Тверской
государственный
технический
университет

Интеллектуальные информационные системы

Метод опорных векторов

Материалы курса доступны по ссылке:
https://github.com/AndreyShpigar/ML-course

2024 г.

Метод опорных векторов в задачах классификации

Задача бинарной классификации: $X = \mathbb{R}^n$,

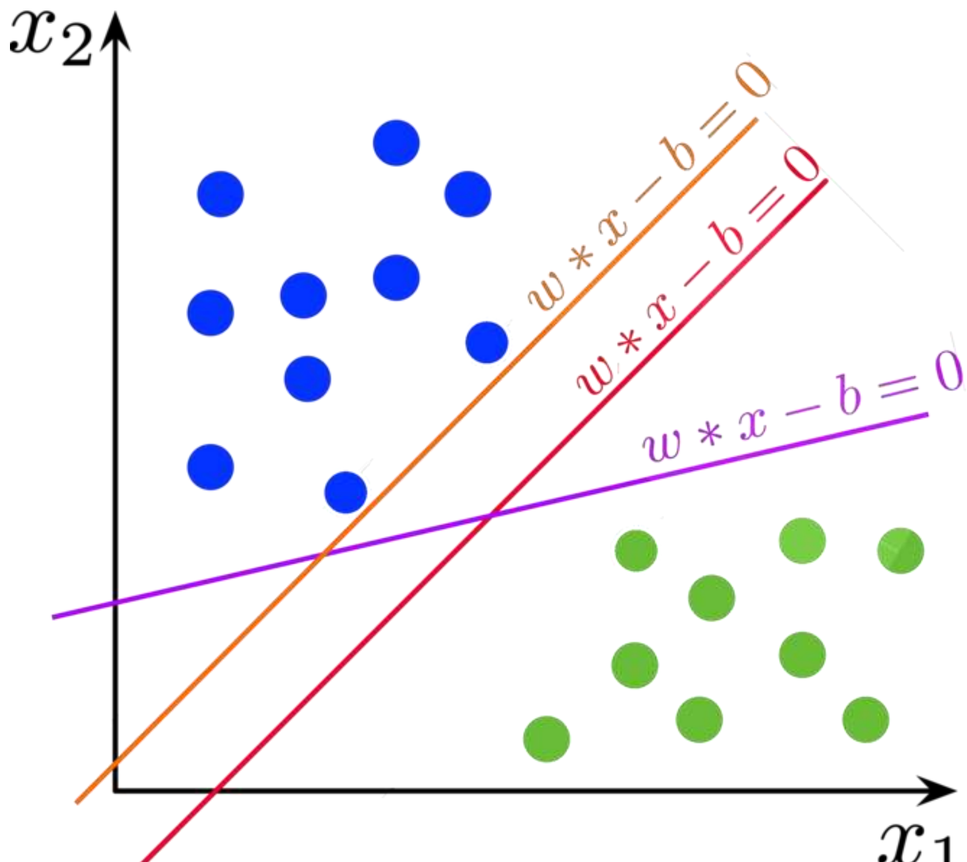
$$Y = \{-1, +1\}$$

Линейный классификатор:

$$a(x) = sign\left(\sum_{j=1}^{n} w_j x^j - w_0\right) = sign(\langle w, x \rangle - w_0)$$

