



Попробуем решить задачу



А ты кто?

Перед нами домашнее животное. Кто это — собака или кот?

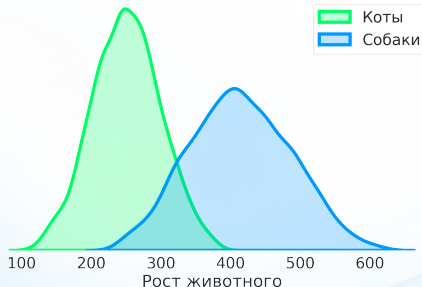




Классификация: собака vs кот

Попробуем сначала извлечь какой-то *признак*.

Построим вероятностные плотности для каждого класса.



При каких-то значениях роста мы уже можем с большой уверенностью сказать ответ.

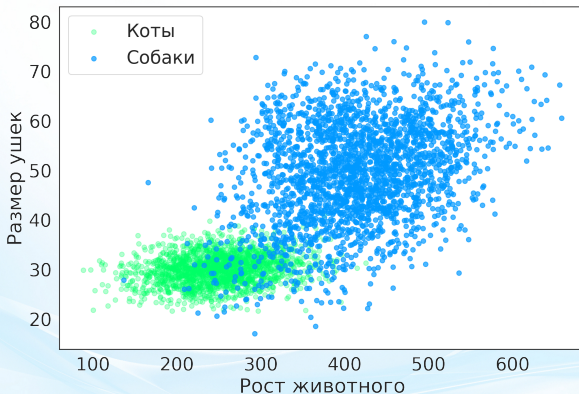
Но есть большое пересечение, это не очень здорово.



Классификация: собака vs кот

Извлечем еще один признак — размер ушек.

Теперь классы лучше разделяются.

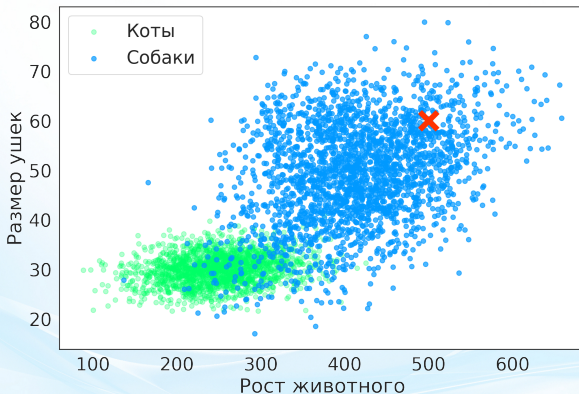




Классификация: собака vs кот

Попробуем классифицировать новое животное.

Кто отмечен красным?

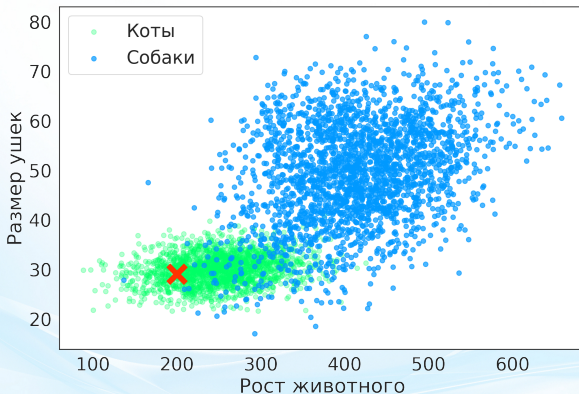




Классификация: собака vs кот

Попробуем классифицировать новое животное.

Кто отмечен красным?

