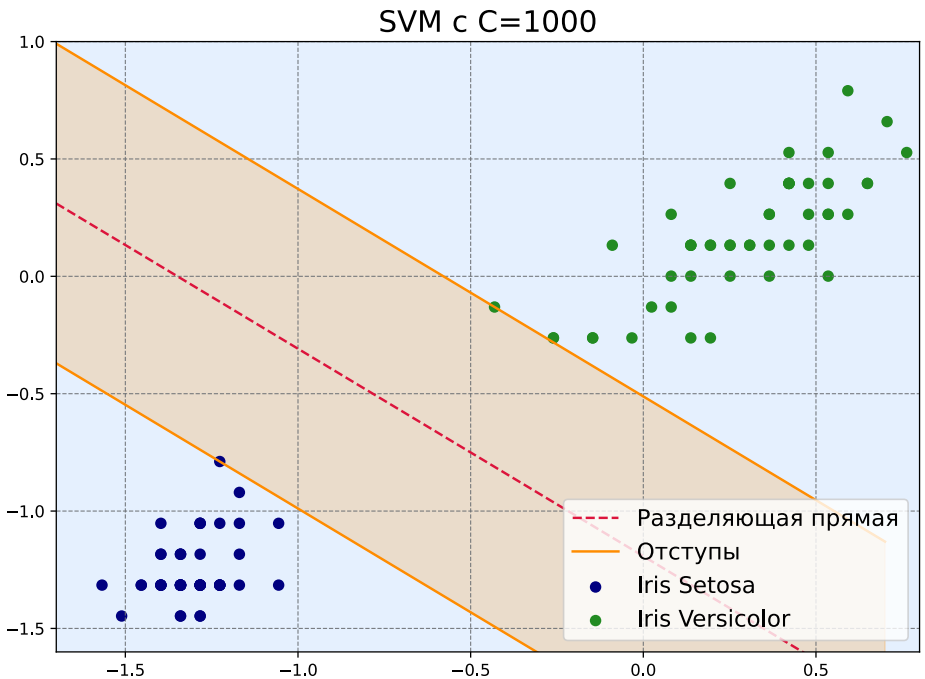
где

$$M_i(w, w_0) = y^i \cdot \left(\langle x^i, w\rangle - w_0\right) - \text{отступ (margin) объекта } x^i$$

Влияние константы C на решение SVM

SVM — аппроксимация и регуляризация эмпирического риска:
$C \to 0$**:** штраф за ошибки **меньше**, ширина разделяющей полосы увеличивается и больше точек попадают в неё.
$C \to \infty$**:** штраф за ошибки **больше**, ширина разделяющей полосы меньше, меньше точек попадают в неё. Модель стремится верно классифицировать **каждую** точку обучающей выборки: это может привести к переобучению.





Переход в пространство более высокой размерности

Вопрос: Как мы можем расширить пространство признаков, чтобы находить не только линейные разделяющие плоскости?

Переход в пространство более высокой размерности

Вопрос: Как мы можем расширить пространство признаков, чтобы находить не только линейные разделяющие плоскости?

Пример: $x^i \xrightarrow{\phi} x^i, (x^i)^2, (x^i)^3$

Тогда формулировка будет следующей:

$$\min_{w,b} \|w\|_2^2 + C \sum_{i=1}^{n} \max\left\{0, (1 - y^i \cdot (\langle w, \phi(x^i)\rangle + b))\right\}. \qquad (8.1)$$

Примечание: Такой подход с раздуванием признаков может работать не только для SVM. По идее, благодаря добавлению таких признаков, можно будет более точно отразить нелинейные зависимости.

Теорема о представителе

Теорема о представителе

Рассмотрим задачу оптимизации

$$f\left(\langle w, \phi(x^1)\rangle, \ldots, \langle w, \phi(x^n)\rangle\right) + R(\|w\|) \tag{8.2}$$

где $f : \mathbb{R}^n \to \mathbb{R}$ – произвольная функция, $R : \mathbb{R}_+ \to \mathbb{R}$ – неубывающая функция. Тогда существуют такие коэффициенты $\alpha_1, \ldots, \alpha_n$, что вектор $w^* = \sum_{i=1}^n \alpha_i \phi(x^i)$ является решением задачи оптимизации.

