Метрическая классификация

Метод ближайших соседей

Skillbox образовательная платформа

Метод k ближайших соседей (kNN — k nearest neighbours)

Классифицируемый объект относится к тому классу, которому принадлежат ближайшие к нему объекты обучающей выборки.

Метод k ближайших соседей (kNN — k nearest neighbours)

Алгоритм:

- вычислить расстояние до каждого из объектов обучающей выборки
- отобрать k объектов обучающей выборки, расстояние до которых минимально
- класс классифицируемого объекта — это класс, наиболее часто встречающийся среди k ближайших соседей

Пример: анализ брака заводской детали

Номер объекта	Вес детали	Диаметр детали	Класс
1	7	7	брак
2	6	4	брак
3	3	4	не брак
4	1	4	не брак
5	4	5	брак

Пусть есть новый объект — деталь весом = 3, диаметром = 5.
Наша задача — определить, к какому классу он относится.

Пример: анализ брака заводской детали

Номер объекта	Вес детали	Диаметр детали	p
1	7	7	$\sqrt{(7-3)^2 + (7-5)^2} = \sqrt{20} = 4{,}47$
2	6	4	$\sqrt{(6-3)^2 + (4-5)^2} = \sqrt{10} = 3{,}16$
3	3	4	$\sqrt{(3-3)^2 + (4-5)^2} = \sqrt{1} = 1{,}0$
4	1	4	$\sqrt{(1-3)^2 + (4-5)^2} = \sqrt{5} = 2{,}24$
5	4	5	$\sqrt{(4-3)^2 + (5-5)^2} = \sqrt{1} = 1{,}00$

Пример: анализ брака заводской детали

Номер объекта	Вес детали	Диаметр детали	p	Порядковый номер при сортировке
1	7	7	4,47	5
2	6	4	3,16	4
3	3	4	1,00	1
4	1	4	2,24	3
5	4	5	1,00	2

Пример: анализ брака заводской детали

Номер объекта	Вес детали	Диаметр детали	ρ	Порядковый номер при сортировке	Входит в 3 ближайших соседа?
1	7	7	4,47	5	нет
2	6	4	3,16	4	нет
3	3	4	1,00	1	да
4	1	4	2,24	3	да
5	4	5	1,00	2	да

K = 3

Пример: анализ брака заводской детали

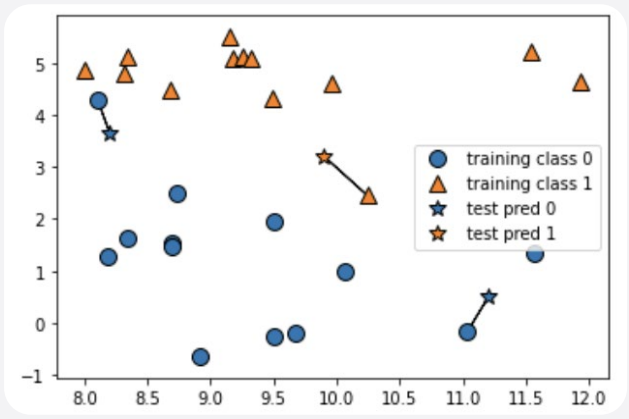
Номер объекта	Вес детали	Диаметр детали	ρ	Порядковый номер при сортировке	Входит в 3 ближайших соседа?	Класс объекта
1	7	7	4,47	5	нет	-
2	6	4	3,16	4	нет	-
3	3	4	1,00	1	да	не брак
4	1	4	2,24	3	да	не брак
5	4	5	1,00	1	да	брак

Объект (3,7) принадлежит классу «не брак»

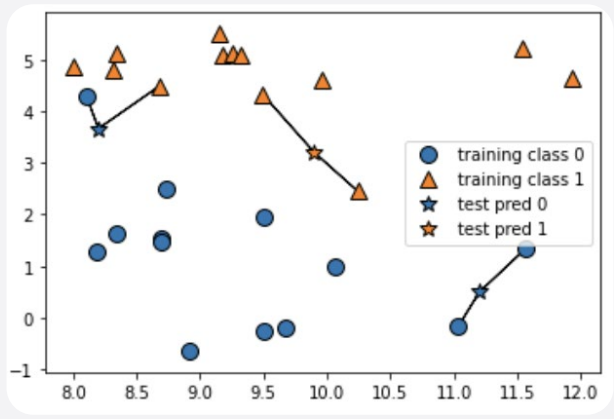
Пример: анализ брака заводской детали

Номер объекта	Вес детали	Диаметр детали	p	Порядковый номер при сортировке	Класс
1	7	7	4,47	5	брак
2	6	4	3,16	4	брак
3	3	4	1,00	1	не брак
4	1	4	2,24	3	не брак
5	4	5	1,00	2	брак

