

Основы машинного обучения

Поляк Марк Дмитриевич

2025

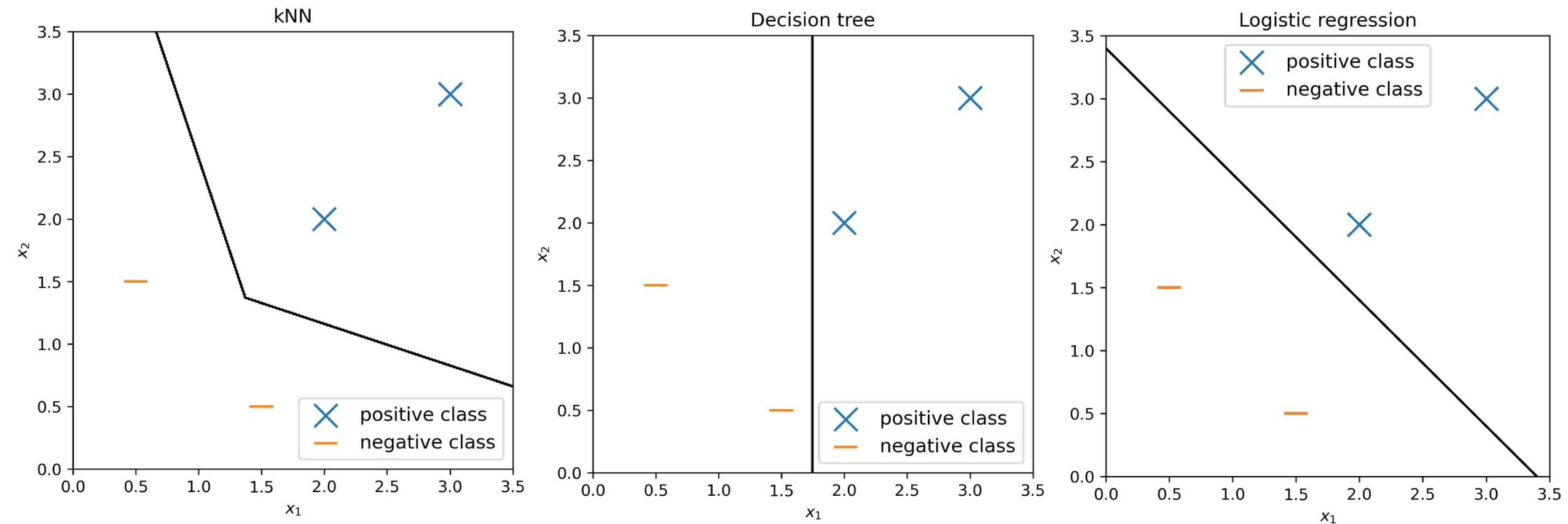
Алгоритмы классификации

Лекция 6

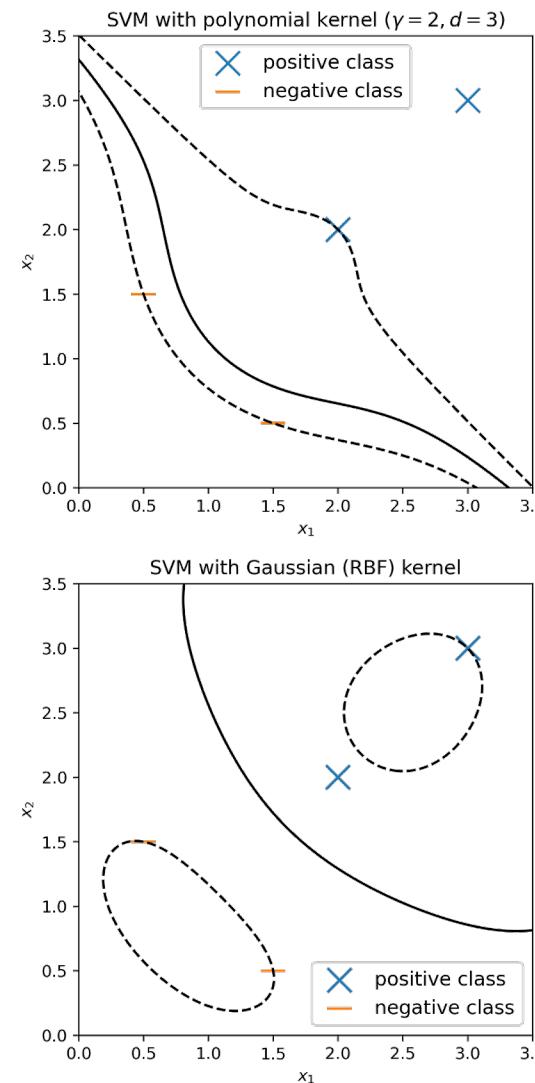
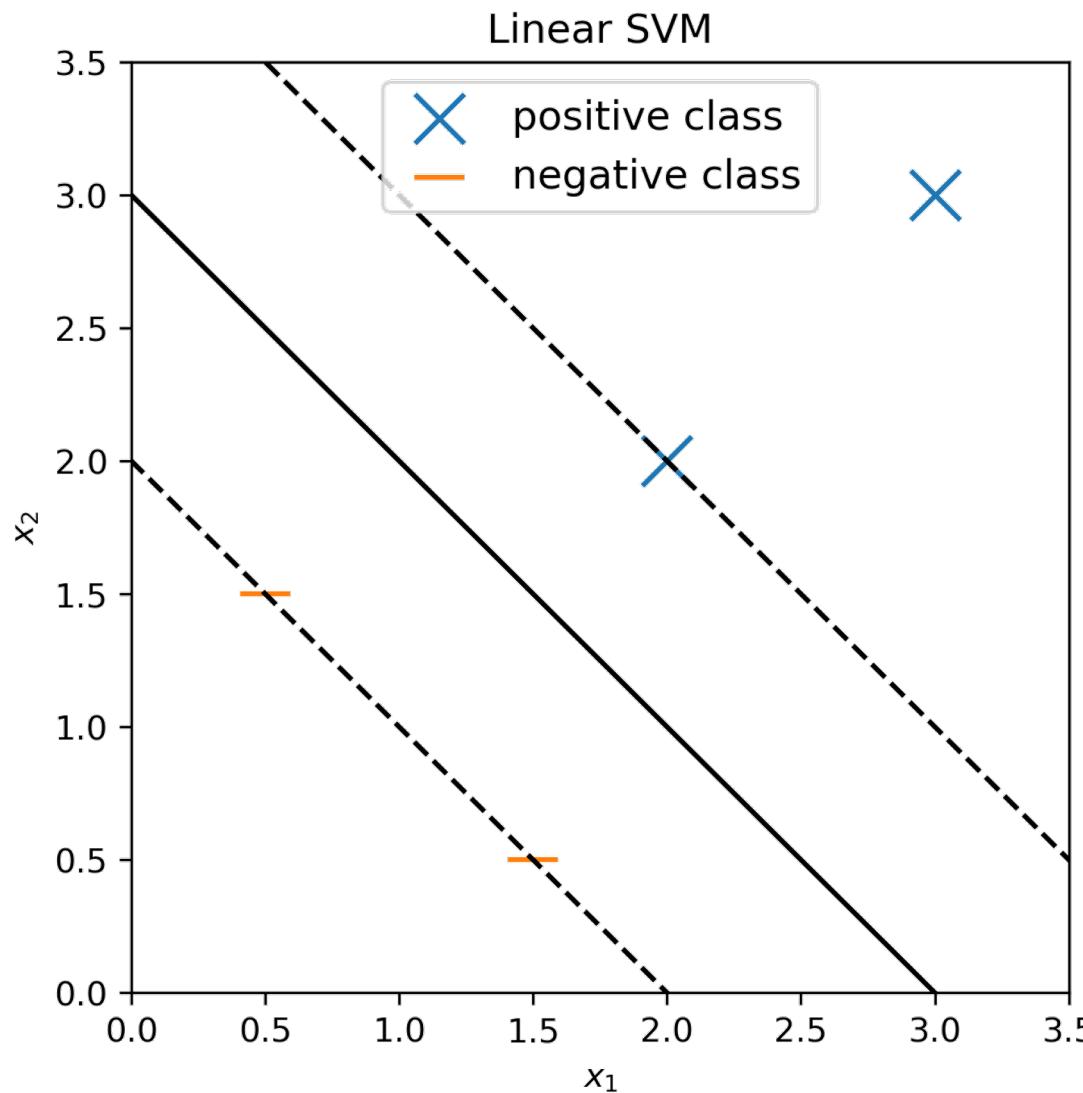
Метод опорных векторов

