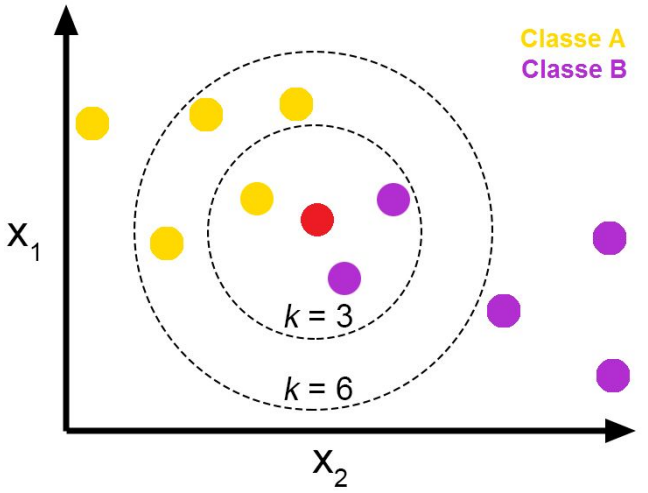
Машинное обучение: часть 2

Алгоритм ближайших соседей

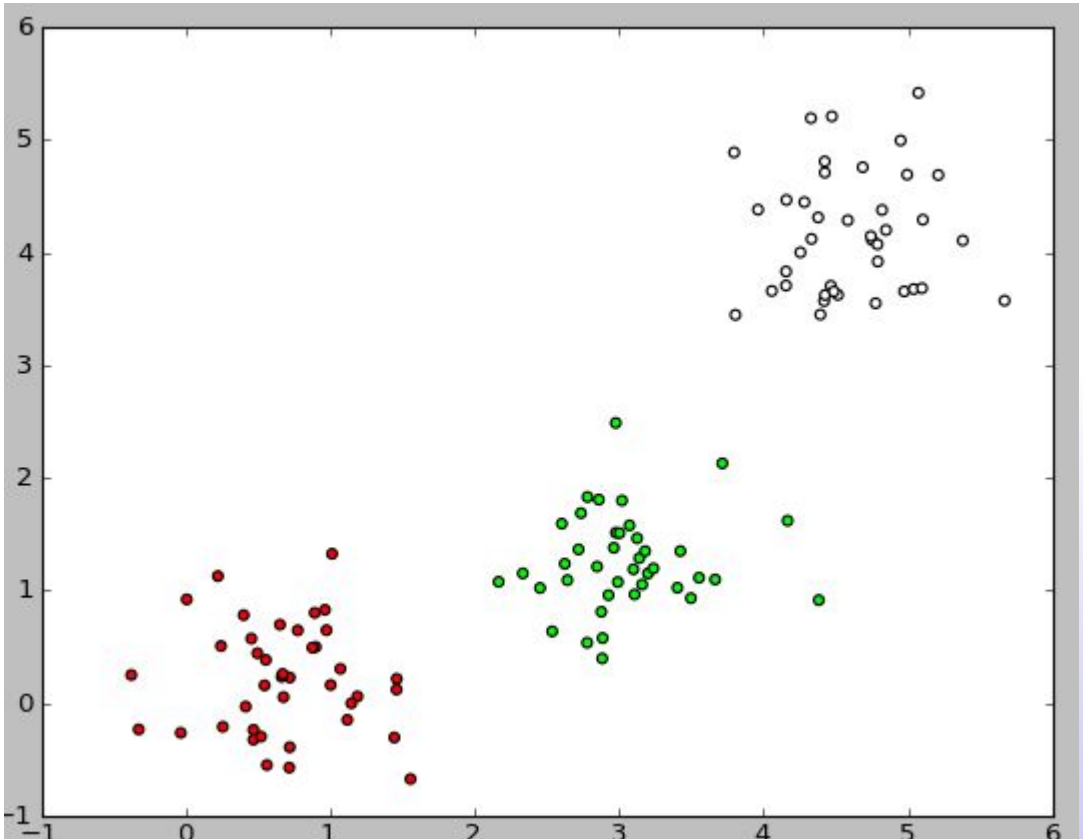
Очень простой алгоритм для классификации – пусть у нас есть наша обучающая и тестовые выборки

