ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ УПРАВЛЕНИЯ
Основан в 1919 году

Методы и технологии машинного обучения

Лекция 6: Машины опорных векторов

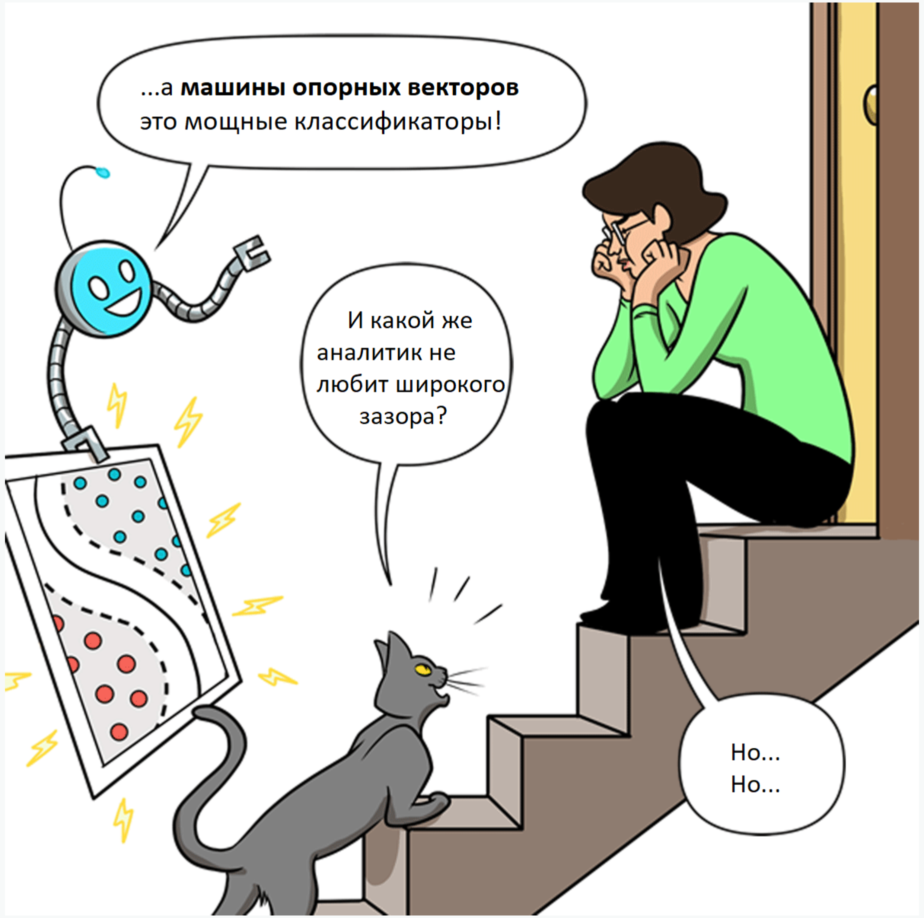
Светлана Андреевна Суязова (Аксюк)
sa_aksyuk@guu.ru

осенний семестр 2021 / 2022 учебного года

План лекции

- Классификатор с максимальным зазором

- Классификаторы на опорных векторах
- Машины опорных векторов



Источник: Learning Machine Learning: An Online Comic from Google AI

Гиперплоскость

В двумерном пространстве (прямая):

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 = 0$$

В p-мерном пространстве:

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p = 0$$

Точки A и B лежат по разные стороны гиперплоскости:

$$\beta_0 + \beta_1 X_1^A + \beta_2 X_2^A + \ldots + \beta_p X_p^A > 0$$
$$\beta_0 + \beta_1 X_1^B + \beta_2 X_2^B + \ldots + \beta_p X_p^B < 0$$

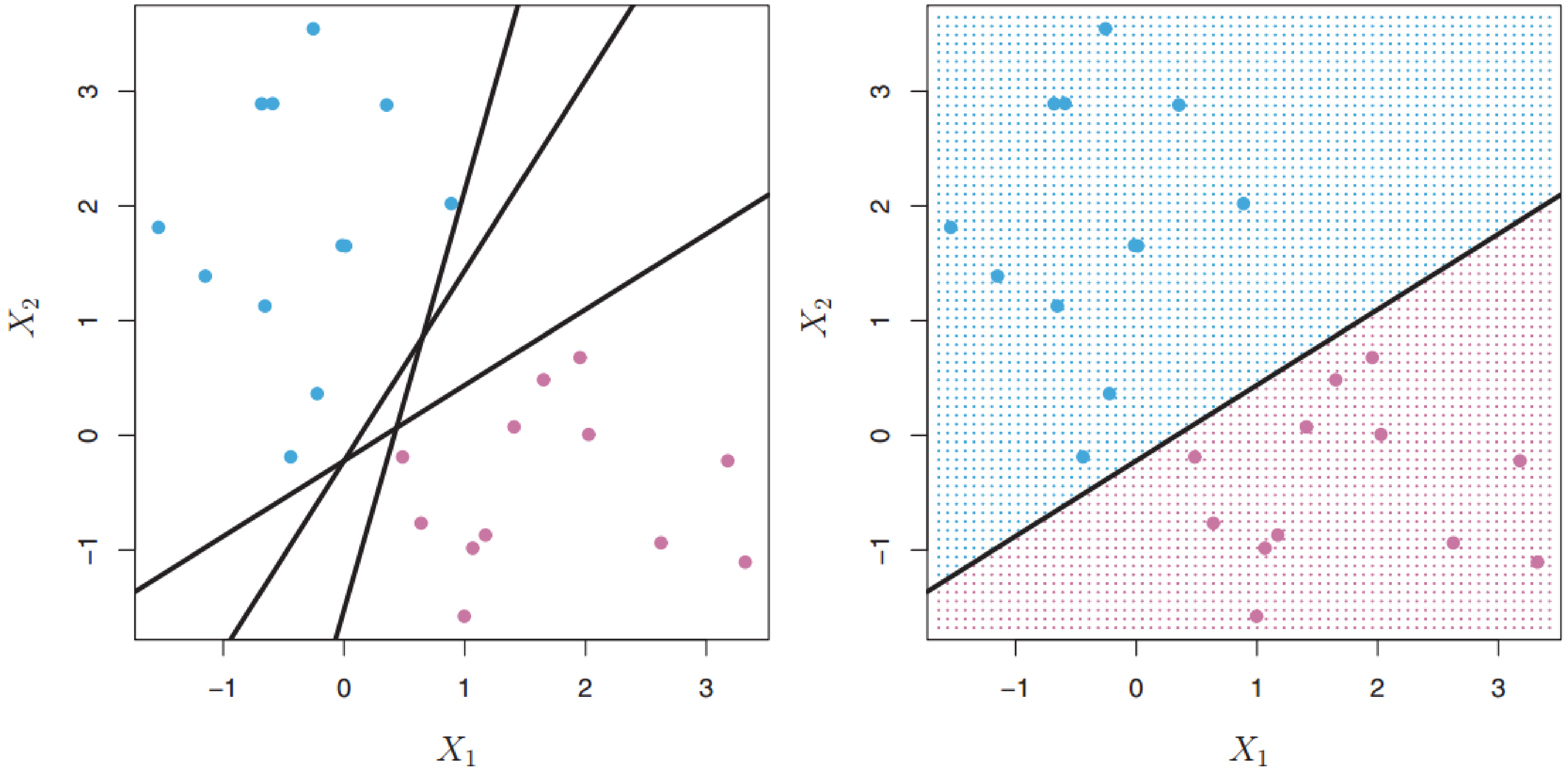
Постановка задачи 1

Обучающая выборка: n наблюдений, p измерений.

