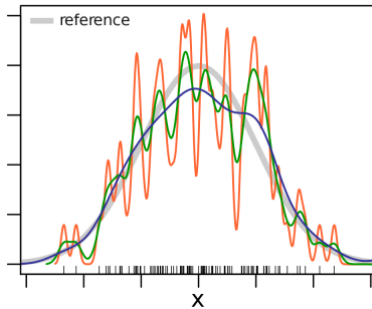


# Ядерная оценка плотности

Виктор Китов

[v.v.kitov@yandex.ru](mailto:v.v.kitov@yandex.ru)



# Содержание

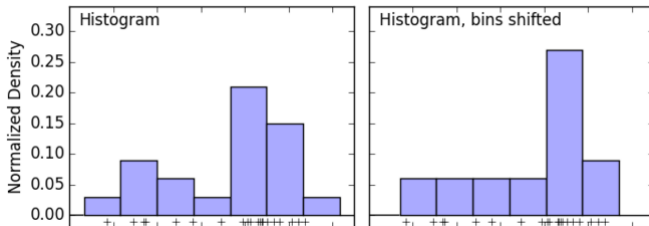
- 1 Одномерный случай
- 2 Многомерный случай
- 3 Правило максимальной апостериорной вероятности

# Непараметрическая оценка плотности

- Нужна непараметрическая оценка плотности.

# Непараметрическая оценка плотности

- Нужна непараметрическая оценка плотности.
- Гистограмма - решение, но она
  - кусочно-постоянная
  - выбор диапазона ячеек влияет на результат:



## Ядерная оценка плотности

Размещаем каждый блок в соответствующем объекте.

Индивидуальный блок

$$K(u) = \frac{1}{2} \mathbb{I}[|u| \leq 1]$$

Имеет свойства

$$K(u) \geq 0, \quad \int_{-\infty}^{+\infty} K(u) du = 1$$

## Ядерная оценка плотности

Размещаем каждый блок в соответствующем объекте.

Индивидуальный блок

$$K(u) = \frac{1}{2} \mathbb{I}[|u| \leq 1]$$

Имеет свойства

$$K(u) \geq 0, \quad \int_{-\infty}^{+\infty} K(u) du = 1$$

### Ядерная оценка плотности

Англ. Kernel density estimation (KDE)

$$\hat{p}(x) = \frac{1}{Nh} \sum_{i=1}^N K\left(\frac{x - x_i}{h}\right)$$

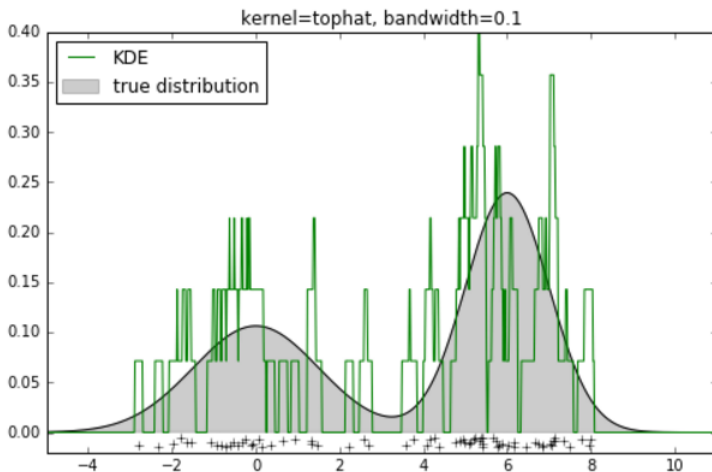
## Комментарии

- $K(u)$  называется *ядром*
- $K(u) = \frac{1}{2}\mathbb{I}[|u| \leq 1]$  - *прямоугольное (tophat) ядро*
- $h$  - параметр ширины окна (*bandwidth*)
- $h$  контролирует гладкость<sup>1</sup>

