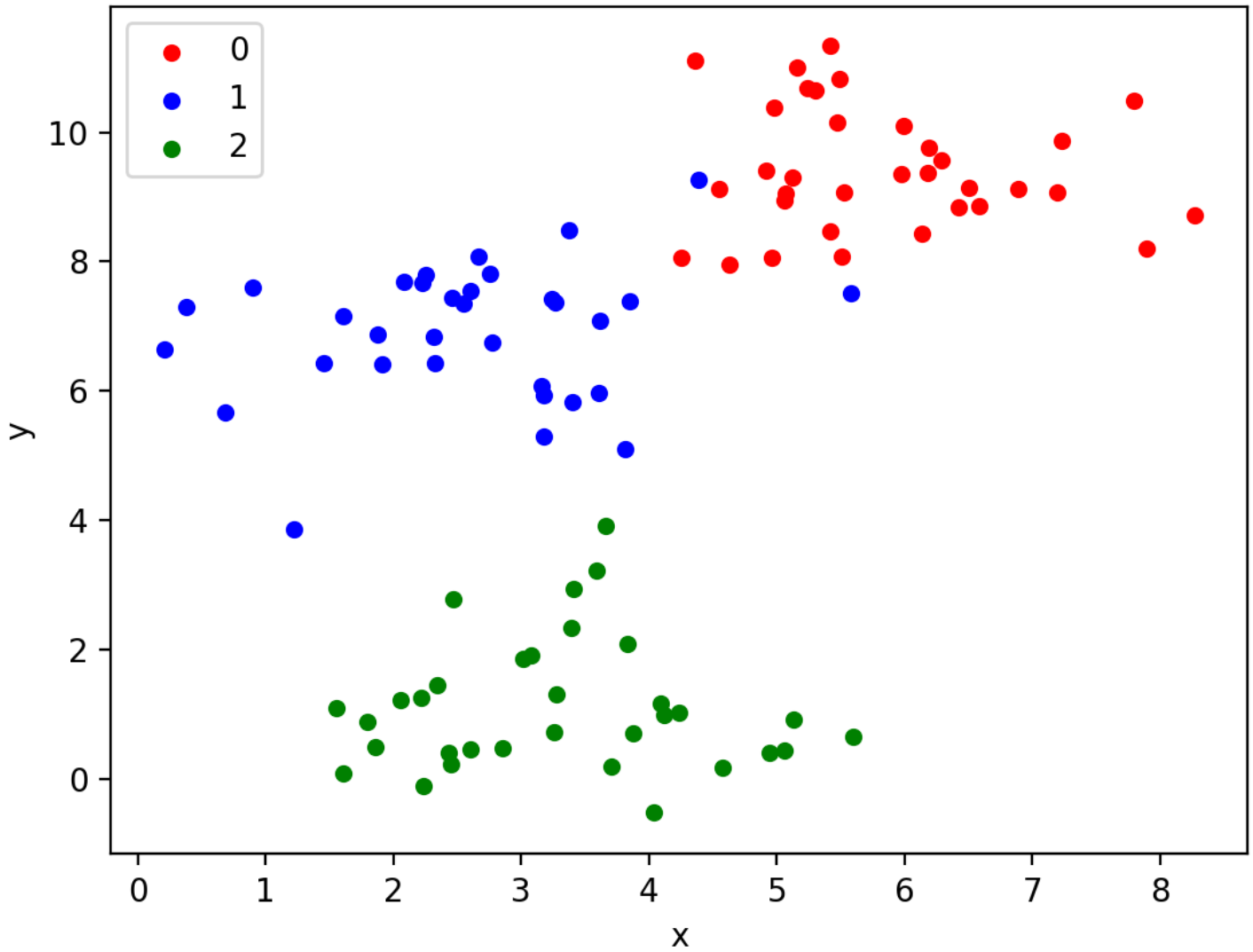
Метрические методы классификации

Патакин Николай

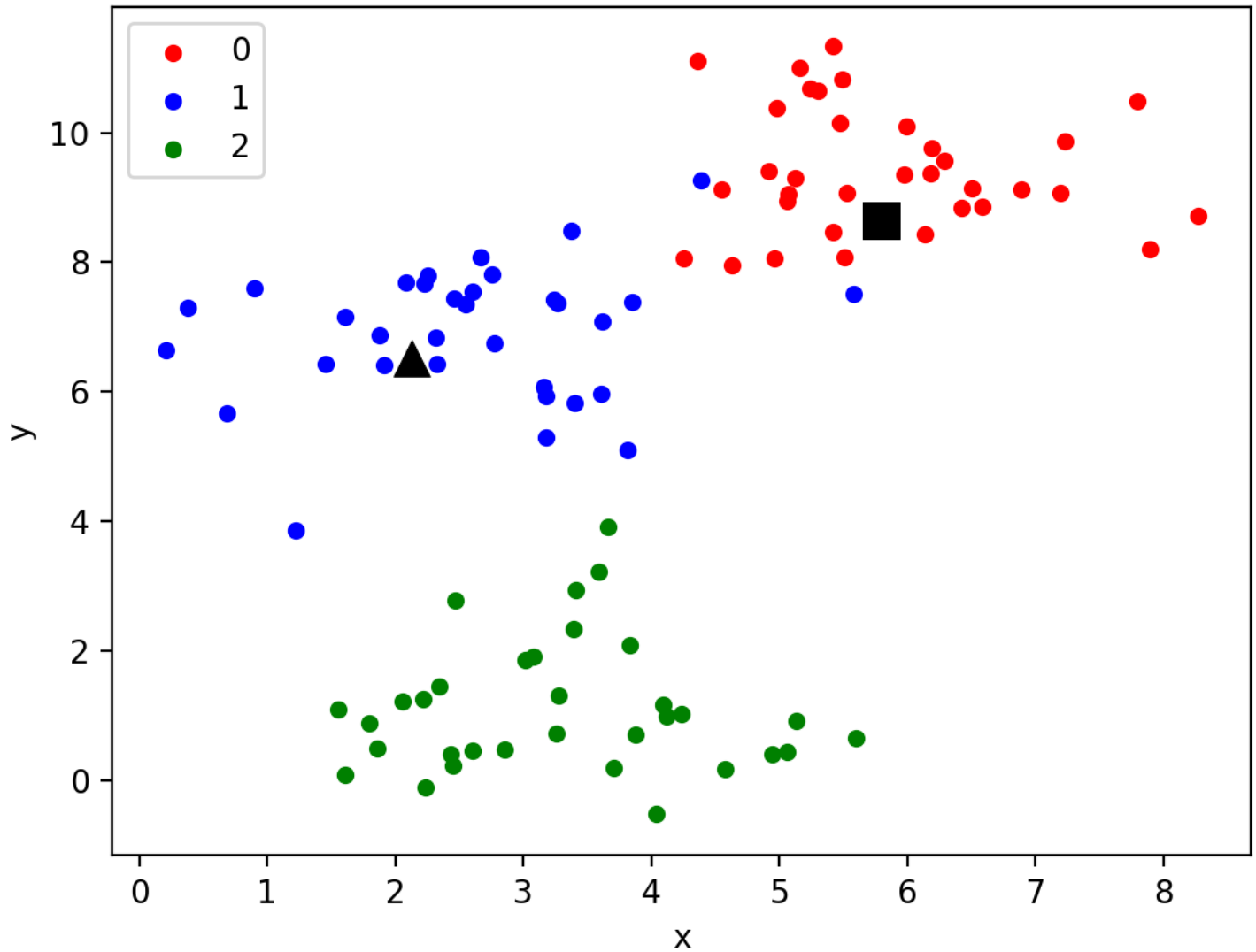
НИУ ВШЭ

2 ноября 2018 г.

Введение



Введение



Case based reasoning

Примеры подобных рассуждений можно найти в

- медицине
- юриспруденции
- геологии

Говорят о «прецендентах» - примерах, схожих по каким-либо критериям с рассматриваемым.

Гипотеза о компактности

В данном признаковом пространстве объекты одного и того же класса находятся близко друг к другу.

Формализация задачи метрической классификации

Имеется пространство объектов X, множество классов Y.