Классификация с помощью машины опорных векторов (Support Vector Machine, SVM)

Классификация с помощью разделяющей поверхности



Оптимальная разделяющая поверхность



$$K(\vec{x}_1, \vec{x}_2) = (\gamma \langle \vec{x}_1, \vec{x}_2 \rangle + r)^d$$
$$K(\vec{x}_1, \vec{x}_2) = \exp(-\gamma \|\vec{x}_1 - \vec{x}_2\|^2)$$

Задача обучения линейного классификатора

Дано:

Обучающая выборка: $X^\ell = (\vec{x}_i, y_i)_{i=1}^\ell$,

\vec{x}_i – объекты, векторы из множества $X = \mathbb{R}^n$,

y_i – метки классов, элементы множества $Y = \{-1, +1\}$

Найти:

Параметры $\vec{\theta} \in \mathbb{R}^n$, $\theta_0 \in \mathbb{R}$ линейной модели классификации

$$g(\vec{x}; \vec{\theta}, \theta_0) = \text{sign} \left(\langle \vec{x}, \vec{\theta} \rangle - \theta_0 \right) = \text{sign} \left(\sum_{j=1}^n \theta_j f_j(\vec{x}) - \theta_0 \right)$$

Критерий – минимизация эмпирического риска:

$$\sum_{i=1}^\ell [g(\vec{x}_i; \vec{\theta}, \theta_0) \neq y_i] = \sum_{i=1}^\ell [M_i(\vec{\theta}, \theta_0) < 0] \rightarrow \min_{\vec{\theta}, \theta_0}$$

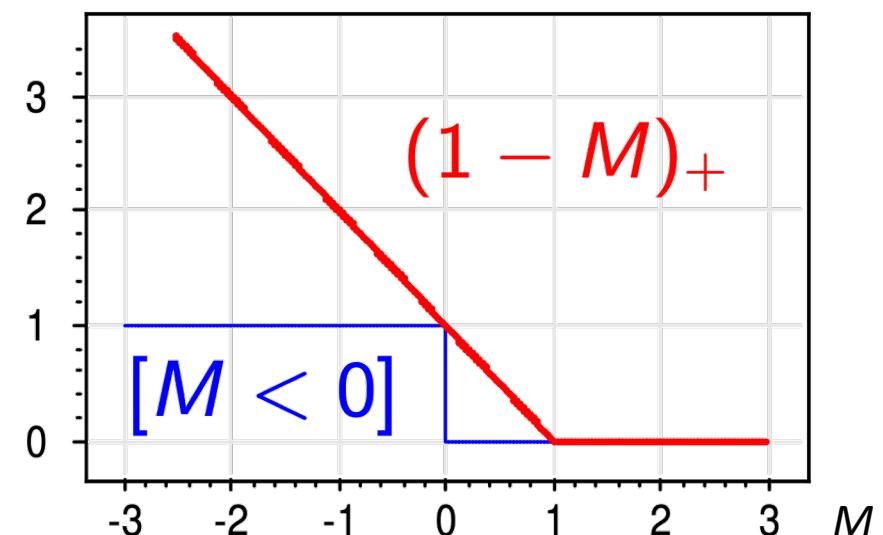
где $M_i(\vec{\theta}, \theta_0) = y_i (\langle \vec{x}_i, \vec{\theta} \rangle - \theta_0)$ – отступ (margin) объекта \vec{x}_i .

Аппроксимация и регуляризация эмпирического риска

- Эмпирический риск – это кусочно-постоянная функция. Заменим его оценкой сверху, непрерывной по параметрам:

$$\begin{aligned} Q(\vec{\theta}, \theta_0) &= \sum_{i=1}^{\ell} [M_i(\vec{\theta}, \theta_0) < 0] \leq \\ &\leq \sum_{i=1}^{\ell} (1 - M_i(\vec{\theta}, \theta_0))_+ + \frac{1}{2C} \|\vec{\theta}\|^2 \rightarrow \min_{\vec{\theta}, \theta_0} \end{aligned}$$

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



Оптимальная разделяющая гиперплоскость

Линейный классификатор: $g(x; \vec{\theta}, \theta_0) = \text{sign}(\langle \vec{x}, \vec{\theta} \rangle - \theta_0)$

Пусть выборка $X^\ell = (\vec{x}_i, y_i)_{i=1}^\ell$ линейно разделима:

$$\exists \vec{\theta}, \theta_0: M_i(\vec{\theta}, \theta_0) = y_i (\langle \vec{x}_i, \vec{\theta} \rangle - \theta_0) > 0, \quad i = 1, \dots, \ell$$

Нормировка (ограничение): $\min_i M_i(\vec{\theta}, \theta_0) = 1$

Разделяющая полоса (разделяющая гиперплоскость посередине):