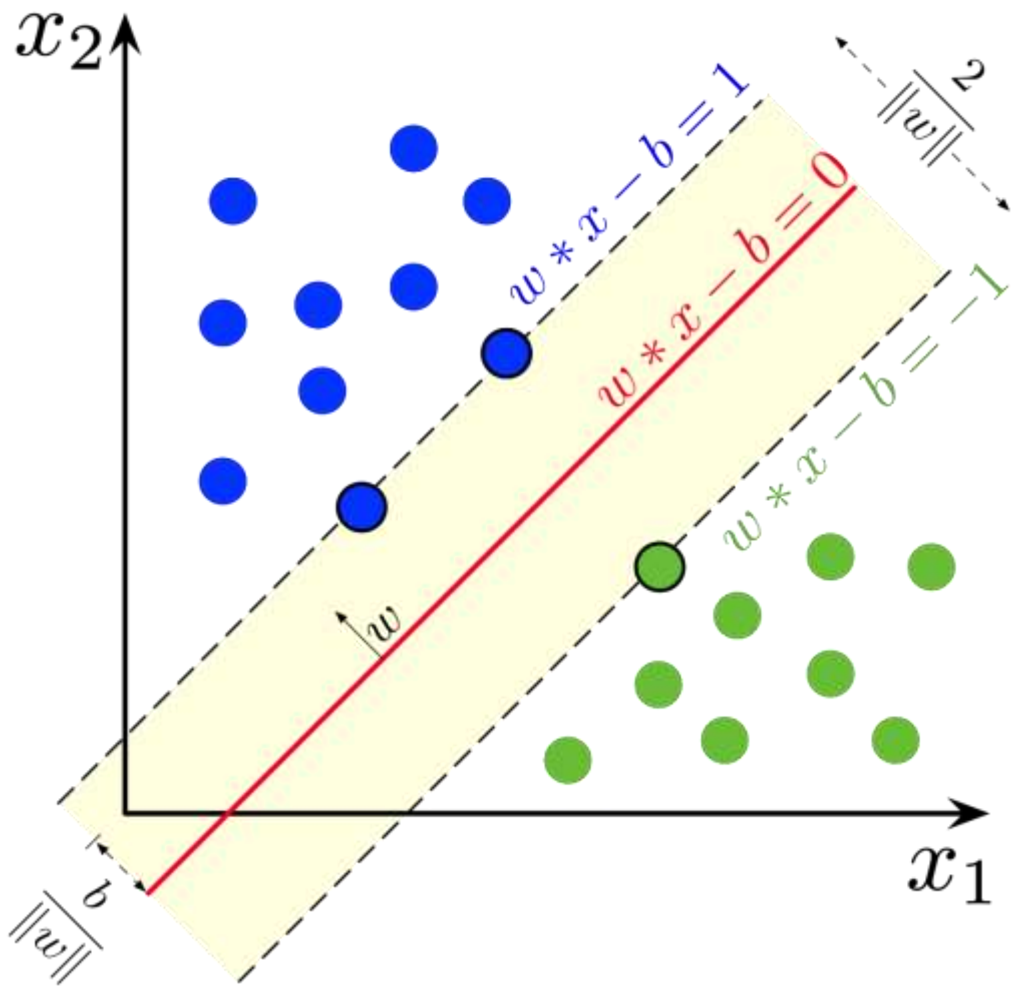
Предположим, что выборка линейно разделима:

$$Q(w, w_0) = \sum_{i=1}^{l} [y_i(\langle w, x_i \rangle - w_0) < 0] = 0$$



Ширина полосы:

$$\left\langle (\boldsymbol{x}_+ - \boldsymbol{x}_-), \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} \right\rangle = \frac{\langle \boldsymbol{w}, \boldsymbol{x}_+ \rangle - \langle \boldsymbol{w}, \boldsymbol{x}_- \rangle}{\|\boldsymbol{w}\|} = \frac{(\boldsymbol{w}_0 + 1) - (\boldsymbol{w}_0 - 1)}{\|\boldsymbol{w}\|} = \frac{2}{\|\boldsymbol{w}\|}$$



Для линейно разделимой выборки:

Классификация с жестким зазором (hard margin)

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

Задача квадратичного программирования - минимизировать квадратичный функционал при линейных ограничениях