Наблюдение	$x_1 = \begin{pmatrix} x_{11} \\ \vdots \\ x_{1p} \end{pmatrix}$	$\cdots$	$x_n = \begin{pmatrix} x_{n1} \\ \vdots \\ x_{np} \end{pmatrix}$
Класс: `$y_j \in \{-1, 1\}$`	y_1	$\cdots$	y_n

Разделяющая гиперплоскость:

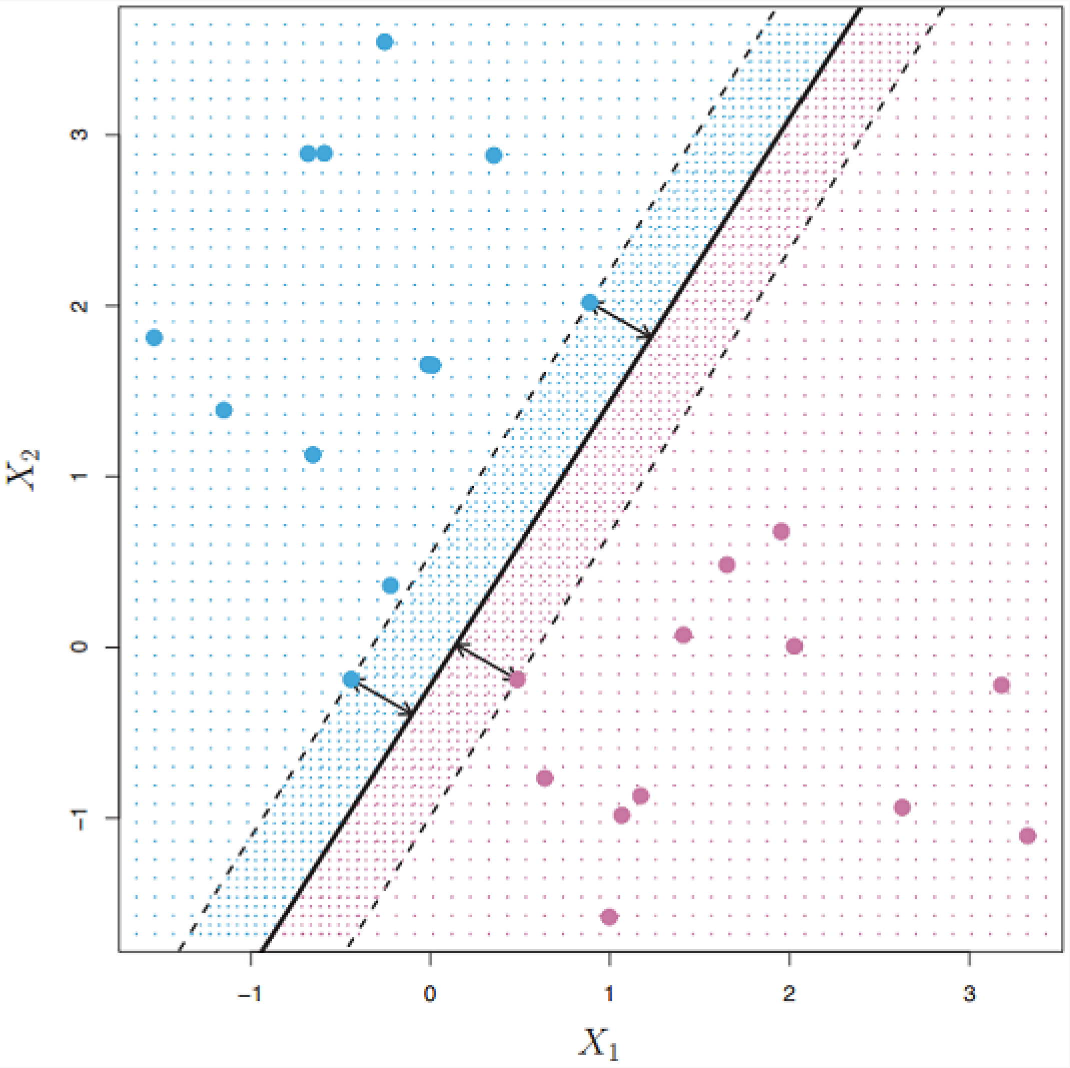
$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} > 0$ если $y_i = 1$,
$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} < 0$ если $y_i = -1$.



Слева: наблюдения двух классов и три разделяющие гиперплоскости из многих возможных
Справа: правило принятия решения показано фоном