Определена метрика $\rho : X \times X \to [0; +\infty)$
Выборка $X^l = \{(x_i, y_i)\}, x_i \in X, y_i \in Y$

Заметим, что относительно произвольного объекта $u \in X$ можно упорядочить объекты выборки:

$$\rho(x_u^{(1)}, u) \leq \rho(x_u^{(2)}, u) \leq \cdots \leq \rho(x_u^{(l)}, u)$$

Тогда:

$x_u^{(i)}$ - i-тый ближайший сосед по отношению к объекту u,
$y_u^{(i)}$ - истинный ответ на нем

Формализация задачи метрической классификации

Классификатор, который определяет принадлежность объекта к классу по формуле

$$a(u) = arg \max_{y \in Y} \sum_{i=1}^{l} [y_u^{(i)} = y] \cdot w(i, u)$$

называется **обобщенным метрическим классификатором**

$w(i, u)$ – вес i-того ближайшего соседа для объекта u

1NN Classifier

Классификатор на основе единственного ближайшего соседа:

$$w(i, u) = [i = 1]$$

Плюсы:

- Интерпретация ответа. Легко можно ответить на вопрос: «Почему?», т.к. в ответ можно предъявить единственный прецедент
- Легкость в написании такого классификатора
- Сравнительно большая скорость работы

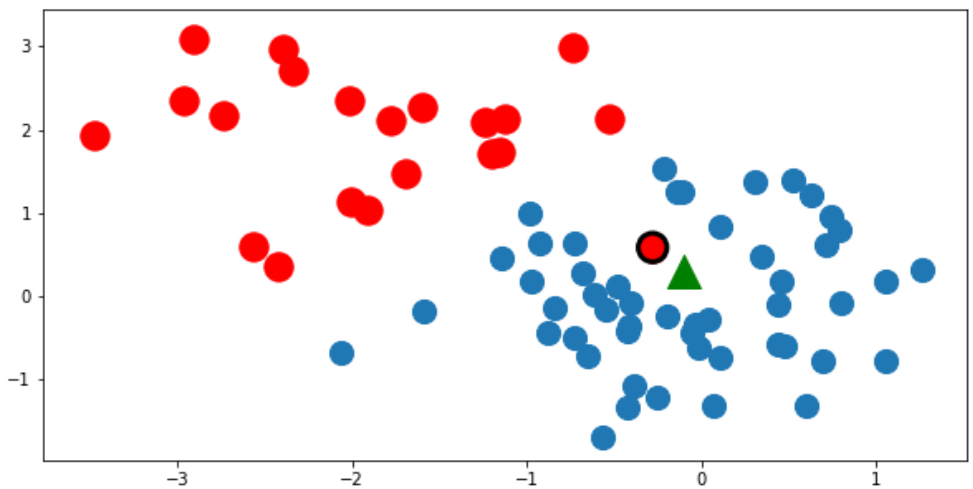
1NN Classifier

Классификатор на основе единственного ближайшего соседа:

$$w(i, u) = [i = 1]$$

Минусы:

- Неустойчивость классификации. Чувствительность к выбросам



$\Rightarrow$ Низкое качество классификации

- Отсутствие настраиваемых параметров

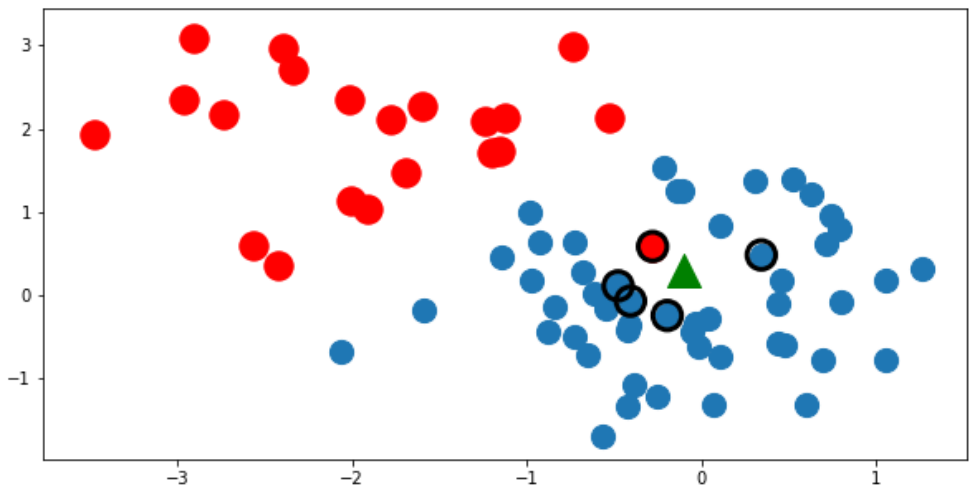
KNN classifier

Введем параметр k – число соседей, на основе которых принимается решение:

$$w(i, u) = [i \leq k]$$

Ответом будет класс, преобладающий по количеству среди k ближайших соседей

При $k = 5$:



$\Rightarrow$ относительная устойчивость к шуму

Проблемы KNN классификатора

- Оказалось, что одинаковое число ближайших соседей из числа k принадлежат двум классам
  Решение: Будем делать k нечетным

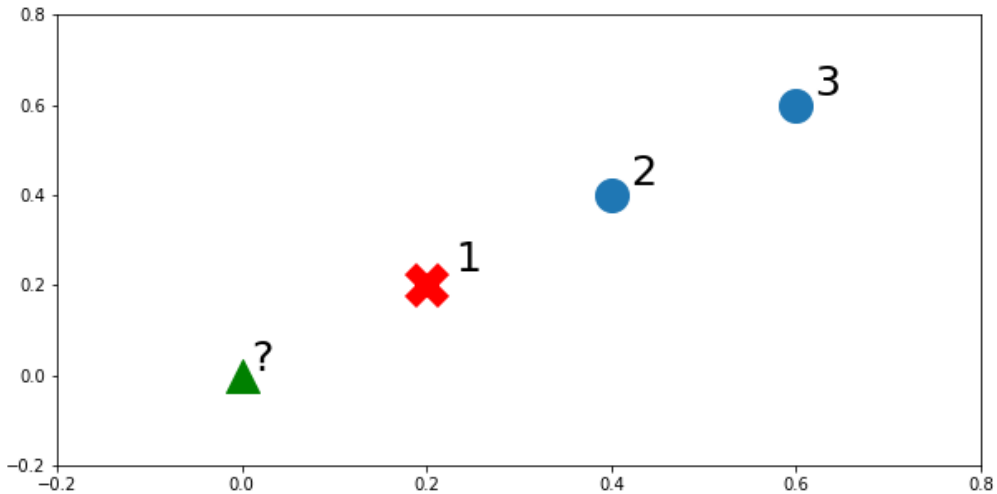
Проблемы KNN классификатора

- *Оказалось, что одинаковое число ближайших соседей из числа k принадлежат двум классам*
  Решение: Будем делать k нечетным
- *А при многоклассовой классификации? Ведь при количестве классов больше двух неопределенности все еще могут возникать*
  Решение: Вместо того, чтобы делать веса еденицами, сделаем их линейно убывающими от номера соседа:

$$w(i, u) = 1 - \frac{i-1}{k}, i \in [1; k]$$

Проблемы KNN классификатора

- Следующая проблема. Опять неопределенность!



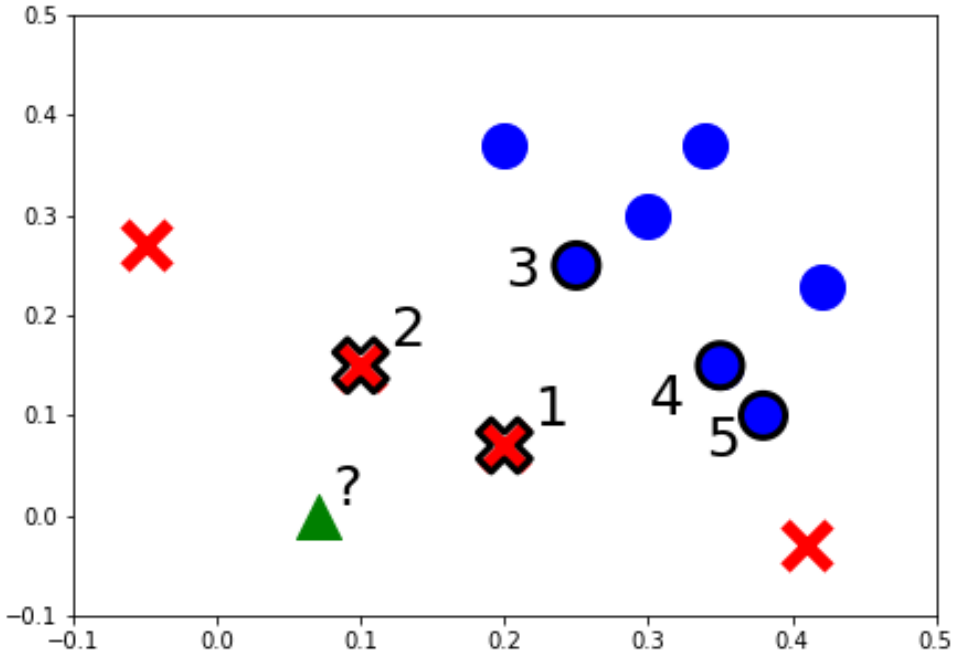
$$a(u) = 0 : \frac{1}{3} + \frac{2}{3} = \frac{3}{3}$$

$$a(u) = 1 : \qquad \frac{3}{3}$$

Решение: Стоит задаться вопросом: А почему веса соседей должны зависеть именно от порядкового номера?

От чего же должны зависеть веса?

Класс «красных» здесь отделяет интересующий нас объект от класса «синих». Иначе говоря, здесь важнее не то, что 3/5 соседей оказались синими, а то что красные объекты ближе.