$$\{\vec{x}: -1 \leq \langle \vec{x}, \vec{\theta} \rangle - \theta_0 \leq 1\}$$

$$\exists \vec{x}_+: \langle \vec{x}_+, \vec{\theta} \rangle - \theta_0 = +1$$

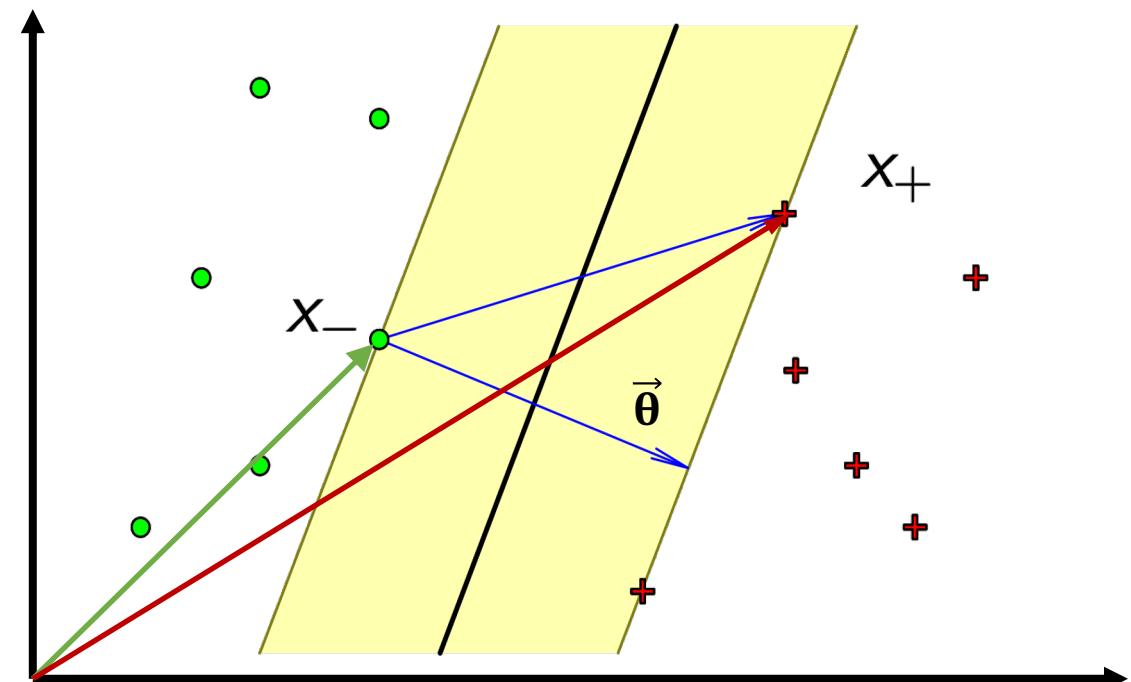
$$\exists \vec{x}_-: \langle \vec{x}_-, \vec{\theta} \rangle - \theta_0 = -1$$

$\vec{\theta}$ – вектор произвольной длины, \perp разделяющей гиперплоскости

Ширина полосы:

$$\frac{\langle \vec{x}_+ - \vec{x}_-, \vec{\theta} \rangle}{\|\vec{\theta}\|} = \frac{2}{\|\vec{\theta}\|} \rightarrow \max$$

Справка: $\frac{\vec{\theta}}{\|\vec{\theta}\|}$ – вектор единичной длины, с тем же направлением, что и $\vec{\theta}$



Справка: $\langle \vec{a}, \vec{b} \rangle = \|\vec{a}\| \|\vec{b}\| \cos \widehat{\vec{a}, \vec{b}}$

Геометрическая интерпретация

Дано:

линейно-разделимое множество объектов двух классов
 $X = X_+ \cup X_- = \mathbb{R}^n$;

$\vec{\theta}$ – вектор, перпендикулярный к разделяющей гиперплоскости;

\vec{x} – объект (вектор), класс которого неизвестен, $\vec{x} \in X$.

Найти:

класс, к которому относится \vec{x} .

Решение:

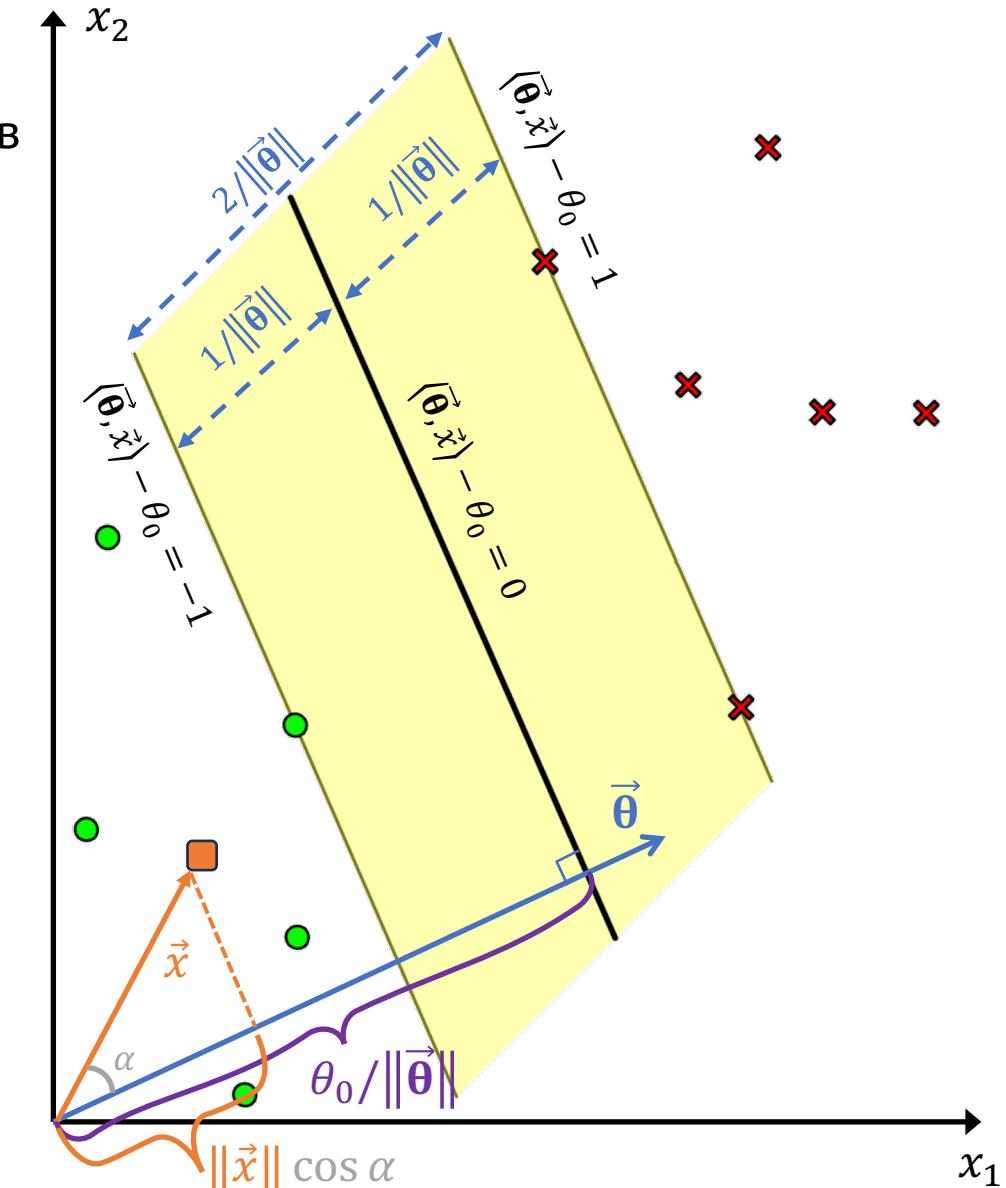
Чтобы узнать, по какую сторону от разделяющей гиперплоскости находится объект \vec{x} , достаточно спроектировать вектор \vec{x} на вектор $\vec{\theta}$ и сравнить длину получившейся проекции с пороговым значением θ_0 :

$$\langle \vec{\theta}, \vec{x} \rangle \geq \theta_0 \Leftrightarrow \|\vec{x}\| \cos \alpha \geq \theta_0 / \|\vec{\theta}\|$$

Решающее правило:

$$\langle \vec{\theta}, \vec{x} \rangle - \theta_0 \geq 0 \Rightarrow \vec{x} \in X_+$$

$$\langle \vec{\theta}, \vec{x} \rangle - \theta_0 < 0 \Rightarrow \vec{x} \in X_-$$