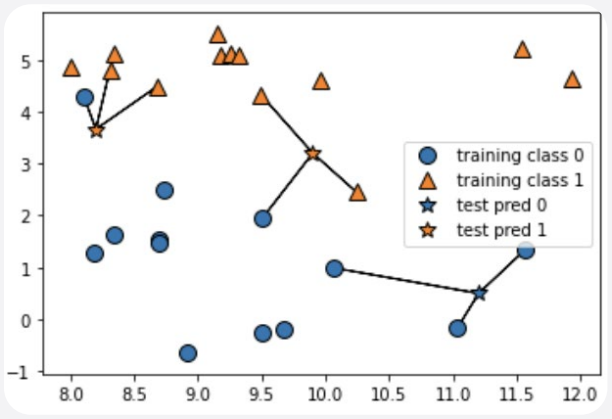
k=1



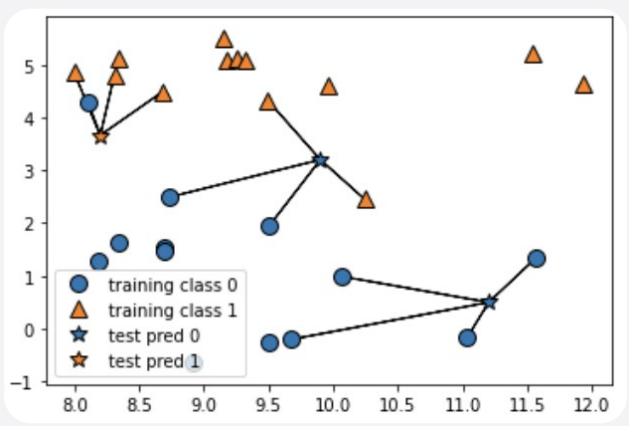
k=2



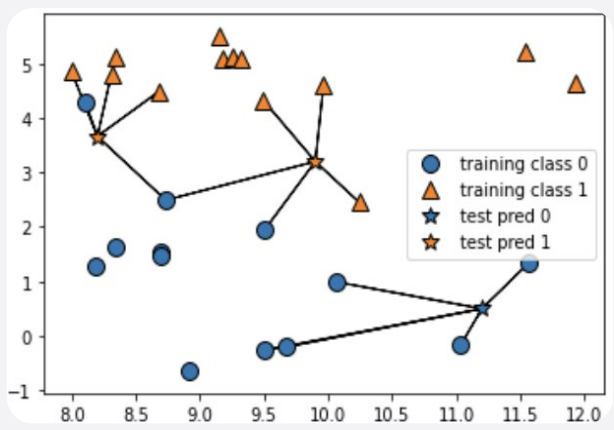
k=3



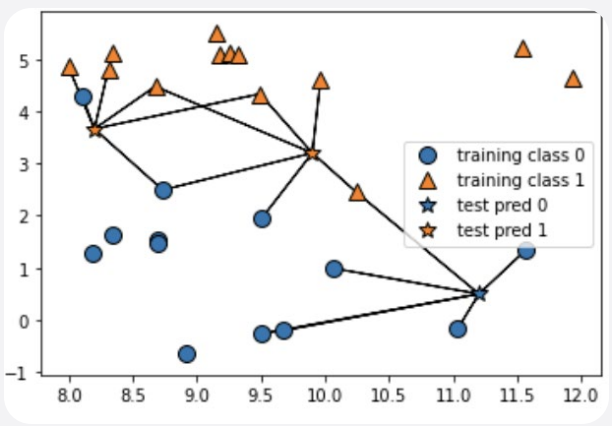
k=4



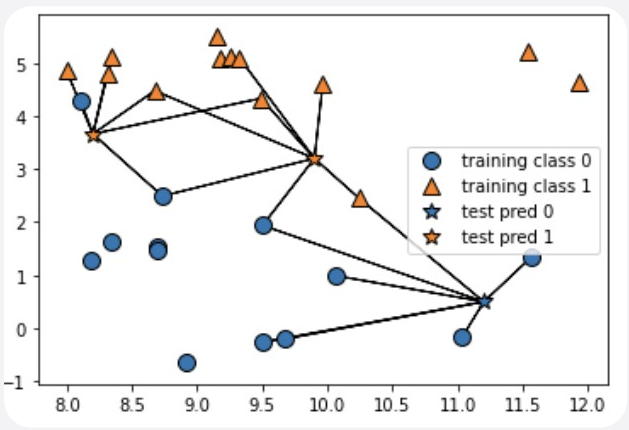
k=5



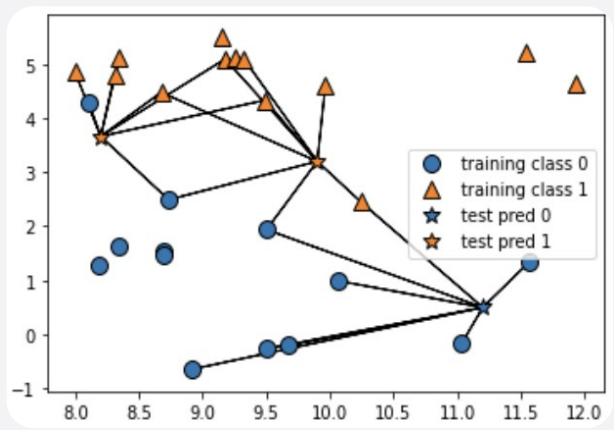
k=6



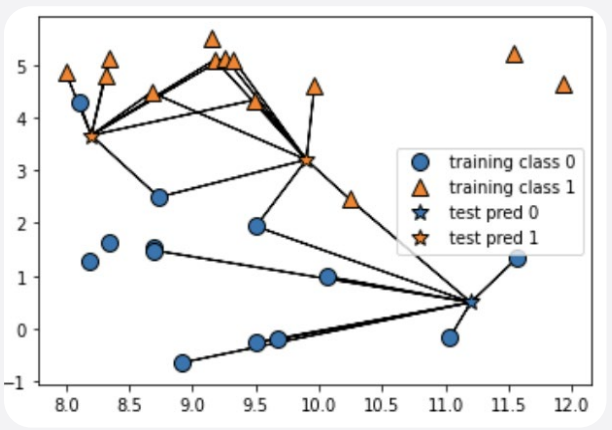
k=7



k=8

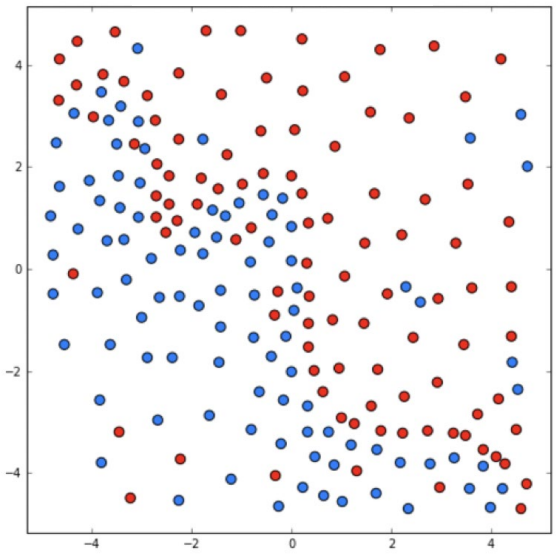


k=9