$\Rightarrow$ вес должен зависеть от **значения метрики**

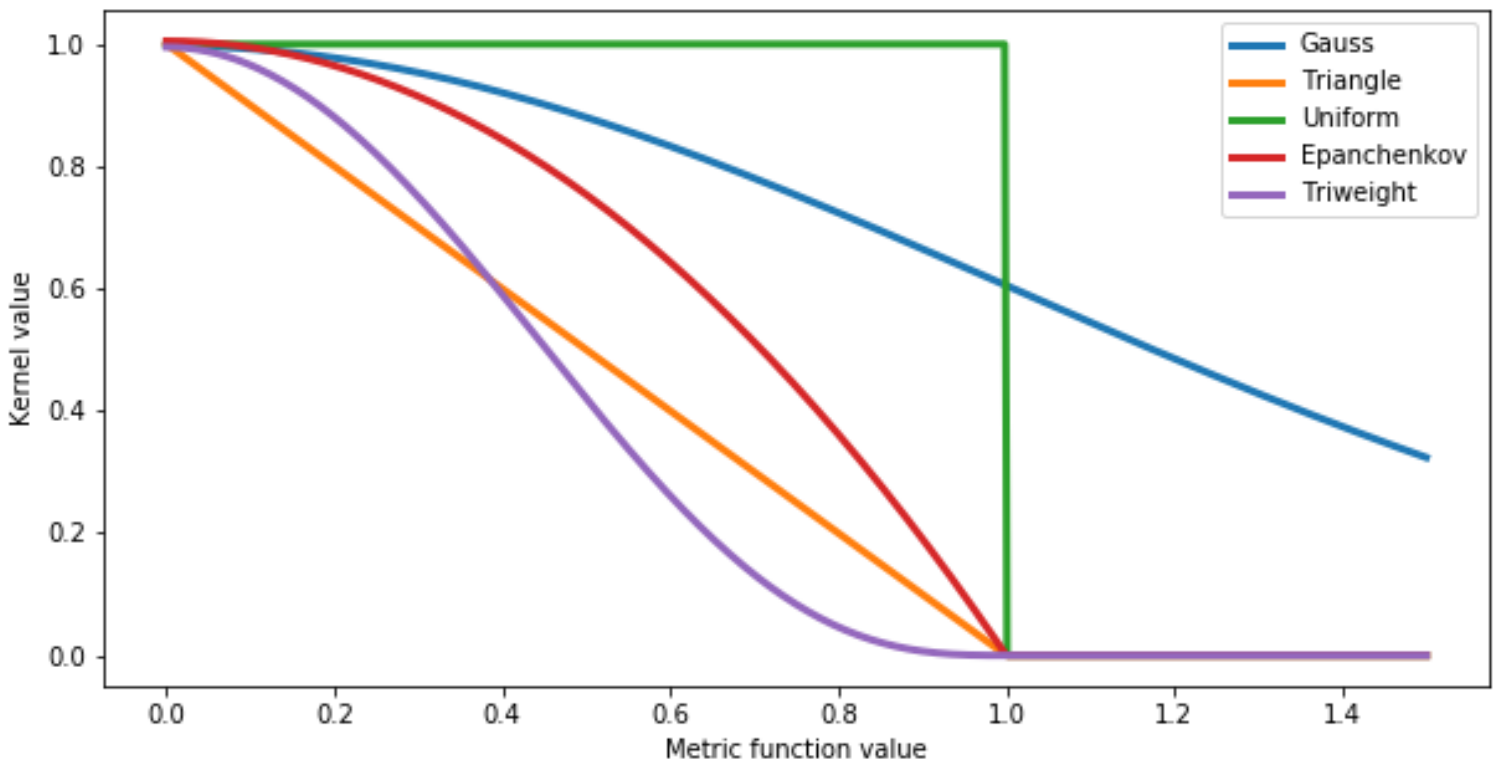
Метод парзеновского окна

$$w(i,u) = [i \leq k] \cdot K\left(\frac{\rho(x_u^{(i)}, u)}{h}\right) \qquad (1)$$

где k – количество значимых соседей
$K(r)$ – ядерная функция
$\rho(x_u^{(i)}, u)$ – расстояние от объекта до его i-того соседа
h – ширина окна

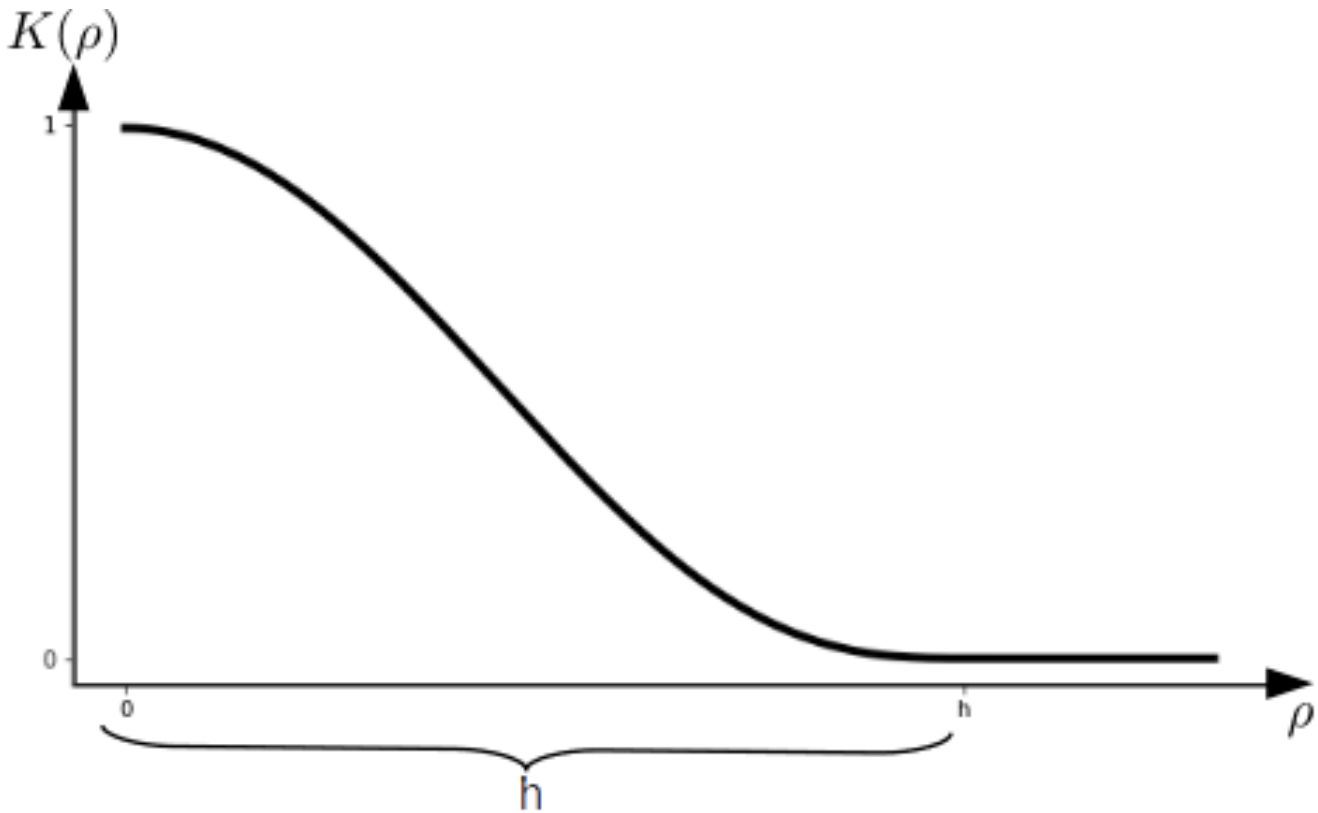
Выбор ядерной функции

Для большинства задач метрической классификации чаще всего достаточно, чтобы выбранное ядро было неотрицательной веществennозначной невозрастающей функцией.



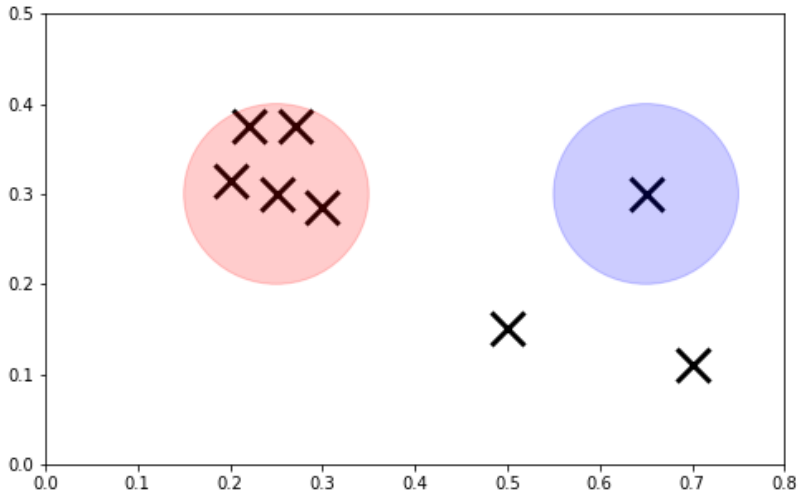
Влияние параметра h

В случае, если ядерная функция проходит через точку $(1; 0)$, то h – определяет максимальное значение метрики, значимое для классификатора. Иначе говоря растягивает ядерную функцию, задает «ширину».



Какой выбрать ширину окна?