Обоснование кусочно-линейной функции потерь

Задача обучения классификатора – максимизация ширины полосы между классами:

$$\frac{2}{\|\vec{\theta}\|} \rightarrow \max \Leftrightarrow \frac{1}{\|\vec{\theta}\|} \rightarrow \max \Leftrightarrow \|\vec{\theta}\| \rightarrow \min \Leftrightarrow \frac{1}{2} \|\vec{\theta}\|^2 \rightarrow \min$$

Линейно-разделимая выборка:

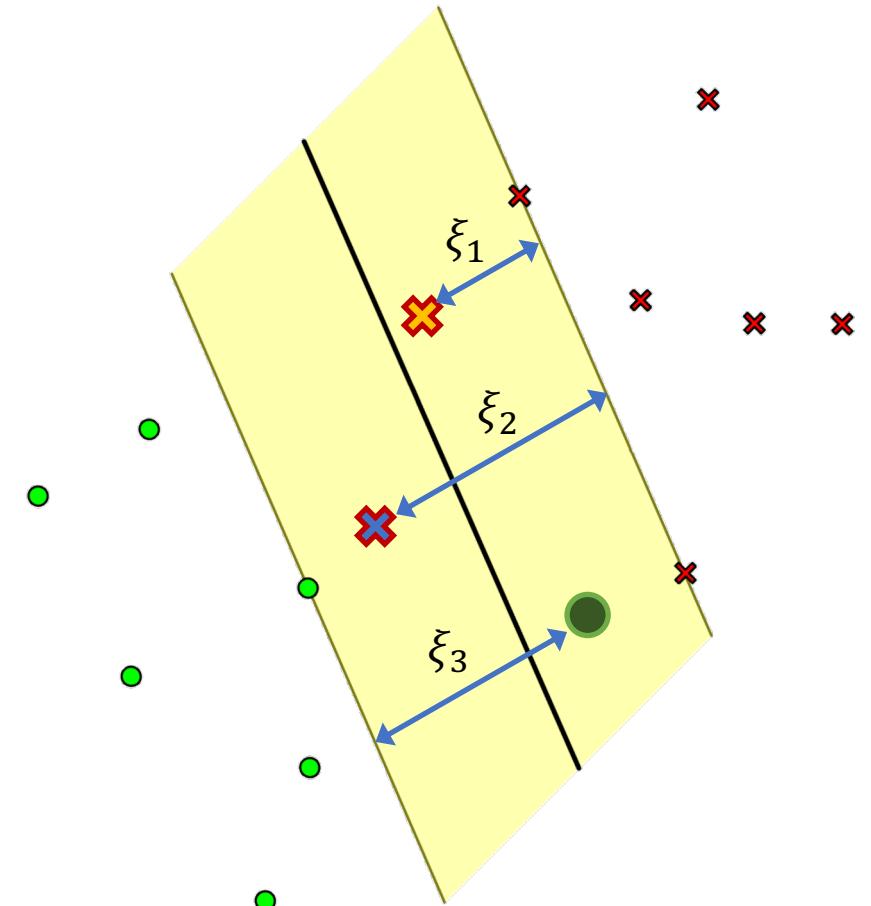
$$\begin{cases} \frac{1}{2} \|\vec{\theta}\|^2 \rightarrow \min; \\ M_i(\vec{\theta}, \theta_0) \geq 1, \quad i = 1, \dots, \ell \end{cases}$$

Переход к линейно-неразделимой выборке (эвристика):

$$\begin{cases} \frac{1}{2} \|\vec{\theta}\|^2 + C \sum_{i=1}^{\ell} \xi_i \rightarrow \min; \\ M_i(\vec{\theta}, \theta_0) \geq 1 - \xi_i, \quad i = 1, \dots, \ell \\ \xi_i \geq 0, \quad i = 1, \dots, \ell \end{cases}$$

Эквивалентная задача безусловной минимизации:

$$C \sum_{i=1}^{\ell} (1 - M_i(\vec{\theta}, \theta_0))_+ + \frac{1}{2} \|\vec{\theta}\|^2 \rightarrow \min_{\vec{\theta}, \theta_0}$$



Справка. Условия Каруша – Куна – Таккера (ККТ)

Метод ККТ – обобщение метода множителей Лагранжа

- Задача математического программирования:

$$\begin{cases} f(x) \rightarrow \min_x; \\ g_i(x) \leq 0, \quad i = 1, \dots, m; \\ h_j(x) = 0, \quad j = 1, \dots, k. \end{cases}$$

- Необходимые условия. Если x – точка локального минимума, то существуют множители Лагранжа $\mu_i, i = 1, \dots, m, \lambda_j, j = 1, \dots, k$:

$$\begin{cases} \frac{\partial L}{\partial x} = 0, \quad L(x; \mu, \lambda) = f(x) + \sum_{i=1}^m \mu_i g_i(x) + \sum_{j=1}^k \lambda_j h_j(x) \\ g_i(x) \leq 0; \quad h_j(x) = 0; \quad (\text{исходные ограничения}) \\ \mu_i \geq 0; \quad (\text{двойственные ограничения}) \\ \mu_i g_i(x) = 0; \quad (\text{условие дополняющей нежесткости}) \end{cases}$$

где $L(x; \mu, \lambda)$ – функция Лагранжа

Применение условий ККТ к задаче SVM

Ограничений типа равенств нет, $k = 0$. Есть два набора ограничений типа неравенств, $m = 2\ell$, поэтому примем $\mu = \{\lambda_i, \eta_i\}$, $i = 1, \dots, \ell$.

Функция Лагранжа $L(\vec{\theta}, \theta_0, \vec{\xi}; \lambda, \eta) =$

$$= \frac{1}{2} \|\vec{\theta}\|^2 - \sum_{i=1}^{\ell} \lambda_i (M_i(\vec{\theta}, \theta_0) - 1) - \sum_{i=1}^{\ell} \xi_i (\lambda_i + \eta_i - C)$$

λ_i – переменные, двойственные к ограничениям $M_i \geq 1 - \xi_i$;

η_i – переменные, двойственные к ограничениям $\xi_i \geq 0$.

$$\begin{cases} \frac{\partial L}{\partial \vec{\theta}} = 0, & \frac{\partial L}{\partial \theta_0} = 0, & \frac{\partial L}{\partial \xi_i} = 0; \\ \xi_i \geq 0, & \lambda_i \geq 0, & \eta_i \geq 0, & i = 1, \dots, \ell \\ \lambda_i = 0 \text{ либо } M_i(\vec{\theta}, \theta_0) = 1 - \xi_i, & & & i = 1, \dots, \ell \\ \eta_i = 0 \text{ либо } \xi_i = 0, & & i = 1, \dots, \ell \end{cases}$$

Необходимые условия седловой точки функции Лагранжа

Функция Лагранжа $L(\vec{\theta}, \theta_0, \vec{\xi}; \lambda, \eta) =$

$$= \frac{1}{2} \|\vec{\theta}\|^2 - \sum_{i=1}^{\ell} \lambda_i \left(\textcolor{red}{y_i (\langle \vec{x}_i, \vec{\theta} \rangle - \theta_0)} - 1 \right) - \sum_{i=1}^{\ell} \xi_i (\lambda_i + \eta_i - C)$$