---

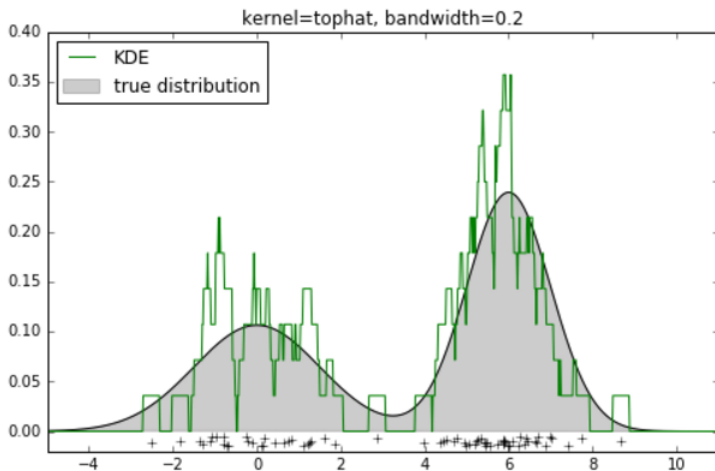
<sup>1</sup>каким образом?

## Пример: ядерная оценка с tophat ядром

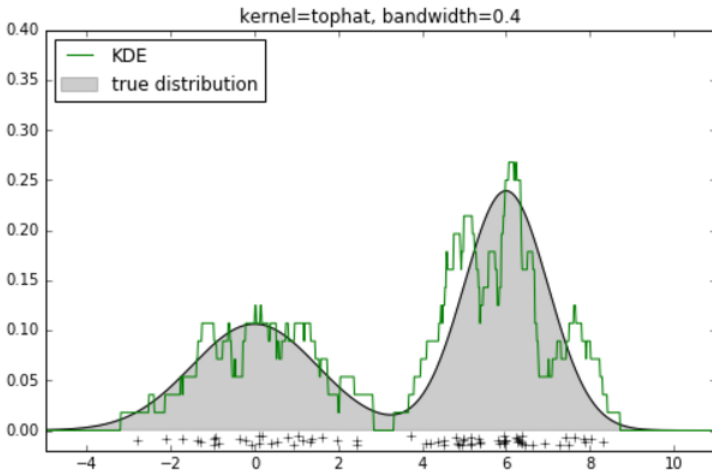




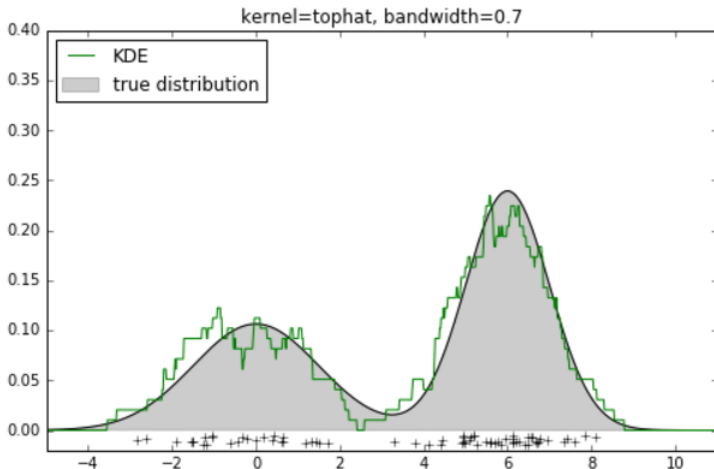
## Пример: ядерная оценка с tophat ядром



