Метод ближайших соседей

Виктор Владимирович Китов

Содержание

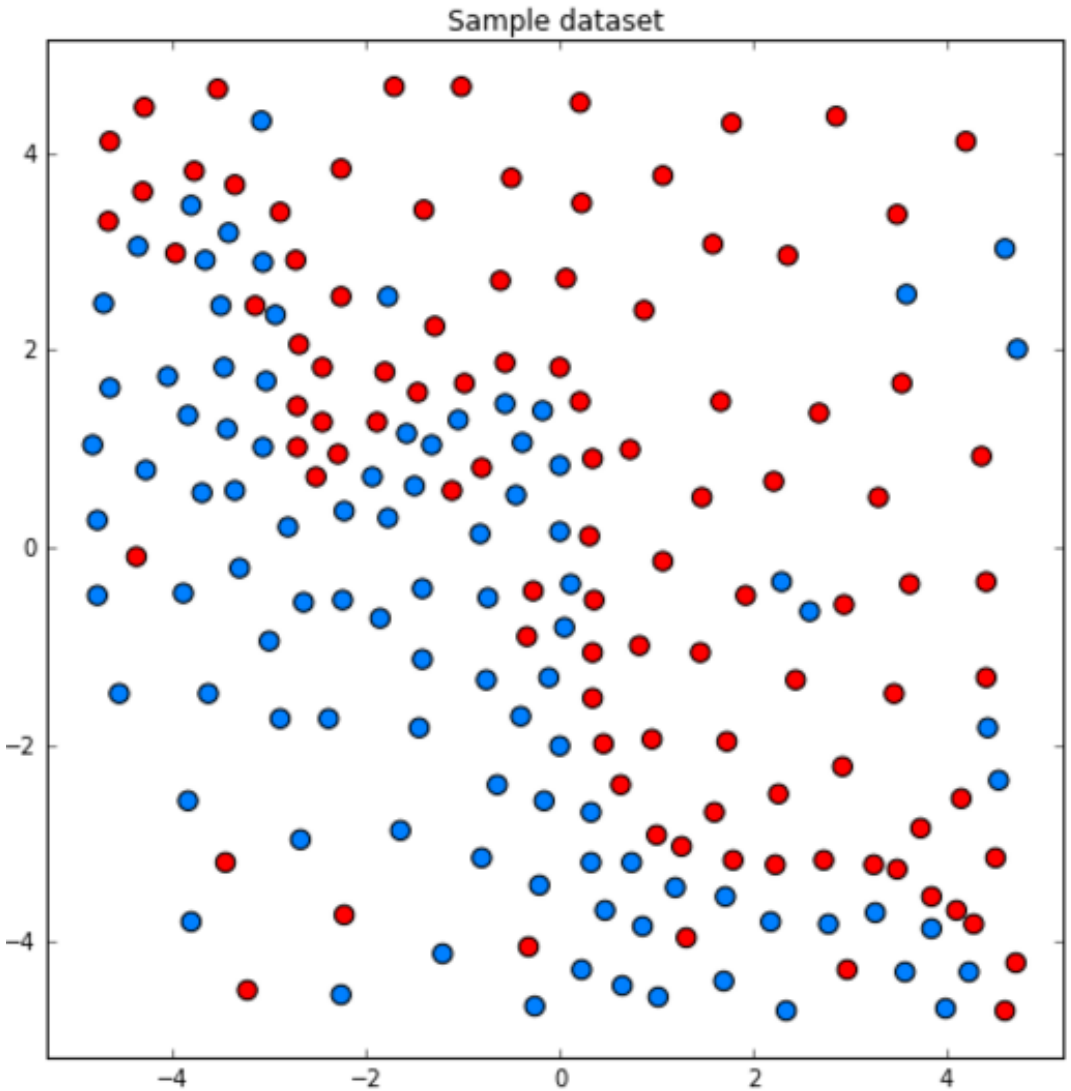
Метод K-ближайших соседей (K-nearest neighbours)

Классификация методом K ближайших соседей

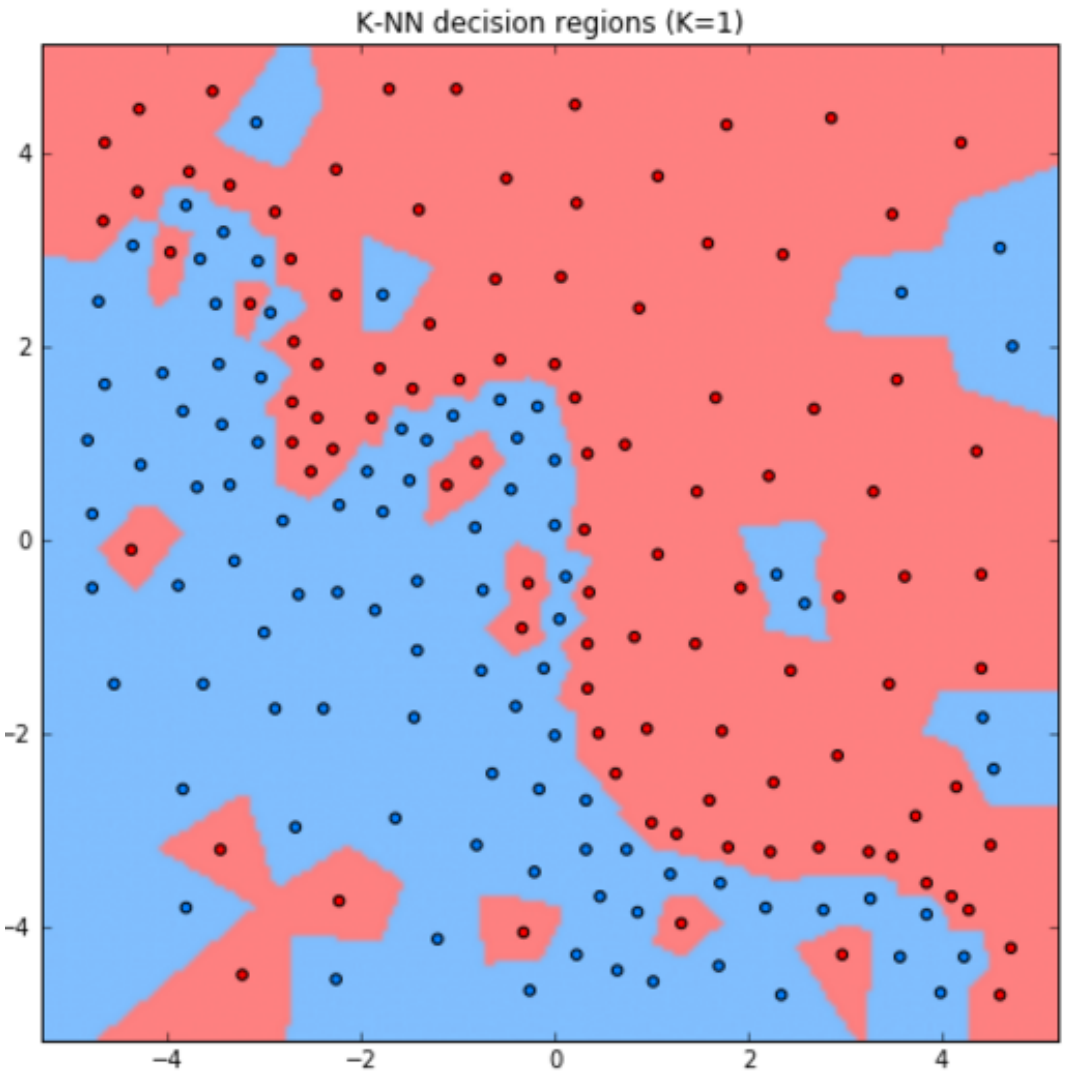
1. Найти K ближайших объектов к интересующему объекту x в обучающей выборке.
2. Сопоставить x самый часто встречающийся класс среди K соседей.