Чёрная прямая: гиперплоскость с максимальным зазором

Стрелками показаны опорные векторы



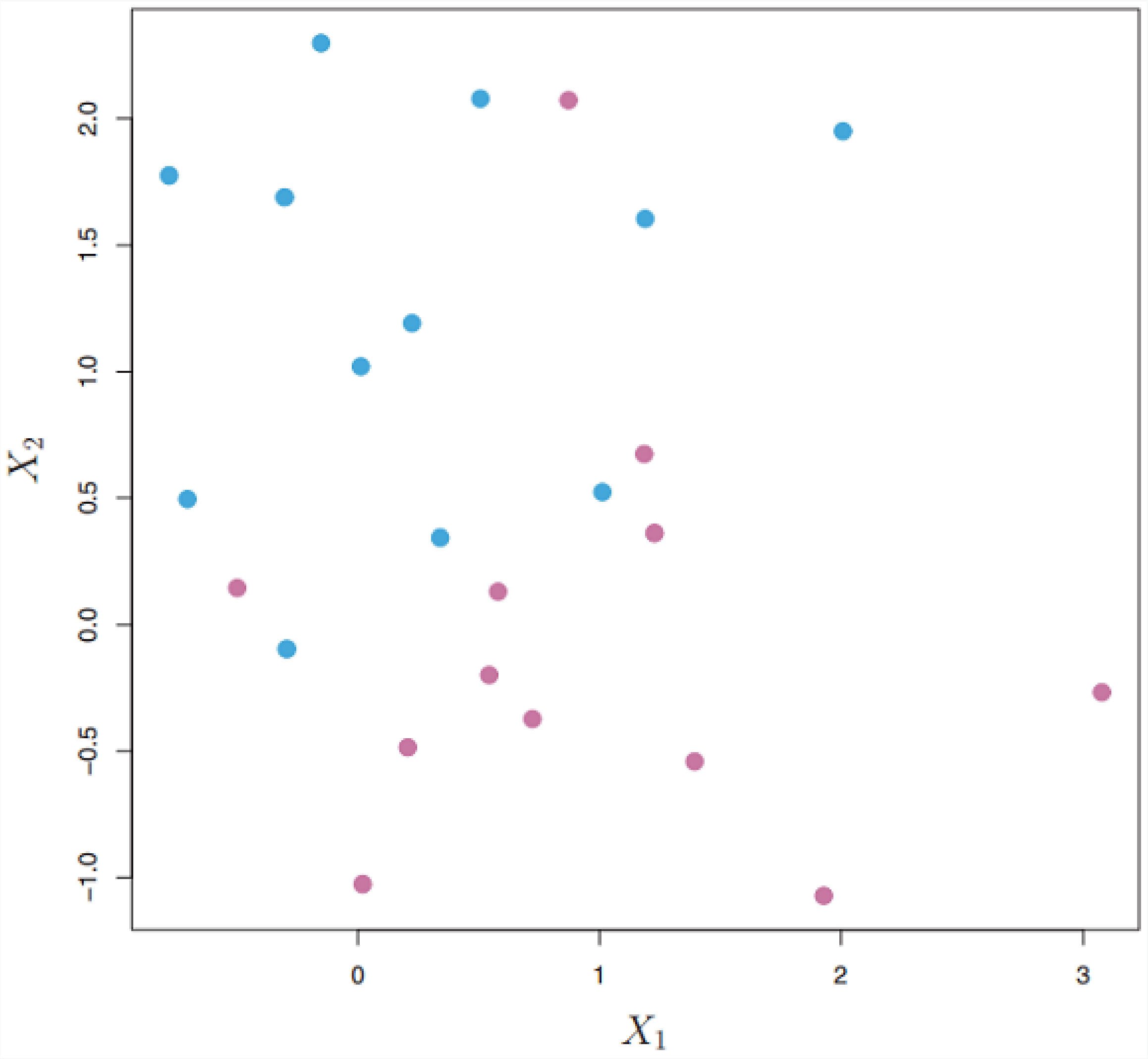
Метод решения задачи 1

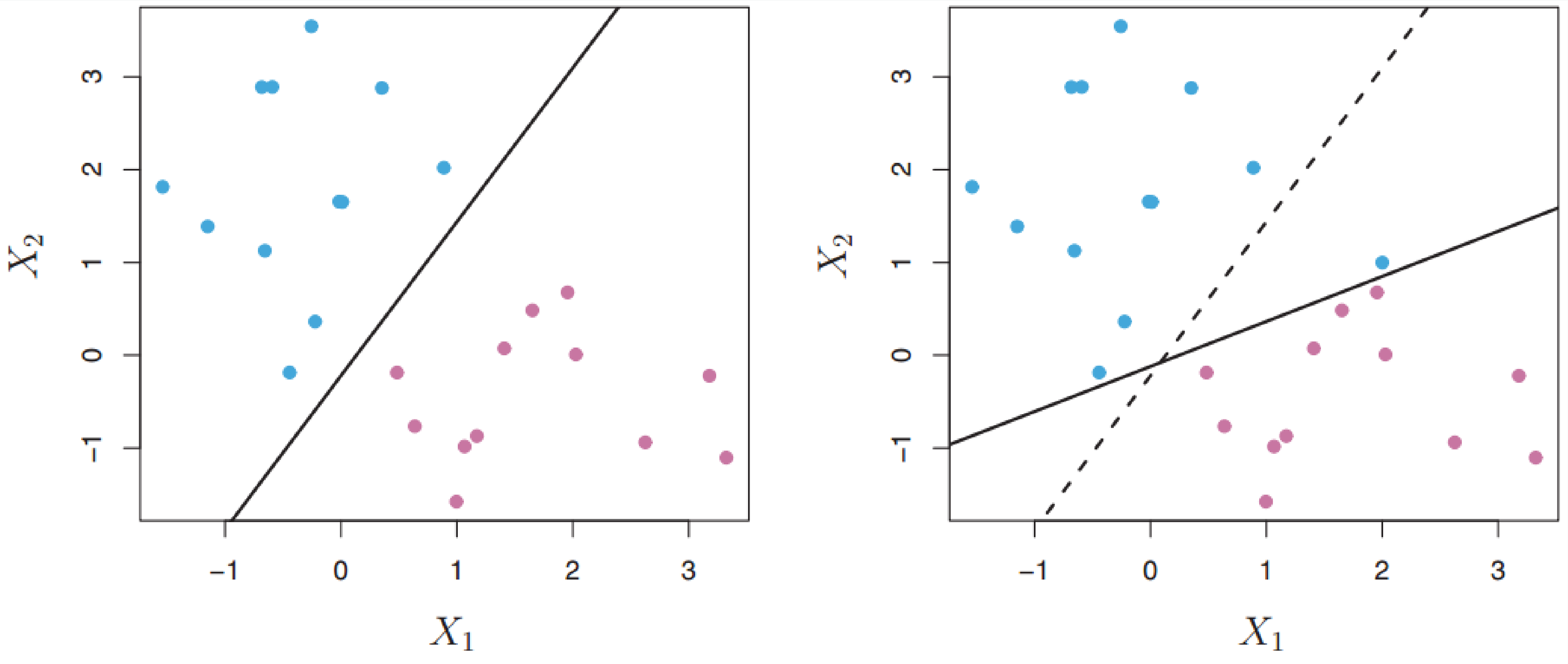
$$\max_{\beta_0,\beta_1,\dots,\beta_p} M$$

$$\begin{cases} y_i\Big(\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\dots+\beta_p x_{ip}\Big)\geq M \\ \qquad\qquad\qquad\qquad \forall i=1,\dots,n, \\ \sum\limits_{j=1}^{p}\beta_j^2=1. \end{cases}$$

Смысл ограничений: все наблюдения лежат по правильную сторону от гиперплоскости и не ближе чем на расстоянии M от неё.

Граница между классами нелинейна, классификатор с максимальным зазором неприменим.





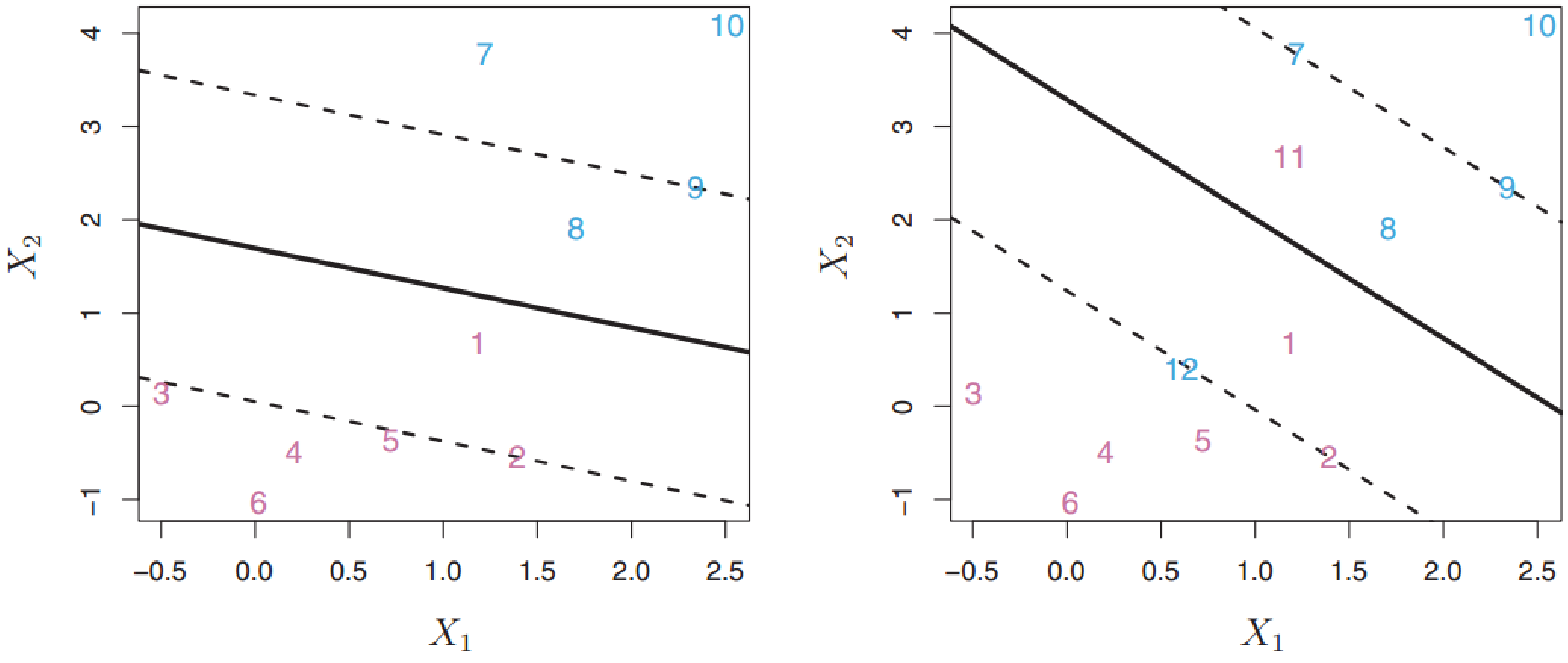
Метод неустойчив к изменению входных данных и склонен к переобучению.
Справа: смещение исходной гиперплоскости (пунктир) при добавлении одного наблюдения синего класса. Минимальный зазор стал гораздо меньше.