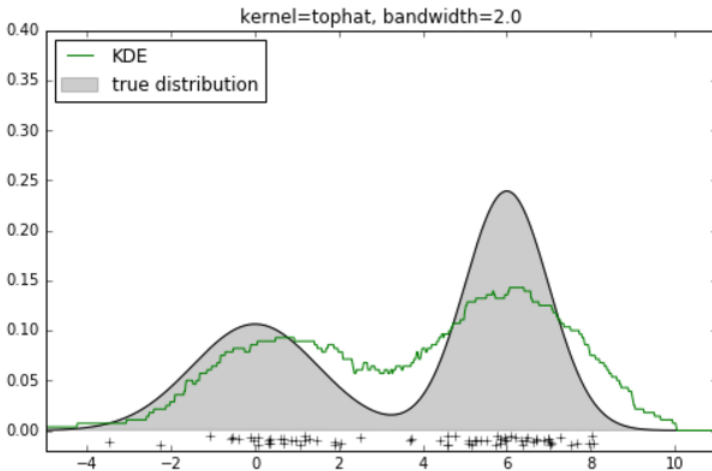
# Пример: ядерная оценка с tophat ядром



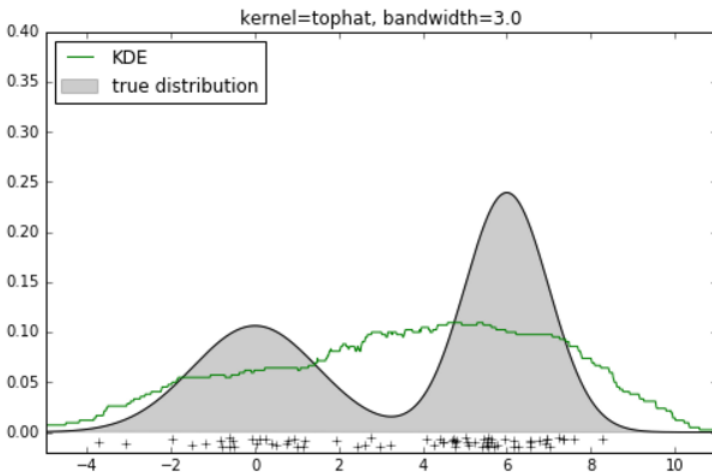
# Пример: ядерная оценка с tophat ядром



## Пример: ядерная оценка с tophat ядром



## Пример: ядерная оценка с tophat ядром



## Другие функции ядра

Проблемы top-hat ядра:

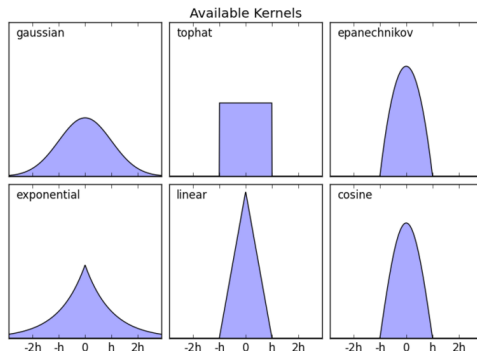
- Результирующая оценка  $\hat{p}(x) = \frac{1}{Nh} \sum_{i=1}^N K\left(\frac{x-x_i}{h}\right)$  кусочно-постоянна.
- Влияние точки  $x_i$  не изменяется с  $\rho(x, x_i)$  вблизи  $x$ .

## Другие функции ядра

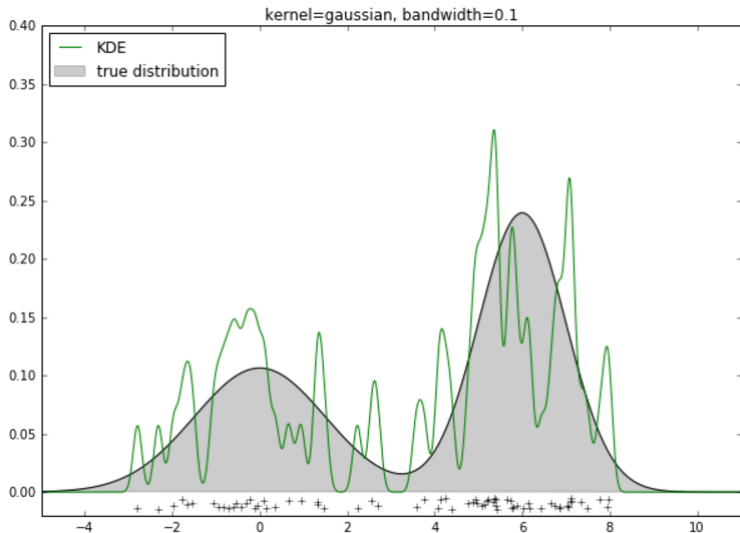
Проблемы top-hat ядра:

- Результирующая оценка  $\hat{p}(x) = \frac{1}{Nh} \sum_{i=1}^N K\left(\frac{x-x_i}{h}\right)$  кусочно-постоянна.
- Влияние точки  $x_i$  не изменяется с  $\rho(x, x_i)$  вблизи  $x$ .

Можем использовать гладкие унимодальные ядра<sup>2</sup>:

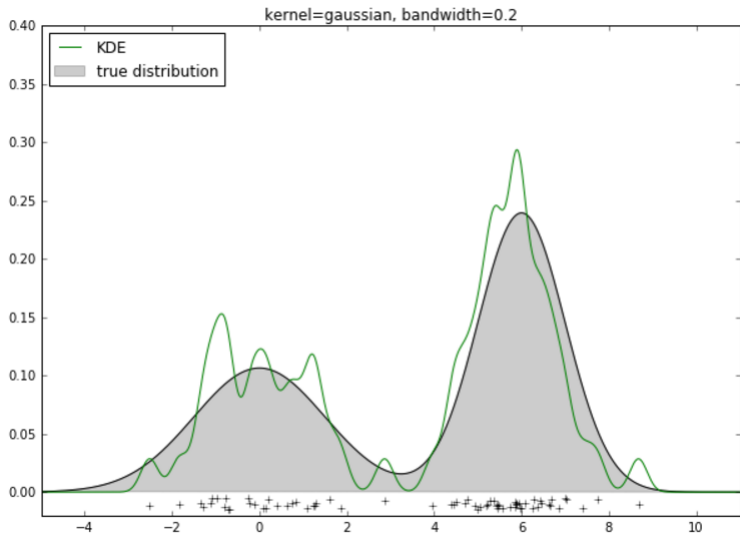


# Пример: Гауссово ядро

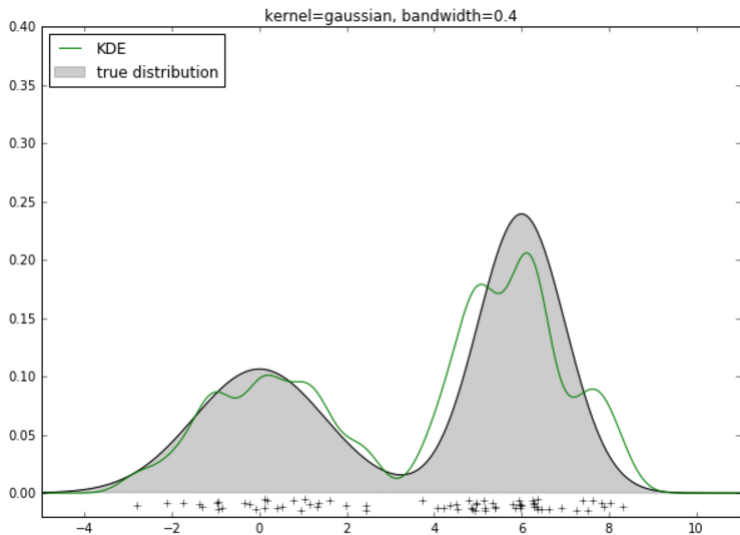




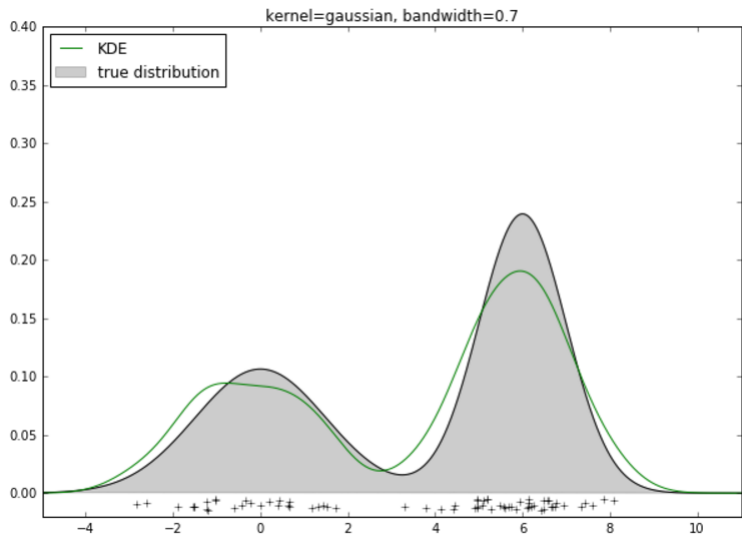
# Пример: Гауссово ядро



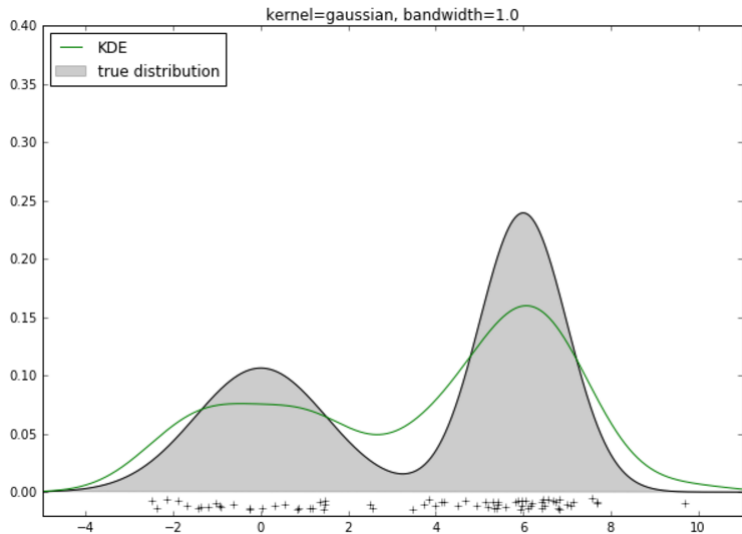