Замечание: Если взять

$$f\left(\langle w, \phi(x^1)\rangle, \ldots, \langle w, \phi(x^n)\rangle\right) = C\sum_{i=1}^{n} \max\left\{0, (1 - y^i \cdot (\langle w, \phi(x^i)\rangle + b))\right\},$$

а $R(\|w\|) = \|w\|^2$, то получим формулировку задачи вида (8.1).

Теорема о представителе (док-во)

Доказательство.

1. Рассмотрим произвольное решение w^* задачи 8.2, представимое как $w^* = \sum_{i=1}^{n} \alpha_i \phi(x^i) + u$, где $u \neq 0$ и $\langle u, \phi(x^i) \rangle = 0$ для всех $1 \leq i \leq n$.
2. Рассмотрим вектор $w = w^* - u = \sum_{i=1}^{n} \alpha_i \phi(x^i)$. Заметим, что

$$\|w^*\|^2 = \|w + u\|^2 = \|w\|^2 + \|u\|^2 + 2\langle w, u \rangle = \|w\|^2 + \|u\|^2.$$

3. Тогда $\|w^*\| > \|w\|$ (т.к. $\|u\| \neq 0$).
4. Воспользуемся неубыванием R: $R(\|w^*\|) \geq R(\|w\|)$.
5. Но при этом верно, что

$$f\left(\langle w, \phi(x^1)\rangle, \ldots, \langle w, \phi(x^n)\rangle\right) = f\left(\langle w^*, \phi(x^1)\rangle, \ldots, \langle w^*, \phi(x^n)\rangle\right),$$

значит w также является решением задачи (8.2).

Следствия из теоремы о представителе

- Решение задачи (8.1) можно искать в виде $w = \sum_{i=1}^{n} \alpha_i \phi(x^i)$.
- Предсказание для объекта x будет осуществляться по правилу $\operatorname{sign}\left(\sum_{i=1}^{n} \alpha_i \langle \phi(x^i), \phi(x) \rangle\right) + b)$.

Следствия из теоремы о представителе

- Решение задачи (8.1) можно искать в виде $w = \sum_{i=1}^{n} \alpha_i \phi(x^i)$.
- Предсказание для объекта x будет осуществляться по правилу $\operatorname{sign}\left(\sum_{i=1}^{n} \alpha_i \langle \phi(x^i), \phi(x) \rangle\right) + b)$.
- Заметим, что для вычисления скалярных произведений $\langle \phi(x), \phi(x') \rangle$ - достаточно уметь вычислять только скалярные произведения $\langle \phi(x), \phi(x') \rangle$ для любой пары точек $x, x' \in X$, но не сами $\phi(x)$.

Опишем такие «хорошие» ϕ. Будем называть **ядром** функцию вида

$$K(x, x') = \langle \phi(x), \phi(x') \rangle.$$

Вопрос: как понять, что $K(x, x')$ – ядро? Очевидно, что не все функции $K(x, x')$ являются ядрами.

Теорема Мерсера

Функция $K(x, x')$ называется **симметричным положительно определённым** на $\mathcal{X} \times \mathcal{X}$, если для любого натурального n и любых $x^1, \ldots, x^n \in \mathcal{X}$ матрица $K = (K(x^i, x^j))_{1 \leq i,j \leq n}$ является симметричной положительно полуопределённой.

Матрица K **положительно полуопределённая**, если $\forall c \in \mathbb{R}^n$ выполнено $c^\top K c \geq 0$.

Теорема Мерсера

Функция $K(x, x')$ называется **симметричным положительно определённым** на $\mathcal{X} \times \mathcal{X}$, если для любого натурального n и любых $x^1, \ldots, x^n \in \mathcal{X}$ матрица $K = (K(x^i, x^j))_{1 \le i,j \le n}$ является симметричной положительно полуопределённой.