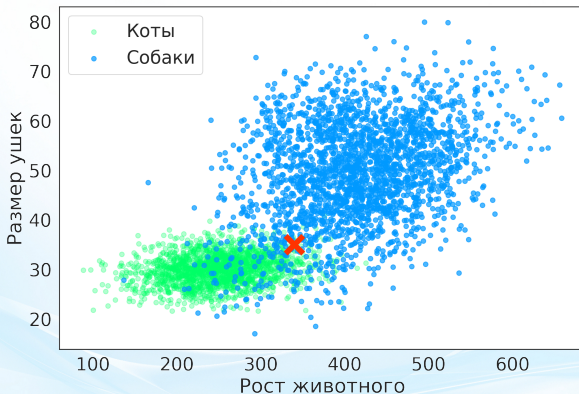




Классификация: собака vs кот

Попробуем классифицировать новое животное.

Кто отмечен красным?



На основе чего вы сделали все выводы?



Метод ближайших соседей (kNN)

Дано:

X_1, \dots, X_n — набор размеченных объектов.

Y_1, \dots, Y_n — соответствующие метки класса.

Задача:

Пусть x — исследуемый объект. Какого он класса?

Решение:

Будем смотреть на свойства k ближайших соседей.

$x_{(1)}, \dots, x_{(k)}$ — k его соседей в порядке удаления от x .

Y_1, \dots, Y_k — соответствующие им классы.

Ответ — наиболее часто встречаемый класс среди $x_{(1)}, \dots, x_{(k)}$.

Свойства:

1. k — гиперпараметр модели;
2. Не редко на практике показывает хорошие результаты.

3. Дорогое применение:

для каждого x результат вычисляется за $O(n)$.*

* С оптимизацией за $O(\ln(n))$



Взвешенный метод ближайших соседей

Пусть x — исследуемый объект.

$x_{(1)}, \dots, x_{(k)}$ — k его соседей в порядке удаления от x .

Y_1, \dots, Y_k — соответствующий отклик.

w_1, \dots, w_k — вклад k -го соседа, определяемый пользователем.

Способы определения веса:

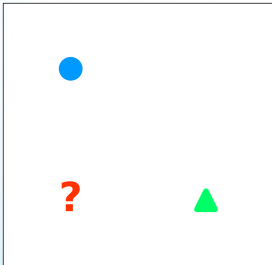
- ▶ $w_j = 1 - j/k$ — зависящий от номера соседа;
- ▶ $w_j = \|x - x_{(j)}\|^{-1}$ — зависящий от расстояния до соседа.

$$\hat{y}(x) =_y \sum_{j=1}^k w_j I\{Y_j = k\} \text{ — классификация}$$



Особенности

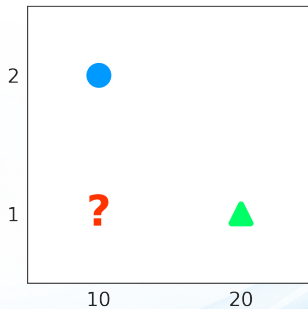
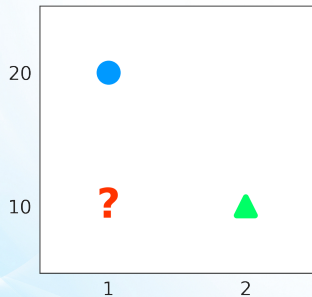
Классифицируйте объект "?".





Особенности

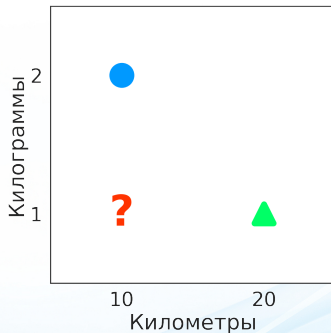
Классифицируйте объект "?".