- Сделать её константной



$\Rightarrow$

Проблема:
Участки с низкой плотностью могут не содержать достаточное количество соседей в области с фиксированным радиусом

- Выбрать ее так, чтобы в нее помещалось ровно k соседей

$$w(i,u) = [i \leq k] \cdot K\left(\frac{\rho(u, x_u^{(i)})}{\rho(u, x_u^{(k+1)})}\right)$$

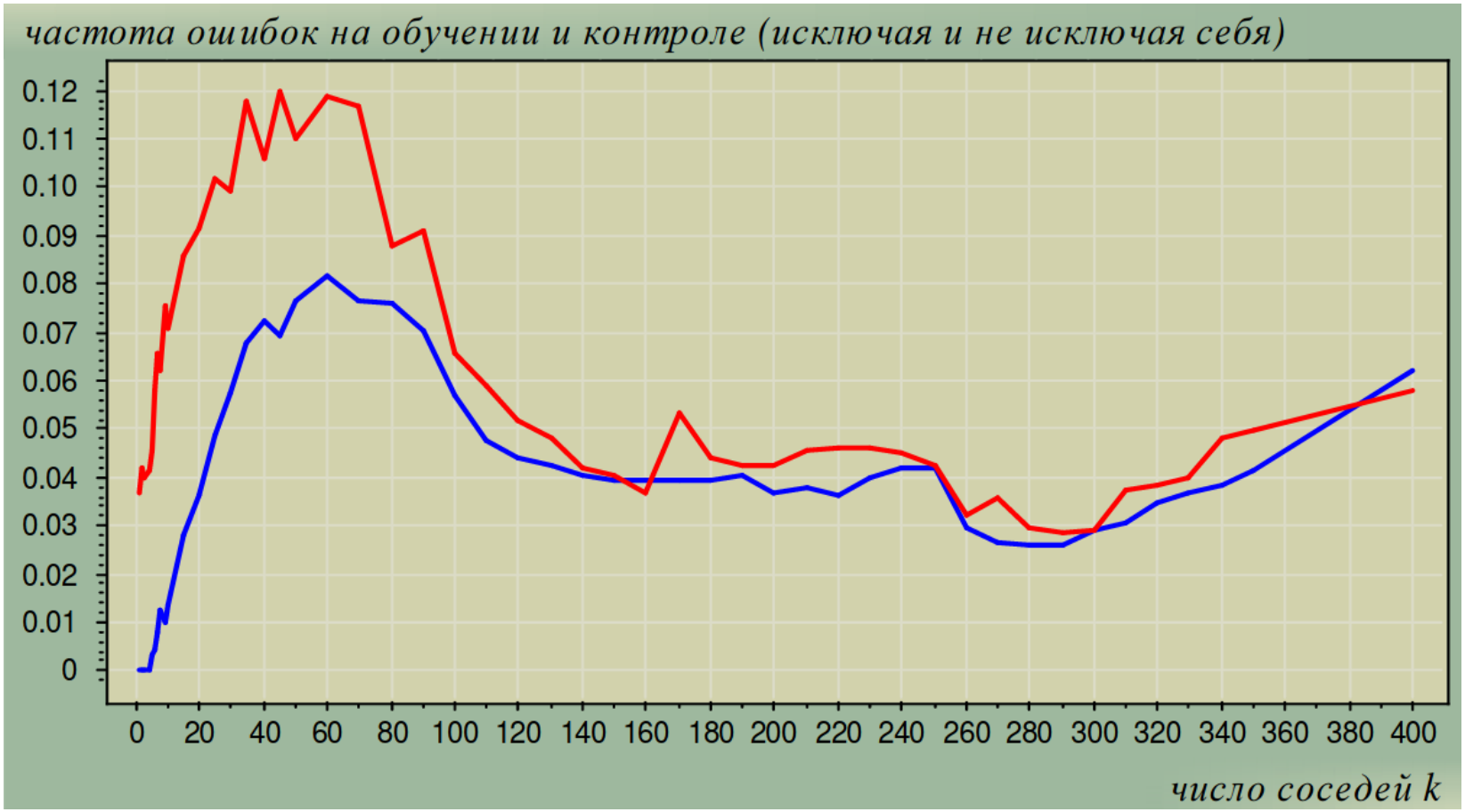
Метод потенциальных функций

$$w(i,u) = [i \leq k] \cdot \psi_{(i)} \cdot K\left(\frac{\rho(x_u^{(i)},u)}{h_{(i)}}\right) \qquad (2)$$

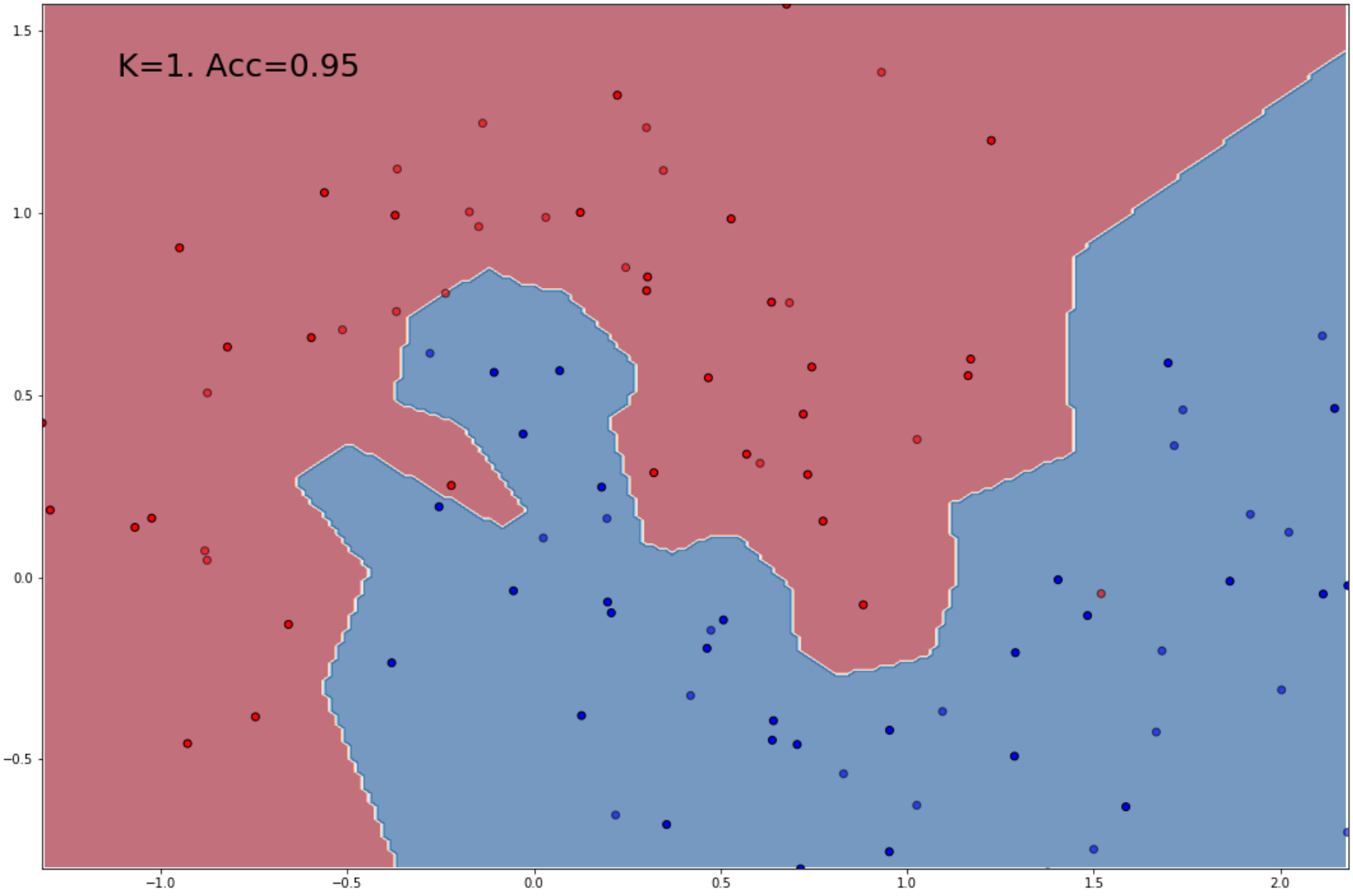
где k – количество значимых соседей
$K(r)$ – ядерная функция
$\rho(x_u^{(i)}, u)$ – расстояние от объекта до его i-того соседа
$h_{(i)}$ – ширина окна для i-того элемента
$\psi_{(i)}$ – вес i-того элемента

Веса $\psi_{(i)}$ можно назначать в соответствии с достоверностью элемента выборки. Таким образом, получаем метод, устойчивый к шуму.