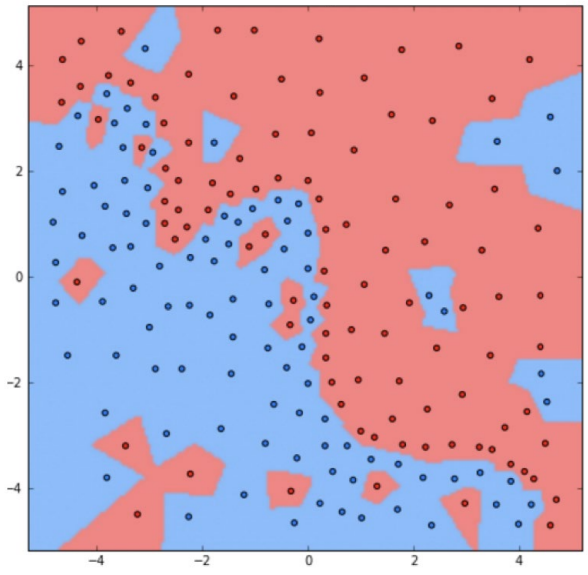


Пример классификации

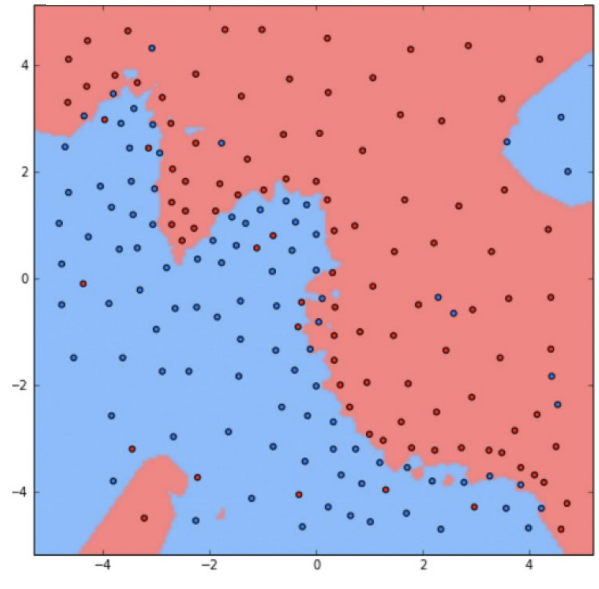
Sample dataset



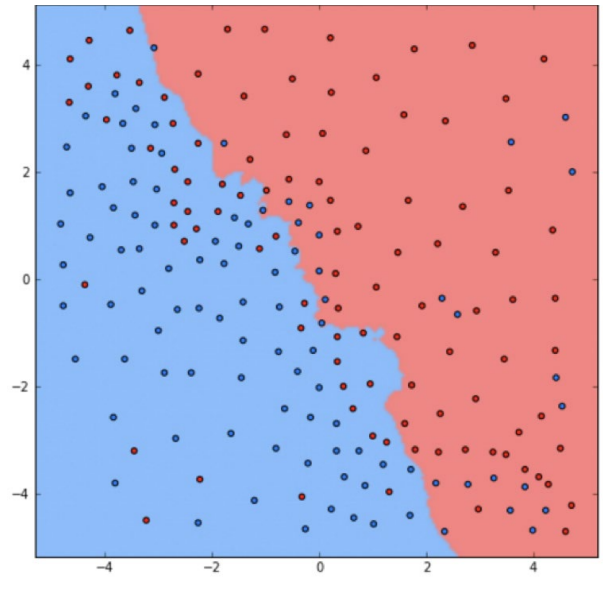
K-NN decision regions (K = 1)



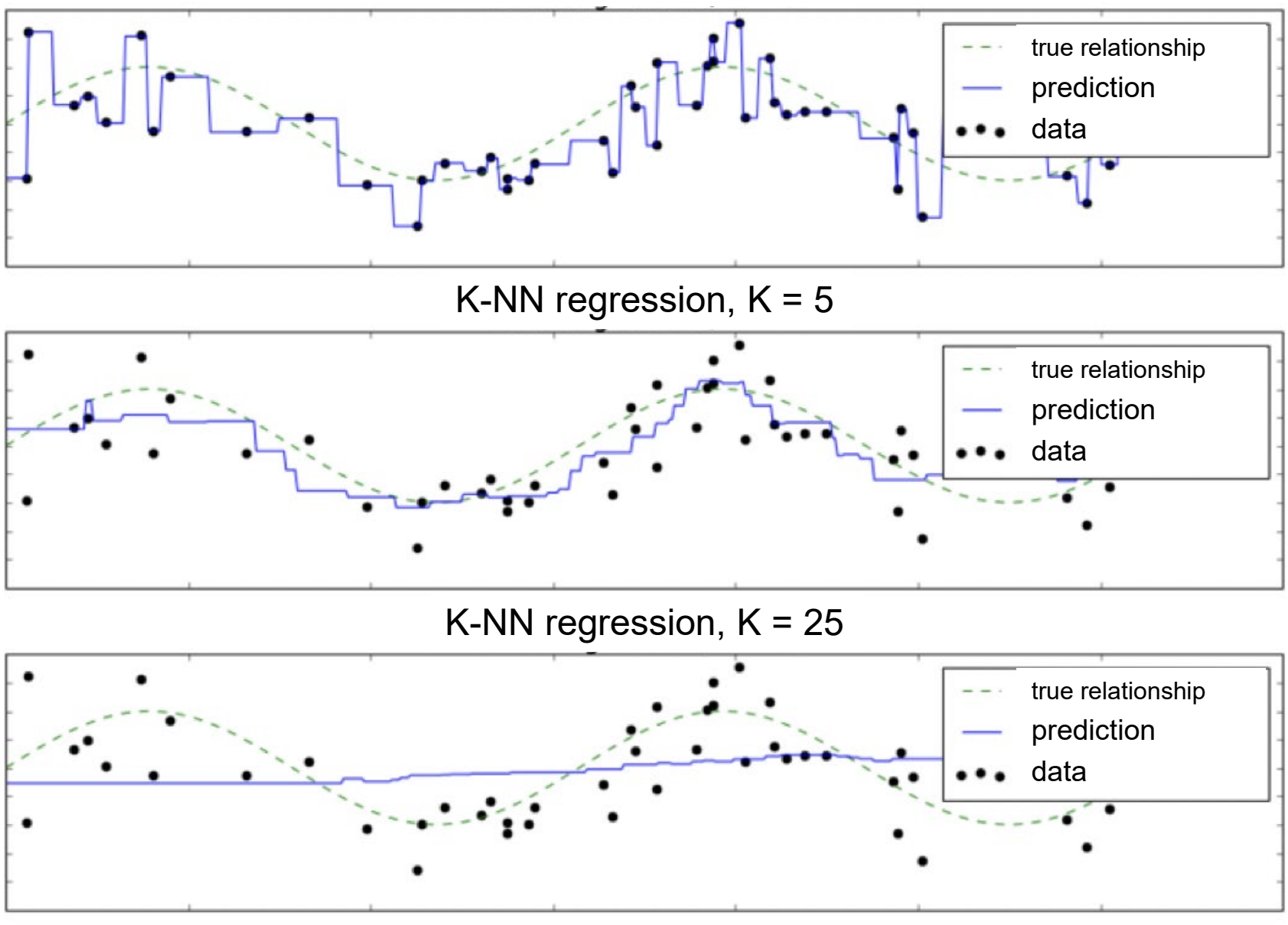
K-NN decision regions (K = 5)



K-NN decision regions (K = 100)



Пример регрессии



leave-one-out

Контроль leave-one-out — это проверка обобщающей способности алгоритма.