Особенности

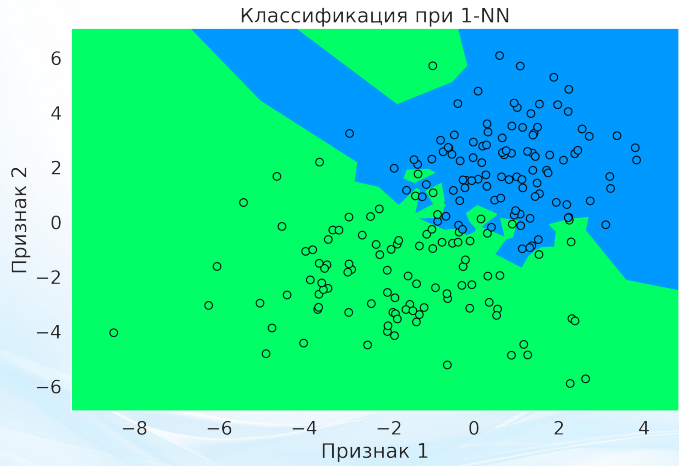
Классифицируйте объект "?".



Вывод: результат сильно зависит от используемой метрики между точками в пространстве. Не складывайте *кг* с *км*!

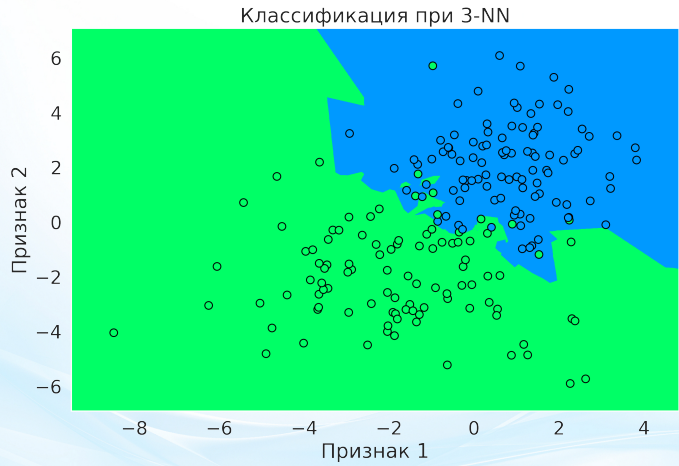


Что происходит при разных k ?



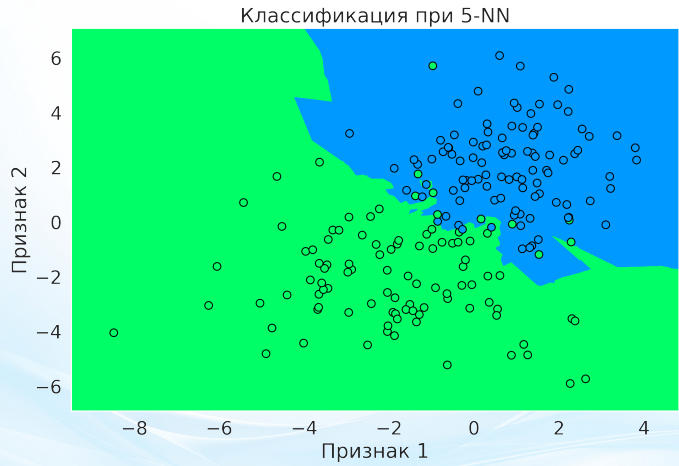


Что происходит при разных k ?



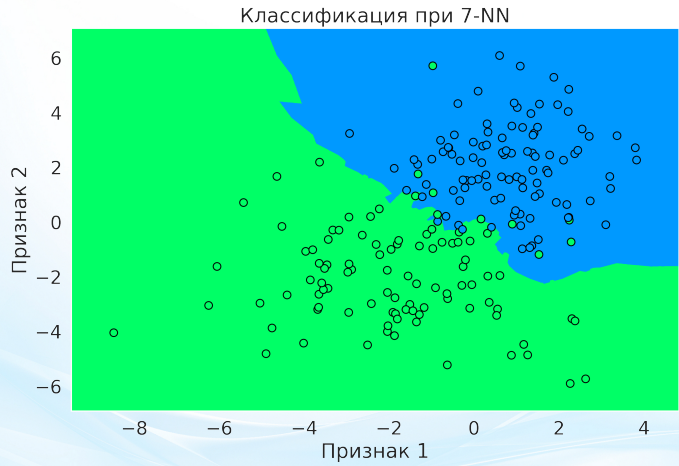


Что происходит при разных k ?