- Случай $K = 1$: алгоритм ближайшего соседа (nearest neighbour)
- Случай $K = N$: константный прогноз наиболее частым классом в выборке.
- В случае регрессии нужно усреднить характеристики k ближайших соседей.
- Основное предположение метода:
  - близкие объекты выдают похожие ответы

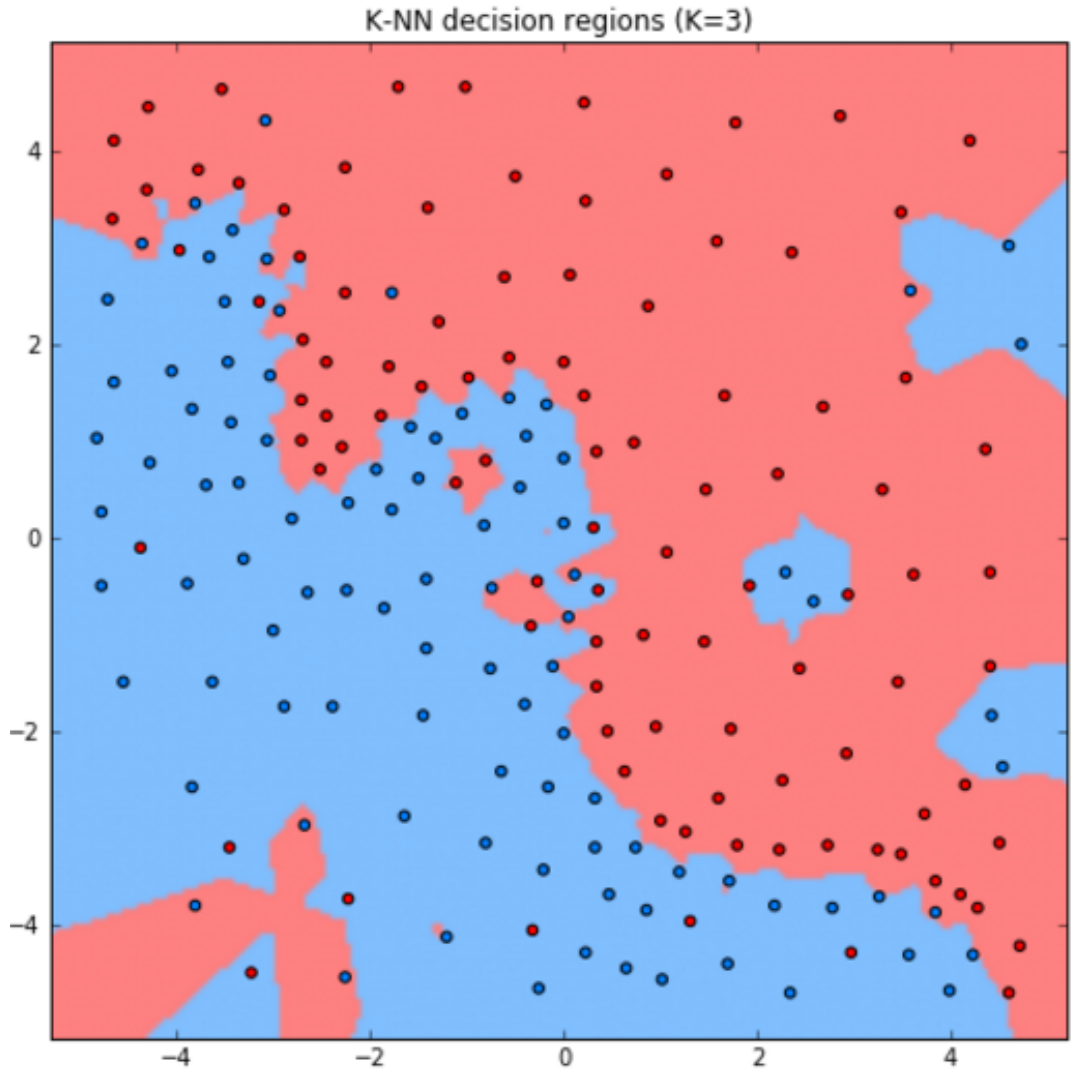
Пример: обучающая выборка



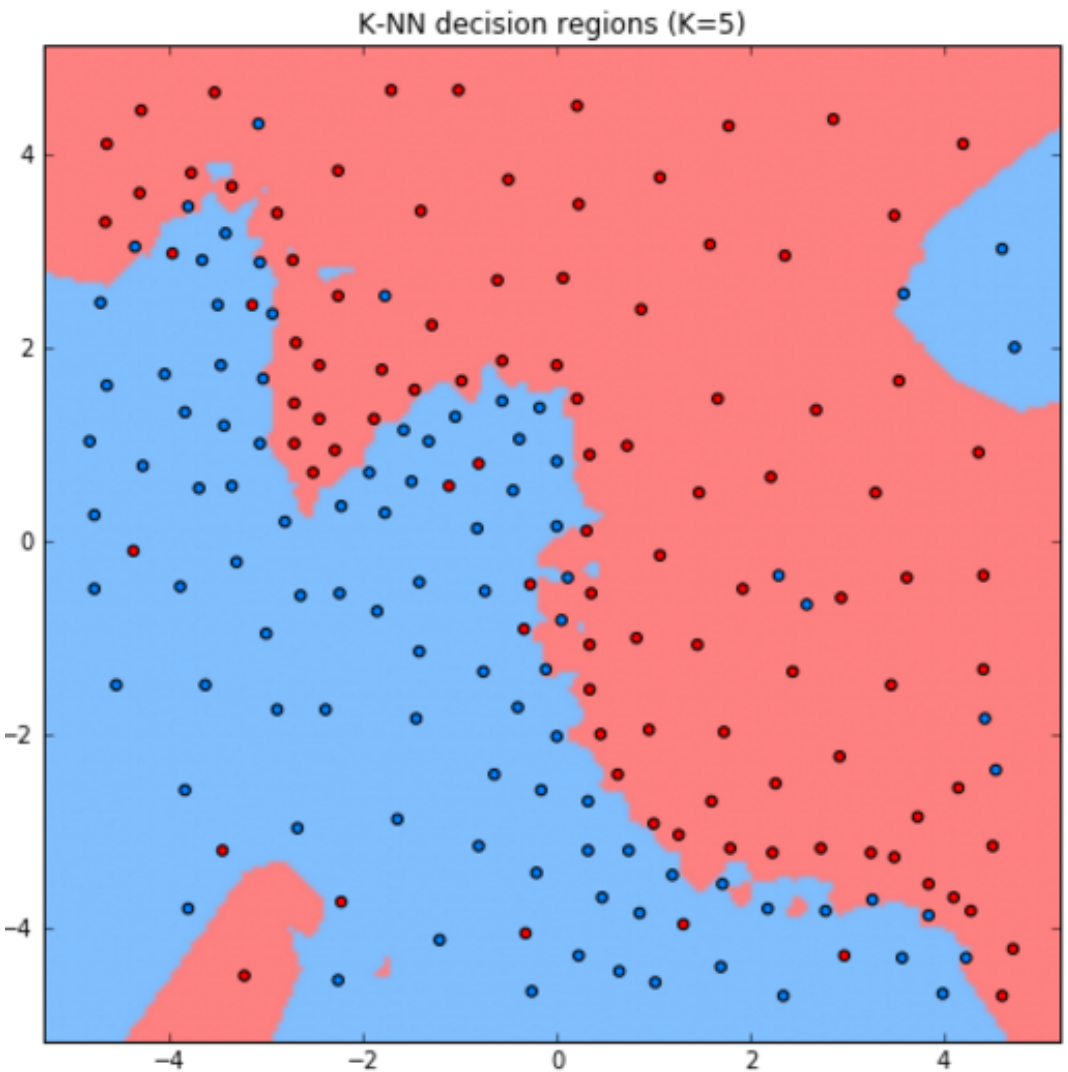
Пример: K-NN, классификация



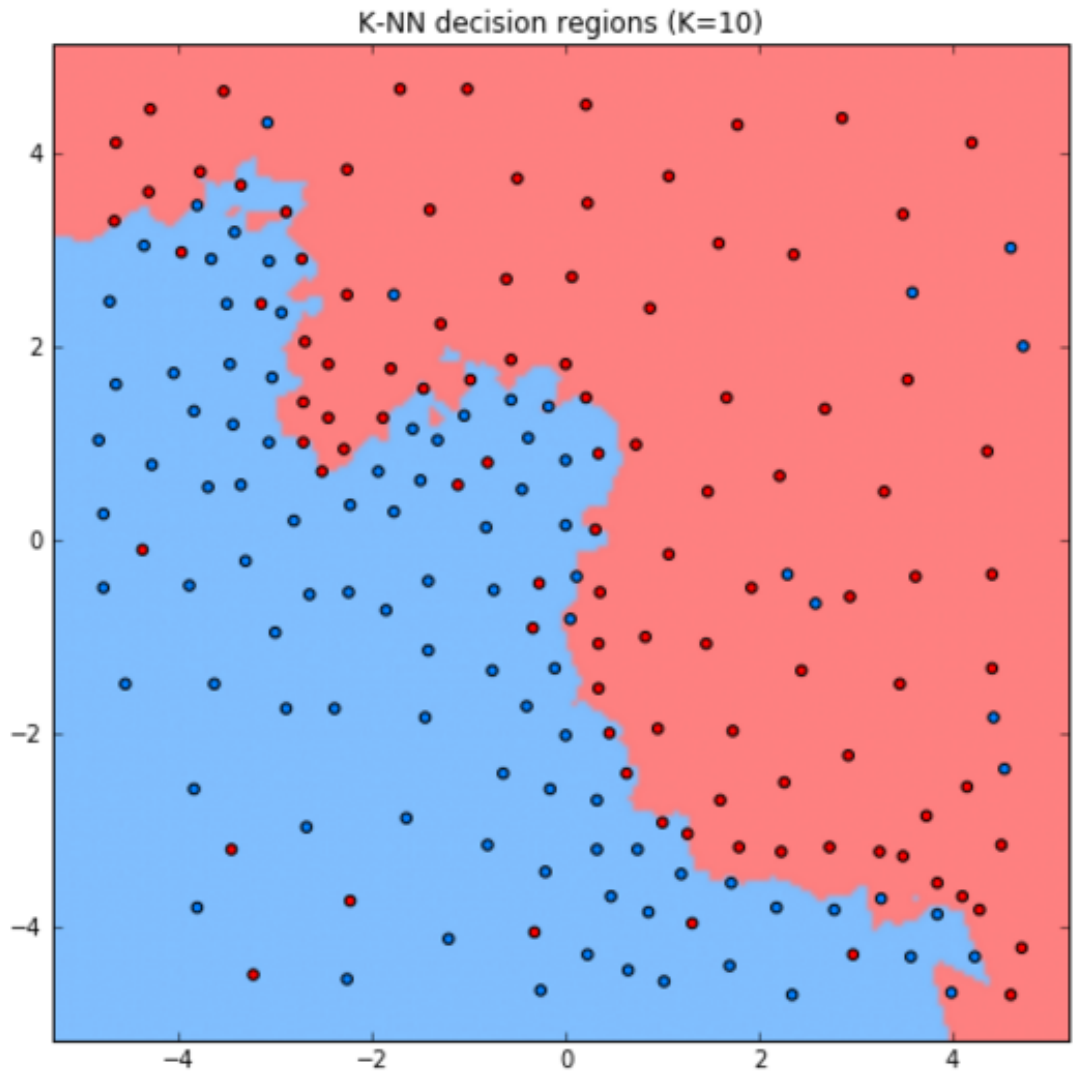
Пример: K-NN, классификация



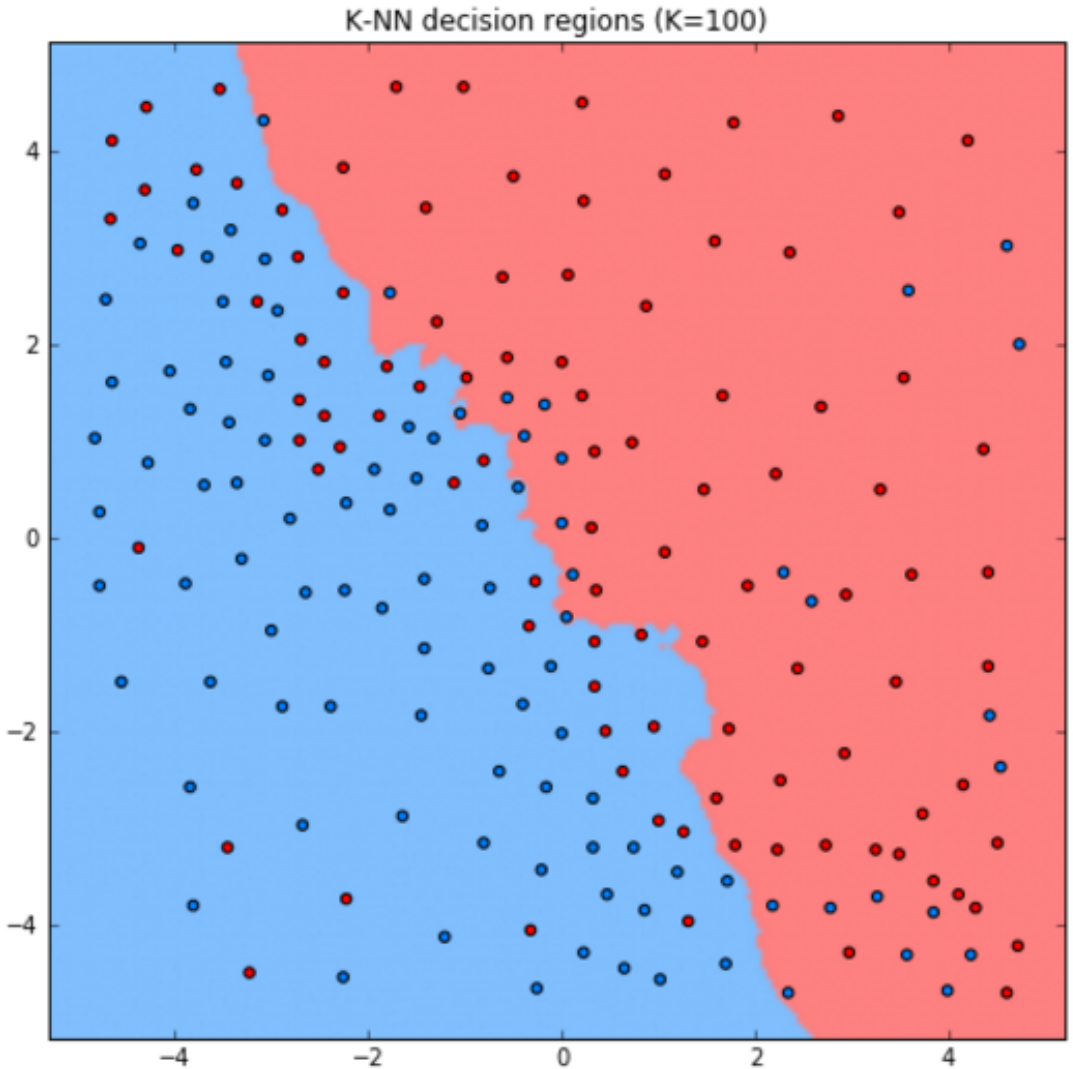
Пример: K-NN, классификация



