IT-Academy

SUPPORT VECTOR MACHINE

КУХАЛЬСКИЙ НИКОЛАЙ ГЕННАДЬЕВИЧ

Вопросы занятия

1. Линейная и нелинейная классификация SVM

2. Регрессионная задача SVM.

3. Пример решения задачи регрессии через SVM: практика;

4. Пример решения задачи классификации через SVM: практика.

В конце занятия научимся:

- будете знать линейный алгоритм SVM;
- будем знать нелинейный алгоритм SVM;
- реализуете в коде задачу классификации и регрессии с помощью алгоритма SVM.

МЕТОД ОПОРНЫХ ВЕКТОРОМ (SVM)

IT-Academy

- мощная и довольно гибкая ML модель
- Поддерживает линейную и нелинейную классификацию, регрессию, поиск выбросов
- Лучше всего подходит для классификации сложных датасетов среднего или малого размера

МЕТОД ОПОРНЫХ ВЕКТОРОМ (SVM)

IT-Academy

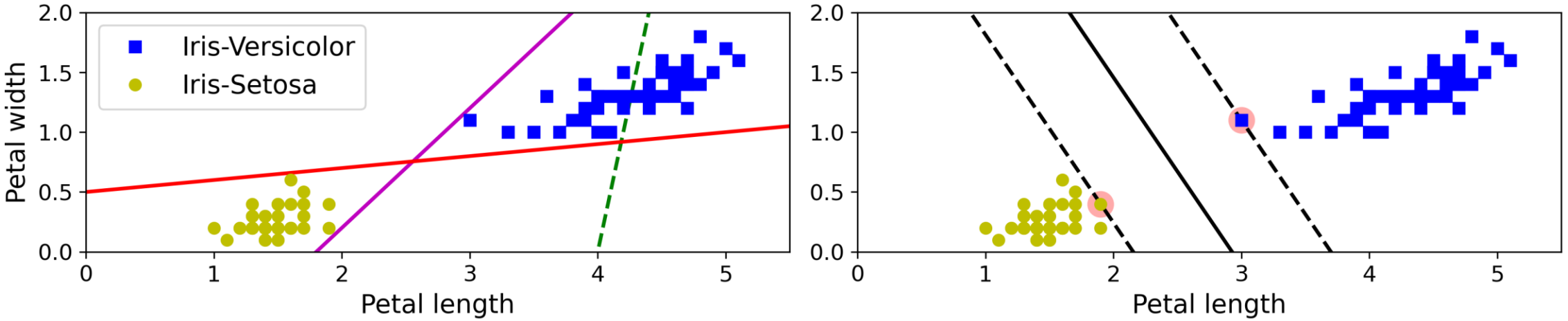
THE IRIS DATASET

- Атрибуты
  - Тип ириса (Setosa, Versicolour, Virginica)
  - Ширина и длина чашелистника (sepal)
  - Ширина и длина лепестка (petal)
- 150 экземпляров

МЕТОД ОПОРНЫХ ВЕКТОРОМ (SVM)

THE IRIS DATASET

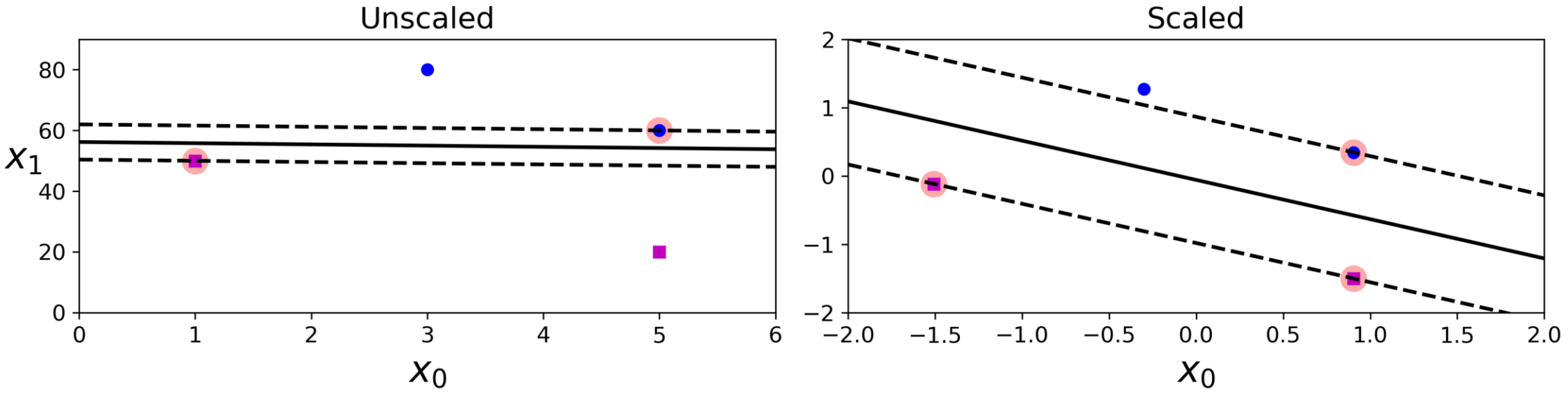
- Рассмотрим двухмерные линейно разделяемые данные
- Граница принятия решения: экземпляры двух классов были максимально удалены от границы
- Точки на границе - опорные вектора



МЕТОД ОПОРНЫХ ВЕКТОРОМ (SVM)

Методы SVM чувствительны к масштабам признаков, график слева имеет масштаб по вертикали, намного превышающий масштаб по горизонтали, поэтому самая широкая полоса близка к горизонтали.

После масштабирования признаков граница решений выглядит гораздо лучше.