Модель, качество

Исходные данные

Обучающая выборка

Тестовая выборка

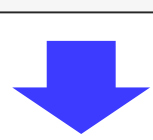
60–70 %

40–30 %

Модель

Качество

Исходные данные N объектов



Обучающая выборка

Тестовая выборка

N-1 объект

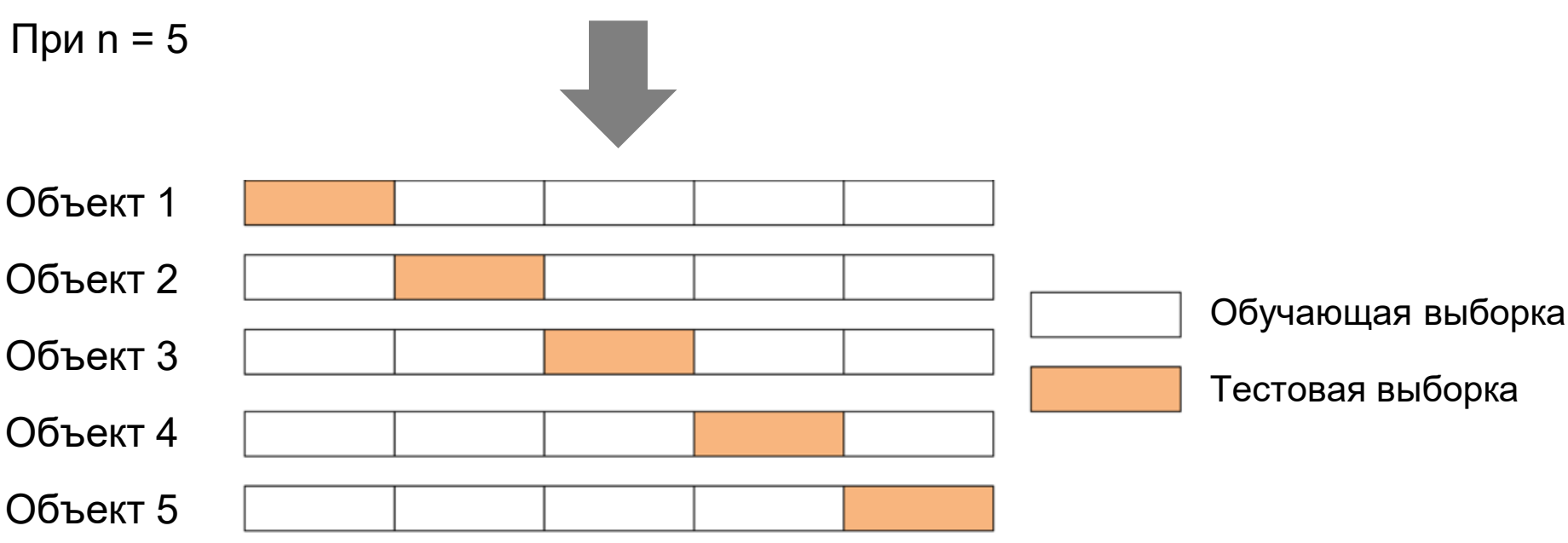
1 объект

Модель

Качество