Необходимые условия седловой точки функции Лагранжа:

$$\frac{\partial L}{\partial \vec{\theta}} = \vec{\theta} - \sum_{i=1}^{\ell} \lambda_i y_i \vec{x}_i = 0 \quad \Rightarrow \quad \vec{\theta} = \sum_{i=1}^{\ell} \lambda_i y_i \vec{x}_i$$

$$\frac{\partial L}{\partial \theta_0} = - \sum_{i=1}^{\ell} \lambda_i y_i = 0 \quad \Rightarrow \quad \sum_{i=1}^{\ell} \lambda_i y_i = 0$$

$$\frac{\partial L}{\partial \xi_i} = -\lambda_i - \eta_i + C = 0 \quad \Rightarrow \quad \lambda_i + \eta_i = C, \quad i = 1, \dots, \ell$$

Понятие опорного вектора и типизация объектов

Система условий ККТ:

$$\begin{cases} \vec{\theta} = \sum_{i=1}^{\ell} \lambda_i y_i \vec{x}_i; \quad \sum_{i=1}^{\ell} \lambda_i y_i = 0; \quad M_i(\vec{\theta}, \theta_0) = 1 - \xi_i; \\ \xi_i \geq 0, \quad \lambda_i \geq 0, \quad \eta_i \geq 0, \quad \eta_i + \lambda_i = C; \\ \lambda_i = 0 \text{ либо } M_i(\vec{\theta}, \theta_0) = 1 - \xi_i; \\ \eta_i = 0 \text{ либо } \xi_i = 0; \end{cases}$$

Определение. Объект \vec{x}_i называется **опорным**, если $\lambda_i \neq 0$.

Типизация объектов \vec{x}_i , $i = 1, \dots, \ell$:

1. $\lambda_i = 0; \eta_i = C; \xi_i = 0; M_i \geq 1$ – периферийный.
2. $0 < \lambda_i < C; 0 < \eta_i < C; \xi_i = 0; M_i = 1$ – **опорный**-границы.
3. $\lambda_i = C; \eta_i = 0; \xi_i > 0; M_i < 1$ – **опорный**-нарушитель.

Оптимизационная задача

Найдем значение функции Лагранжа $L(\vec{\theta}, \theta_0, \vec{\xi}; \lambda, \eta) =$

$$= \frac{1}{2} \|\vec{\theta}\|^2 - \sum_{i=1}^{\ell} \lambda_i \left(y_i (\langle \vec{x}_i, \vec{\theta} \rangle - \theta_0) - 1 \right) - \sum_{i=1}^{\ell} \xi_i (\lambda_i + \eta_i - C)$$

в седловой точке $\vec{\theta} = \sum_{i=1}^{\ell} \lambda_i y_i \vec{x}_i; \sum_{i=1}^{\ell} \lambda_i y_i = 0; \lambda_i + \eta_i = C, i = 1, \dots, \ell.$

$$L(\dots) = \frac{1}{2} \left\langle \sum_{i=1}^{\ell} \lambda_i y_i \vec{x}_i, \sum_{j=1}^{\ell} \lambda_j y_j \vec{x}_j \right\rangle - \sum_{i=1}^{\ell} \left(\left\langle \lambda_i y_i \vec{x}_i, \sum_{j=1}^{\ell} \lambda_j y_j \vec{x}_j \right\rangle - \lambda_i y_i \theta_0 - \lambda_i \right) - 0 =$$

$$= \frac{1}{2} \left\langle \sum_{i=1}^{\ell} \lambda_i y_i \vec{x}_i, \sum_{j=1}^{\ell} \lambda_j y_j \vec{x}_j \right\rangle - \left\langle \sum_{i=1}^{\ell} \lambda_i y_i \vec{x}_i, \sum_{j=1}^{\ell} \lambda_j y_j \vec{x}_j \right\rangle + \theta_0 \sum_{i=1}^{\ell} \lambda_i y_i + \sum_{i=1}^{\ell} \lambda_i$$

$$L(\vec{\theta}, \theta_0, \vec{\xi}; \lambda, \eta) = \sum_{i=1}^{\ell} \lambda_i - \frac{1}{2} \sum_{i=1}^{\ell} \sum_{j=1}^{\ell} \lambda_i \lambda_j y_i y_j \langle \vec{x}_i, \vec{x}_j \rangle$$

Минимизируемая функция зависит только от скалярного произведения $\langle \vec{x}_i, \vec{x}_j \rangle$

Построение классификатора

Решение оптимизационной задачи:

$$\begin{cases} \vec{\theta} = \sum_{i=1}^{\ell} \lambda_i y_i \vec{x}_i \\ \theta_0 = \langle \vec{x}_i, \vec{\theta} \rangle - y_i, \quad \forall i: \lambda_i > 0, M_i = 1 \end{cases}$$

Линейный классификатор с признаками $f_i(\vec{x}) = \langle \vec{x}_i, \vec{x} \rangle$:

$$g(\vec{x}) = \text{sign} \left(\sum_{i=1}^{\ell} \lambda_i y_i \langle \vec{x}_i, \vec{x} \rangle - \theta_0 \right)$$

Линейный классификатор с признаками $f_i(\vec{x}) = K(\vec{x}_i, \vec{x})$ (нелинейное обобщение с ядром K):

$$g(\vec{x}) = \text{sign} \left(\sum_{i=1}^{\ell} \lambda_i y_i K(\vec{x}_i, \vec{x}) - \theta_0 \right)$$

Нелинейное обобщение SVM

Идея: заменить $\langle x, x' \rangle$ нелинейной функцией $K(x, x')$.

Переход к спрямляющему пространству, как правило более высокой размерности: $\psi: X \rightarrow H$, т.е. ψ – функция для преобразования из X в H .

Определение:

Функция $K: X \times X \rightarrow \mathbb{R}$ – ядро, если $K(x, x') = \langle \psi(\vec{x}), \psi(\vec{x}') \rangle$ при некотором $\psi: X \rightarrow H$, где H – гильбертово пространство.

Теорема:

Функция $K(x, x')$ является ядром тогда и только тогда, когда она

- симметрична: $K(x, x') = K(x', x)$;
- и неотрицательно определена: $\int_X \int_X K(x, x') g(x)g(x') dx dx' \geq 0$ для любой $g: X \rightarrow \mathbb{R}$.

Примеры ядер

1. Линейное ядро

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

2. Квадратичное ядро

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

3. Полиномиальное ядро с произведениями (одночленами) степени d

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

4. Полиномиальное ядро с одночленами степени $\leq d$

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

5. Нейросеть с сигмоидными функциями активации

$$K(x, x') = \text{th}(k_1 \langle x, x' \rangle - k_0), \quad k_0, k_1 \geq 0$$

6. Сеть радиальных базисных функций (RBF ядро, гауссовское ядро)

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

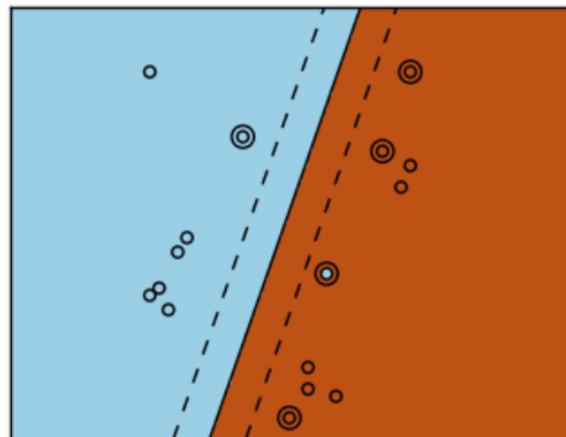
Классификация с различными ядрами

Гиперплоскость в спрямляющем пространстве соответствует нелинейной разделяющей поверхности в исходном.