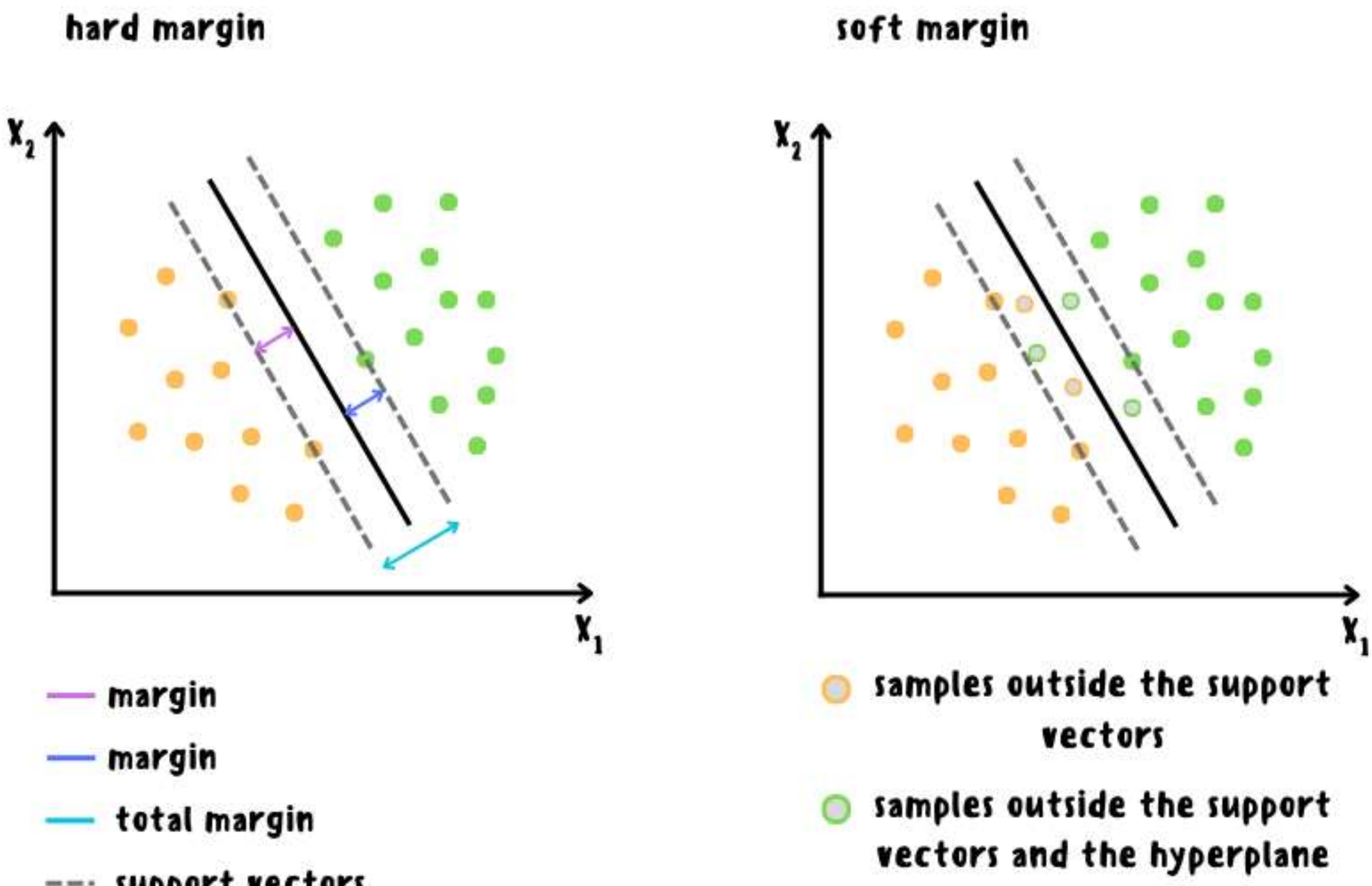
Обобщение для линейно неразделимой выборки:

Классификация с мягким зазором (soft margin)

$$\begin{cases} \frac{1}{2}\|w\|^2 + C\sum\limits_{i=1}^{l} \xi_i \to \min\limits_{w, w_0, \xi} \\ M_i(w, w_0) \geq 1 - \xi_i, \qquad i = 1, \dots, l \\ \xi_i \geq 0, \qquad i = 1, \dots, l \end{cases}$$