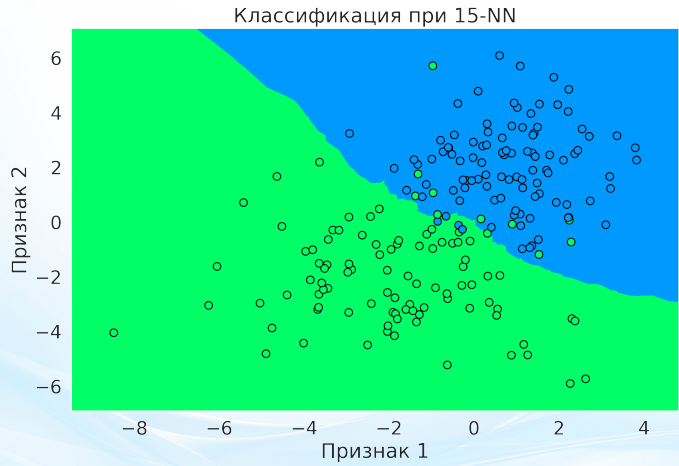


Что происходит при разных k ?



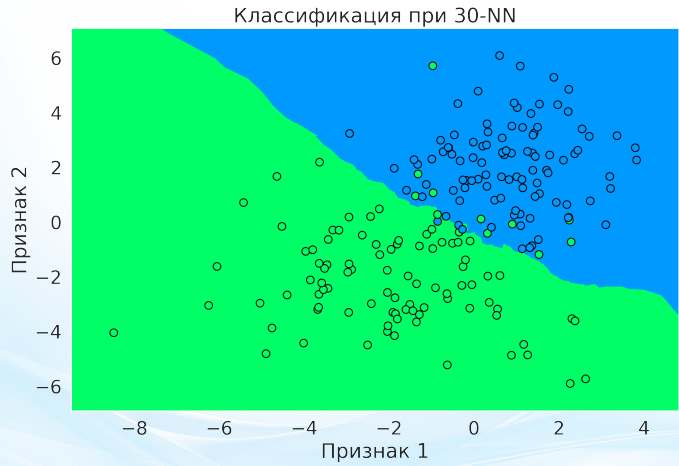


Что происходит при разных k ?





Что происходит при разных k ?





Как оценить качество классификации?

Пусть $\hat{y}(x)$ — оценка класса для объекта x .

Можем посчитать **точность** — доля правильно угаданных классов

$$A = \frac{1}{n} \sum_{i=1}^n I\{Y_i = \hat{y}(x_i)\}$$

Оценка качества называется **метрикой** (не путать с метр. пр-вами).

Какое число соседей оптимизирует эту метрику?

Ответ: $k = 1$, т.к. при вычислении $\hat{y}(x_i)$ берем сам Y_i .

Поэтому данные делят случайно на **две непересекающиеся части**:

1. на одной определяют правило классификации,
2. на другой — считают оценку качества классификации.

Точность 90% это много или мало?

Кажется, круто. А если в данных 85% котов? Тогда отвечая всегда "кот" сможем добиться точности 85%, и 90% уже не так круто...



А что если по картинке?

Хорошо, но что если объект — изображение кота или собаки?
Изображение 100×100 состоит из 10^4 пикселей,
в каждом по 3 числа. Какой размерности получается объект?
Ответ: $100 \times 100 \times 3 = 30\,000$ чисел в одной картинке.

Проблема:

в пр-ве больших размерностей расстояния неинформативны.

Например, среди фиксированного количества случайных точек
в единичном кубе в пространстве большой размерности
почти все точки будут лежать около границы куба.