Примеры с различными ядрами $K(x, x')$

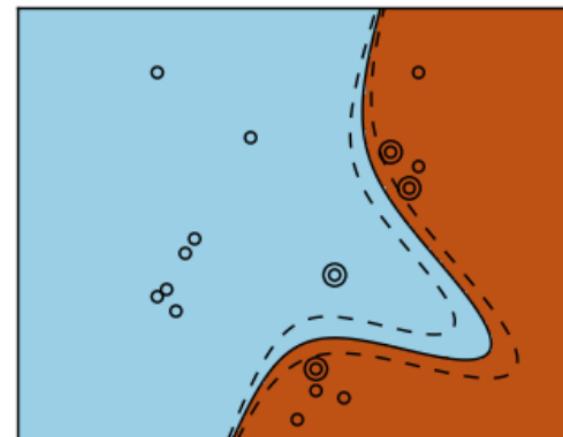
линейное

$$\langle x, x' \rangle$$



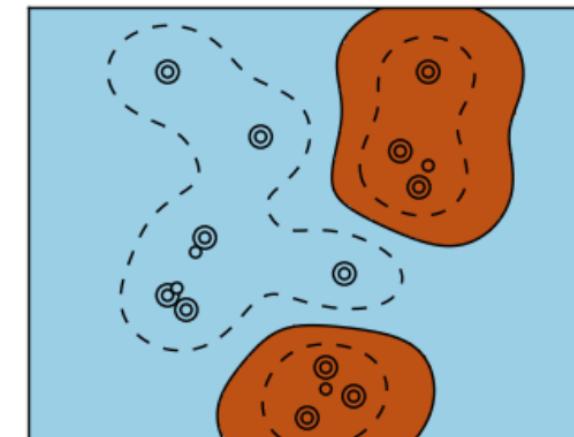
полиномиальное

$$(\langle x, x' \rangle + 1)^d, \quad d=3$$



гауссовское (RBF)

$$\exp(-\gamma \|x - x'\|^2)$$

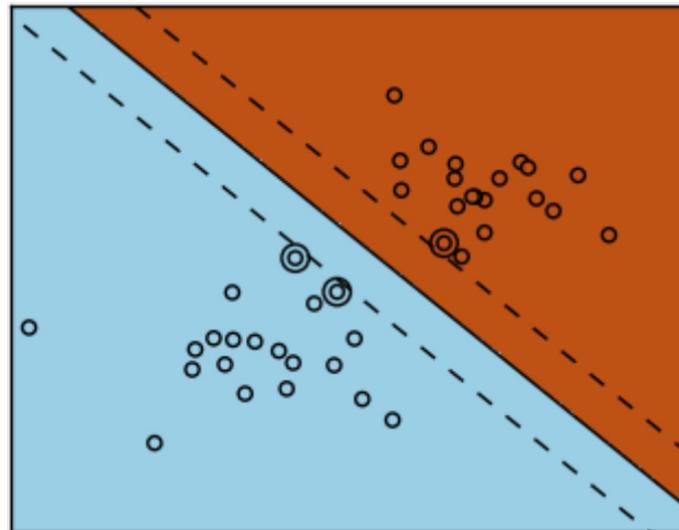


Влияние константы C на решение SVM

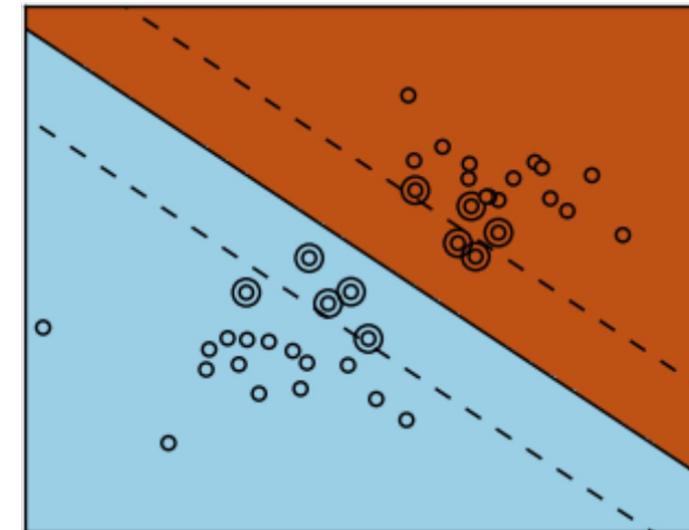
SVM – аппроксимация и регуляризация эмпирического риска:

$$\sum_{i=1}^{\ell} \left(1 - M_i(\vec{\theta}, \theta_0) \right)_+ + \frac{1}{2C} \|\vec{\theta}\|^2 \rightarrow \min_{\vec{\theta}, \theta_0}$$