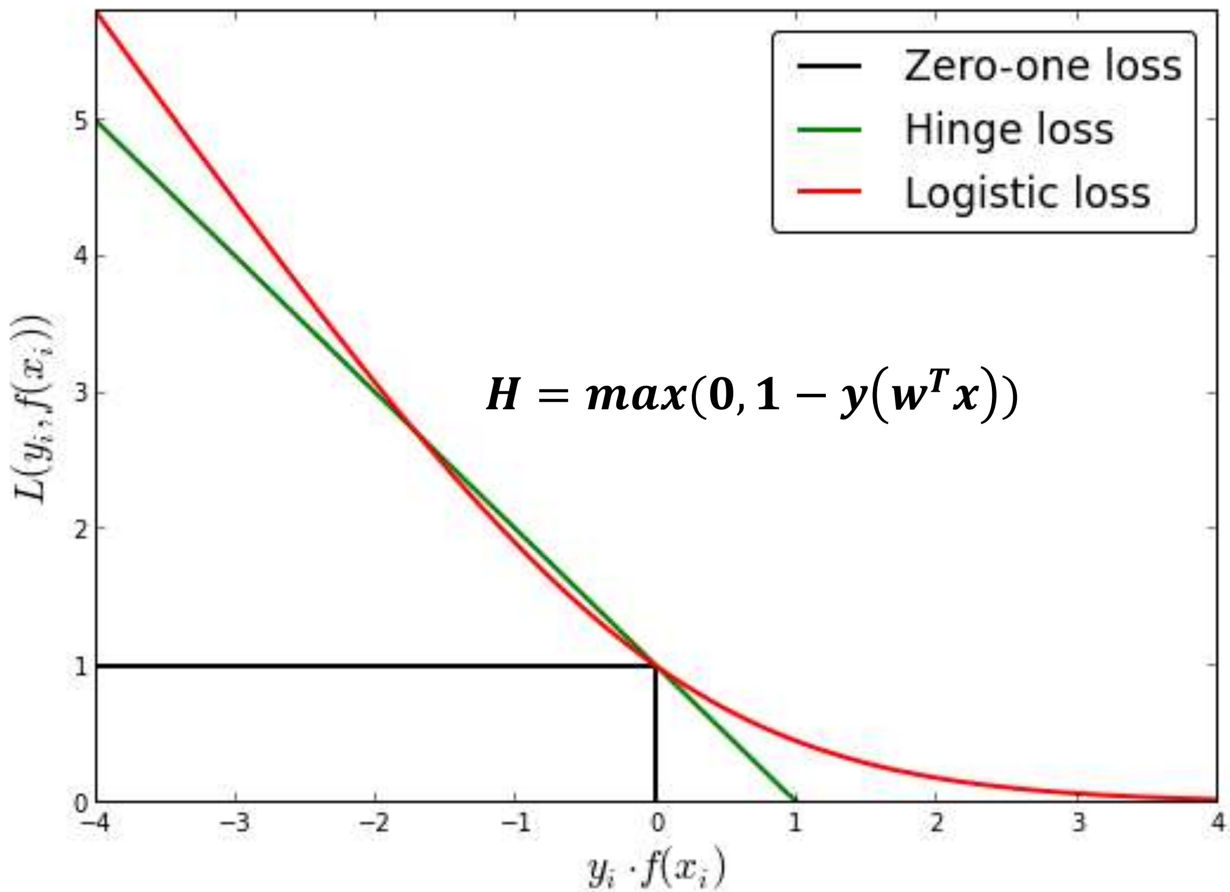
ξ_i – величина ошибки на объектах x_i

C – штраф за суммарную ошибку


hard margin
soft margin
X_2
X_1
X_2
X_1
margin
margin
total margin
support vectors
samples outside the support vectors
samples outside the support vectors and the hyperplane

Эквивалентная задача безусловной оптимизации:

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



Инструмент для решения задачи оптимизации с ограничениями равенства и неравенства –

условия Каруша-Куна-Таккера

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

По теореме ККТ эта задача эквивалентна двойственной задаче поиска седловой точки функции Лагранжа

Если x – точка локального минимума, то существуют множители $\mu_i, i = 1, \ldots, m$ и $\lambda_i, j = 1, \ldots, k$:

$$\begin{cases} \dfrac{\partial \mathcal{L}}{\partial x} = 0, \qquad \mathcal{L}(x; \mu, \lambda) = f(x) + \sum_{i=1}^{m} \mu_i g_i(x) + \sum_{j=1}^{k} \lambda_j h_j(x) \\ g_i(x) \leq 0;\ h_j(x) = 0 \\ \mu_i \geq 0 \\ \mu_i \vee g_i(x) = 0 \end{cases}$$