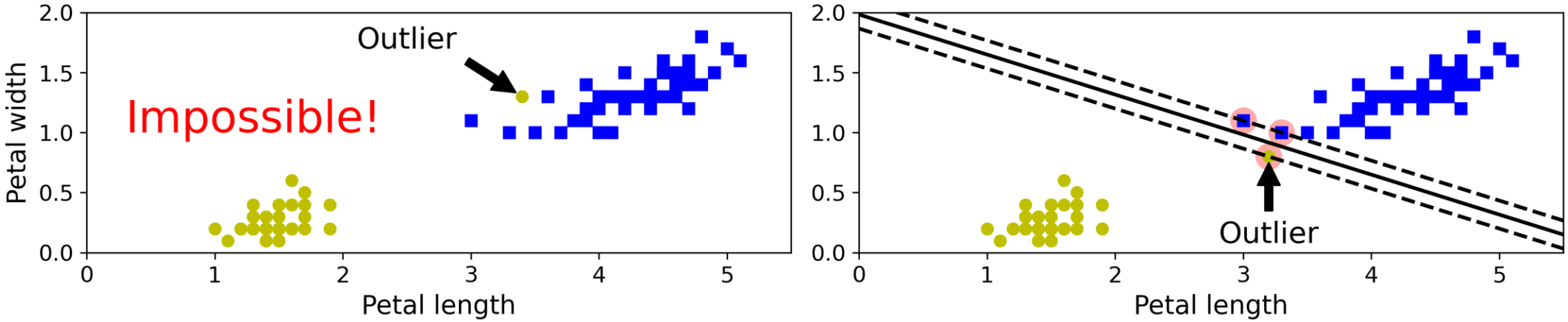


МЕТОД ОПОРНЫХ ВЕКТОРОМ (SVM)

HARD MARGIN CLASSIFICATION

- Экземпляры одного класса находятся по одну сторону от разделяющей поверхности

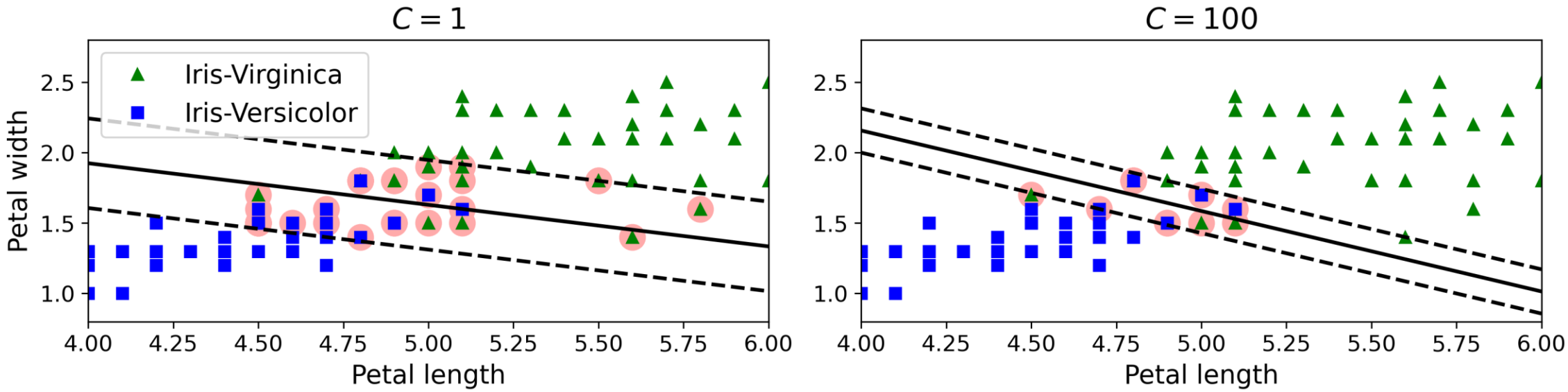
- Если в данных есть выбросы, то SVM не всегда находит оптимальную границу разделения



МЕТОД ОПОРНЫХ ВЕКТОРОМ (SVM)

SOFT MARGIN CLASSIFICATION

- Допускает нарушение границы разделения
- Допустимое нарушение регулируется параметром C



МЕТОД ОПОРНЫХ ВЕКТОРОМ (SVM)