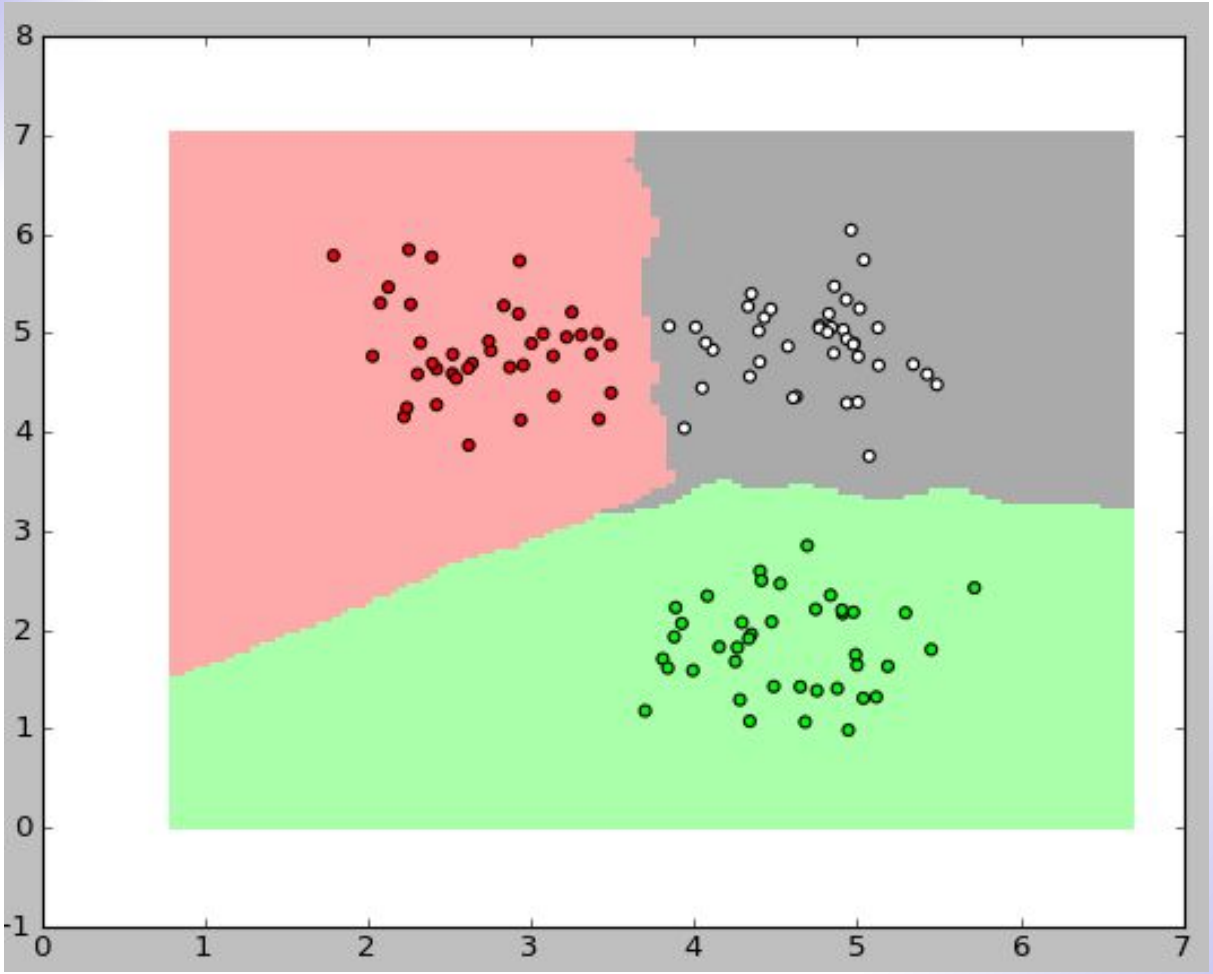
Для классификации каждого из объектов тестовой выборки необходимо последовательно выполнить следующие операции:

- 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

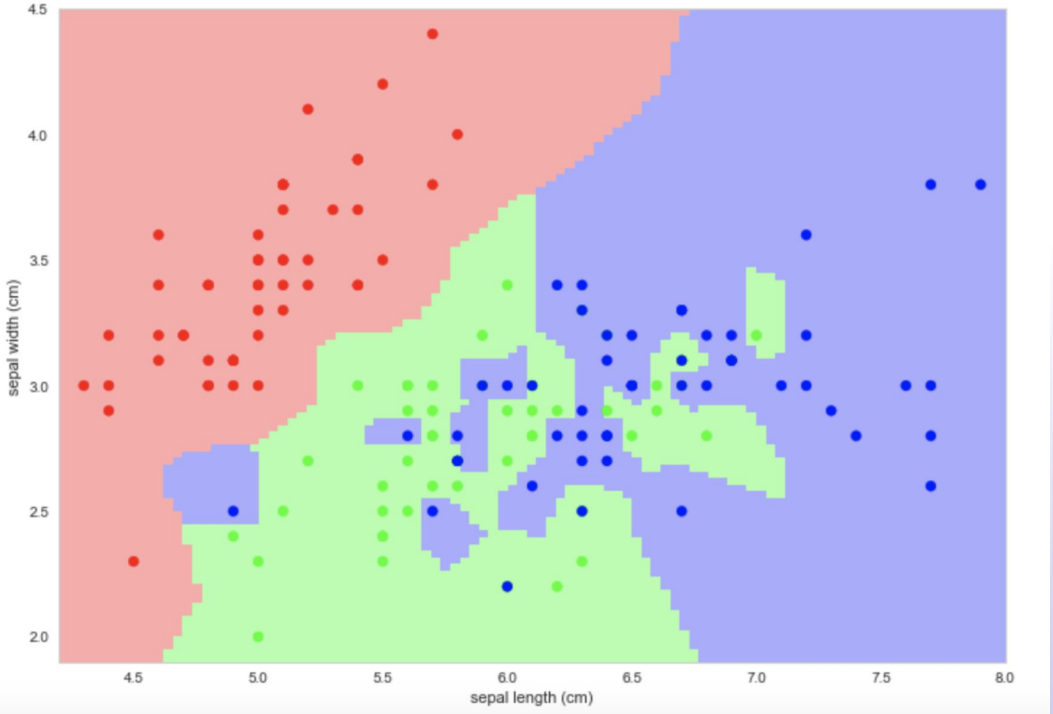


Идеальный случай разделения

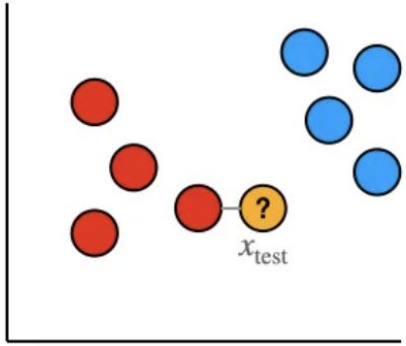



8
7
6
5
4
3
2
1
0
-1
0
1
2
3
4
5
6
7

Понятно, что в реальном мире так случается довольно редко

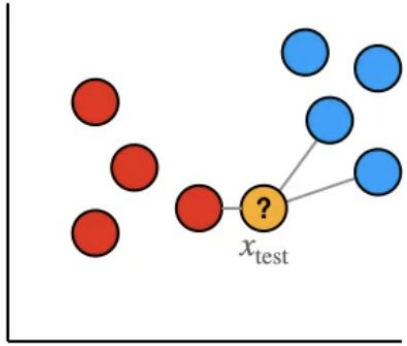


Какое k оптимально?



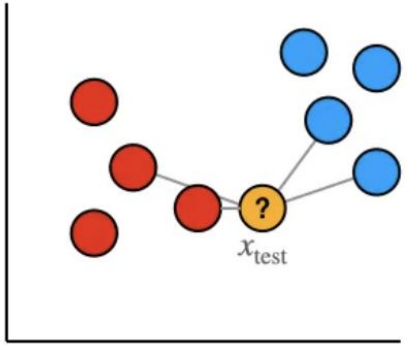
$k = 1$

Nearest point is red, so x_{test} classified as red



$k = 3$

Nearest points are {red, blue, blue} so x_{test} classified as blue



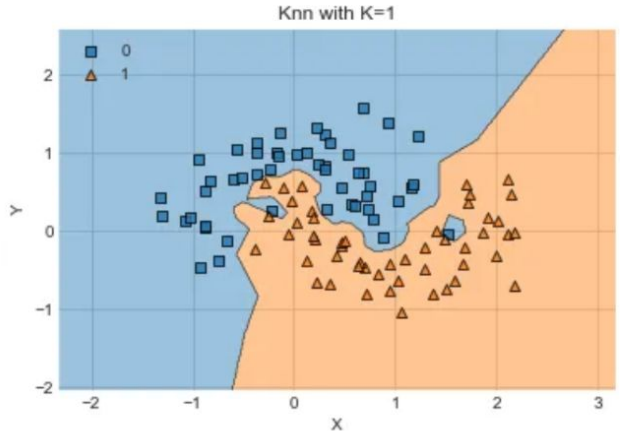
$k = 4$

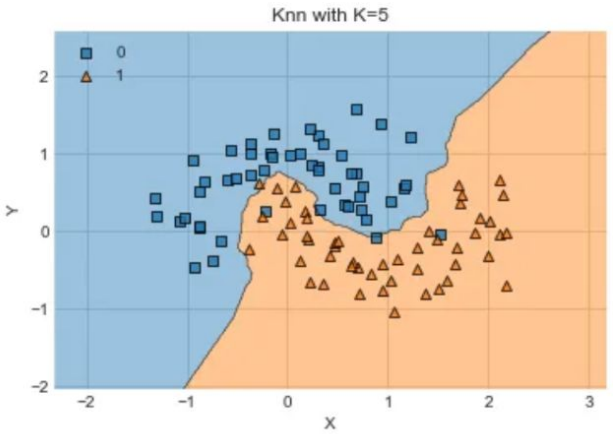
Nearest points are {red, red, blue, blue} so classification of x_{test} is not properly defined