План лекции

- Классификатор с максимальным зазором
- Классификаторы на опорных векторах
- Машины опорных векторов

Классификатор с мягким зазором

- Не стремится чётко классифицировать все наблюдения
- Более высокое качество классификации *большинства* наблюдений
- Более устойчив к отдельным наблюдениям



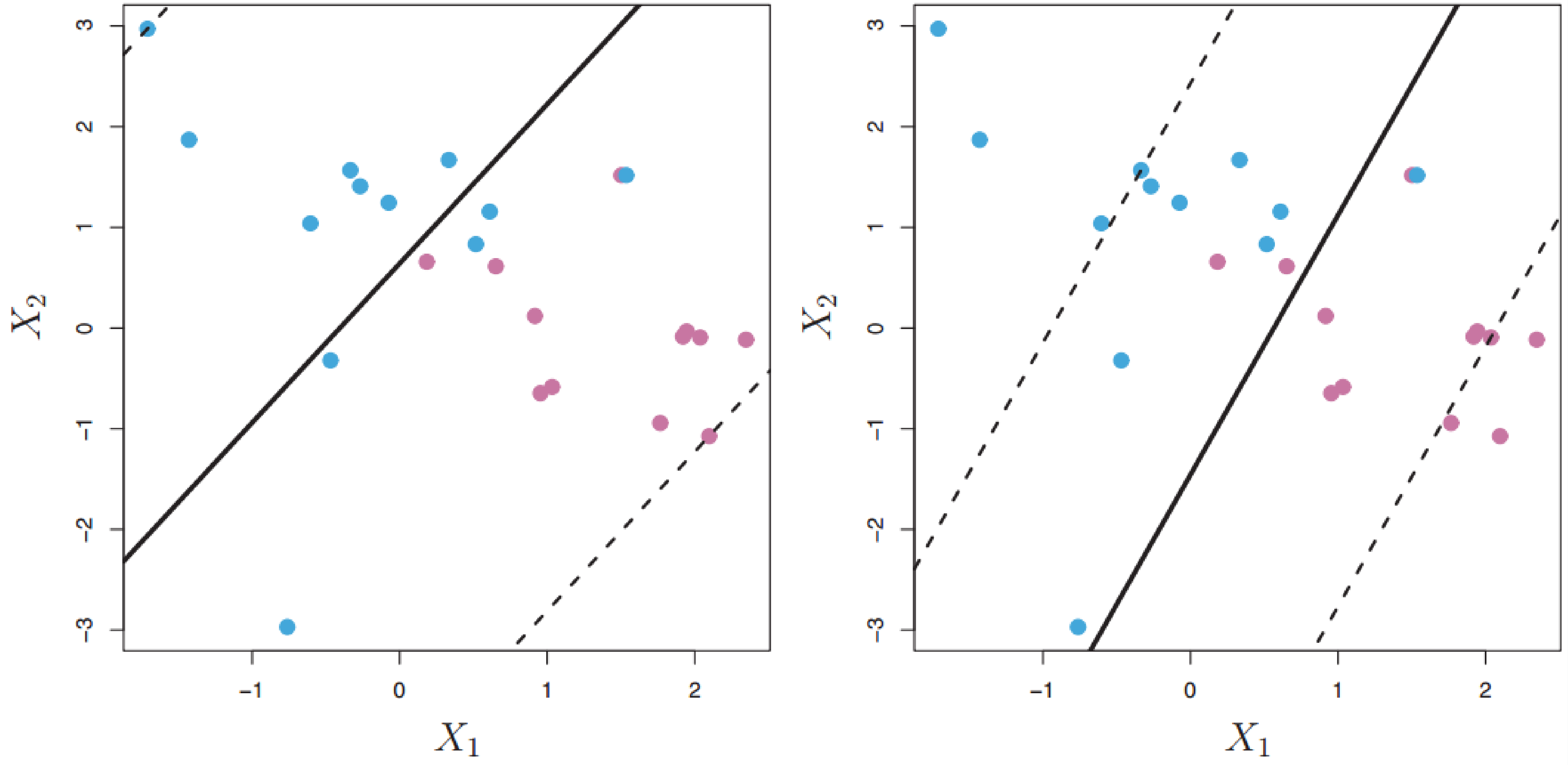
Классификатор с мягким зазором: сплошная линия – гиперплоскость, пунктир – границы зазора.
Слева: наблюдения 1, 8 лежат на неправильной стороне относительно зазора.
Справа: наблюдения 11, 12 лежат на неправильной стороне гиперплоскости (неверная классификация).

Классификатор с мягким зазором

$$\max_{\beta_0,\beta_1,\ldots,\beta_p,\epsilon_1,\ldots,\epsilon_n} M$$

$$\begin{cases} y_i\Big(\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}\Big) \geq M(1-\epsilon_i) \\ \qquad\qquad\qquad\qquad \forall i = 1,\ldots,n, \\ \epsilon_i \geq 0,\ \ \sum_{i=1}^{n}\epsilon_i \leq C,\ \ \sum_{j=1}^{p}\beta_j^2 = 1. \end{cases}$$

ϵ_i – фиктивные переменные: $\epsilon_i = 0 \Rightarrow$ верная сторона по гиперплоскости и зазору; $\epsilon_i > 0 \Rightarrow$ зазор нарушен, $\epsilon_i > 1 \Rightarrow$ гиперплоскость нарушена. Параметр C – "параметр стоимости" (*cost parameter*), допустимое число нарушений.



Классификатор на опорных векторах с разными значениями гиперпараметра: слева C больше, зазор шире.

На неправильной стороне гиперплоскости могут находиться не более C наблюдений.