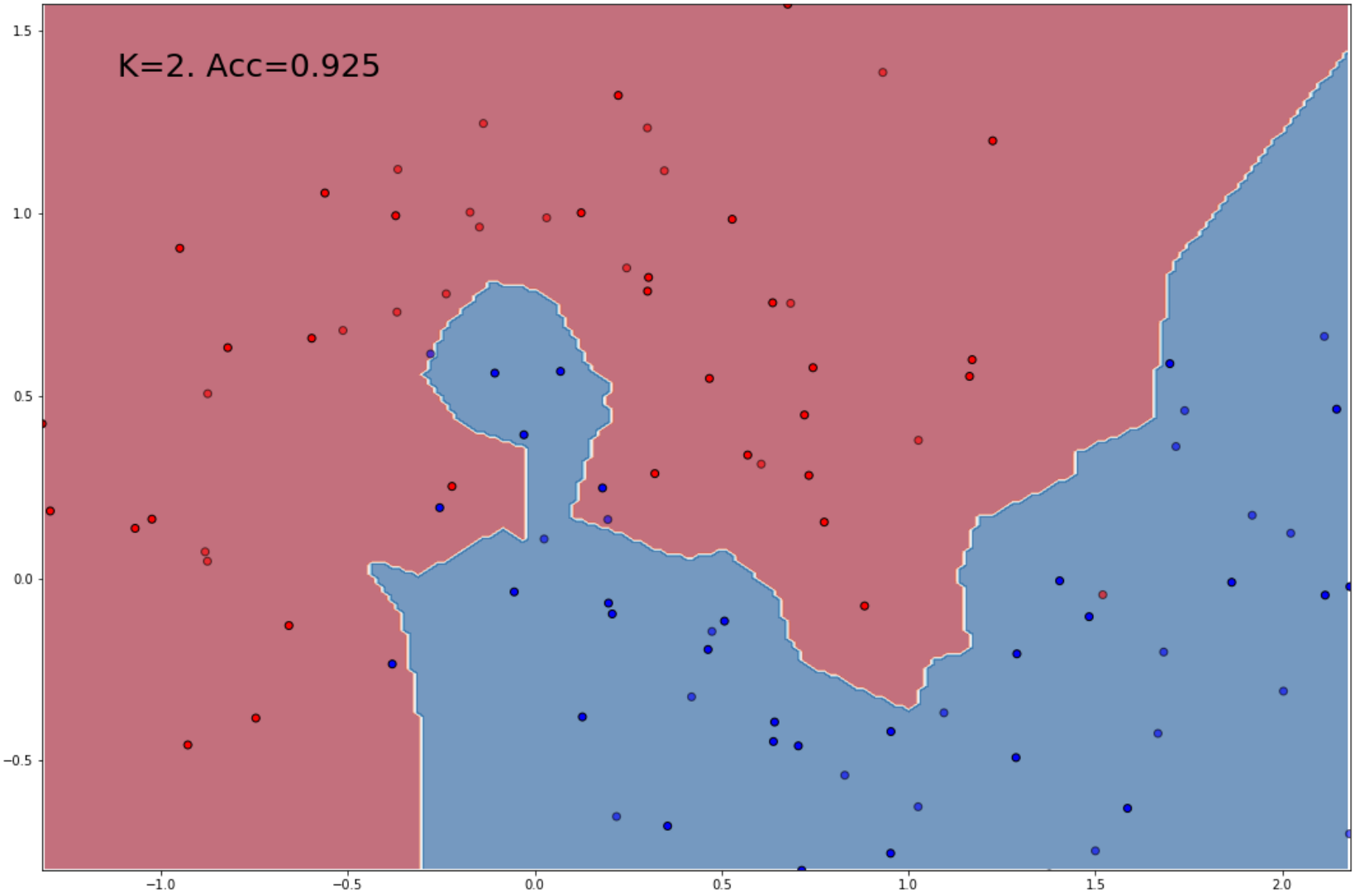
Как выбирать число соседей - k?