IT-Academy

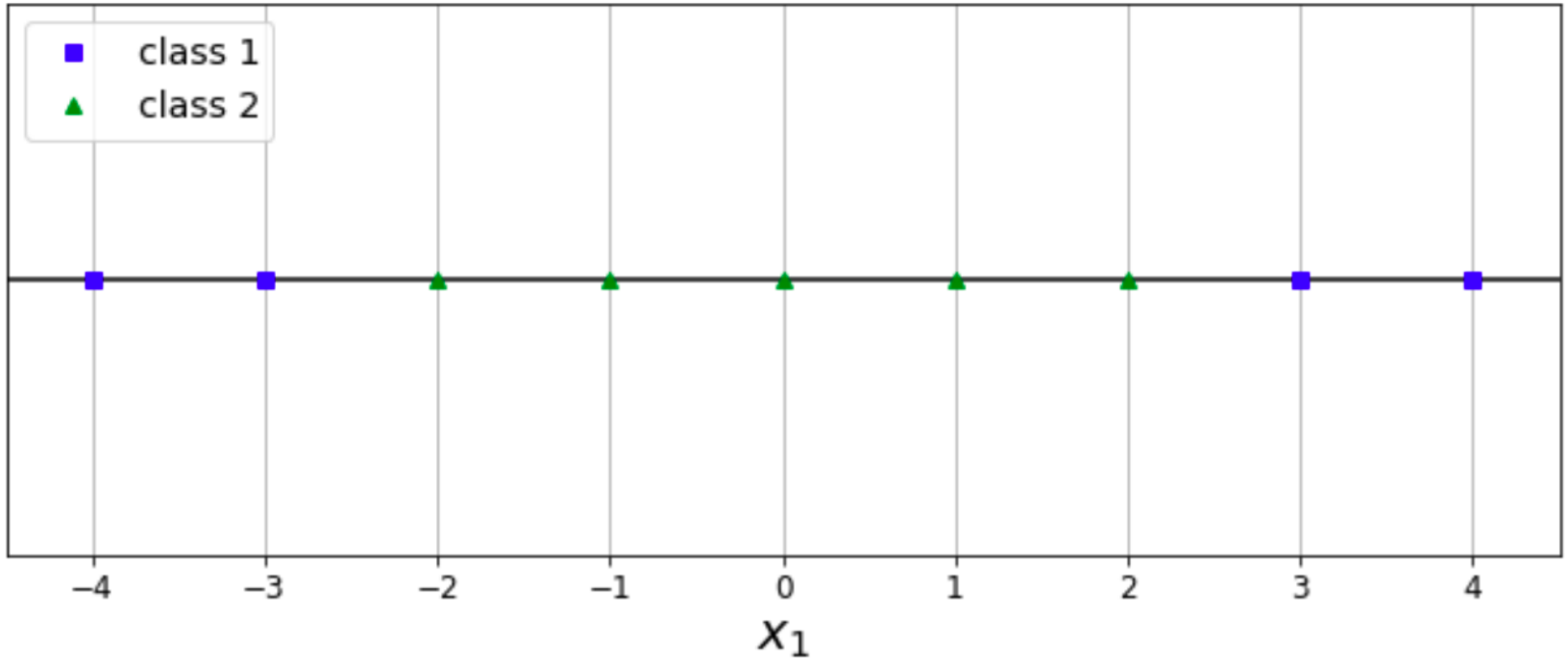
```
from sklearn.svm import LinearSVC
from sklearn.svm import SVC
from sklearn.linear_model import SGDClassifier
```

- LinearSVC в отличие от классификаторов, основанных на логистической регрессии, не выдает вероятности для каждого класса.
- SVC намного медленнее, особенно с крупными обучающими наборами данных.
- SGDClassifier - применяется стохастический градиентный спуск. Он не сходится настолько быстро, как класс LinearSVC, но может быть полезным для обработки гигантских наборов данных.

МЕТОД ОПОРНЫХ ВЕКТОРОМ (SVM)

НЕЛИНЕЙНЫЙ SVM

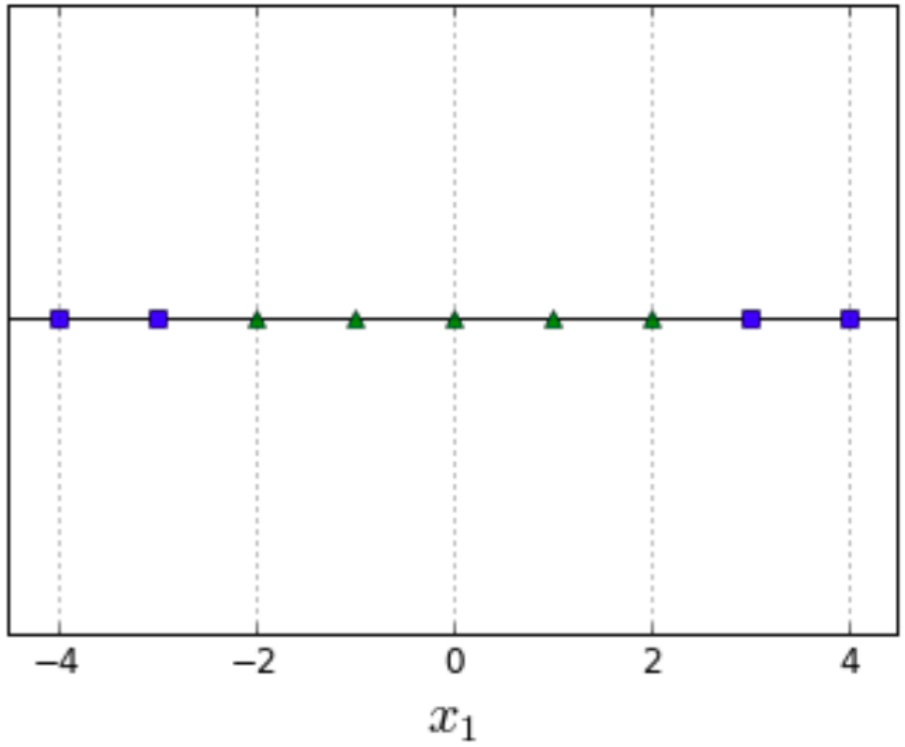
Данные не могут быть разделены линейно. Что делать?