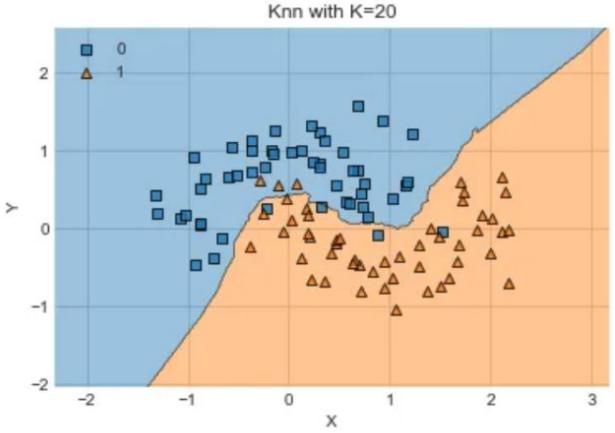
Какое k оптимально?

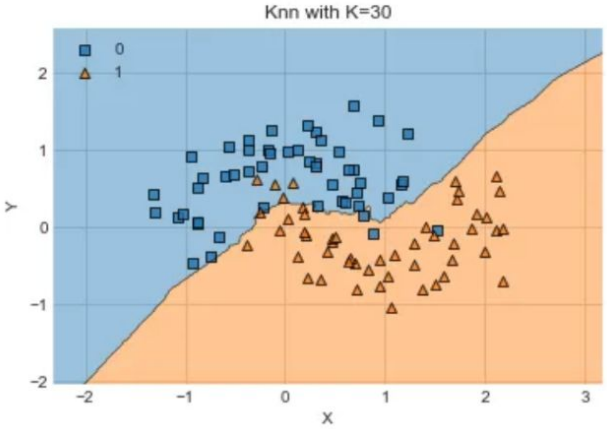
Если выбрать k слишком маленьким, алгоритм может быть слишком чувствителен к шуму в данных, что может привести к переобучению.

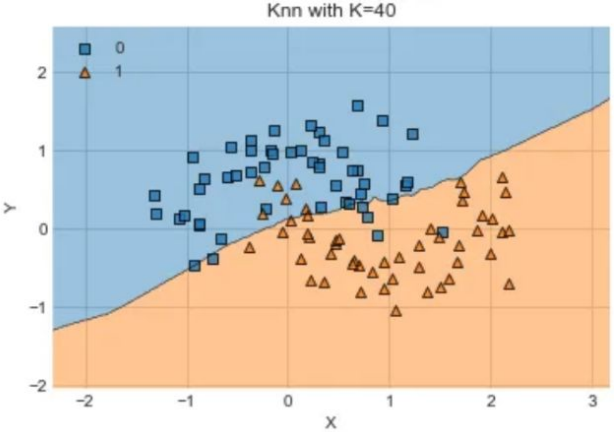
С другой стороны, если k слишком велико, алгоритм может чрезмерно упростить границы по которым разделяются классы и не уловить важные закономерности в данных, что может привести к ухудшению качества.

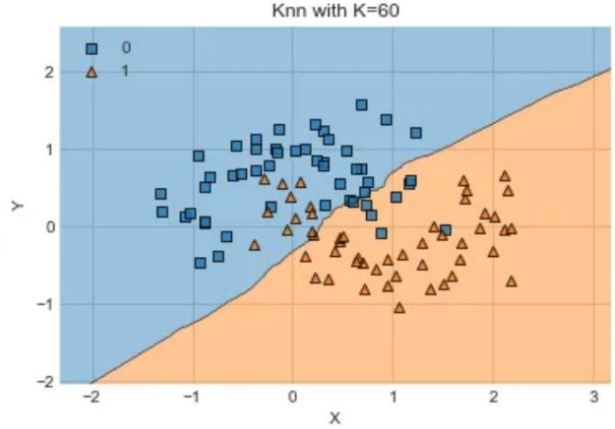

Knn with K=1
0
1
X
Y


Knn with K=5
0
1
X
Y


Knn with K=20
0
1
X
Y


Knn with K=30
0
1
X
Y


Knn with K=40
0
1
X
Y


Knn with K=60
0
1
X
Y

Преимущества и недостатки KNN

Преимущества:

- простота в реализации и интерпретации;
- применяется во многих задачах, особенно в рекомендательных системах;
- высокая точность прогнозов при правильном подборе k и метрики расстояния.

Недостатки:

- большое потребление памяти и низкая скорость работы из-за хранения и вычисления расстояний между всеми обучающими и тестовыми образцами;
- чувствительность к выбросам и шуму, а также к несбалансированным