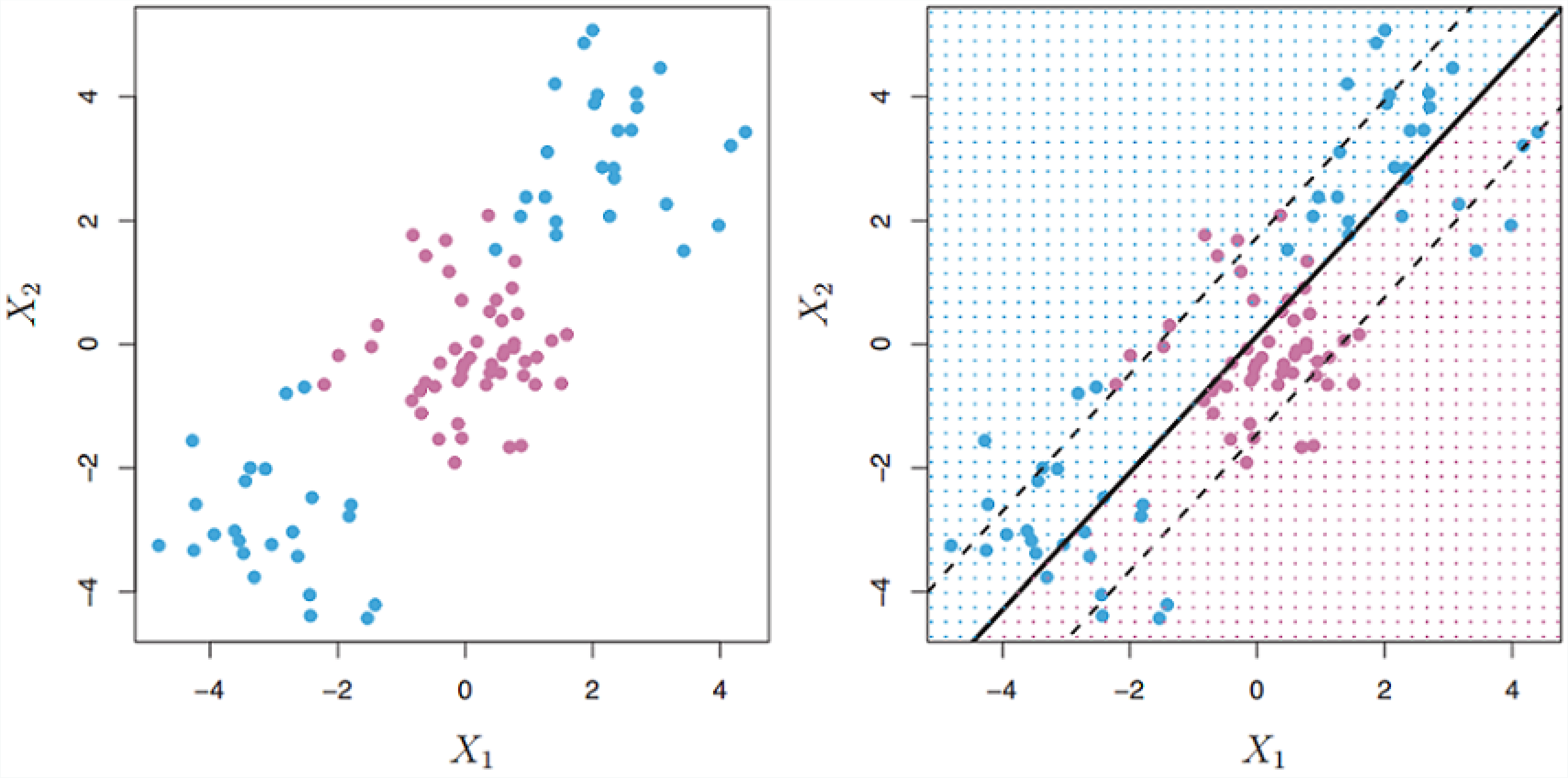
Классификатор с мягким зазором

- C подбираем перекрёстной проверкой
- при небольших C зазор узкий, граница нарушается редко, $\Rightarrow$ хорошая аппроксимация, но высокая дисперсия
- при больших C низкая дисперсия, но плохая аппроксимация
- на гиперплоскость влияют только наблюдения, которые лежат на границах зазора или нарушают их – **опорные вектора** $\Rightarrow$ метод устойчив к наблюдениям, далёким от гиперплоскости
- задача оптимизации решается на скалярных произведениях наблюдений: $\langle x_i, x_{i'} \rangle = \sum_{j=1}^{n} x_{ij} x_{i'j}$

План лекции

- Классификатор с максимальным зазором
- Классификаторы на опорных векторах
- Машины опорных векторов

Классификатор с нелинейной границей



Можно расширить пространство предикторов с помощью полиномов. Например, для $X_1, X_2, \ldots, X_n$ построить классификатор по $X_1, X_1^2, X_2, X_2^2, \ldots, X_p, X_p^2$

Машина опорных векторов (SVM)

- пространство векторов расширяется за счёт **ядерных функций** различной формы
- вычислительная сложность классификатора с мягким зазором резко возрастает: можем получить бесконечную размерность пространства признаков
- SVM требует вычислить значение ядерной функции для скалярных произведений только уникальных пар наблюдений

Машина опорных векторов (SVM)

Линейный классификатор на опорных векторах:

$$f(x) = \beta_0 + \sum_{i=1}^{n} \alpha_i \langle x, x_i \rangle$$

α_i – по одному параметру на каждое обучающее наблюдение x_i
$\alpha_i \neq 0$ только для опорных векторов с индексами из множества S:

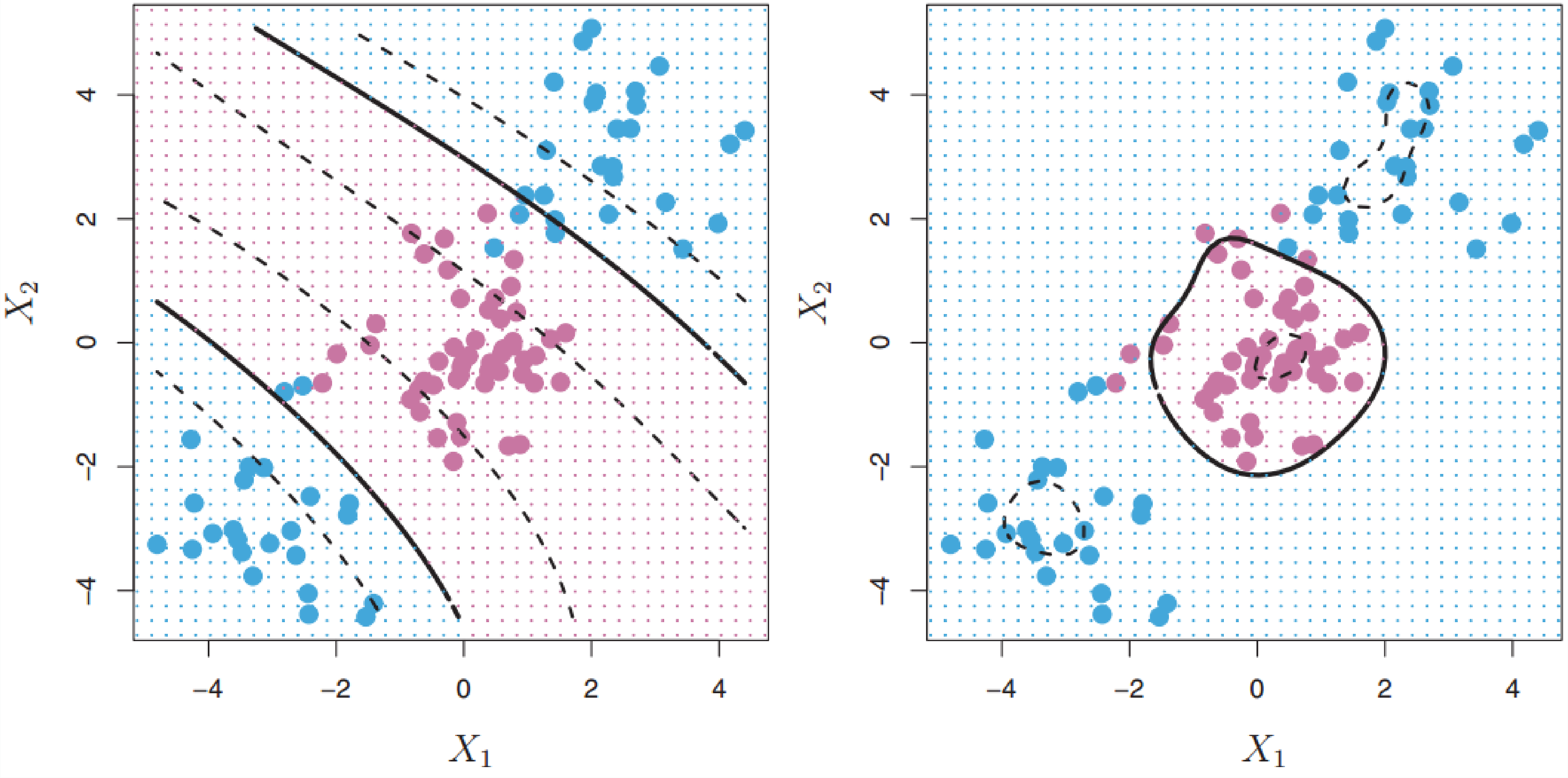
$$f(x) = \beta_0 + \sum_{i \in S}^{n} \alpha_i \langle x, x_i \rangle$$