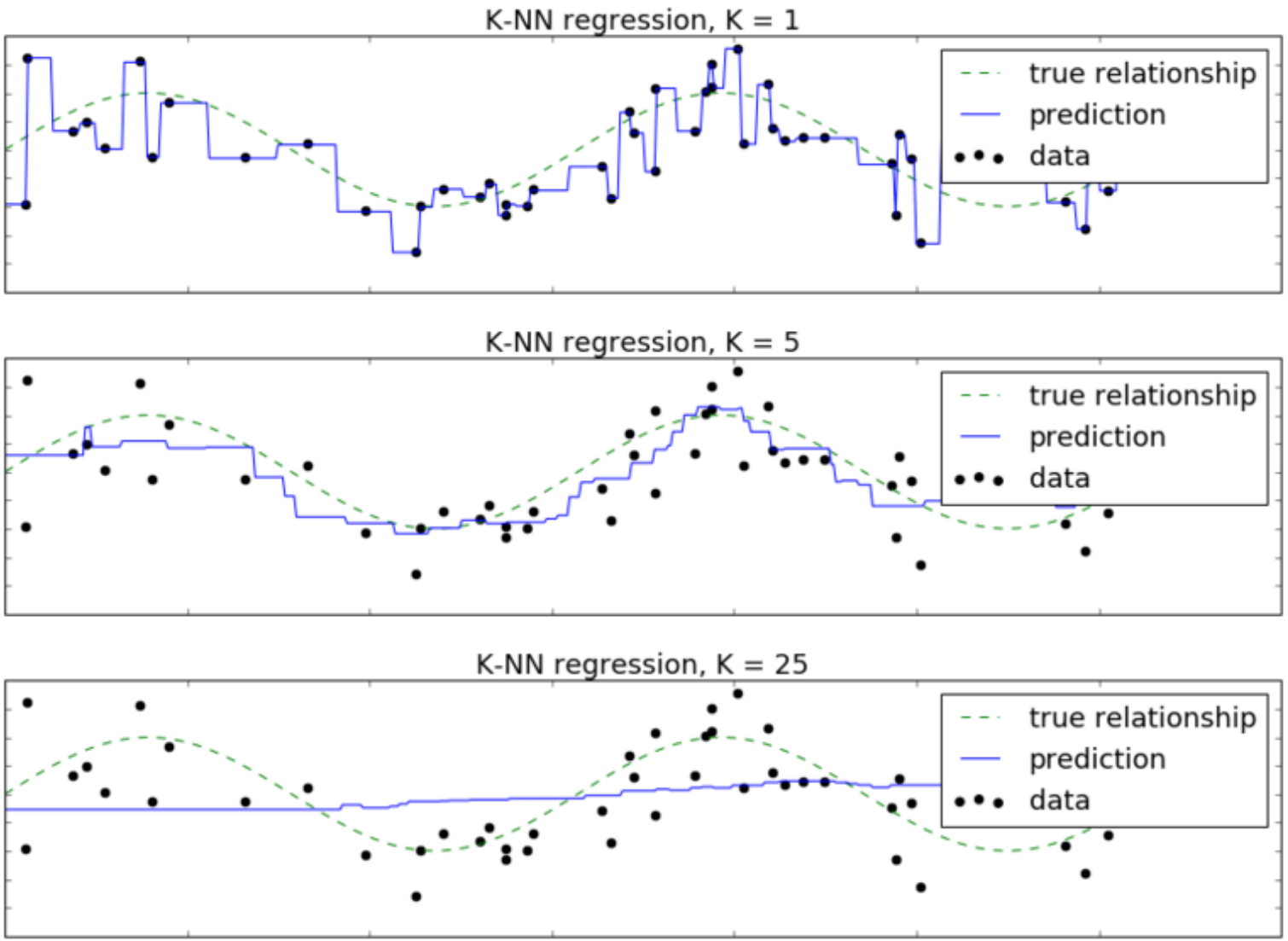
Пример: K-NN, классификация



Пример: K-NN, классификация



Пример: K-NN, регрессия



Параметры метода

- Параметры:
  - число соседей K
  - функция близости $\rho(x, y)$
- Вариант метода с адаптивным K по точности ближайших соседей.