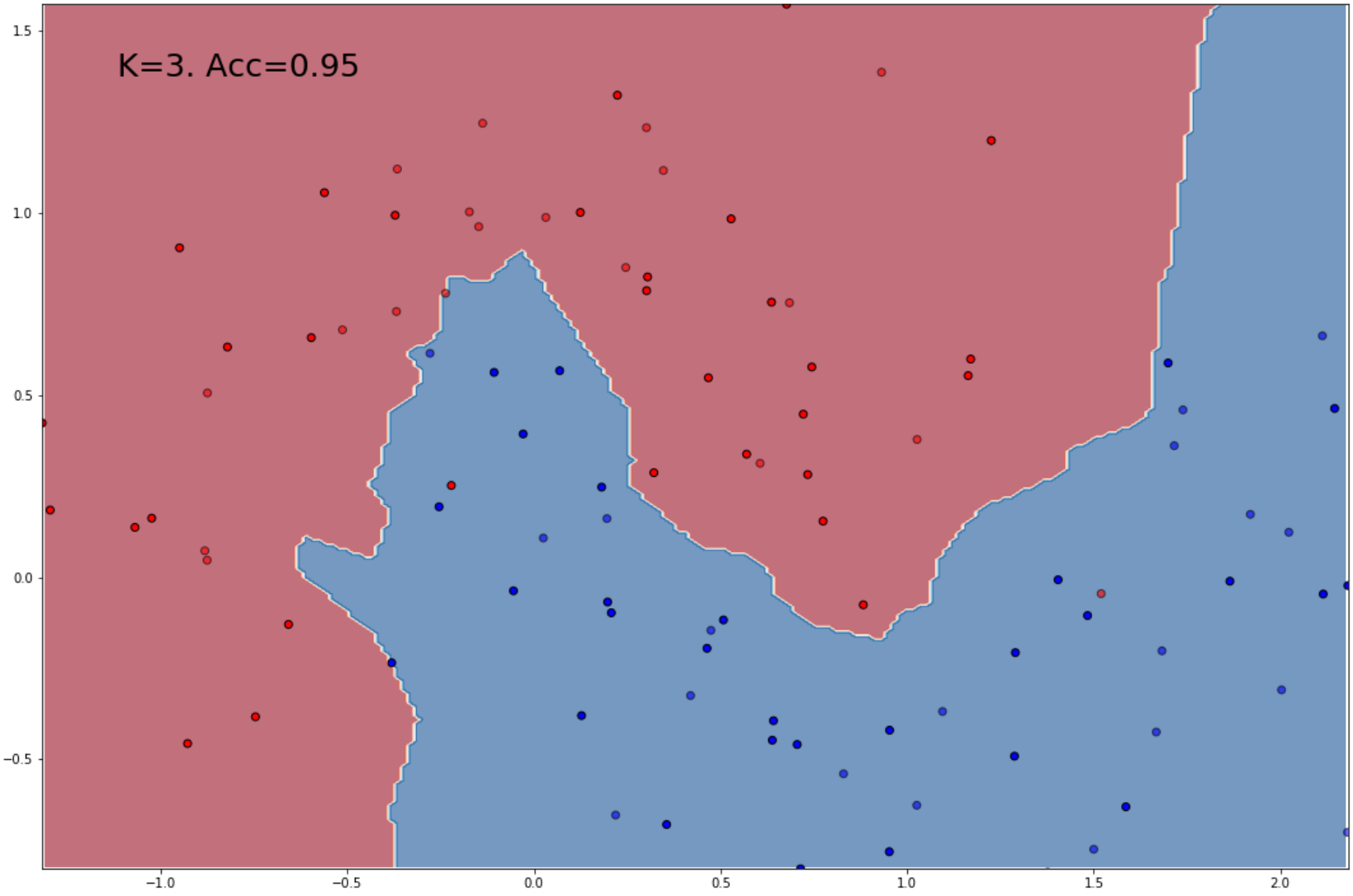
Визуализация разделяющих областей



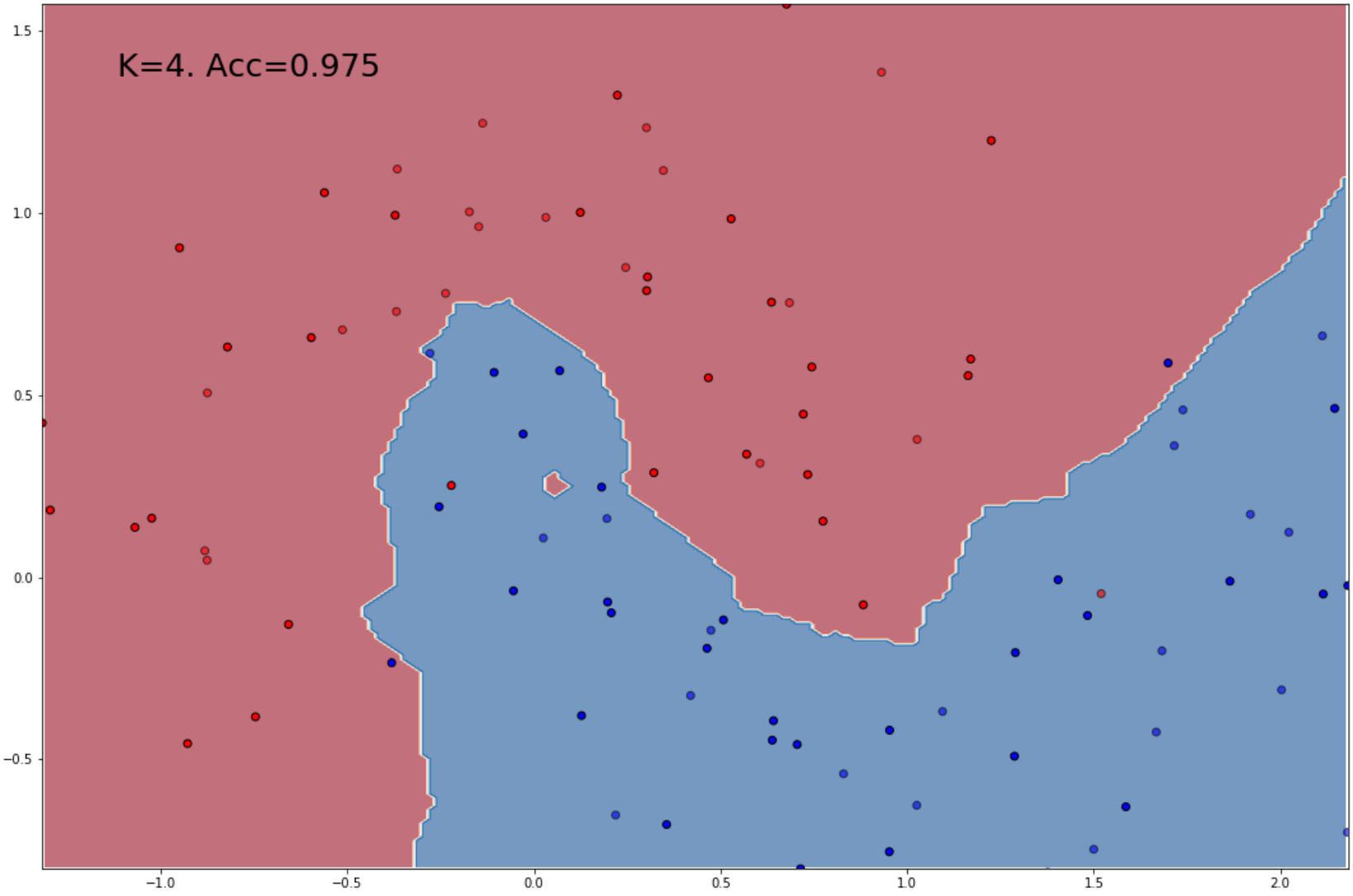
Визуализация разделяющих областей



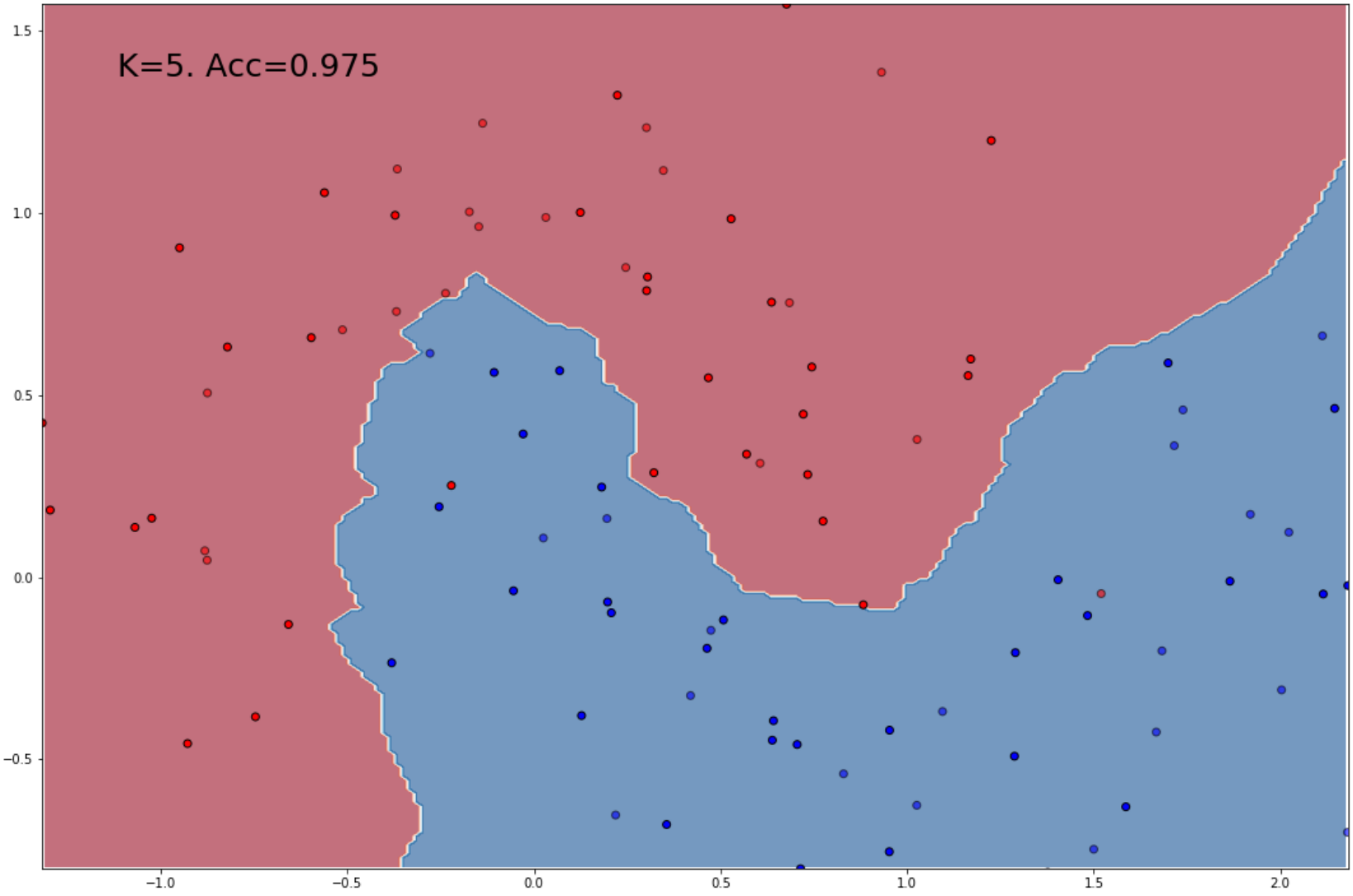
Визуализация разделяющих областей



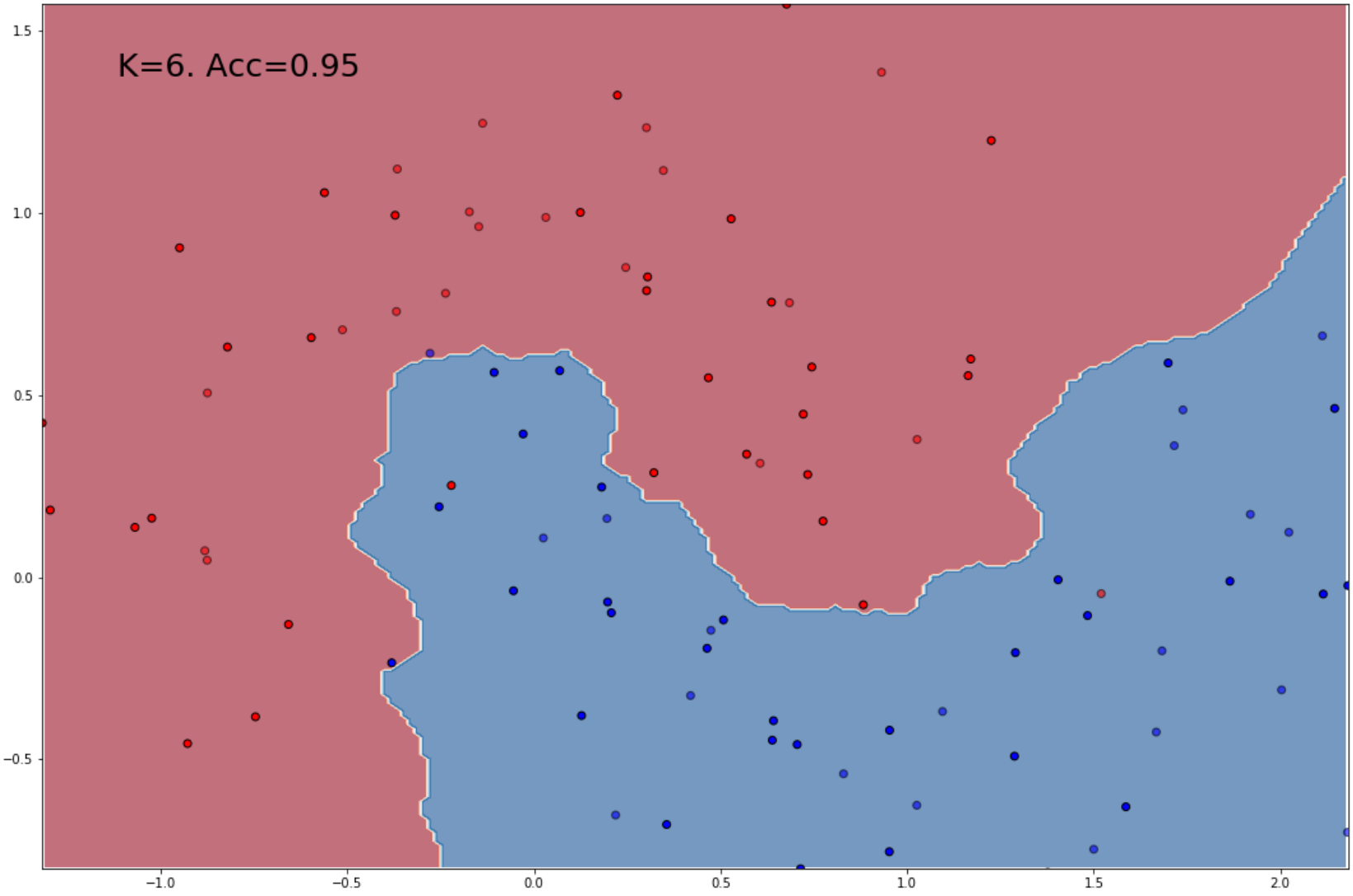
Визуализация разделяющих областей



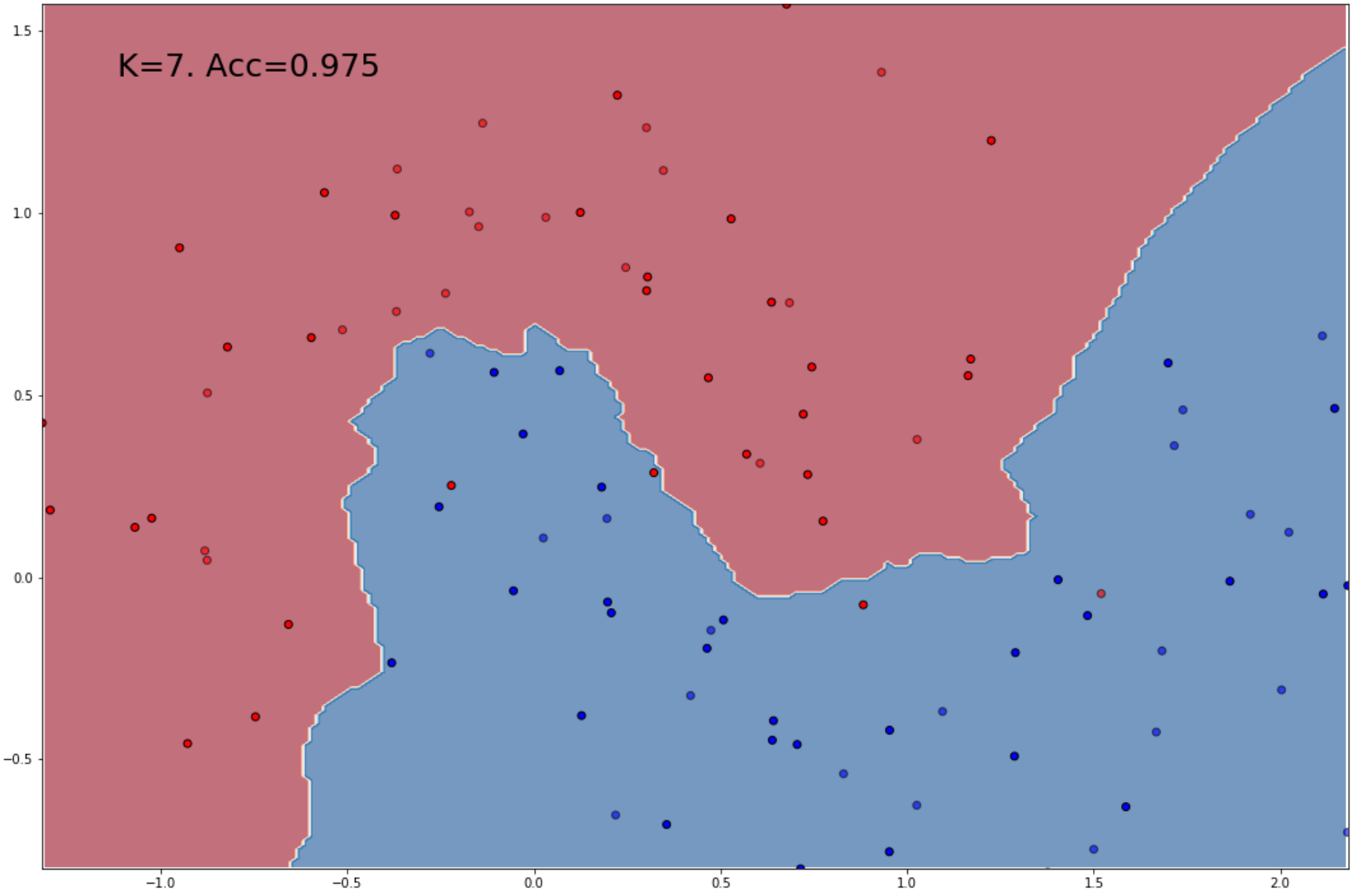
Визуализация разделяющих областей