Машина опорных векторов (SVM)

Заменим скалярное произведение ядерной функцией (ядром):

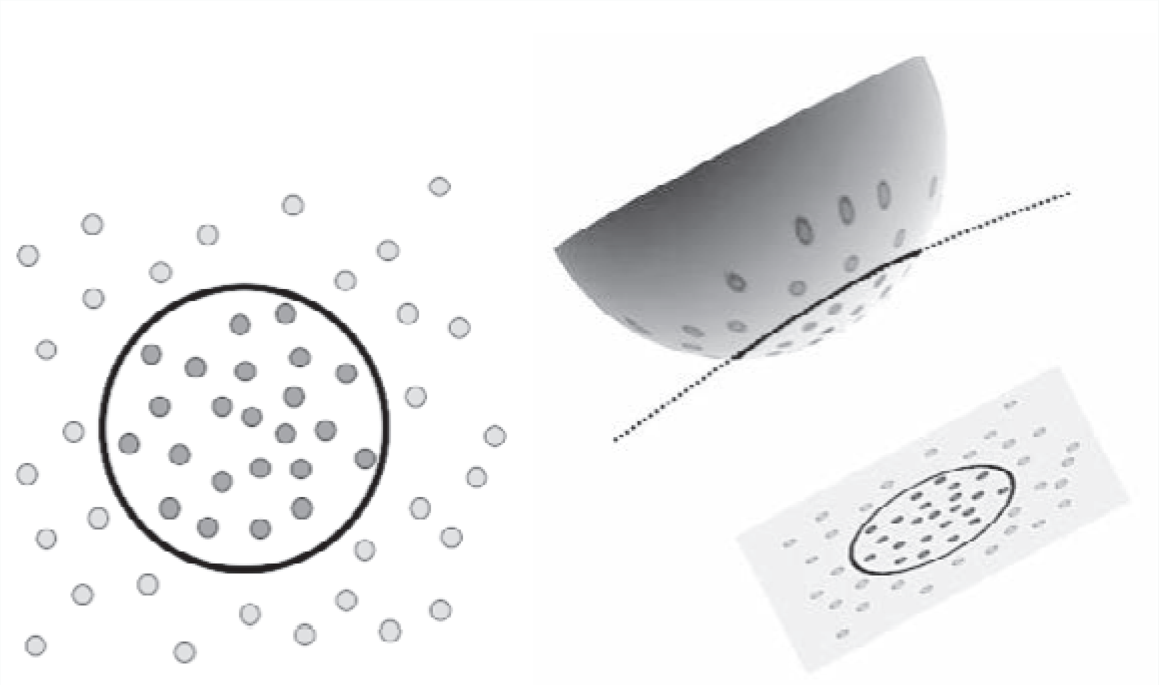
$$K(x_i, x_{i'})$$

- линейное ядро: $K(x_i, x_{i'}) = \sum_{j=1}^{p} x_{ij} x_{i'j}$
- полиномиальное: $K = \left(1 + \sum_{j=1}^{p} x_{ij} x_{i'j}\right)^d$
- радиальное: $K = \exp\left(-\gamma \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2\right)$, где $\gamma > 0$



Слева: SVM-классификатор на основе полиномиального ядра 3 степени
Справа: SVM-классификатор с радиальным ядром

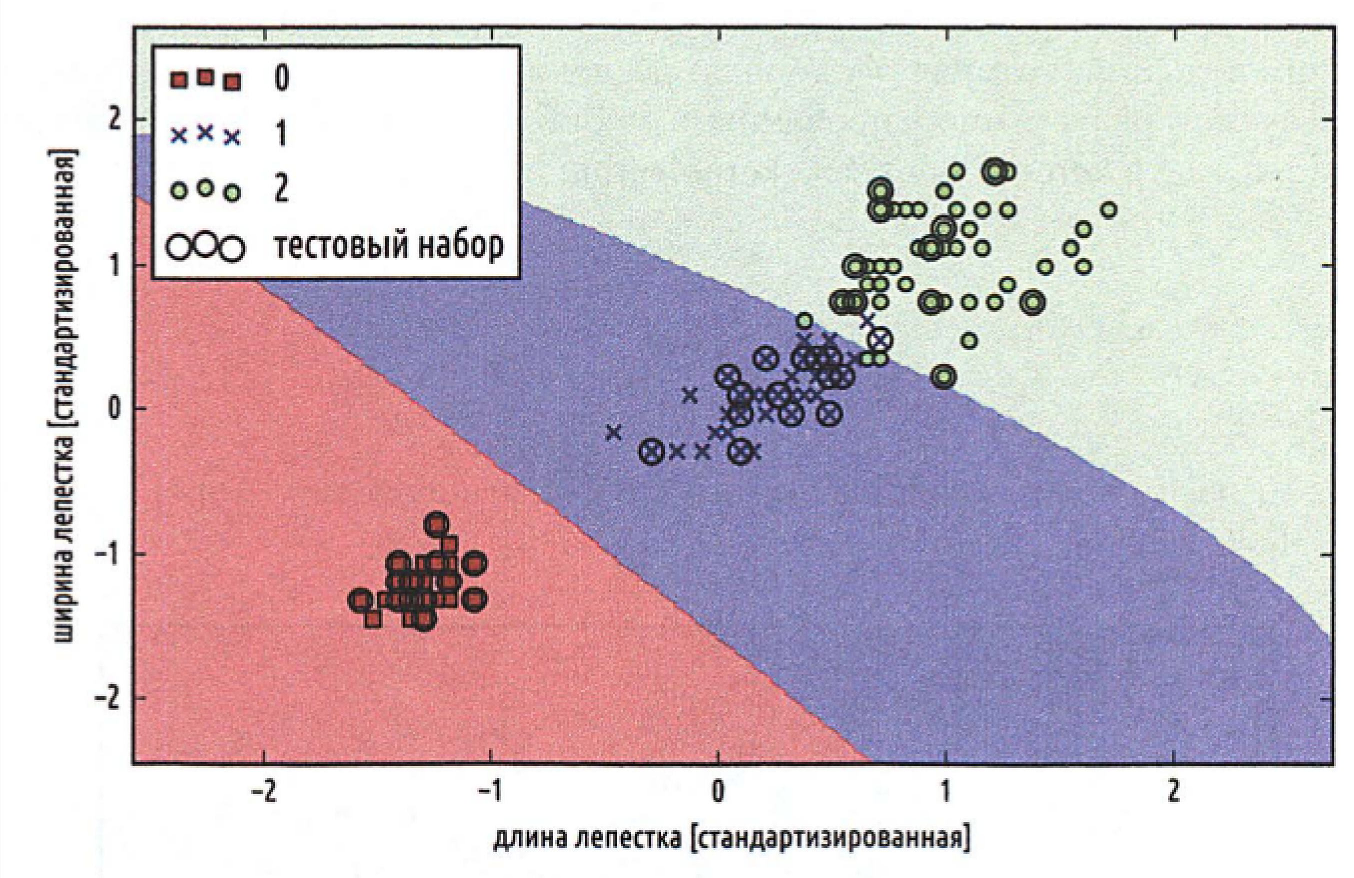
Иллюстрация роли ядерной функции



Ядерная функция проецирует наблюдения на пространство, в котором разделяющая граница между классами является прямой