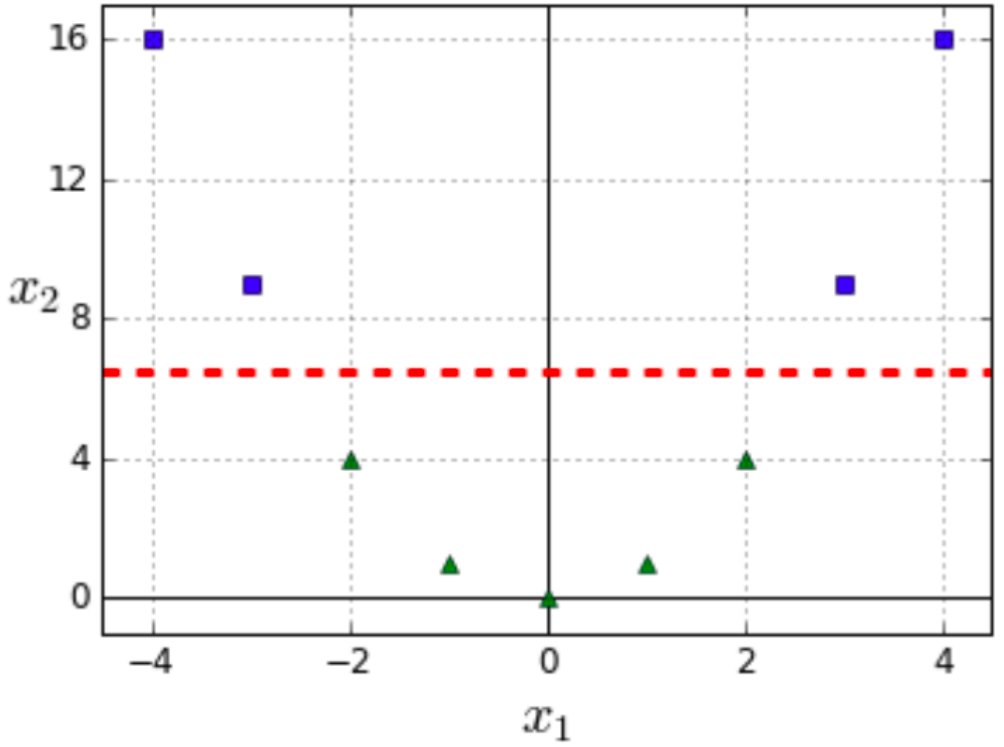


МЕТОД ОПОРНЫХ ВЕКТОРОМ (SVM)

НЕЛИНЕЙНЫЙ SVM

Добавим новую фичу $x_2 = x_1^2$

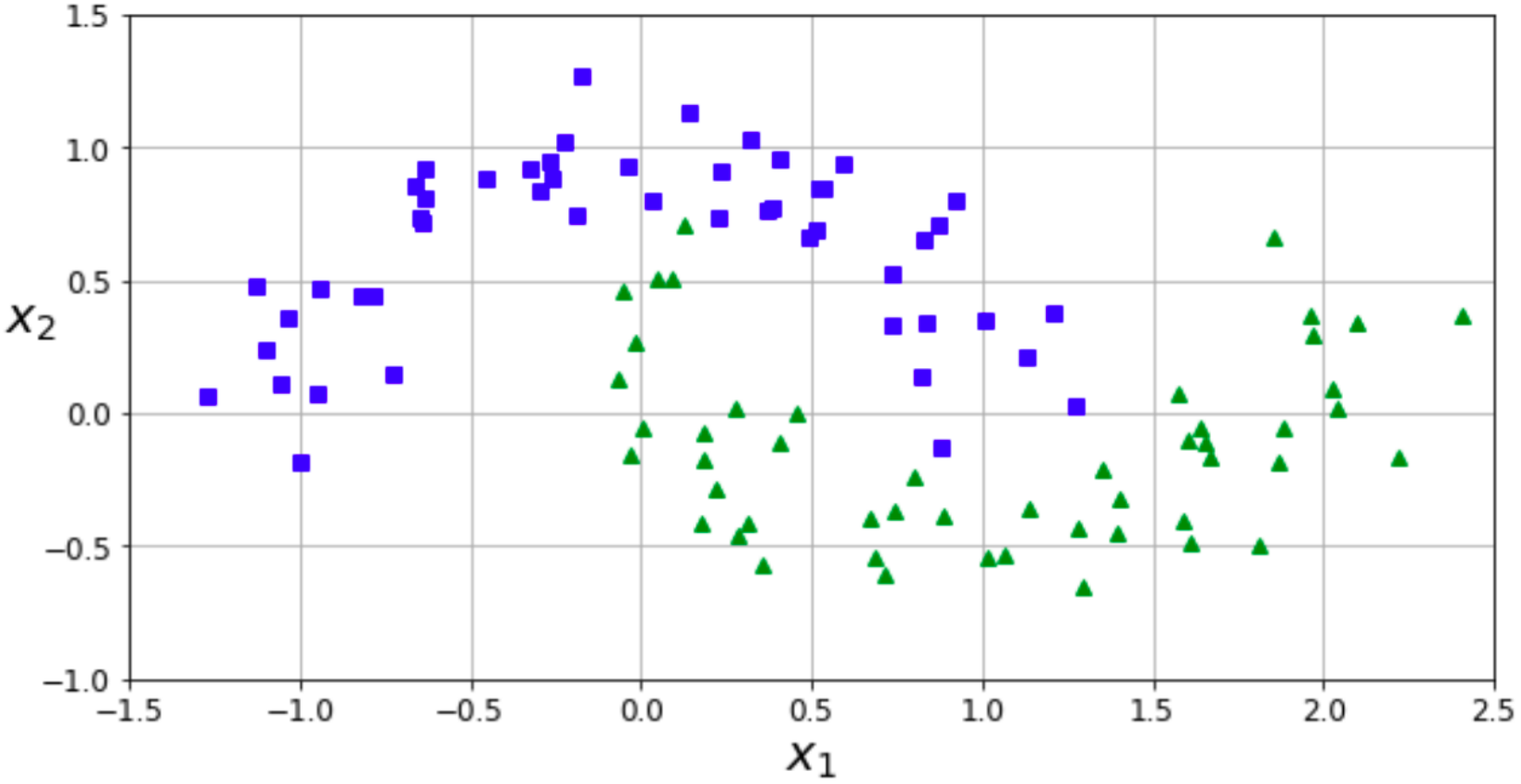




МЕТОД ОПОРНЫХ ВЕКТОРОМ (SVM)

НЕЛИНЕЙНЫЙ SVM

А теперь что делать?