Свойства

- Преимущества:
  - нужно знать только ф-цию близости между объектами, сами признаки не нужны.
    - может быть применен к объектам любой сложности, если задана ф-ция близости
  - простая логика работы, легко объяснить и реализовать
  - интерпретируемость (case based reasoning)
  - не требует обучения
    - может применяться в online случаях.
    - K-CV можно заменить на LOO оценивание.
- Недостатки:
  - медленная классификация со сложностью $O(N)$
  - требования по памяти тоже $O(N)$, т.к. нужно хранить всю выборку
  - точность ухудшается с ростом размерности пространства

Содержание

Нормализация признаков

- Чаще всего используется Евклидова метрика близости:

$$\rho(x,z) = \sqrt{\sum_{d=1}^{D}(x^d - z^d)^2}$$

- Необходимо нормализовывать признаки.
  - Определим μ_j, σ_j, L_j, U_j как среднее значение, стандартное отклонение, минимальное и максимальное значение j-го признака.

Преобразование	**Свойства результата**
$x'_j = \frac{x_j - \mu_j}{\sigma_j}$	среднее=0, дисперсия=1.
$x'_j = \frac{x_j - L_j}{U_j - L_j}$	принадлежит интервалу $[0,1]$.

Нормализация признаков

- Нелинейные трансформации для признаков, допускающих редкие большие значения:
  - $\tilde{x}^i = \log(x^i)$
  - $\tilde{x}^i = \left(x^i\right)^p, 0 \leq p < 1$
- Для $F_i(\alpha) = P(x^i \leq \alpha)$ преобразование $\tilde{x}^i \to F_i(x^i)$ даст признак, равномерно распределенный на $[0, 1]$.

Выбор функции расстояния

Metric	$\boldsymbol{d(x,z)}$
Евклидова	$\sqrt{\sum_{i=1}^{D}(x^i-z^i)^2}$
L_p	$\sqrt[p]{\sum_{i=1}^{D}(x^i-z^i)^p}$
L_∞	$\max_{i=1,2,...D}\lvert x^i-z^i\rvert$
L_1	$\sum_{i=1}^{D}\lvert x^i-z^i\rvert$
Canberra	$\frac{1}{D}\sum_{i=1}^{D}\frac{\lvert x^i-z^i\rvert}{x^i+z^i}$
Ланса-Уильямса	$\frac{\sum_{i=1}^{D}\lvert x^i-z^i\rvert}{\sum_{i=1}^{D}x^i+z^i}$
косинусная мера	$1-\frac{1}{\pi}cos^{-1}\left(\frac{\sum_{i=1}^{D}x^iz^i}{\sqrt{\sum_{i=1}^{D}\left(x^i\right)^2}\sqrt{\sum_{i=1}^{D}\left(z^i\right)^2}}\right)$

Содержание

Взвешенный учет

Определим $x_{i_1}, x_{i_2}, ...x_{i_N}$ как переупорядочивание $x_1, x_2, ...x_N$ по расстоянию до x: $\rho(x, x_{i_1}) \leq \rho(x, x_{i_2}) \leq ... \leq \rho(x, x_{i_N})$.
Определим $z_1 = x_{i_1}$, $z_2 = x_{i_2}, ...z_K = x_{i_K}$.
Метод ближайших соседей можно определить через C дискриминантных функций:

$$g_c(x) = \sum_{k=1}^{K} \mathbb{I}[z_k \in \omega_c], \quad c = 1, 2, ...C.$$

Взвешенный учет ближайших соседей:

$$g_c(x) = \sum_{k=1}^{K} w(k, \rho(x, z_k))\mathbb{I}[z_k \in \omega_c], \quad c = 1, 2, ...C.$$

Часто выбираемые веса

Независимые от x:

$$w_k = \alpha^k, \quad \alpha \in (0,1)$$

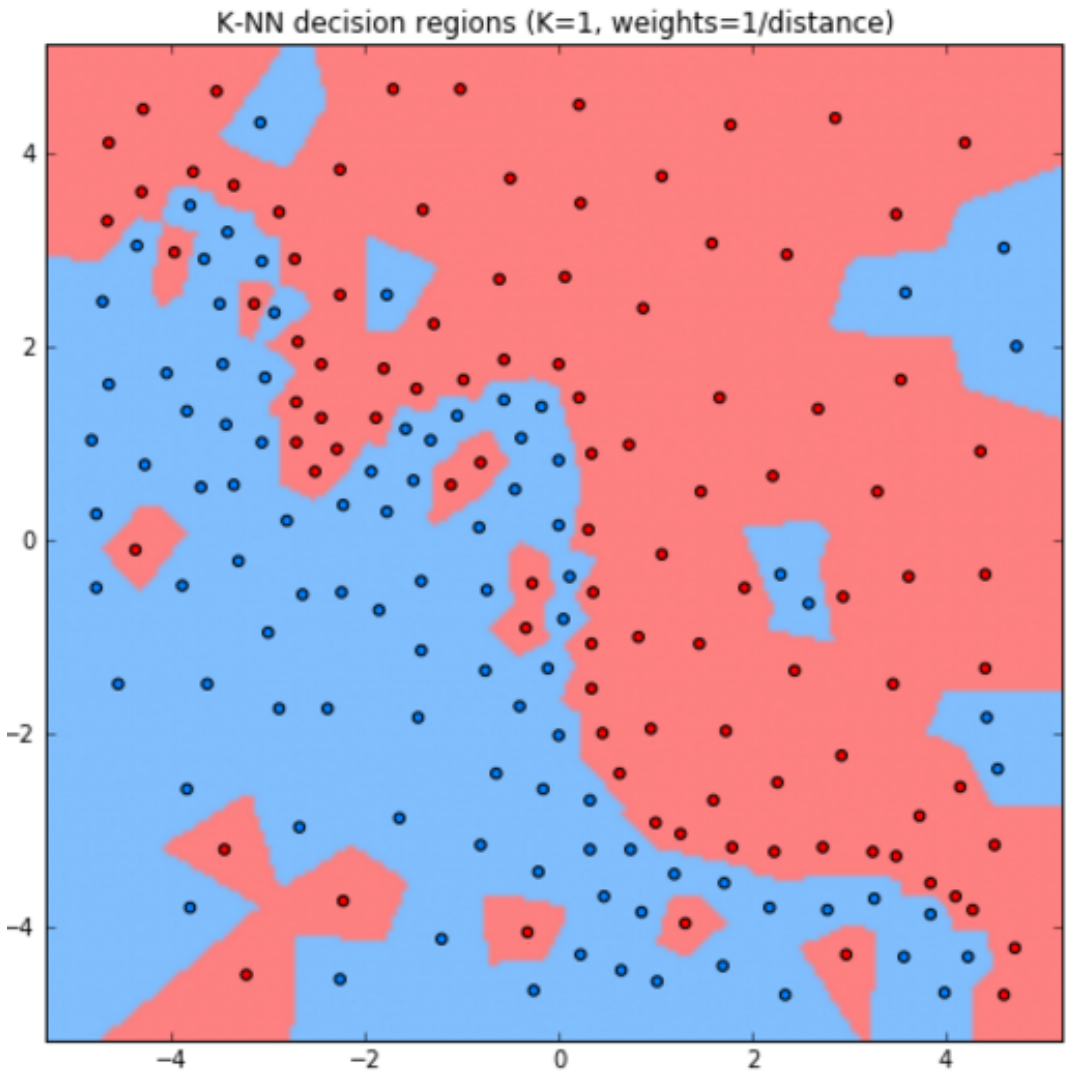
$$w_k = \frac{K+1-k}{K}$$

Зависимые от x:

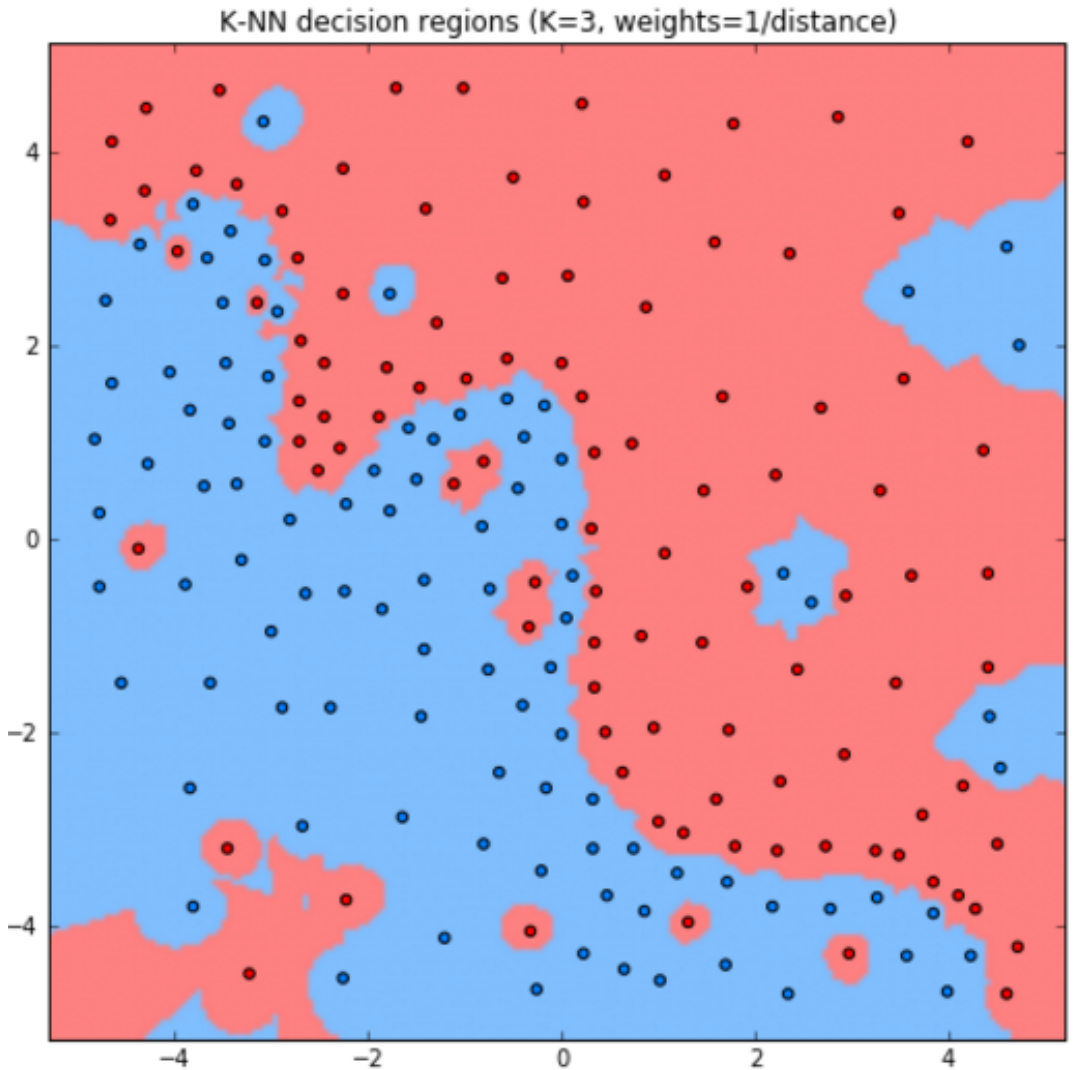
$$w_k = \begin{cases} \frac{\rho(z_K,x)-\rho(z_k,x)}{\rho(z_K,x)-\rho(z_1,x)}, & \rho(z_K,x) \neq \rho(z_1,x) \\ 1 & \rho(z_K,x) = \rho(z_1,x) \end{cases}$$