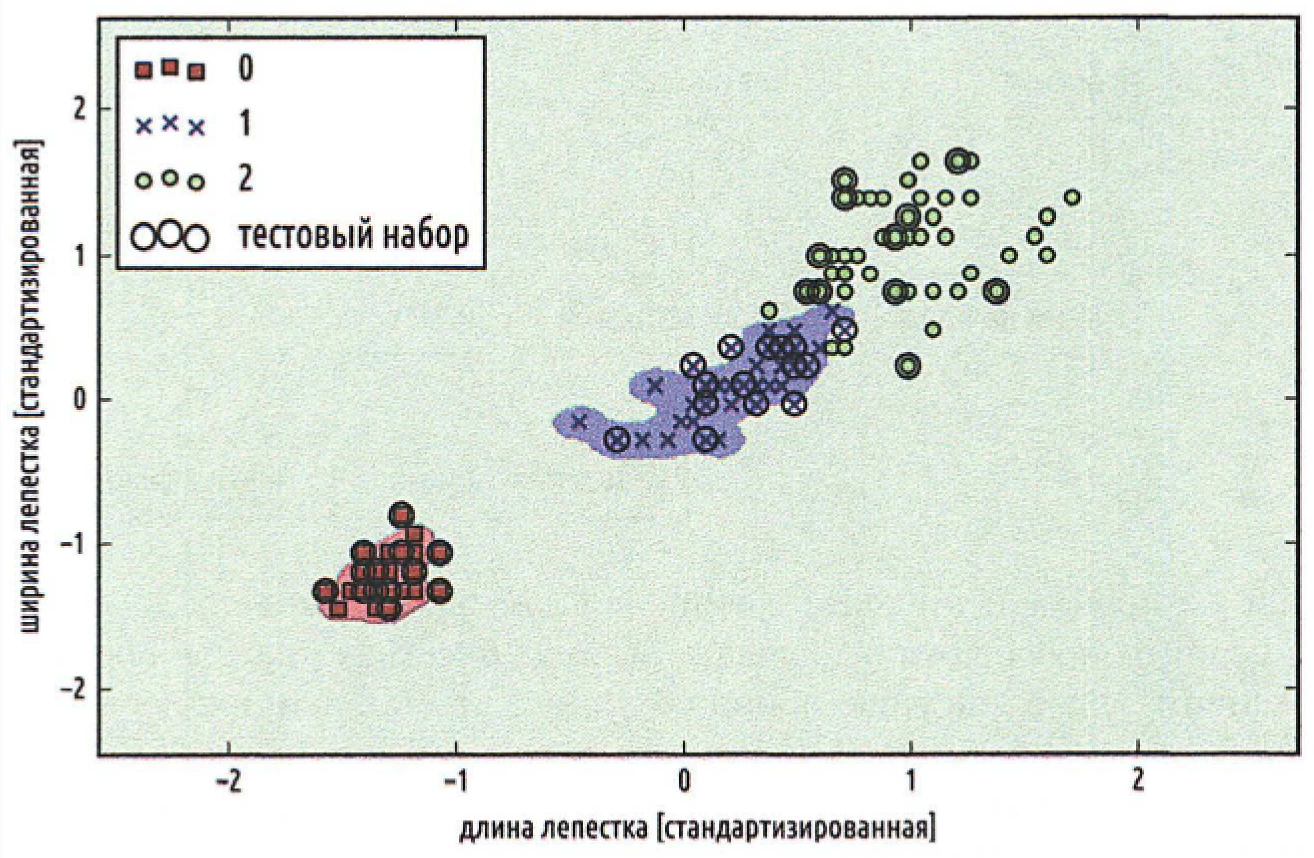
Данные iris
SVM с радиальной ядерной функцией

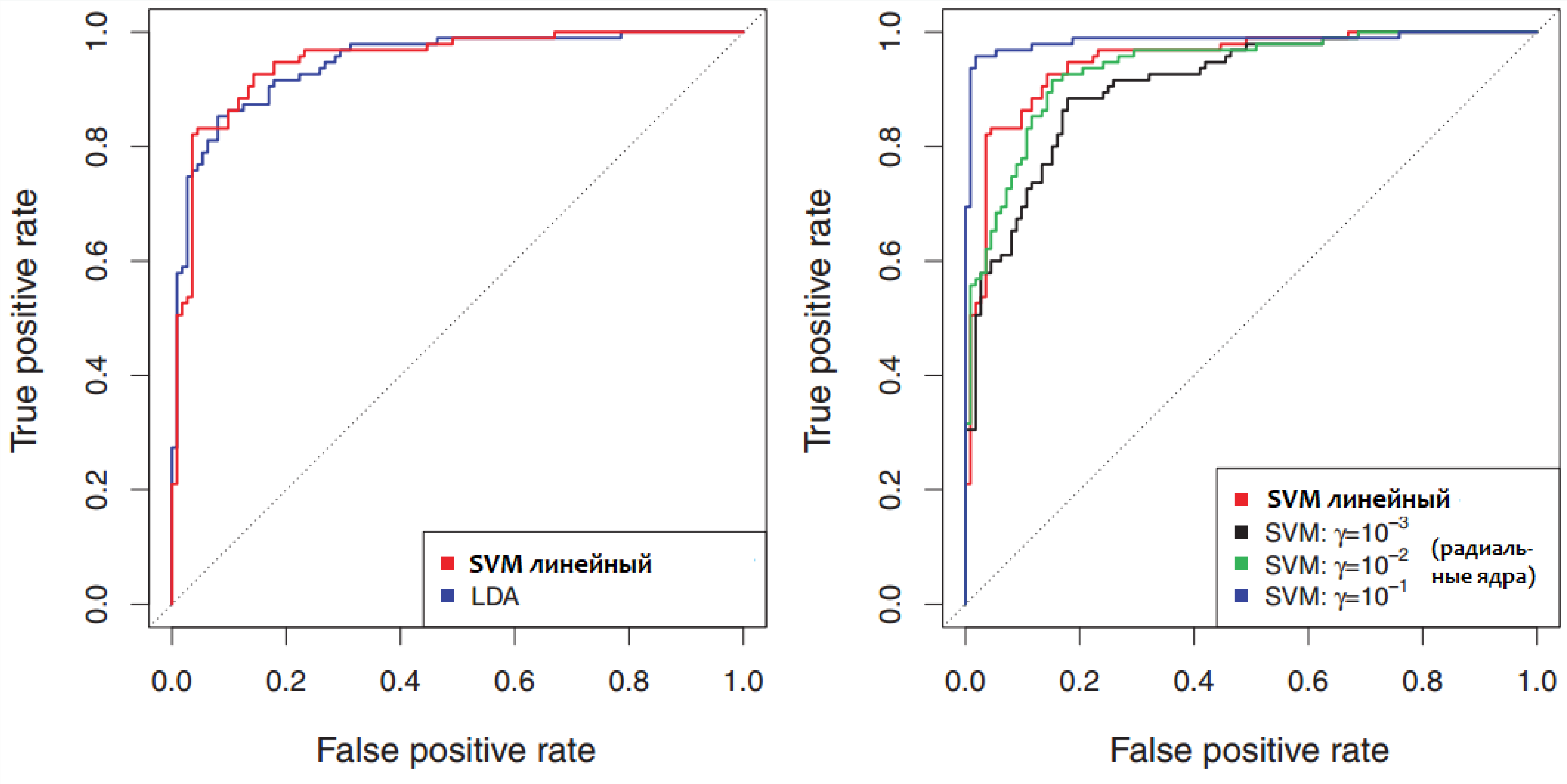
$\gamma = 0.2$ даёт мягкую границу между классами