Исходные данные N объектов

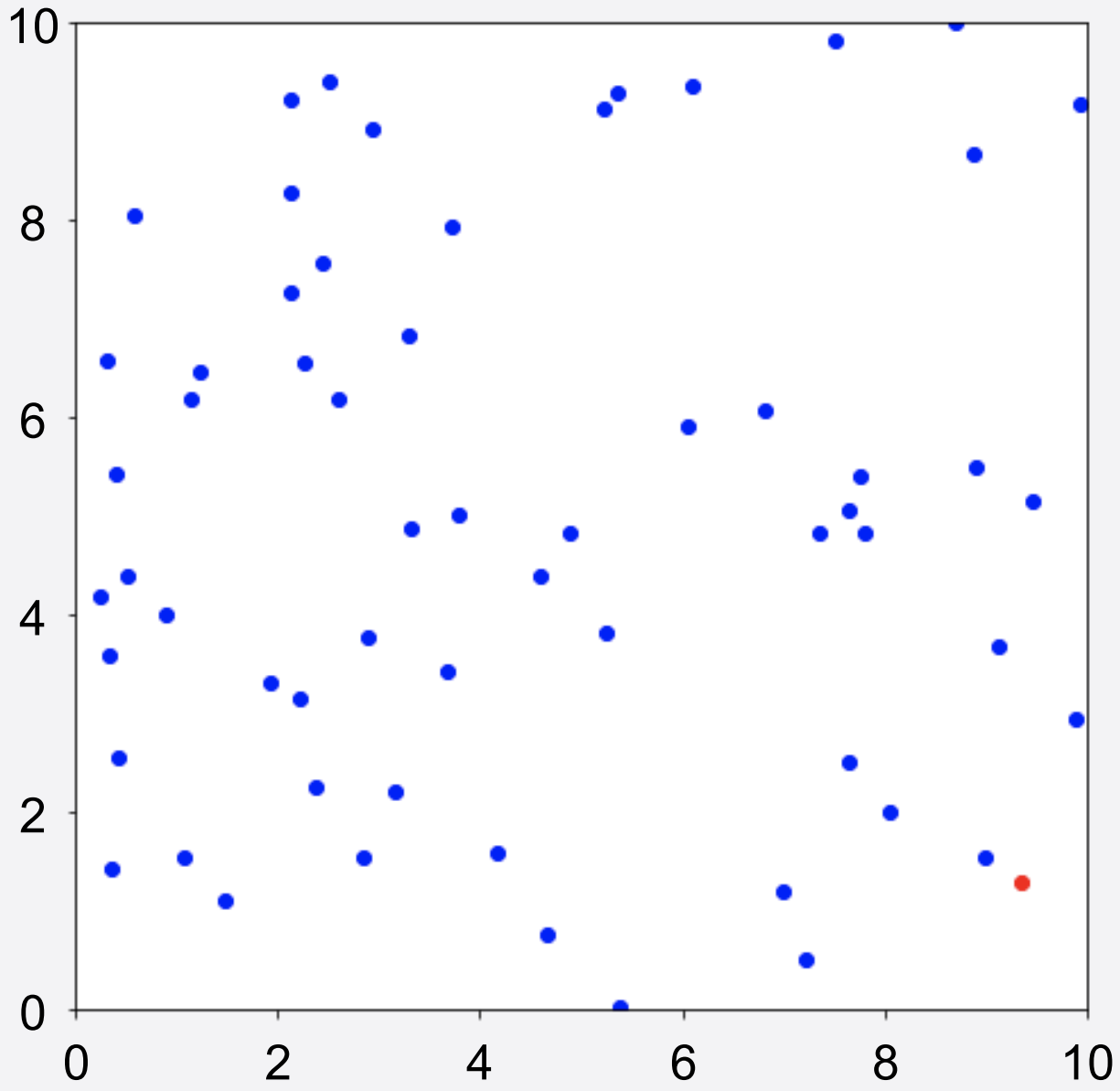
При n = 5



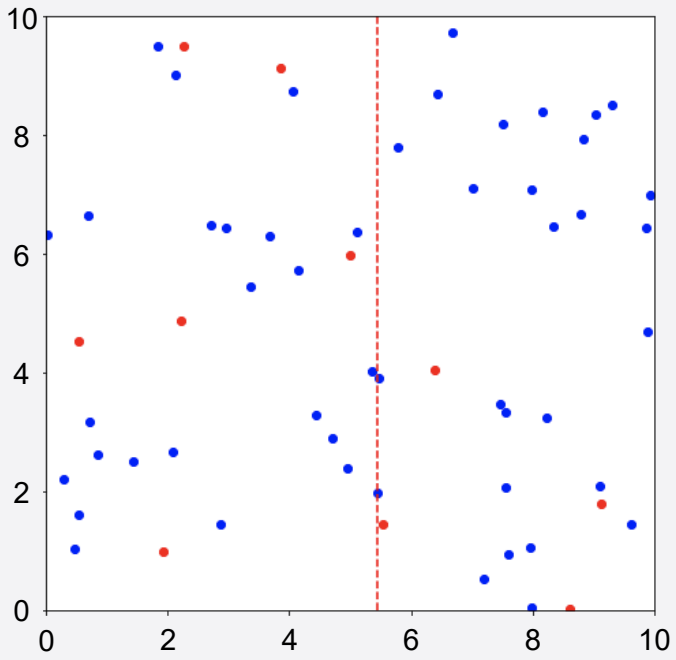
Алгоритмы

1. BallTree
2. KDTree
3. Алгоритм полного перебора

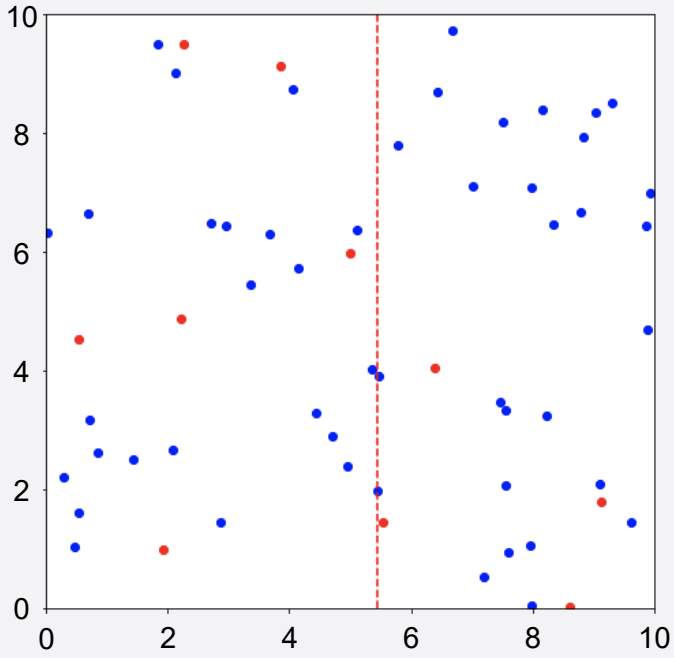
Алгоритм полного перебора

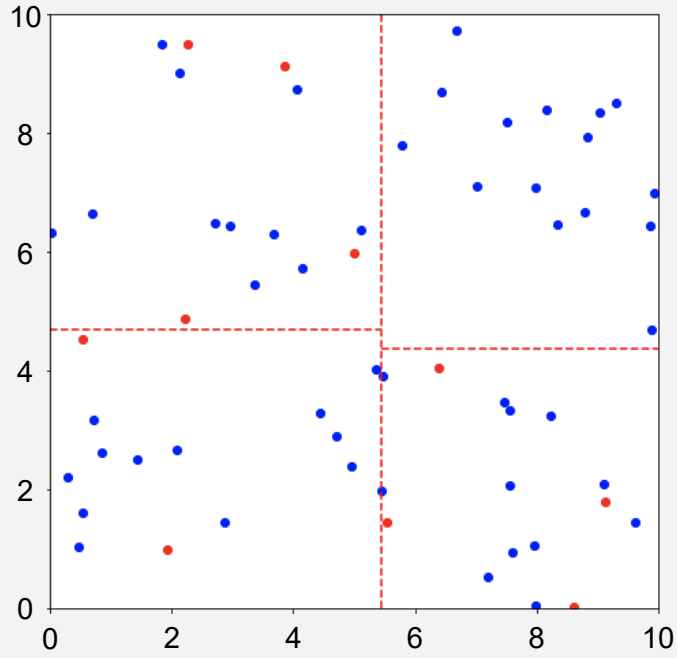


K-D Tree