# Пример: Гауссово ядро



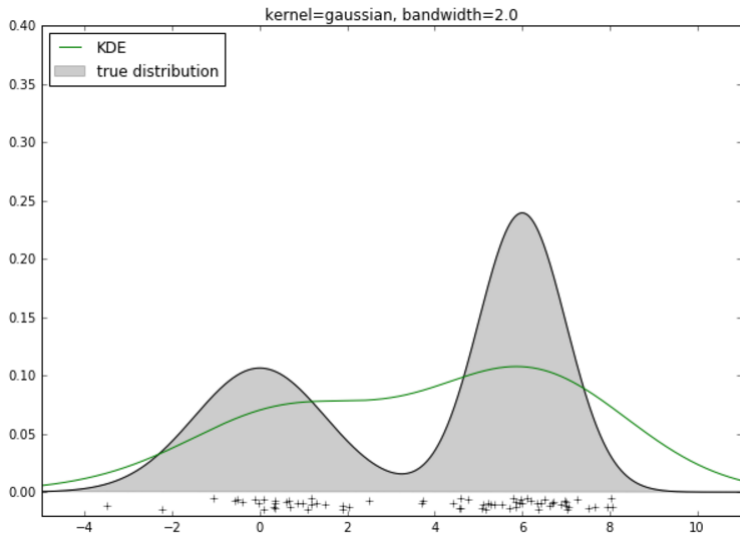
# Пример: Гауссово ядро



# Пример: Гауссово ядро



# Пример: Гауссово ядро



## Формулы основных ядер

название	формула $K(u)$
top-hat (прямоугольное)	$\frac{1}{2}\mathbb{I}[ u  \leq 1]$
Гауссово	$\frac{1}{\sqrt{2\pi}}e^{-\frac{u^2}{2}}$
квартическое	$\propto [(1 - u^2)]_+^2$
треугольное	$[1 -  u ]_+$
Епанечникова	$\propto [1 - u^2]_+$