большой C
слабая регуляризация

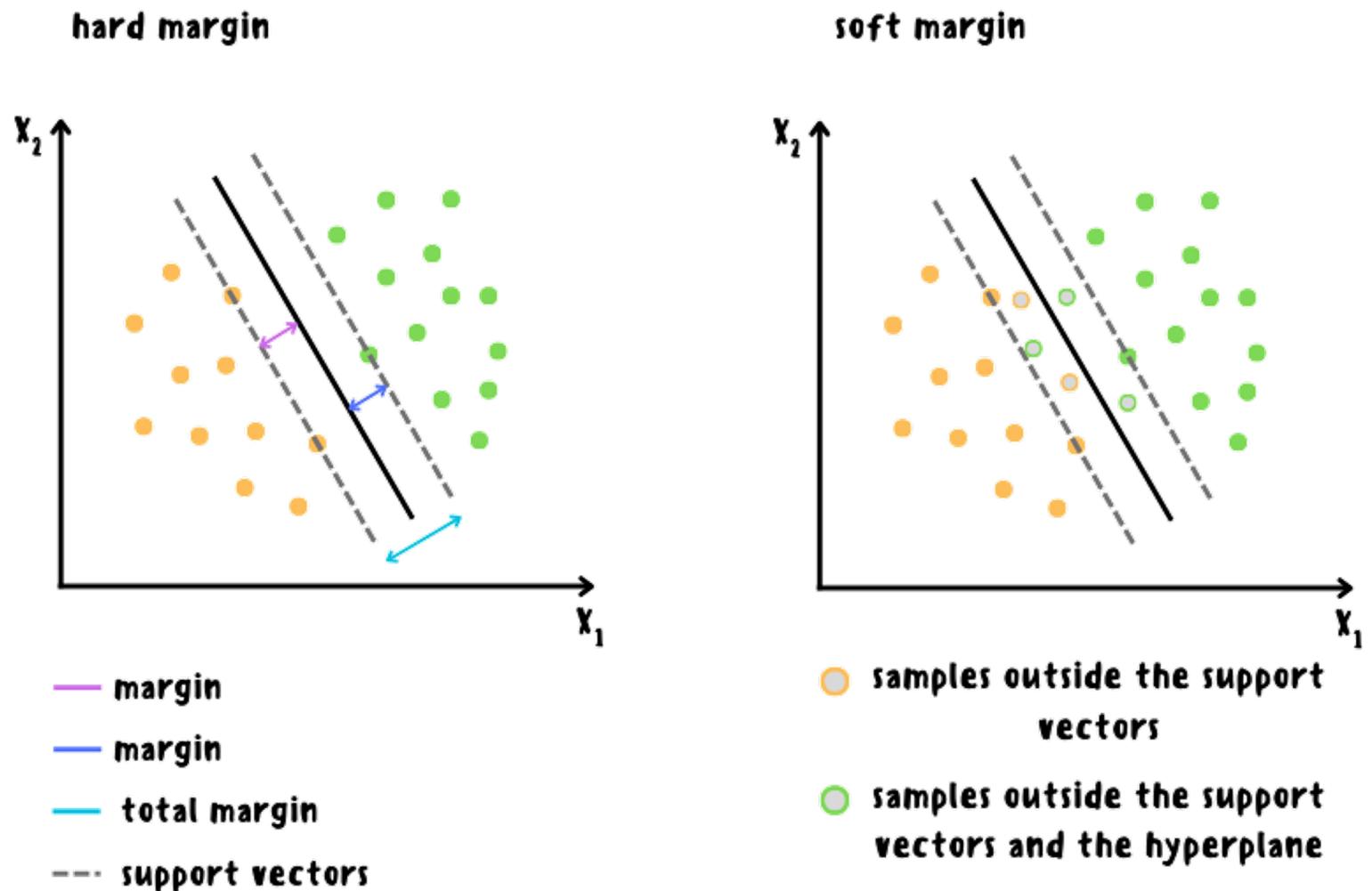


малый C
сильная регуляризация



Терминология. Жесткий и мягкий зазор

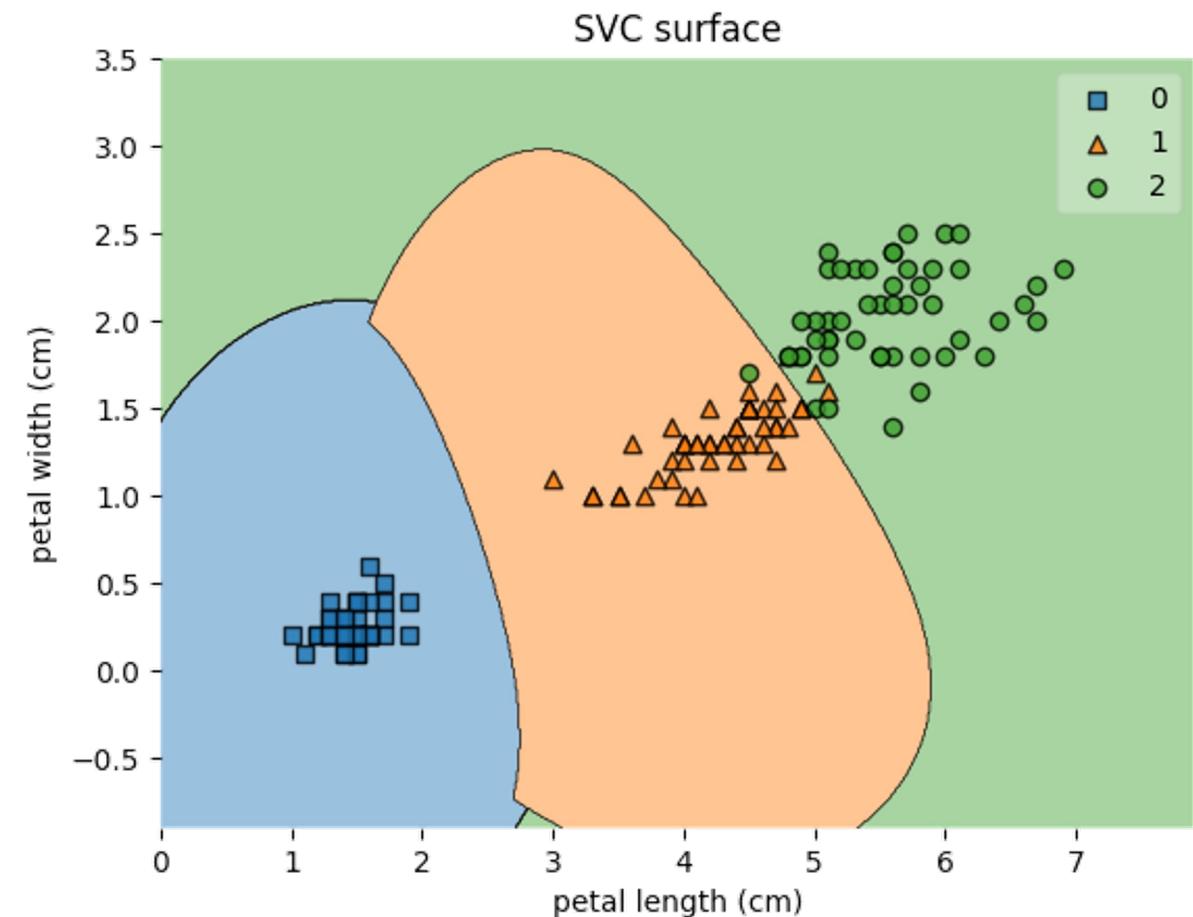
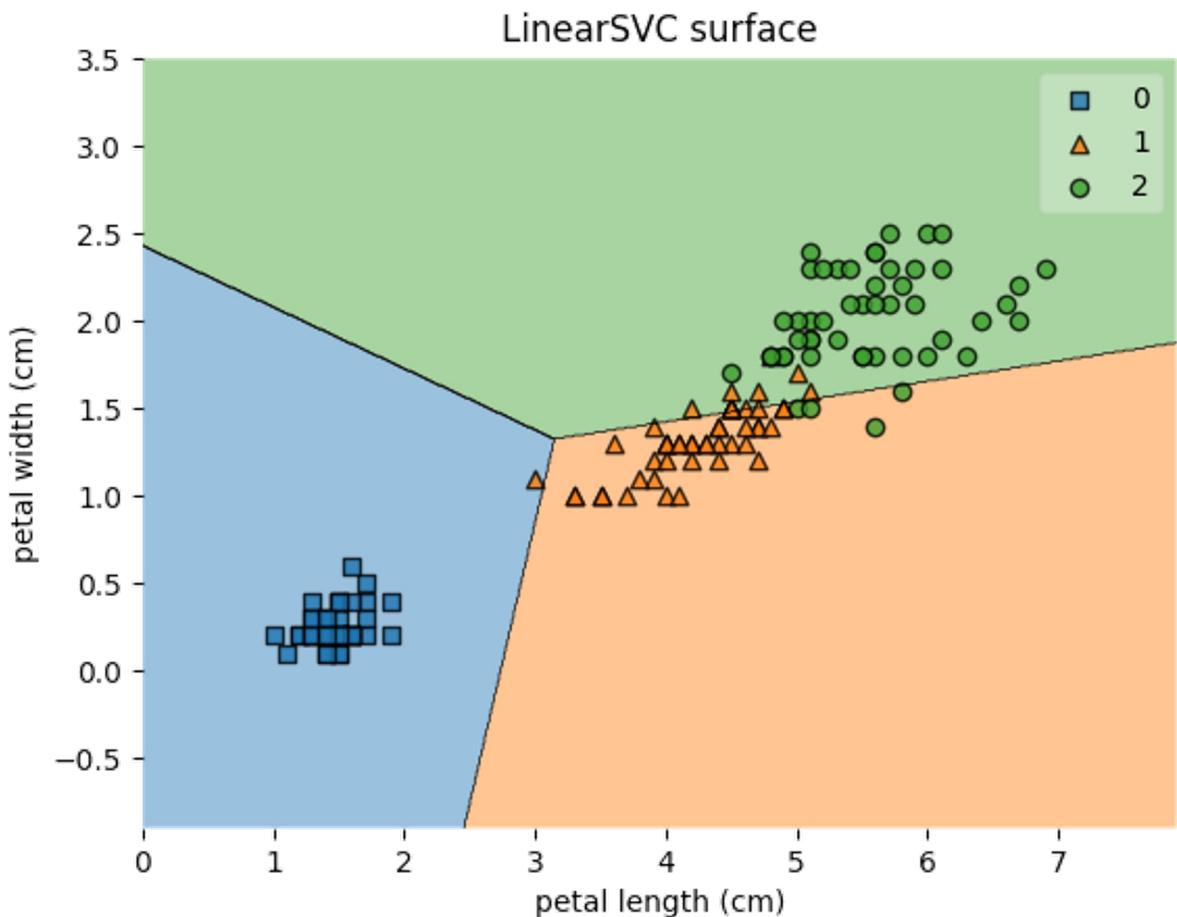
- Классификация с **жестким отступом (зазором)**: все обучающие образцы должны быть правильно классифицированы, т.е. $\forall \xi_i = 0, i = 1, \dots, \ell$
- Классификация с **мягким отступом (зазором)**: некоторые образцы могут нарушать условие правильной классификации или попадать в разделяющую полосу, т.е. $\exists \xi_i > 0, i = 1, \dots, \ell$