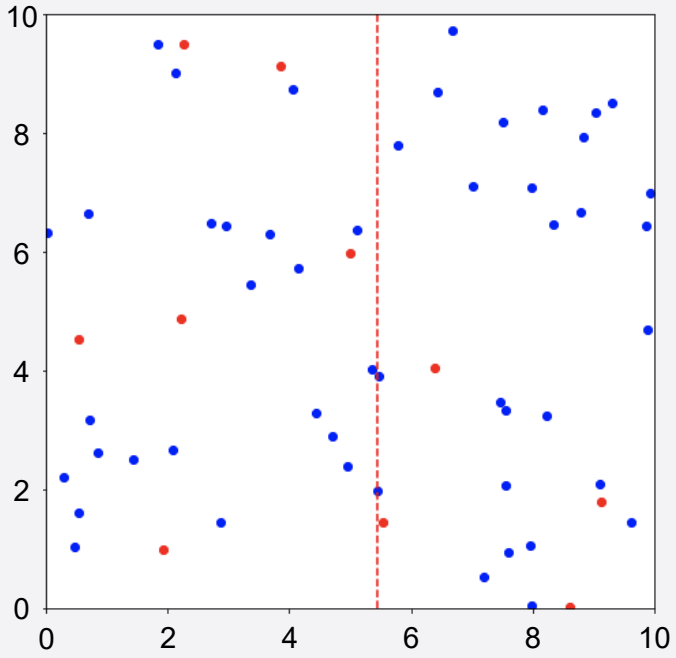


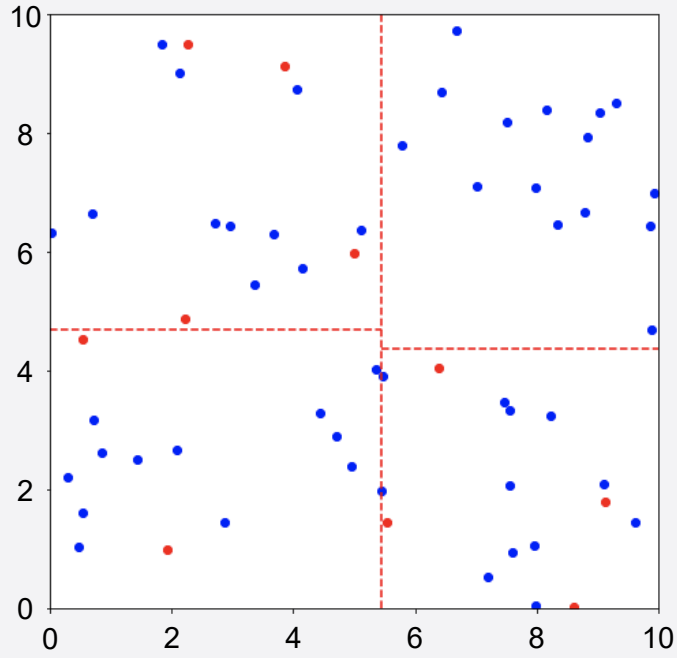
K-D Tree

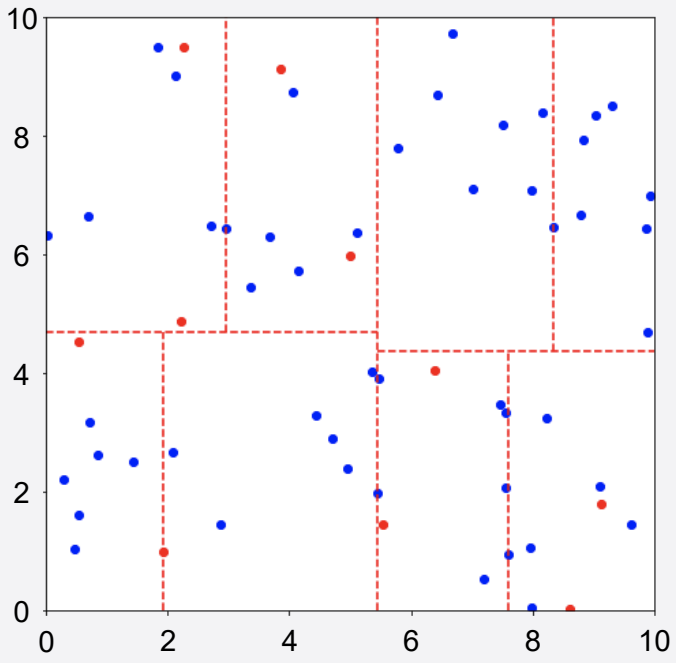




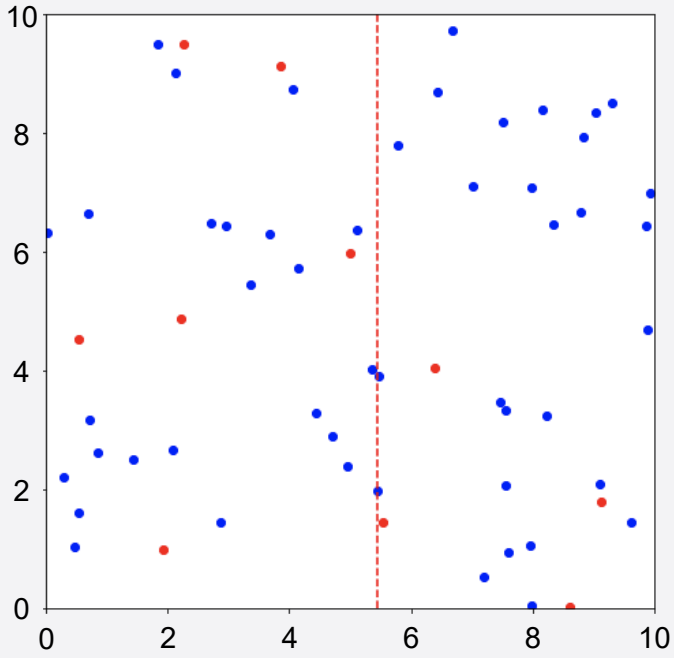
K-D Tree

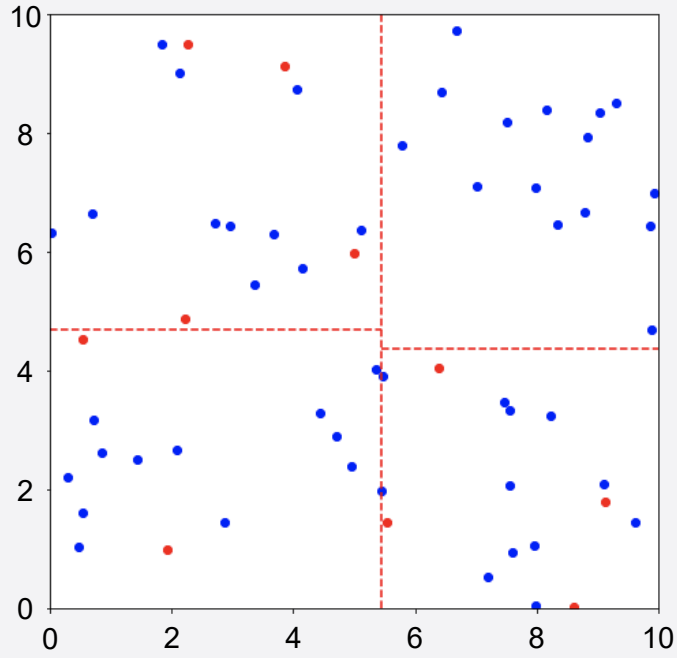


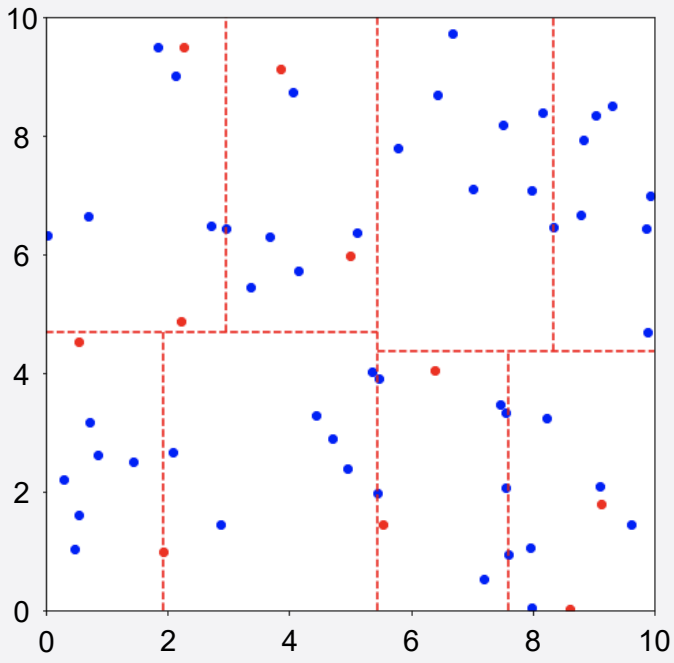


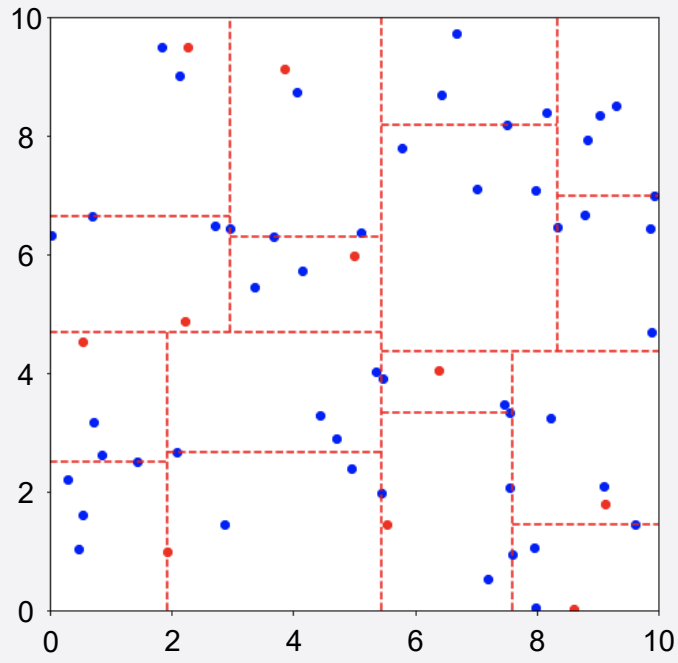