Матрица K **положительно полуопределённая**, если $\forall c \in \mathbb{R}^n$ выполнено $c^\top K c \ge 0$.

Теорема Мерсера

Симметричная функция $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ задаёт скалярное произведение в некотором гильбертовом пространстве тогда и только тогда, когда она является неотрицательно определённой.

Вывод: Если функция удовлетворяет теореме, то его можно использовать в качестве ядра/скалярного произведения.

Свойства положительно определённых ядер

- **Воспроизводящее свойство ядра**: $\forall f(x) = \sum_{i=1}^{l} \alpha_i K(x, x_i)$ выполнено $f(x) = \sum_{i=1}^{l} \alpha_i K(x, x_i) = \langle f, \phi_x \rangle$.

Свойства положительно определённых ядер

- **Воспроизводящее свойство ядра**: $\forall f(x) = \sum_{i=1}^{l} \alpha_i K(x, x_i)$ выполнено $f(x) = \sum_{i=1}^{l} \alpha_i K(x, x_i) = \langle f, \phi_x \rangle$.
- Если K и K' — симметричные положительно определённые ядра, то $K + K'$ — симметричное положительно определённое ядро;
- Если K и K' — симметричные положительно определённые ядра, то $K \cdot K'$ — симметричное положительно определённое ядро;

Свойства положительно определённых ядер

- **Воспроизводящее свойство ядра**: $\forall f(x) = \sum_{i=1}^{l} \alpha_i K(x, x_i)$ выполнено $f(x) = \sum_{i=1}^{l} \alpha_i K(x, x_i) = \langle f, \phi_x \rangle$.
- Если K и K' — симметричные положительно определённые ядра, то $K + K'$ — симметричное положительно определённое ядро;
- Если K и K' — симметричные положительно определённые ядра, то $K \cdot K'$ — симметричное положительно определённое ядро;
- Если $\{K_m\}_{m=1}^{\infty}$ — симметричные положительно определённые ядра и для всех x и x' существует поточечный предел $\lim_{m\to\infty} K_m(x, x') = K(x, x')$, то K — симметричное положительно определённое ядро;

Свойства положительно определённых ядер

- **Воспроизводящее свойство ядра**: $\forall f(x) = \sum\limits_{i=1}^{l} \alpha_i K(x, x_i)$ выполнено $f(x) = \sum_{i=1}^{l} \alpha_i K(x, x_i) = \langle f, \phi_x \rangle$.
- Если K и K' — симметричные положительно определённые ядра, то $K + K'$ — симметричное положительно определённое ядро;
- Если K и K' — симметричные положительно определённые ядра, то $K \cdot K'$ — симметричное положительно определённое ядро;
- Если $\{K_m\}_{m=1}^{\infty}$ — симметричные положительно определённые ядра и для всех x и x' существует поточечный предел $\lim_{m\to\infty} K_m(x, x') = K(x, x')$, то K — симметричное положительно определённое ядро;
- Если K — симметричное положительно определённое ядро, $\forall x, x' \quad |K(x, x')| < \rho$, и ряд $\sum_{m=0}^{\infty} a_m x^m, \quad a_m \geq 0$, имеет радиус сходимости ρ, то $\sum_{m=0}^{\infty} a_m K^m$ — симметричное положительно определённое ядро.

Конструктивные методы синтеза ядер

1. $K(x, x') = \langle x, x' \rangle$ — ядро;
2. Константа $K(x, x') = 1$ — ядро;
3. Произведение ядер $K(x, x') = K_1(x, x')K_2(x, x')$ — ядро;
4. $\forall \psi : X \to \mathbb{R}$, произведение $K(x, x') = \psi(x)\psi(x')$ — ядро;
5. $K(x, x') = \alpha_1 K_1(x, x') + \alpha_2 K_2(x, x')$ при $\alpha_1, \alpha_2 > 0$ — ядро;
6. Для любого отображения $\varphi : X \to X$, если K_0 — ядро, то $K(x, x') = K_0(\varphi(x), \varphi(x'))$ — ядро;
7. Если $s : X \times X \to \mathbb{R}$ — симметричная интегрируемая функция, то $K(x, x') = \int_X s(x, z)s(x', z)\, dz$ — ядро;
8. Если K_0 — ядро и функция $f : \mathbb{R} \to \mathbb{R}$ представима в виде сходящегося степенного ряда с неотрицательными коэффициентами, то $K(x, x') = f(K_0(x, x'))$ — ядро.