МЕТОД ОПОРНЫХ ВЕКТОРОМ (SVM)

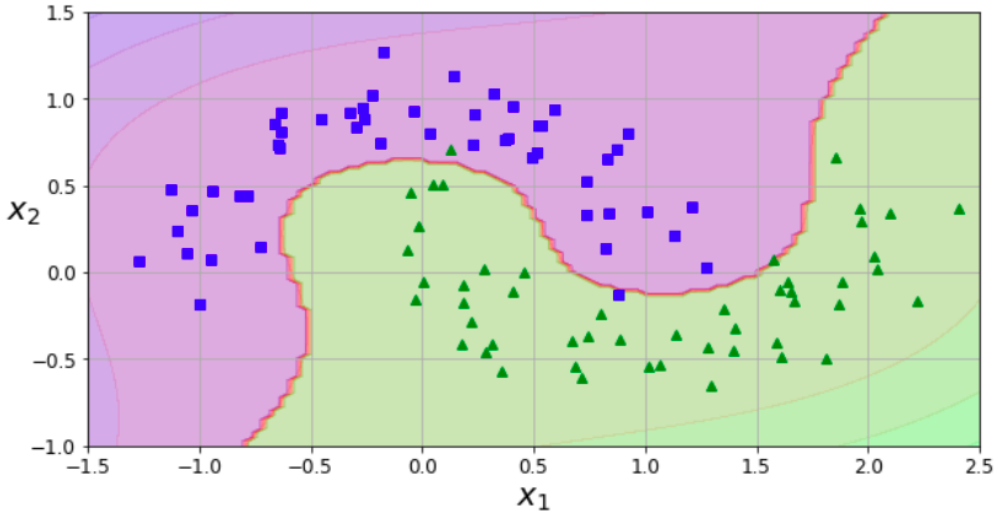
НЕЛИНЕЙНЫЙ SVM

```
from sklearn.preprocessing import PolynomialFeatures
```

- Полином третьей степени

$$[x_1, x_2] \Rightarrow [1, x_1, x_2, x_1x_2, x_1^2, x_2^2, x_1^2x_2, x_1x_2^2, x_1^3, x_2^3]$$

- Размерность X:
  - было (N, 2)
  - стало (N, 10)



МЕТОД ОПОРНЫХ ВЕКТОРОМ (SVM)

НЕЛИНЕЙНЫЙ SVM

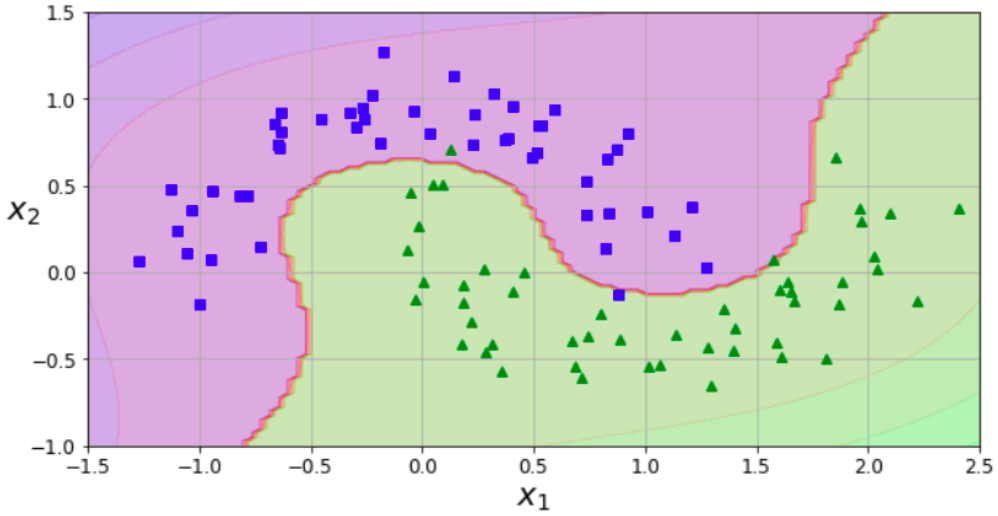
```
from sklearn.preprocessing import PolynomialFeatures
```

- Полином третьей степени

$$[x_1, x_2] \Rightarrow [1, x_1, x_2, x_1x_2, x_1^2, x_2^2, x_1^2x_2, x_1x_2^2, x_1^3, x_2^3]$$

- Размерность X:
  - было (N, 2)
  - стало (N, 10)

Что делать?



НЕЛИНЕЙНЫЙ SVM

Проблема:

- Большая полиномиальная степень ⇒ большое количество фич;
- Много фич:
  - Нужно больше памяти;
  - Модель медленная.

Хотелось бы:

- Математическое преобразование, которое позволит получить аналогичные результаты без “физического” создания фич.