K-D Tree

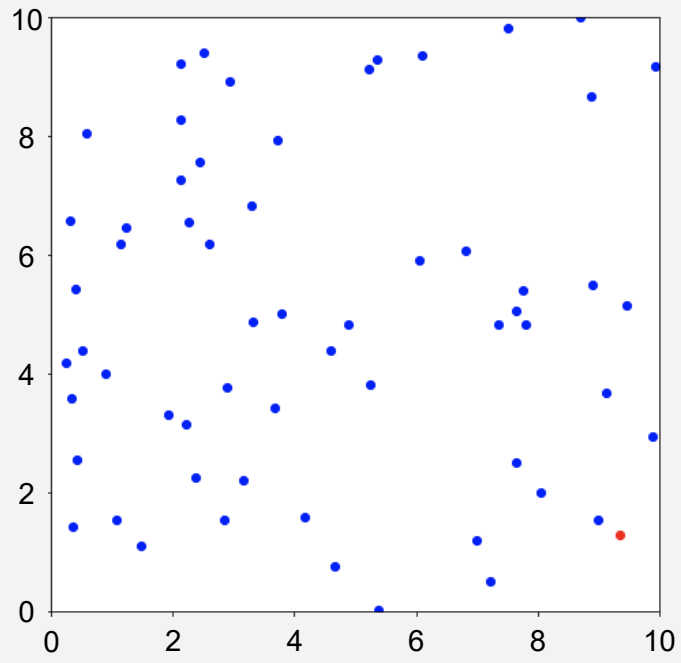


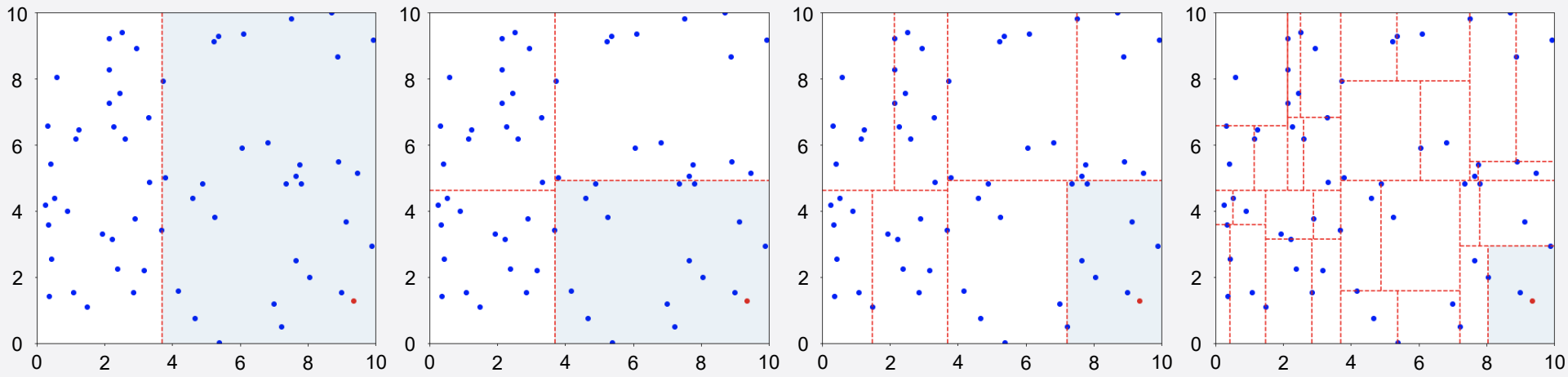




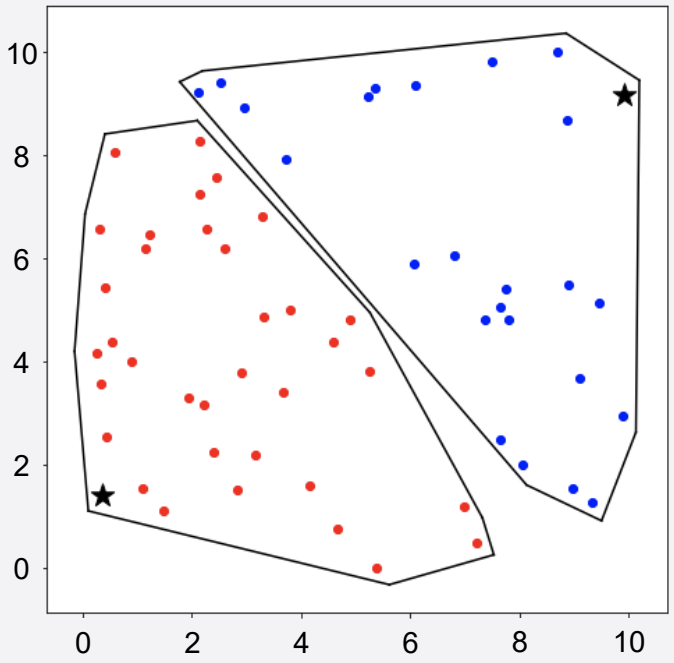


K-D Tree

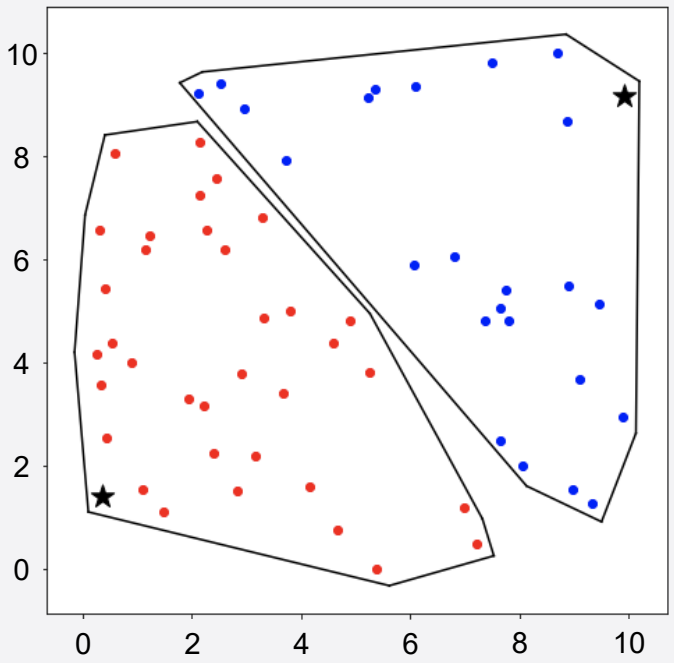


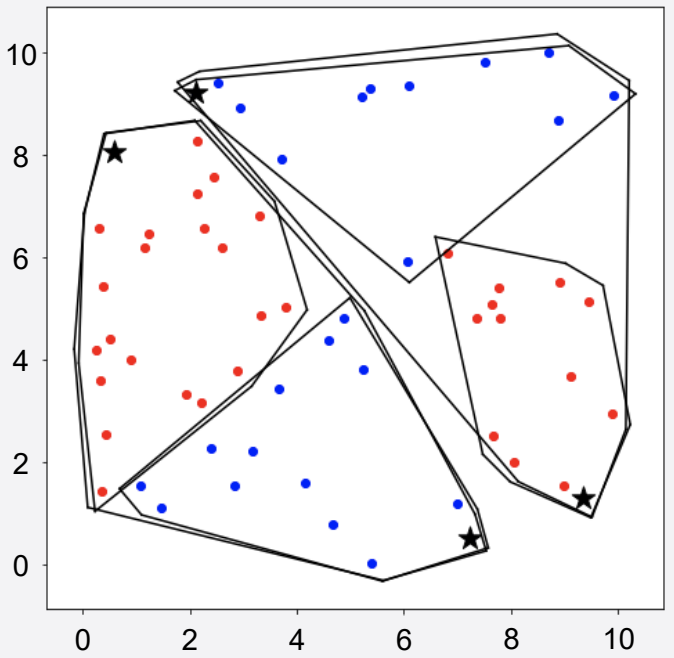


ball tree

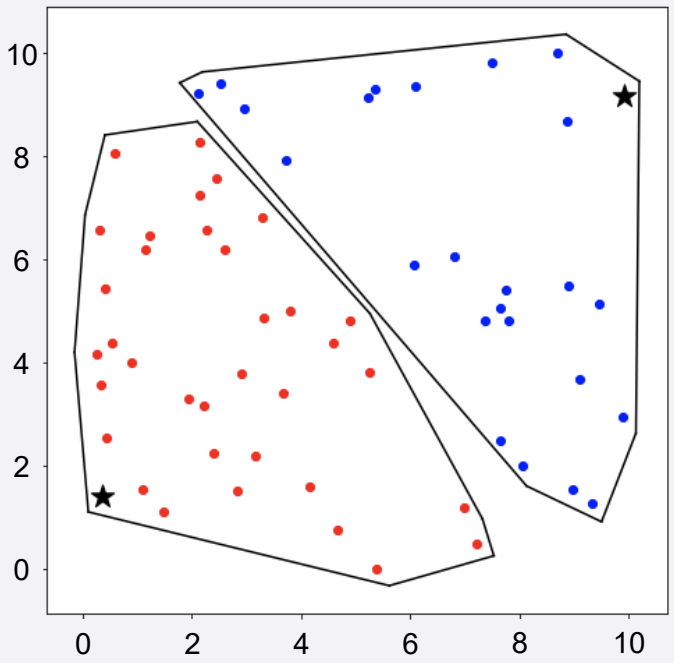