Пример: спрямляющее пространство для квадратичного ядра

Пусть $X = \mathbb{R}^2$, $K(u, v) = \langle u, v\rangle^2$, где $u = (u_1, u_2)$, $v = (v_1, v_2)$.
Задача: найти пространство H и отображение $\psi : X \to H$, такие что $K(x, x') = \langle \psi(x), \psi(x')\rangle_H$.
Разложим квадрат скалярного произведения:

$$\begin{aligned} K(u, v) = \langle u, v\rangle^2 &= \langle (u_1, u_2), (v_1, v_2)\rangle^2 = \\ &= (u_1 v_1 + u_2 v_2)^2 = u_1^2 v_1^2 + u_2^2 v_2^2 + 2 u_1 v_1 u_2 v_2 = \\ &= \langle (u_1^2, u_2^2, \sqrt{2} u_1 u_2), (v_1^2, v_2^2, \sqrt{2} v_1 v_2)\rangle. \end{aligned}$$

Таким образом,

$$H = \mathbb{R}^3,\ \psi : (u_1, u_2) \mapsto (u_1^2, u_2^2, \sqrt{2} u_1 u_2),$$

Линейной поверхности в пространстве H соответствует квадратичная поверхность в исходном пространстве X.

Примеры ядер

- Квадратичное ядро, $\dim H = \frac{1}{2}n(n+1)$: $K(x, x') = \langle x, x' \rangle^2$
- Полиномиальное ядро с мономами степени d, $\dim H = C_{n+d-1}^{d}$

$$K(x, x') = \langle x, x' \rangle^d$$

- Полиномиальное ядро с мономами степени не выше d

$$K(x, x') = (\langle x, x' \rangle + 1)^d$$

- «Универсальное» полиномиальное ядро

$$K(x, x') = \frac{1}{1 - \alpha^2 \langle x, x' \rangle}$$

- Нейросеть с сигмоидными функциями активации

$$K(x, x') = \tanh\big(k_1 \langle x, x' \rangle - k_0\big), \quad k_0, k_1 \geq 0$$

- Сеть радиальных базисных функций (RBF-ядро)

$$K(x, x') = \exp\big(-\gamma \|x - x'\|^2\big)$$

Обоснование «универсального» ядра

Поймем, какая ϕ порождает «универсальное» ядро

$$K(x,z) = \frac{1}{1-\alpha^2\langle x,z\rangle}.$$

$$\frac{1}{1-\alpha^2\langle x,z\rangle} = \sum_{i=0}^{+\infty}(\alpha^2\langle x,z\rangle)^i = 1+\alpha^2\langle x,z\rangle+\alpha^4(\langle x,z\rangle)^2+\ldots =$$

$$= 1+\alpha^2(x_1z_1+\ldots+x_nz_n)+\alpha^4\left(\sum c_{i,j}x_ix_jz_iz_j\right)+\ldots = \varphi(x)^T\varphi(z)$$

$$\varphi(x) = \begin{bmatrix}1, & \alpha x_1, & \alpha x_2, & \ldots, & \alpha x_n, & \ldots, & \alpha^2 x_i x_j, & \ldots\end{bmatrix}^T$$

Пространство бесконечномерное, а вычисляем за конечное время!