---

<sup>3</sup>Как его выбрать?

## Формулы основных ядер

название	формула $K(u)$
top-hat (прямоугольное)	$\frac{1}{2}\mathbb{I}[ u  \leq 1]$
Гауссово	$\frac{1}{\sqrt{2\pi}}e^{-\frac{u^2}{2}}$
квартическое	$\propto [(1 - u^2)]_+^2$
треугольное	$[1 -  u ]_+$
Епанечникова	$\propto [1 - u^2]_+$

Комментарии:

- тип ядра влияет на гладкость, но не на точность аппроксимации.
- для точности важен правильный выбор ширины окна<sup>3</sup>.

<sup>3</sup>Как его выбрать?

## Условия сходимости оценки

### Условия сходимости оценки

Ядерная оценка плотности  $\hat{p}(x)$  сходится

$\mathbb{E}[(\hat{p}(x) - p(x))^2] \xrightarrow{N \rightarrow \infty} 0$  для  $\forall x$  при выполнении условий ниже.



## Условия сходимости оценки

### Условия сходимости оценки

Ядерная оценка плотности  $\hat{p}(x)$  сходится

$\mathbb{E}[(\hat{p}(x) - p(x))^2] \xrightarrow{N \rightarrow \infty} 0$  для  $\forall x$  при выполнении условий ниже.

Достаточные условия сходимости:

- Сходимость ширины окна:
- $\lim_{N \rightarrow \infty} h(N) = 0$ 
  - $\lim_{N \rightarrow \infty} Nh(N) = \infty$

## Условия сходимости оценки

### Условия сходимости оценки

Ядерная оценка плотности  $\hat{p}(x)$  сходится

$\mathbb{E}[(\hat{p}(x) - p(x))^2] \xrightarrow{N \rightarrow \infty} 0$  для  $\forall x$  при выполнении условий ниже.

Достаточные условия сходимости:

- Сходимость ширины окна:

- $\lim_{N \rightarrow \infty} h(N) = 0$

- $\lim_{N \rightarrow \infty} Nh(N) = \infty$

- Регулярность ядра:

- $\int |K(u)| du < \infty$

- $\int K(u) du = 1$

- $\sup_u K(u) < \infty$

- $\lim_{u \rightarrow \infty} |uK(u)| = 0$

# Содержание

- 1 Одномерный случай
- 2 Многомерный случай
- 3 Правило максимальной апостериорной вероятности

## Расширение на многомерный случай

Многомерные ядра:

$$\hat{p}(\mathbf{x}) = \frac{1}{Nh^D} \sum_{i=1}^N K\left(\frac{1}{h}(\mathbf{x} - \mathbf{x}_i)\right)$$

название	формула $K(\mathbf{u})$
Гауссово	$\frac{1}{(2\pi)^{D/2}} e^{-\frac{\mathbf{u}^T \mathbf{u}}{2}}$
Епанечникова	$\propto [1 - \mathbf{u}^T \mathbf{u}]_+$
произведение одномерных	$\prod_{d=1}^D K_{1D}\left(\frac{\mathbf{x}^d - \mathbf{x}_n^d}{h}\right)$

## Ядра, зависящие от расстояния

Ядра, основанные на расстоянии:

$$\hat{p}(\mathbf{x}) = \frac{1}{Nh^D} \sum_{i=1}^N K\left(\rho\left(\frac{\mathbf{x}}{h}, \frac{\mathbf{x}_i}{h}\right)\right)$$