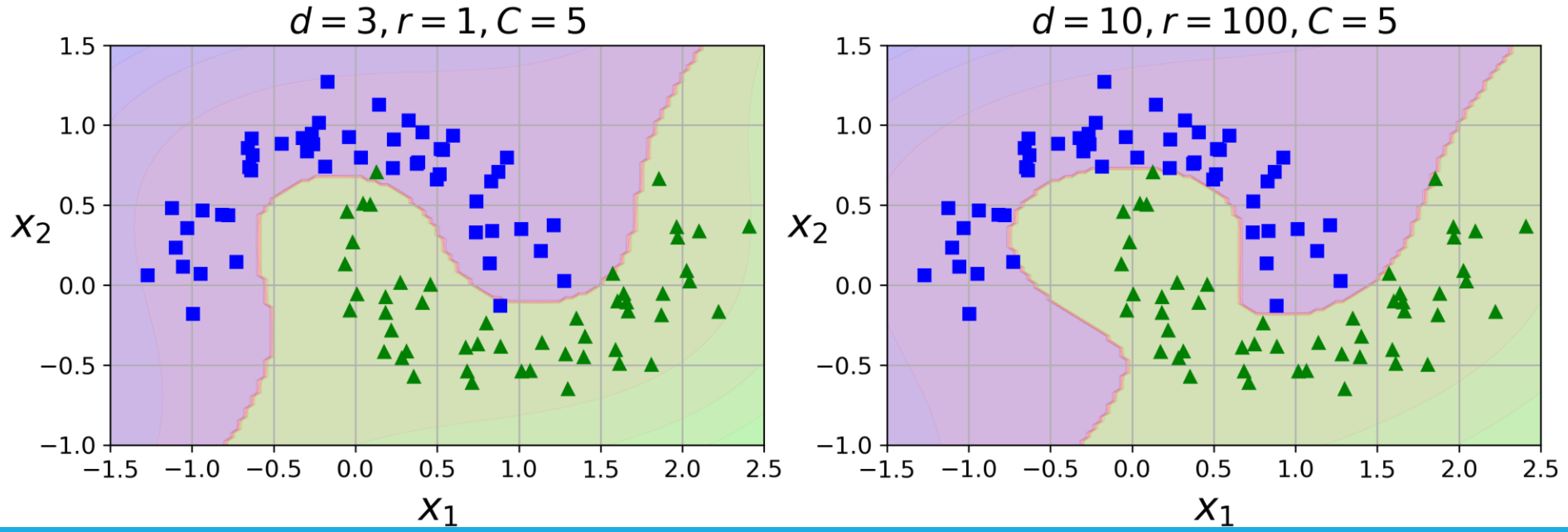
МЕТОД ОПОРНЫХ ВЕКТОРОМ (SVM)

НЕЛИНЕЙНЫЙ SVM

kernel trick

Позволяет получить тот же самый результат, как если бы вы добавили много полиномиальных признаков, даже при полиномах очень высокой степени, без фактического их добавления.

SVC(kernel="poly", degree=3 , coef0=1 , C=5)



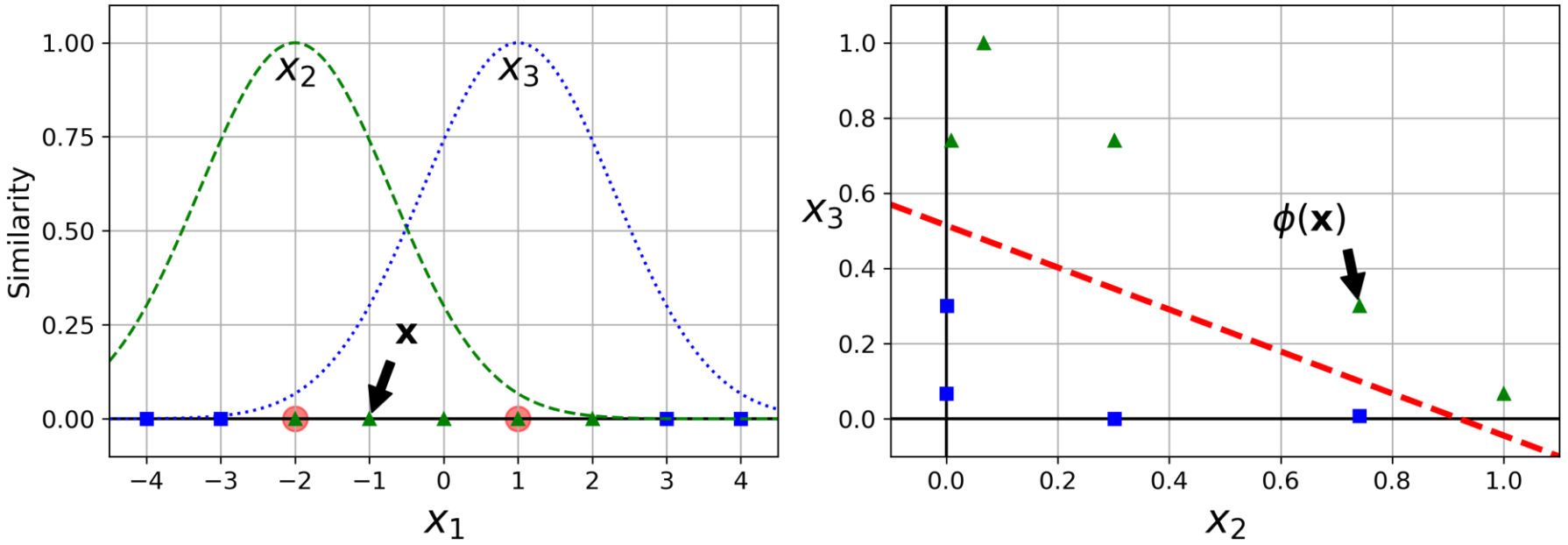
МЕТОД ОПОРНЫХ ВЕКТОРОМ (SVM)

НЕЛИНЕЙНЫЙ SVM

Добавление признаков близости

Функция близости - гауссова радиальная базисная функция (*Radial Basis Function - RBF*)