$$w_k = \frac{1}{\rho(z_k,x)}$$

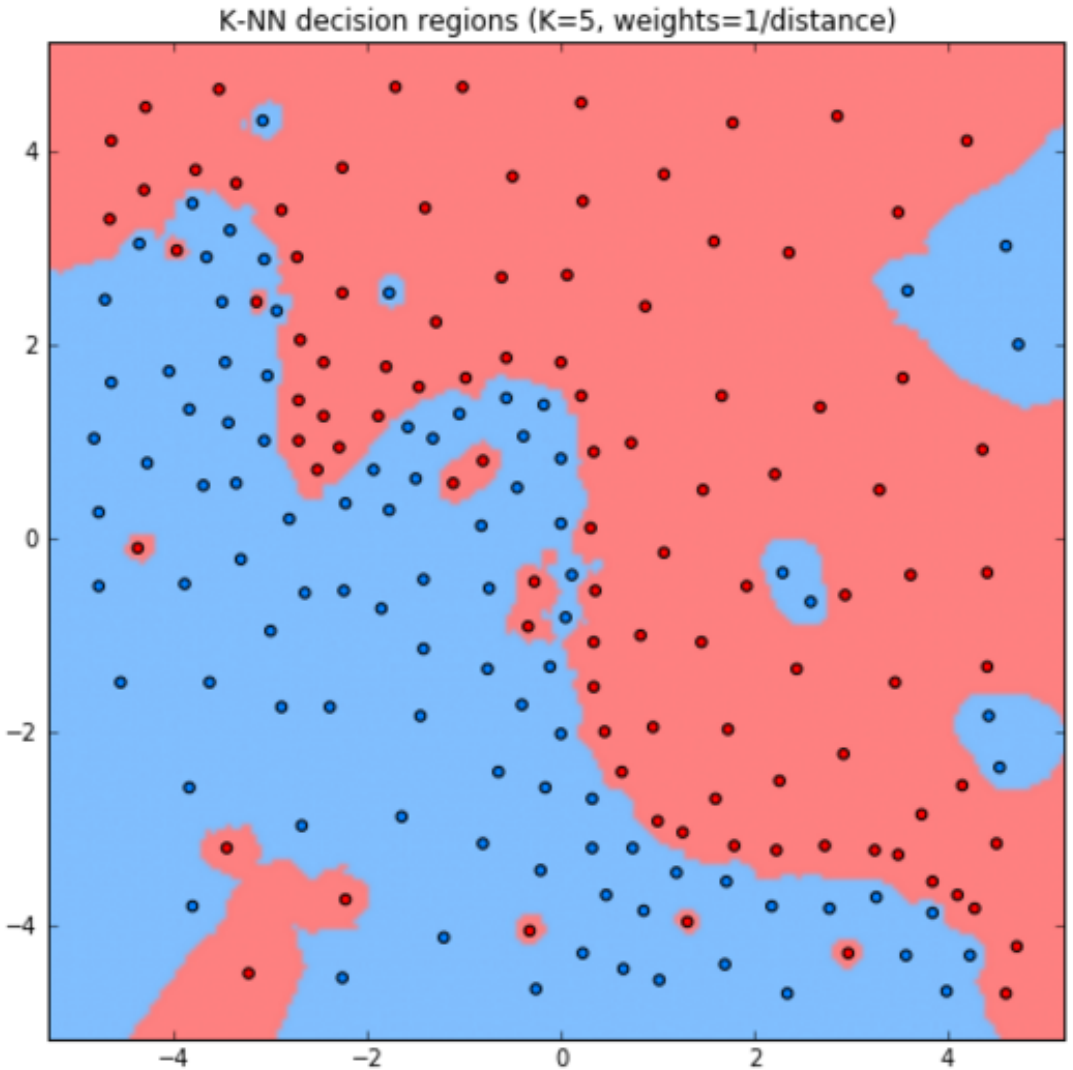
Пример: K-NN, классификация с весами



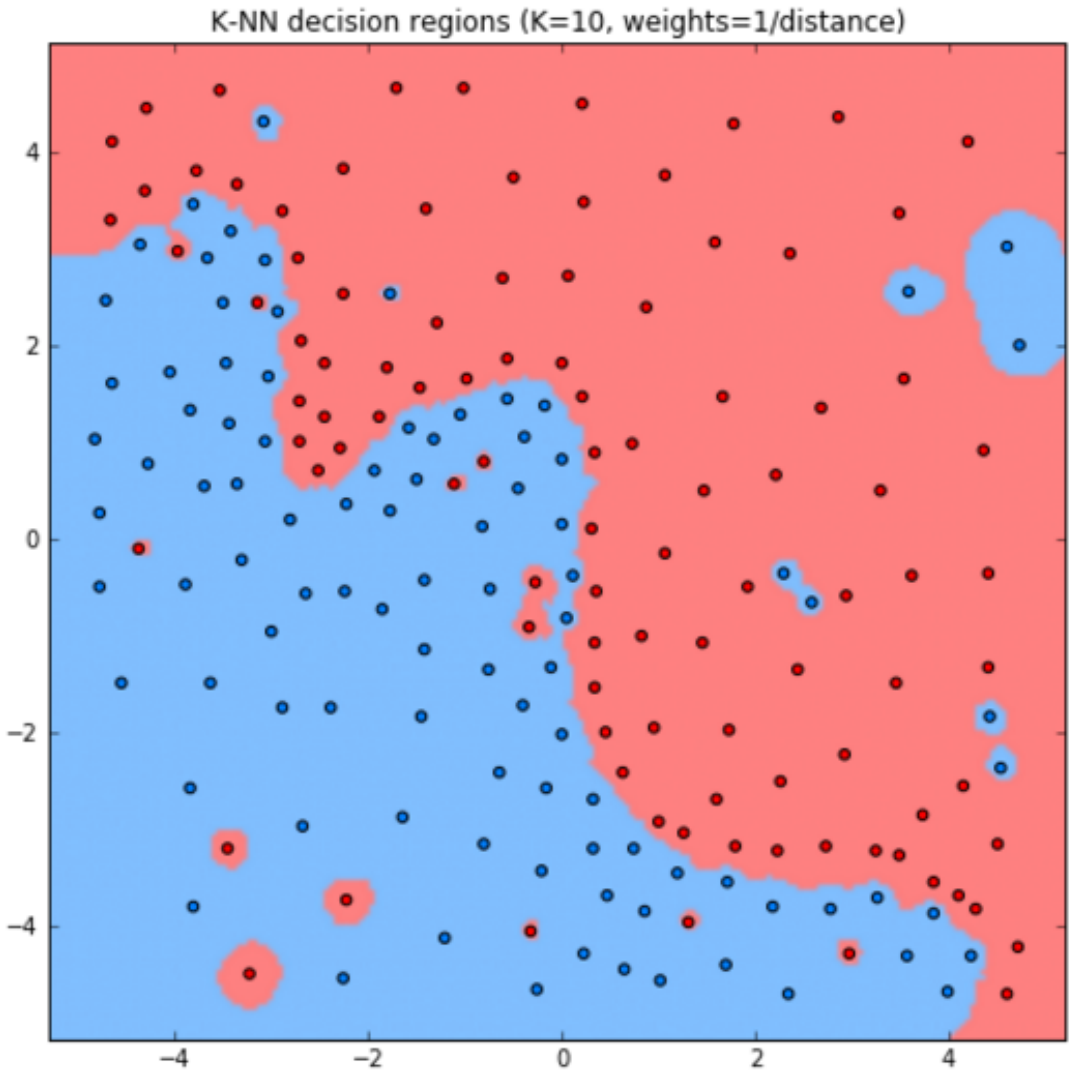
Пример: K-NN, классификация с весами



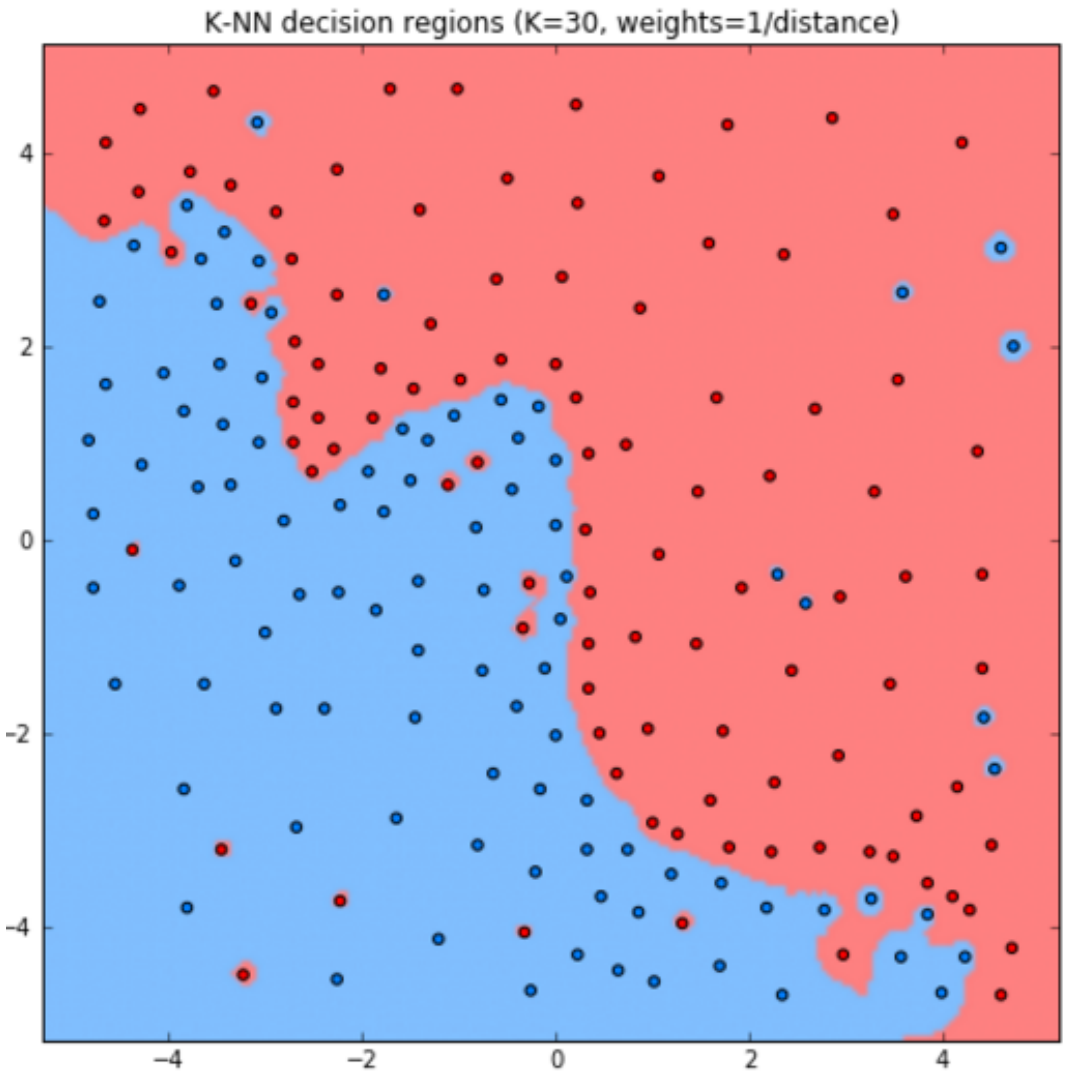