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

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

Продифференцируем функцию Лагранжа и приравняем нулю производные:

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

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

$$\frac{\partial \mathcal{L}}{\partial \xi_i} = 0 \Rightarrow \boxed{\eta_i + \lambda_i = C,} \qquad i = 1, \dots, l$$

Вектор весов $\boldsymbol{w}$ выражается через $\boldsymbol{\lambda_i}$:

$$\boldsymbol{w} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{l}} \boldsymbol{\lambda_i y_i x_i}$$

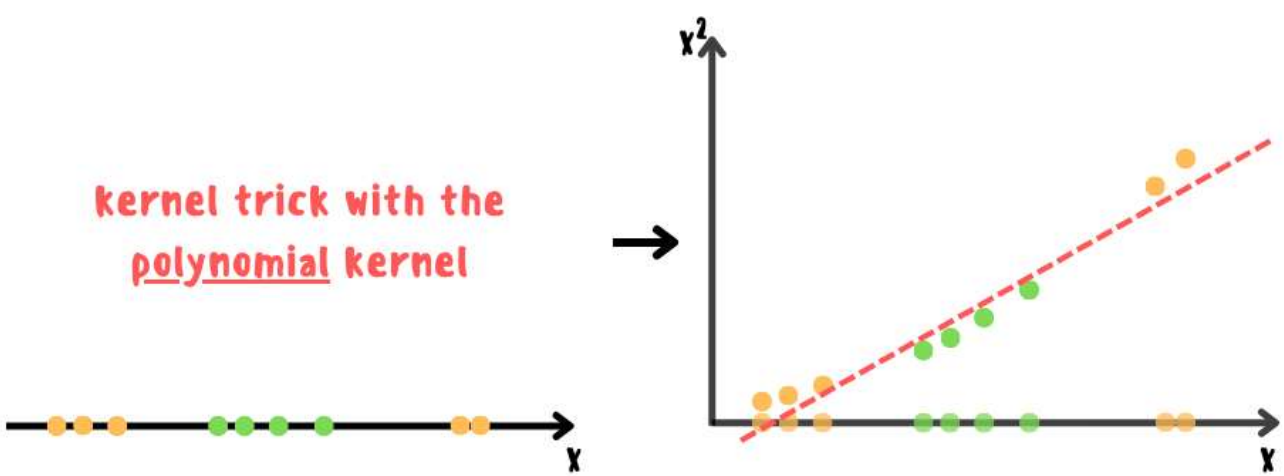
Если $\boldsymbol{\lambda_i = 0}$, то решение задачи никак не зависит от объекта обучающей выборки $\boldsymbol{x_i}$ (эти объекты можно назвать неинформативными)

Объект $\boldsymbol{x_i}$ называется *опорным*, если $\boldsymbol{\lambda_i \neq 0}$

Нелинейное обобщение SVM

<u>*Идея*</u>: переход от исходного пространства признаковых описаний объектов X к новому пространству H (<u>*спрямляющему*</u> пространству) с помощью некоторого преобразования $\boldsymbol{\psi}: \boldsymbol{X} \rightarrow \boldsymbol{H}$

<u>Kernel trick (ядерный трюк)</u> – замена скалярного произведения векторов n-й степени заменяется на их произведение в степени n: $\boldsymbol{\psi}(\boldsymbol{a})^{\boldsymbol{T}}\boldsymbol{\psi}(\boldsymbol{b}) = \left(\boldsymbol{a}^{\boldsymbol{T}}\boldsymbol{b}\right)^{\boldsymbol{n}}$



- Линейное ядро:

$$K(x, x^T) = \langle x, x^T \rangle$$

- Квадратичное ядро:

$$K(x, x^T) = \langle x, x^T \rangle^2$$

- Полиномиальное ядро:

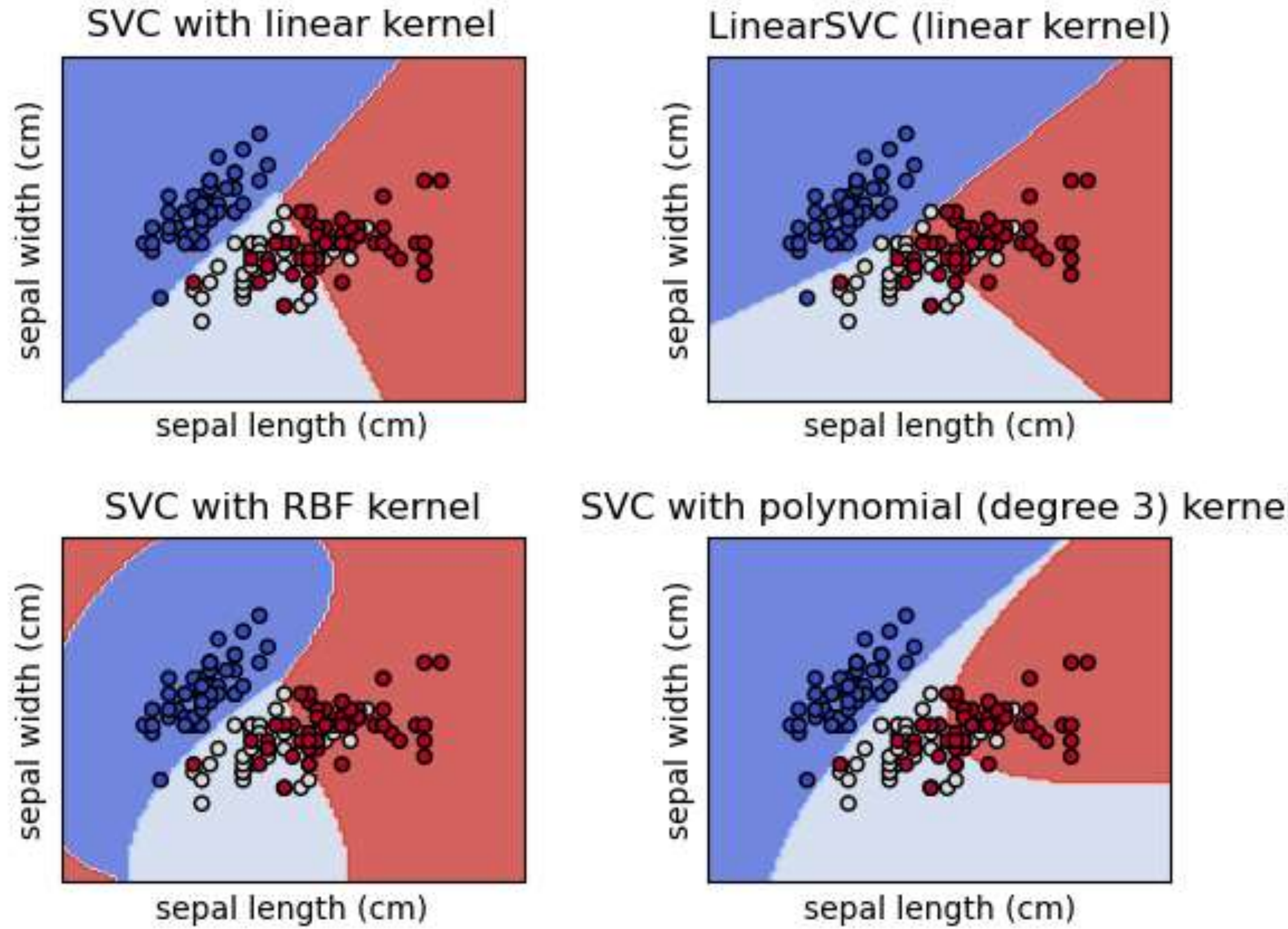
$$K(x, x^T) = (\langle x, x^T \rangle + 1)^d$$

- Гауссовское RBF ядро (радиально-базисная функция):