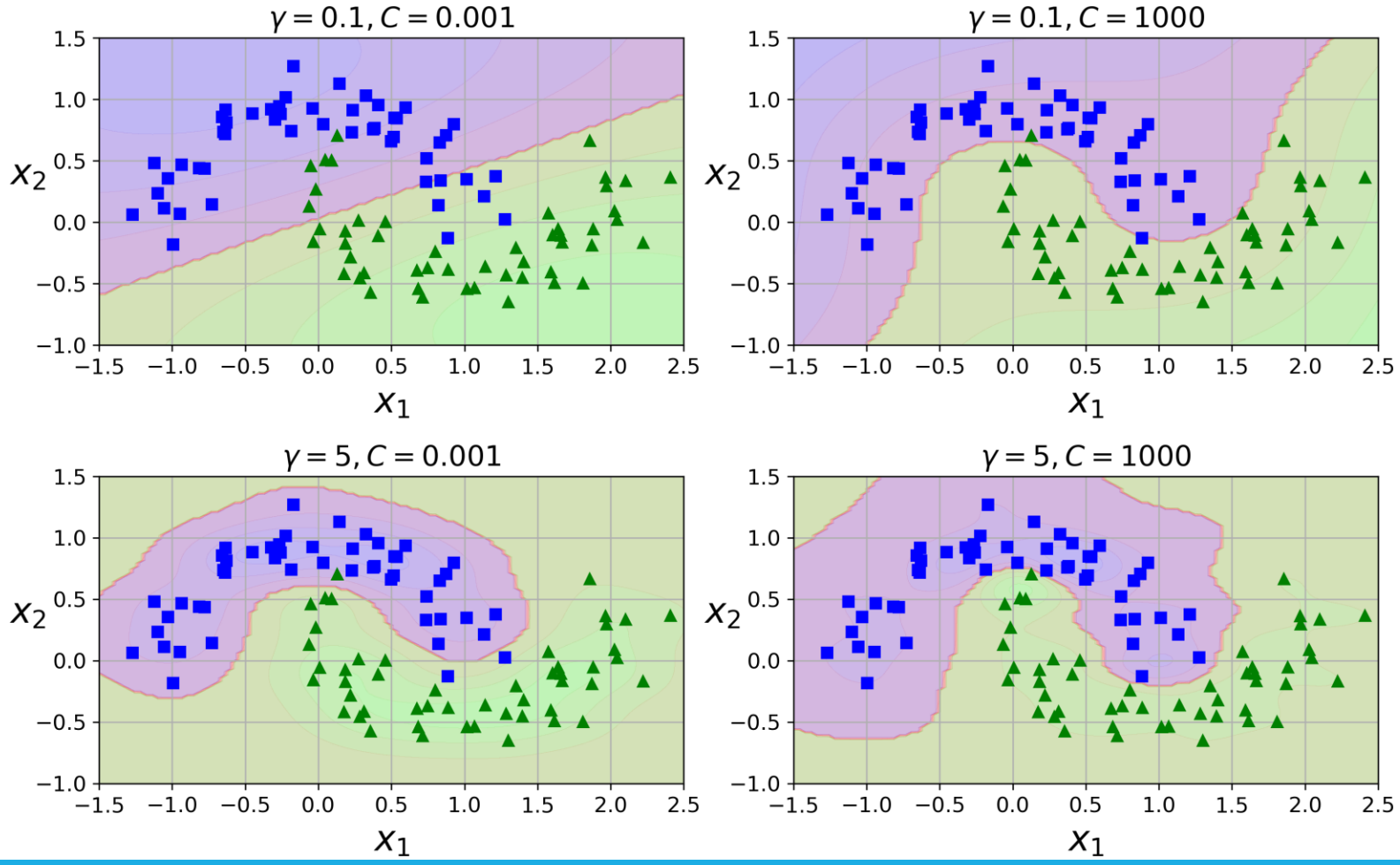
$$\phi_\gamma(\mathbf{x}, \ell) = \exp\left(-\gamma \| \mathbf{x} - \ell \|^2\right)$$



МЕТОД ОПОРНЫХ ВЕКТОРОМ (SVM)

Гауссово ядро RBF

SVC(kernel="rbf", gamma=5 , C=0.001)