Пример: K-NN, классификация с весами



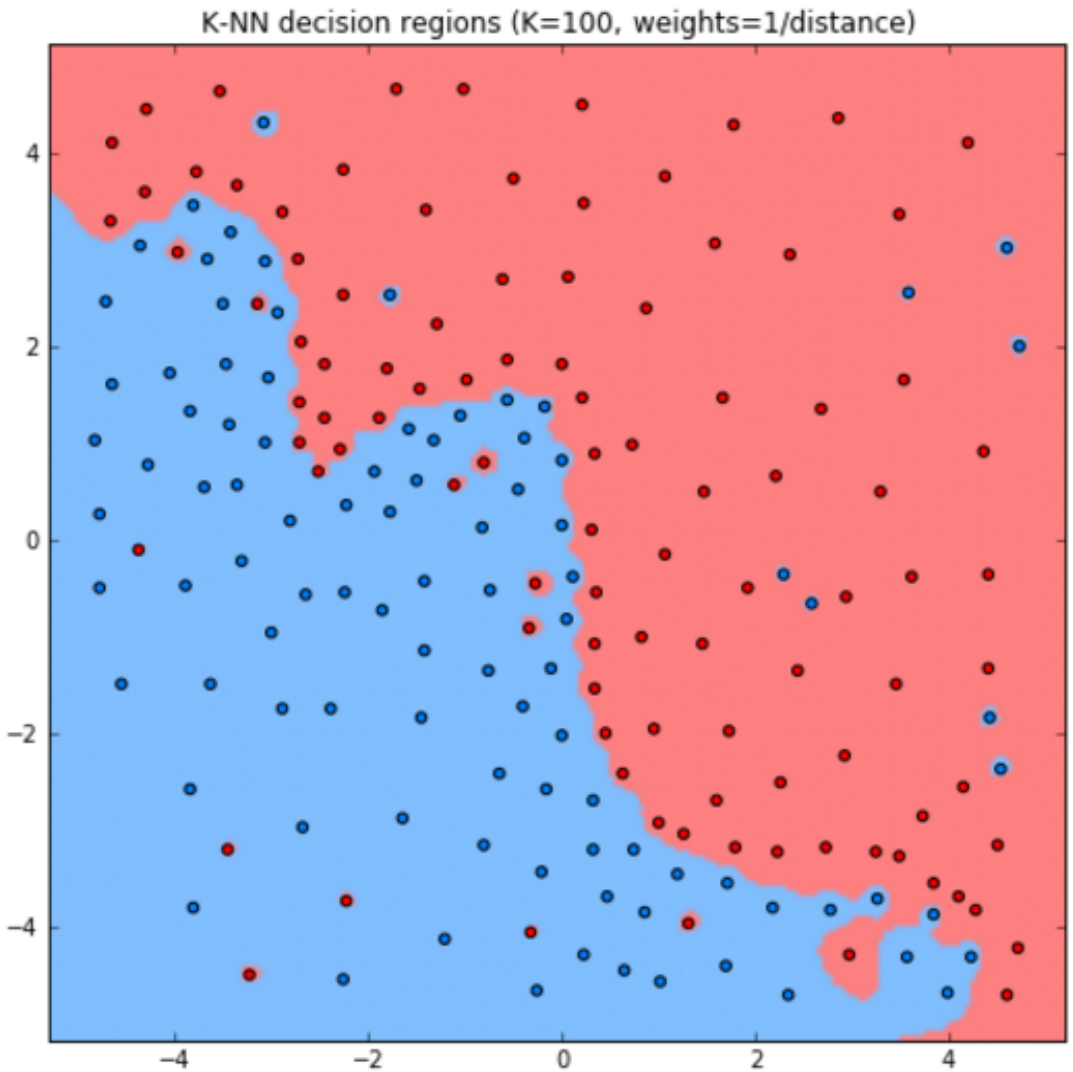
Пример: K-NN, классификация с весами



Пример: K-NN, классификация с весами



Пример: K-NN, классификация с весами



Пример: K-NN, регрессия с весами