Данные iris SVM с радиальной ядерной функцией,

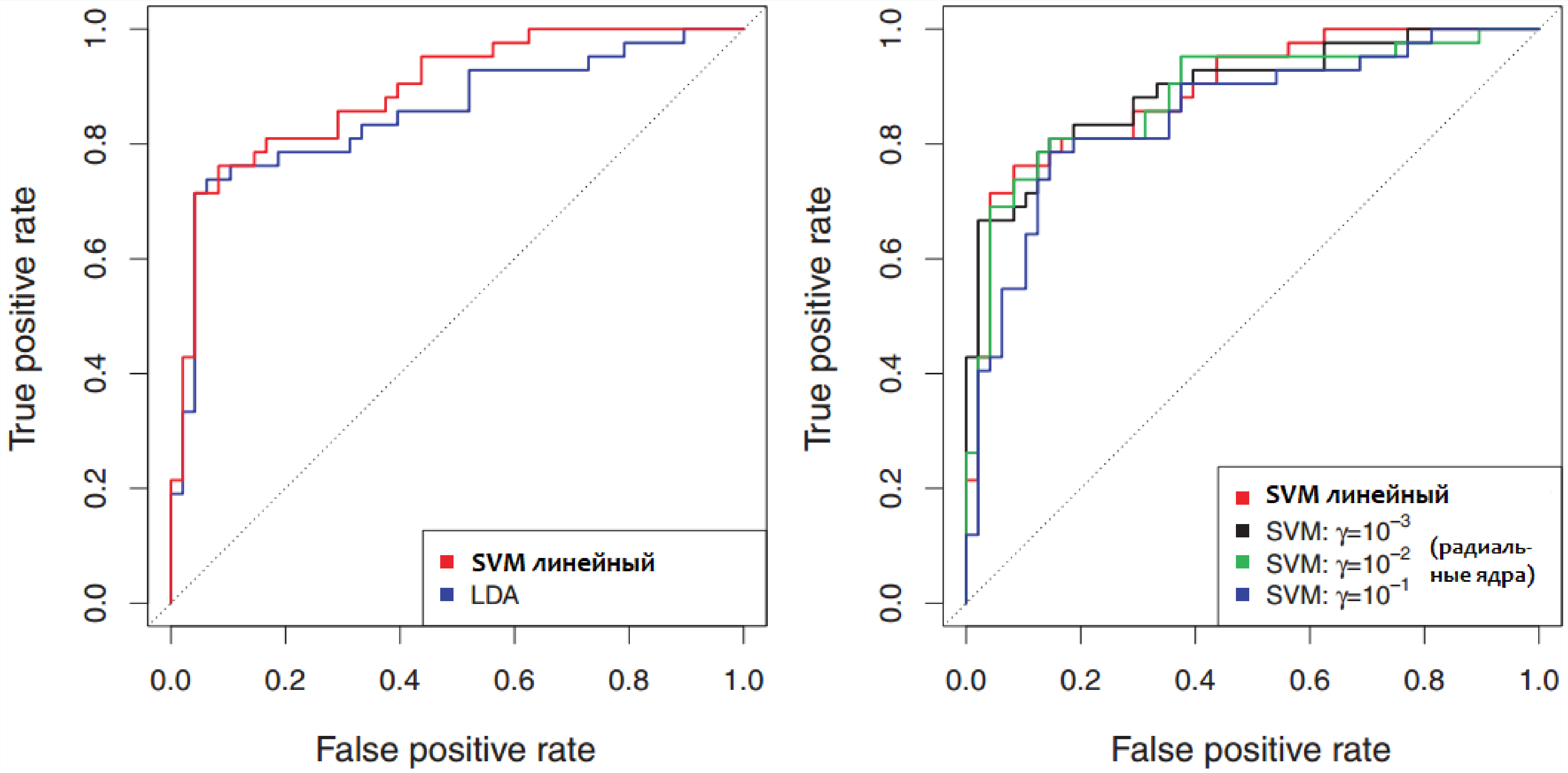
$\gamma = 100.0$ даёт компактную границу





Данные Heart, **обучающая выборка**

На обучающих данных качество модели линейного дискриминантного анализа (LDA) и SVM примерно одинаковое



Данные Heart, **тестовая выборка**

На тестовой выборке более гибкий SVM ведёт себя хуже, чем на обучающей

Сходства с другими методами

- Постановку задачи оптимизации SVM можно переписать в форме целевой функции со штрафом, близкой к гребневой регрессии
- SVM и логистическая регрессия ведут себя похожим образом на наблюдениях, далёких от решающей границы
- Исторически сложилось, что нелинейные ядра применяют в SVM, однако это возможно и в регрессионных методах

Ограничения SVM

- Гарантированное переобучение на малых наборах данных
- При сильном смешении классов результаты сомнительны, т.к. метод не даёт информации о вероятности ошибочной классификации для элемента данных
- Сложности с мультиклассовой классификацией (т.к. классификация происходит на две группы за раз)

SVM с несколькими классами

Наиболее популярные подходы:

- **один против одного** – строим $K(K-1)/2$ машин, сравнивая все классы попарно; наблюдение относится к классу, который чаще всего присваивался во всех парных сравнениях

- **один против всех** – строим K машин, каждый раз сравнивая класс K с остальными $K-1$ классами; наблюдение относится к классу с наибольшим значением ядерной функции

Сеточный поиск (grid search)

– настройка модели путём перебора параметров.

1. Выбрать оценку качества модели (AUC, AIC, $MSE_{\text{ТЕСТ}}$, $Acc_{\text{ТЕСТ}}$ и др.)

2. Выбрать алгоритм и определить, какие параметры нужно оптимизировать

3. Задать сетку: интервал изменения и шаг для каждого параметра

4. Для каждой комбинации настроечных параметров рассчитать оценку качества с перекрёстной проверкой

Источники

1. *Джеймс Г., Уиттон Д., Хасти Т., Тибширани Р.* Введение в статистическое обучение с примерами на языке R. Пер. с англ. С.Э. Мастицкого – М.: ДМК Пресс, **2016** – 450 с.

2. *Бринк Х., Ричардс Дж., Феверолф М.* Машинное обучение. – Спб.: Питер, **2018**. – 336 с.

3. *Анналин Ын, Кеннет Су* Теоретический минимум по Big Data. Всё, что нужно знать о больших данных. – Спб.: Питер, **2019**. – 208 с.

4. Данные Heart, iris.