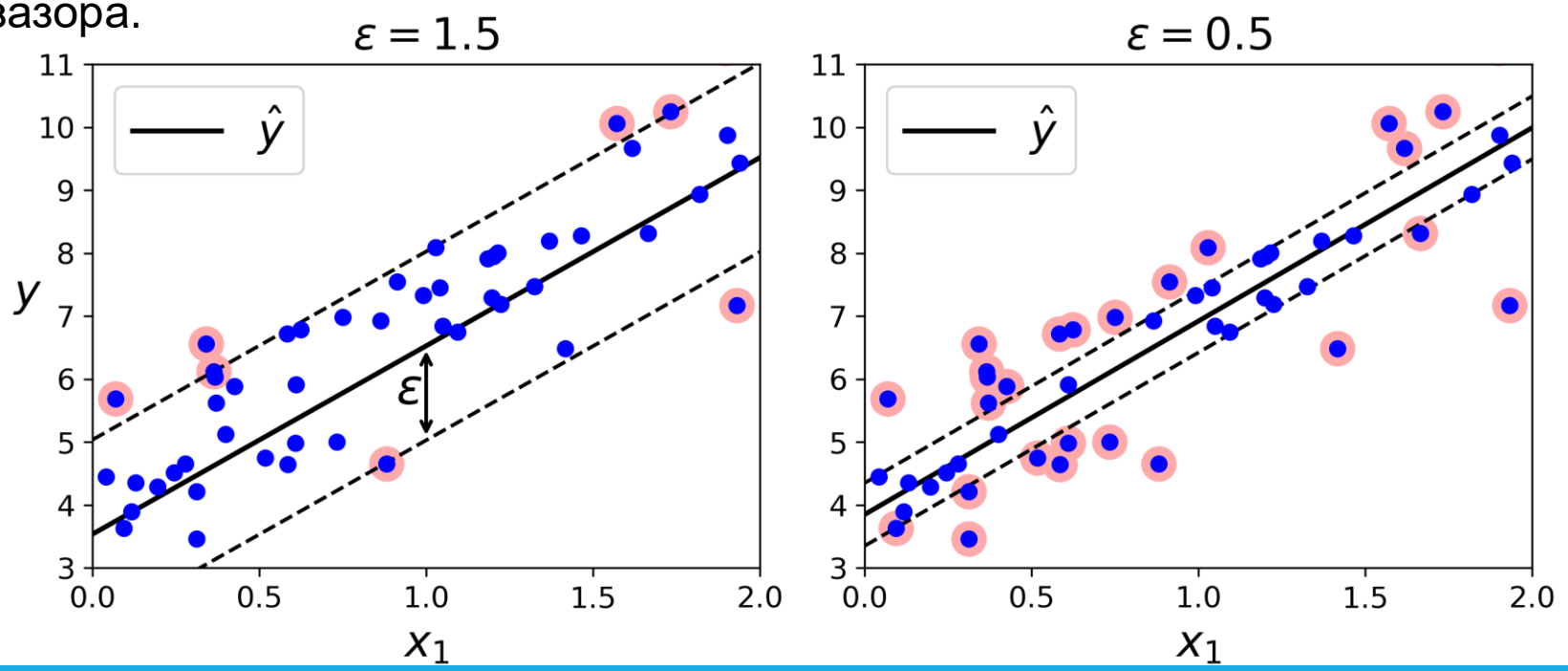


МЕТОД ОПОРНЫХ ВЕКТОРОМ (SVM)

Регрессия SVM

```
from sklearn.svm import LinearSVR
from sklearn.svm import SVR
```

Прием заключается в инвертировании цели: вместо попытки приспособиться к самой широкой из возможных полосе между двумя классами, одновременно ограничивая нарушения зазора, регрессия SVM пробует уместить как можно больше образцов на полосе наряду с ограничением нарушений зазора.

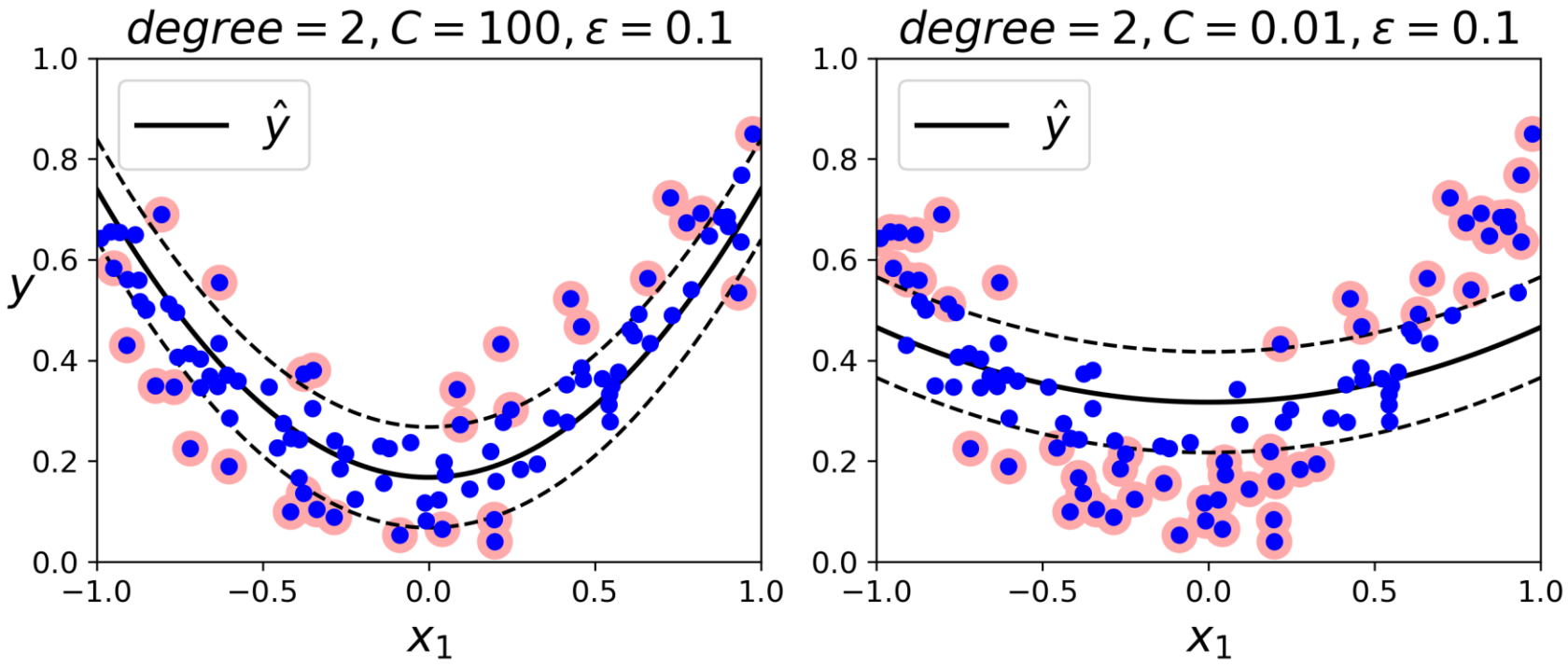


МЕТОД ОПОРНЫХ ВЕКТОРОМ (SVM)

Регрессия SVM

Для решения задач нелинейной регрессии можно применять *параметрически редуцированную (kernelized) модель SVM.*

SVR(kernel="poly", degree=2 , epsilon=0.1 , C=100)



IT-Academy

ПРАКТИКА

Dataframe : cars

IT-Academy

ПРАКТИКА

Dataframe : affair_data

IT-Academy

ПРАКТИКА

Dataframe : Shelter

IT-Academy

SUPPORT VECTOR MACHINE

КУХАЛЬСКИЙ НИКОЛАЙ ГЕННАДЬЕВИЧ