название	формула $K(\rho(\mathbf{x}, \mathbf{x}_i))$
Гауссово	$\frac{1}{(2\pi)^{D/2}} e^{-\frac{\rho^2(\mathbf{x}, \mathbf{x}_i)}{2}}$
Епанечникова	$\propto [1 - \rho^2(\mathbf{x}, \mathbf{x}_i)]_+$

## Выбор ширины окна

### Выбор ширины окна

Чем чаще лежат точки, тем меньше должна быть ширина окна  $h$ .

---

<sup>4</sup>Чему будет равна оценка  $h$  на обучающей выборке?

## Выбор ширины окна

### Выбор ширины окна

Чем чаще лежат точки, тем меньше должна быть ширина окна  $h$ .

#### Постоянная ширина окна:

- $h = \frac{1}{N} \sum_{i=1}^N d_{iK}$ ,  $d_{iK}$ -расстояние от  $x_i$  до  $K$ -го ближайшего соседа
- Можно оценить  $h$  максимизацией правдоподобия на валидационной выборке<sup>4</sup>

#### Переменная ширина окна (для изменяющейся частоты точек):

- $h(x)$  = расстояние до  $K$ -го ближайшего соседа  $x$

---

<sup>4</sup>Чему будет равна оценка  $h$  на обучающей выборке?

# Содержание

- 1 Одномерный случай
- 2 Многомерный случай
- 3 **Правило максимальной апостериорной вероятности**



## Метод Парзеновского окна

Оценим  $p(x|y)$  ядерной оценкой плотности:

$$p(x|y) = \frac{1}{N_y h^D} \sum_{i:y_i=y} K\left(\frac{\rho(x, x_i)}{h}\right)$$

Байесовское решающее мин. ошибки правило дает **метод Парзеновского окна**:

$$\begin{aligned}\hat{y} &= \arg \max_y p(y|x) = \arg \max_y p(y)p(x|y) \\ &= \arg \max_y \frac{N_y}{N} \frac{1}{N_y h^D} \sum_{i:y_i=y} K\left(\frac{\rho(x, x_i)}{h}\right) \\ \hat{y}(x) &= \arg \max_y \sum_{i:y_i=y} K\left(\frac{\rho(x, x_i)}{h}\right)\end{aligned}$$

## Метод K-ближайших соседей

- Строим прогноз для  $x$ :
  - $K(u) = \mathbb{I}[|u| \leq 1]$
  - $h(x)$  = расстояние от  $x$  до  $K$ -го ближайшего соседа.
- Тогда метод Парзеневского окна дает метод  $K$ -ближайших соседей:

$$\hat{y}(x) = \arg \max_y \sum_{i: y_i = y} K \left( \frac{\rho(x, x_i)}{h(x)} \right) = \arg \max_y \sum_{i: \rho(x, x_i) \leq h(x)} \mathbb{I}[y_i = y]$$

- Сравнение подходов выбора  $h$ :
  - фиксированный  $h$ : усредняем по фиксированной окрестности вокруг  $x$ 
    - может попасть много или мало объектов, зато гарантия их близости
  - адаптивный  $h$  ( $K$  ближайших соседей): всегда усредняем по  $\geq K$  ближайшим точкам
    - достаточная статистика для усреднения, но объекты могут оказаться далекими и не репрезентативными

## Заключение

- Ядерная оценка плотности - непараметрический метод оценки плотности.
- Выбор  $\phi$ -ции ядра контролирует непрерывность и дифференцируемость плотности.
- Ширина окна контролирует сложность (гибкость) модели.
- Ширина окна должна быть
  - больше для разреженных областей
  - меньше для густо наполненных областей
- Метод Парзенковского окна и  $K$  ближайших соседей - частные случаи правила минимальной ошибки и ядерной оценки плотности.