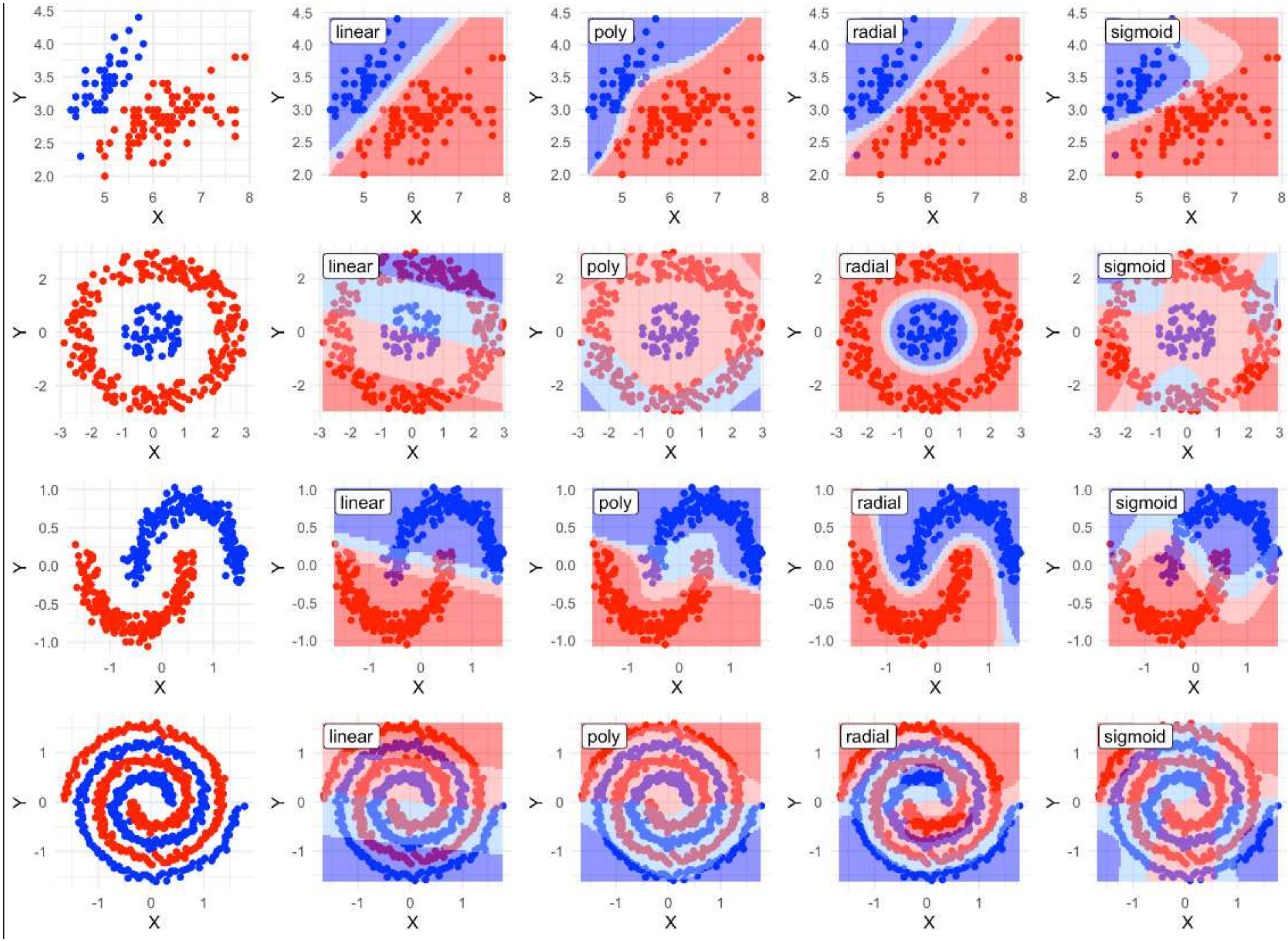
$$K(x, x^T) = \exp(-\gamma \|x - x^T\|^2)$$

- Сигмоидальное ядро:

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



Источник: https://scikit-learn.org/stable/auto_examples/svm/plot_iris_svc.html



Источник: https://rpubs.com/jmsallan/SVMkernels

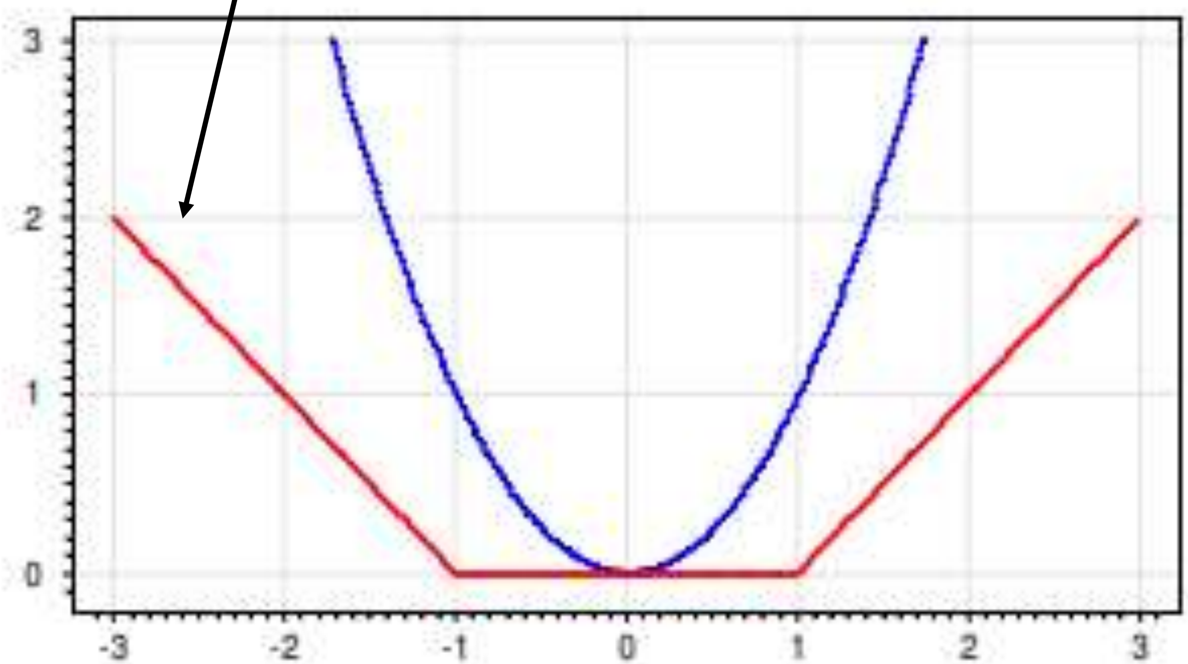
Метод опорных векторов в задачах регрессии

Модель регрессии:

$$a(x) = \langle x, w \rangle - w_0, \qquad w \epsilon \mathbb{R}^n, \;\; w_0 \epsilon \mathbb{R}$$

Функция потерь: $L(\varepsilon) = (|\varepsilon| - \delta)_+$

Кусочно-линейная функция $\boldsymbol{\varepsilon}$*-чувствительности*



Смысл функции потерь – *если мы находимся на расстоянии не более чем* $\boldsymbol{\delta}$ *от правильного ответа, то потери нет, за такой ответ не штрафуем. Если дальше, то штраф увеличивается линейно*.

Постановка задачи (с L2-регуляризацией):

$$\sum_{i=1}^{l}(|\langle w, x_i\rangle - w_0 - y_i| - \delta)_+ + \frac{1}{2C}\|w\|^2 \to \min_{w, w_0}$$

Достоинства SVM:

- ✓ Задача выпуклого квадратичного программирования имеет единственное решение – методы оптимизации существенно более эффективны
- ✓ Принцип оптимальной разделяющей гиперплоскости приводит к максимизации разделяющей полосы между классами – более уверенная классификация
- ✓ Возможность обработки многомерных данных без предварительного преобразования

Недостатки SVM:

- × Неустойчивость к шуму, выбросы в обучающей выборке могут стать опорными объектами-нарушителями и непосредственно влиять на построение разделяющей гиперплоскости
- × Не существует общих методов построения ядер и спрямляющих пространств
- × Необходимо подбирать параметр C в случае линейной неразделимости