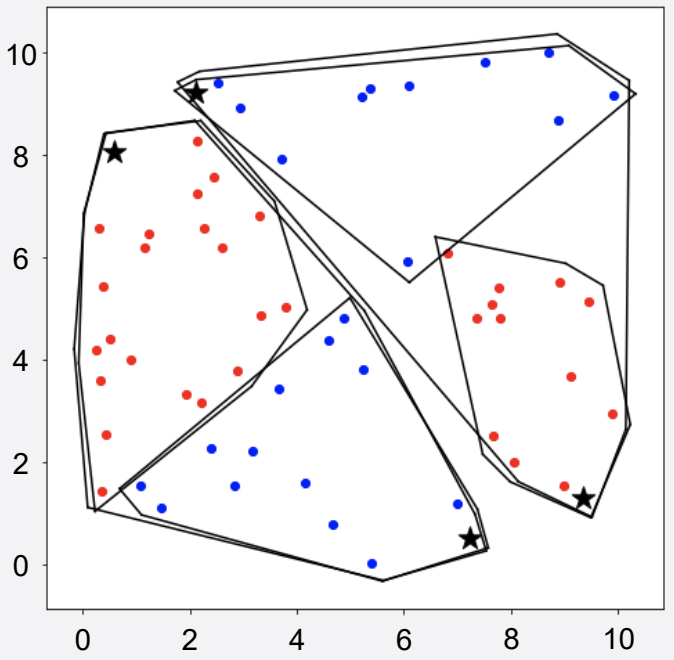
ball tree

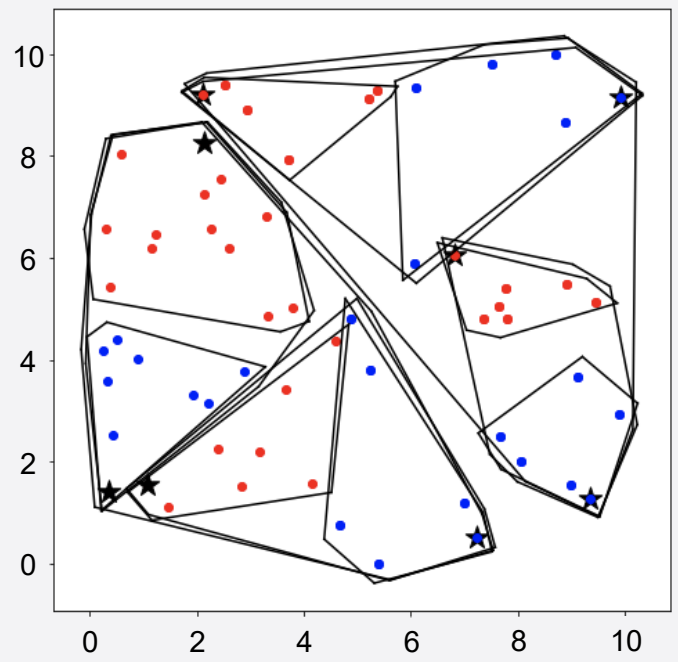




ball tree







Преимущества:

- простота реализации
- интерпретируемость

Недостатки:

- неустойчивость к погрешностям (шуму, выбросам)
- отсутствие настраиваемых параметров
- приходится хранить всю выборку целиком