RBF-ядро

RBF-ядро также соответствует переходу в бесконечномерное пространство

Пусть для простоты $\gamma = 1$, $d = 2$, $x, z \in \mathbb{R}$

$$K(x,z) = \exp\left(-(x-z)^2\right) =$$

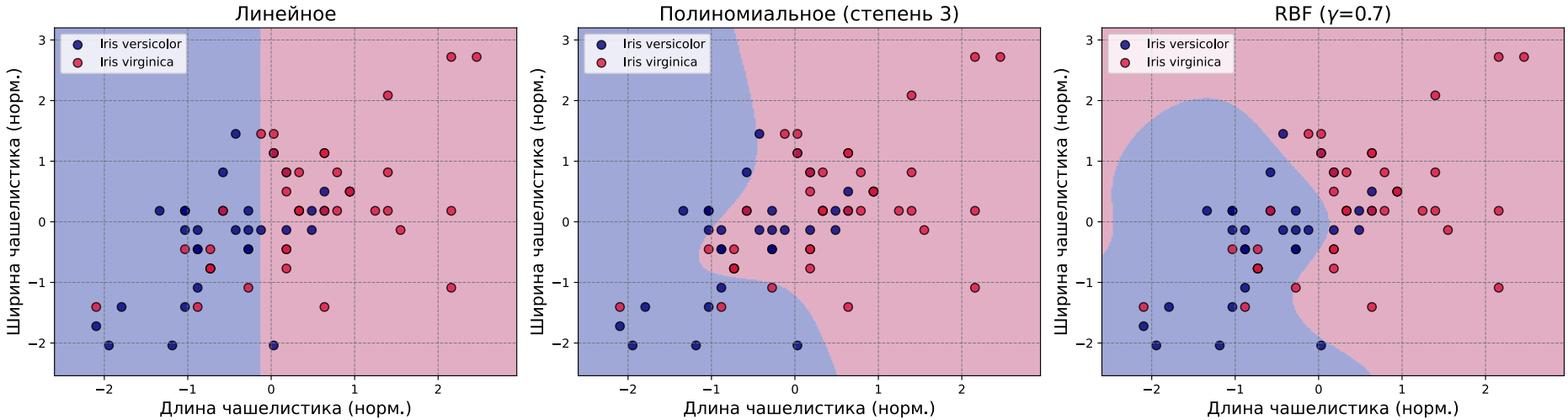
$$= \exp\left(-x^2 + 2xz - z^2\right) =$$

$$= \exp(-x^2)\left(\sum_{k=0}^{\infty} \frac{2^k x^k z^k}{k!}\right)\exp(-z^2) =$$