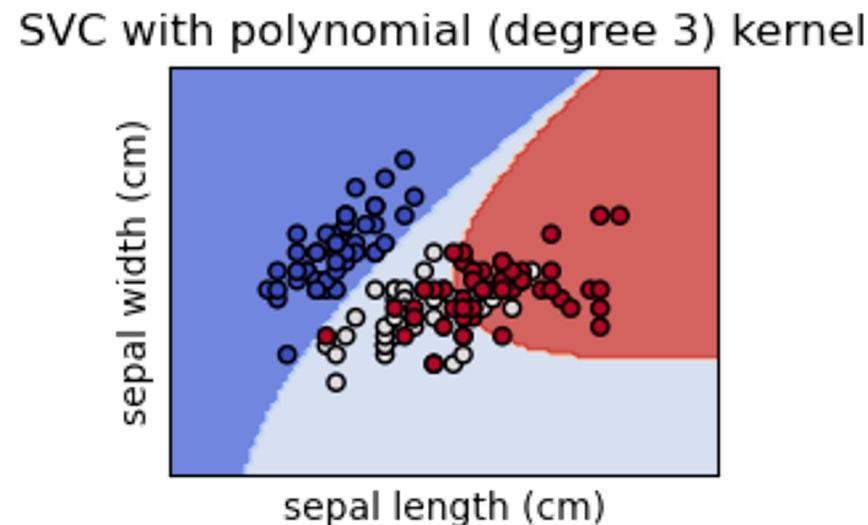
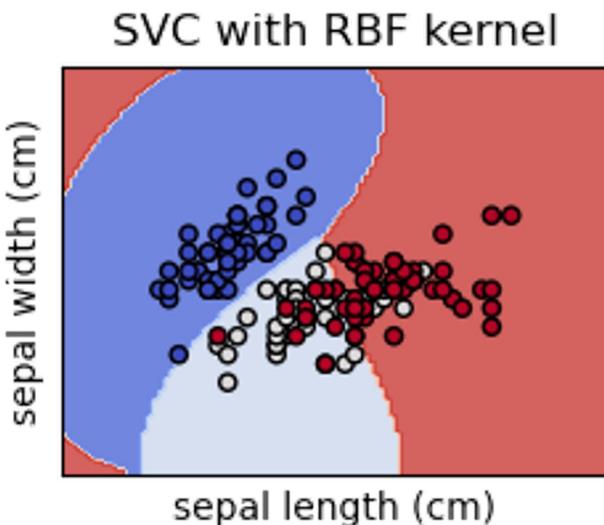
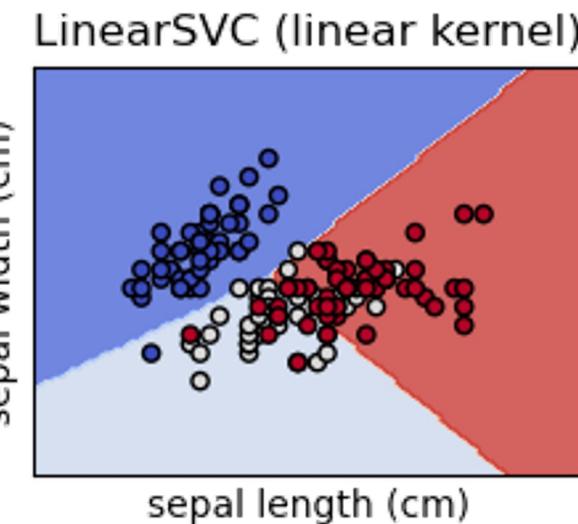
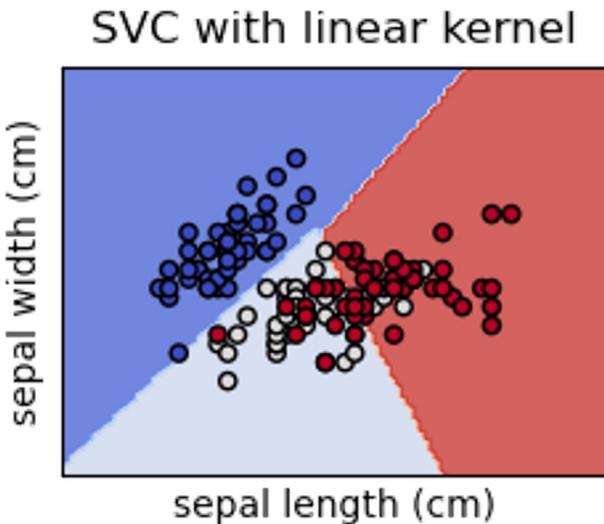
Ирисы Фишера

```
sklearn.svm.SVC(..., kernel='rbf',  
gamma='auto', random_state=0, ...)
```

gamma – коэффициент ядра:
'auto' = 1 / n_features



Ирисы Фишера



Преимущества и недостатки SVM

- 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 данных SVM обучается медленнее SG
 - нет «встроенного» отбора признаков
 - приходится подбирать константу C

Резюме

- SVM – лучший метод линейной классификации
- С помощью ядер (kernel trick) SVM изящно обобщается для нелинейной классификации и нелинейной регрессии
- Апроксимация пороговой функции потерь $\mathcal{L}(M)$ увеличивает зазор и повышает надёжность классификации
- Регуляризация увеличивает зазор, устраняет мультиколлинеарность и уменьшает переобучение
- Негладкость функции потерь приводит к отбору объектов
- Негладкость регуляризатора приводит к отбору признаков

В.Н. Вапник, А. Я. Лerner. Узнавание образов при помощи обобщенных портретов, 1963

C. Cortes, V. Vapnik. Support vector networks, 1995.

Деревья принятия решений

Интерпретируемость алгоритмов машинного обучения, классификация с помощью решающих деревьев