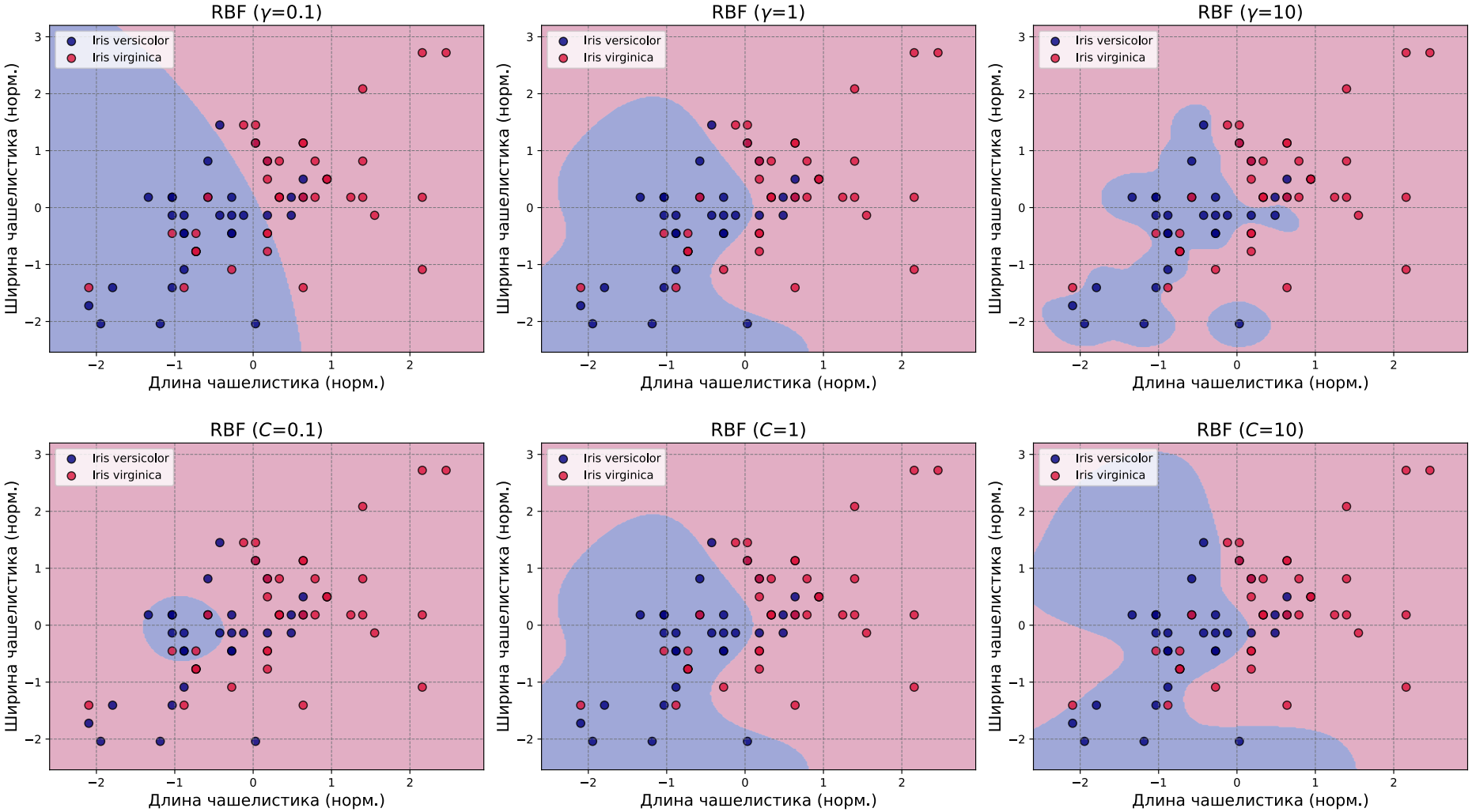
$$= \left(\sum_{k=0}^{\infty} \frac{2^{k/2}\exp(-x^2)x^k}{\sqrt{k!}} \cdot \frac{2^{k/2}\exp(-z^2)z^k}{\sqrt{k!}}\right)$$

SVM с различными ядрами

Гиперплоскость в спрямляющем пространстве соответствует нелинейной разделяющей поверхности в исходном.



SVM с RBF ядрами разной ширины и регуляризацией



Преимущества и недостатки SVM

Таким образом, мы умеем решать сложные задачи линейными методами (пополненим признакового пространства).

Преимущества и недостатки SVM

Таким образом, мы умеем решать сложные задачи линейными методами (пополненим признакового пространства).

Преимущества SVM перед двухслойными нейронными сетями:

- Задача выпуклого квадратичного программирования имеет единственное решение
- Число нейронов скрытого слоя определяется автоматически — это число опорных векторов

Недостатки классического SVM:

- Нет общих подходов к оптимизации $K(x, x')$ под задачу
- На больших объемах данных SVM обучается медленнее регрессии
- Нет «встроенного» отбора признаков
- Приходится подбирать константу C

Резюме по линейным классификаторам

- С помощью **ядер** (kernel trick) SVM изящно обобщается для нелинейной классификации и нелинейной регрессии
- **Аппроксимация пороговой функции потерь** $L(M)$ увеличивает зазор и повышает надёжность классификации
- **Регуляризация** увеличивает зазор, устраняет мультиколлинеарность и уменьшает переобучение
- **Неплавкость функции потерь** приводит к отбору объектов
- **Неплавкость регуляризатора** приводит к отбору признаков