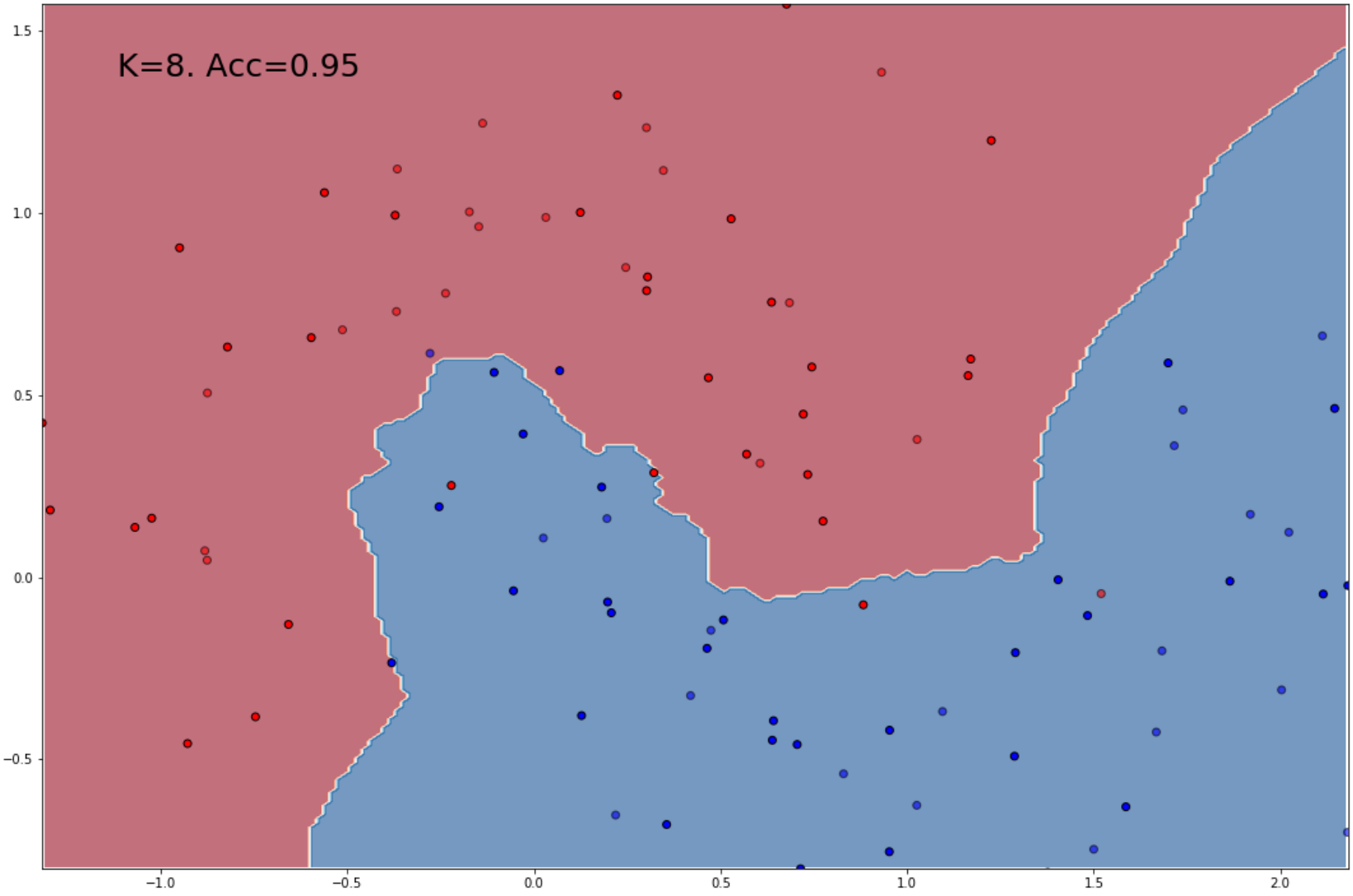
Визуализация разделяющих областей



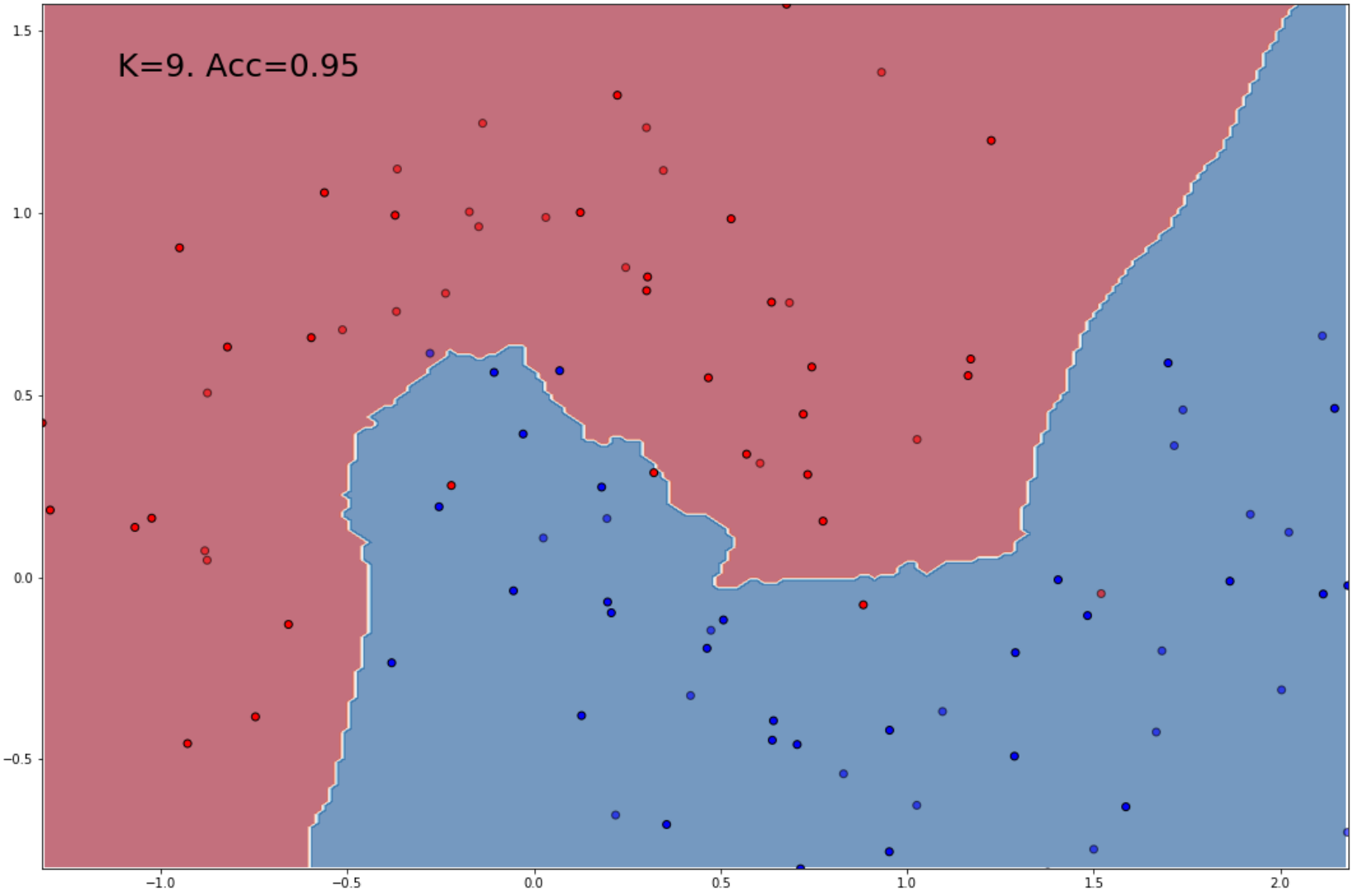
Визуализация разделяющих областей



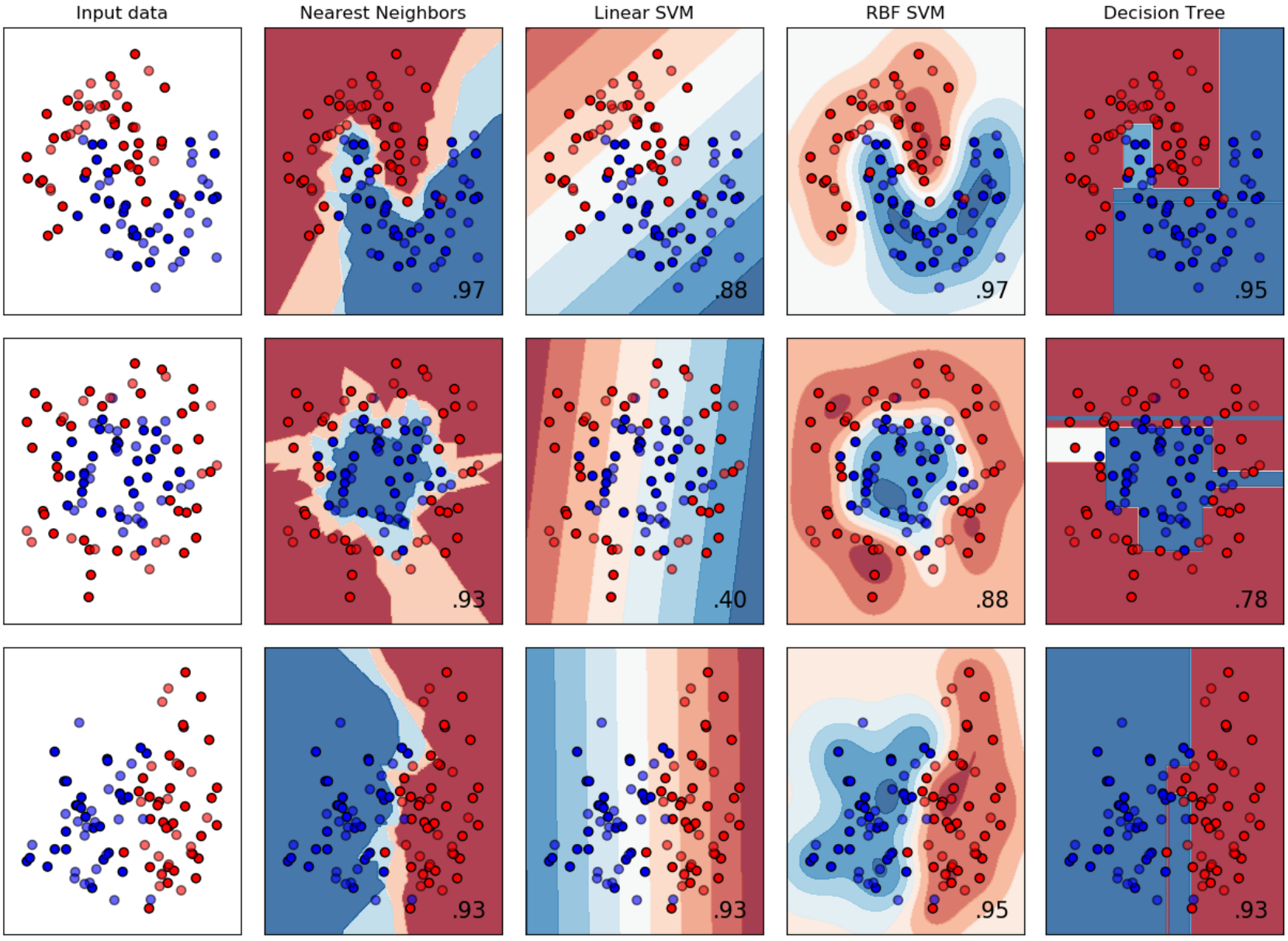
Визуализация разделяющих областей



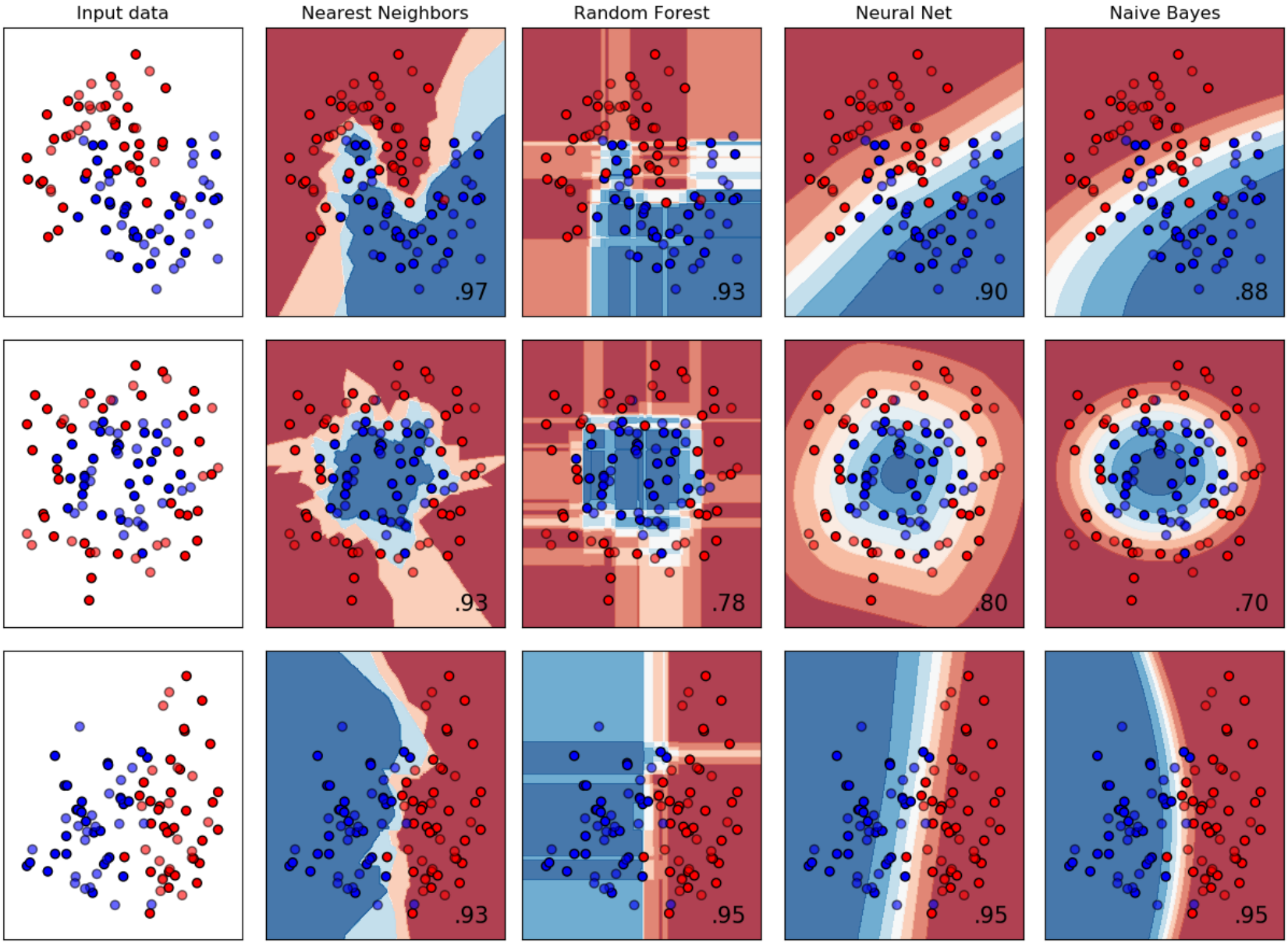
Визуализация разделяющих областей



Сравнение разделяющих областей



Сравнение разделяющих областей



Скорость работы KNN

Одним из главных недостатков алгоритма KNN является низкая скорость работы. Узкое место - поиск ближайших соседей

- Поиск одного ближайшего соседа - $O(l)$
- Если имеем дело с высокой размерностью признаков - $O(l \cdot d)$
- Если искать k ближайших соседей - $O(l \cdot d + l \ log \ l)$

Однако, существуют приближенные алгоритмы поиска ближайших соседей, решающие эту проблему.

KD-tree

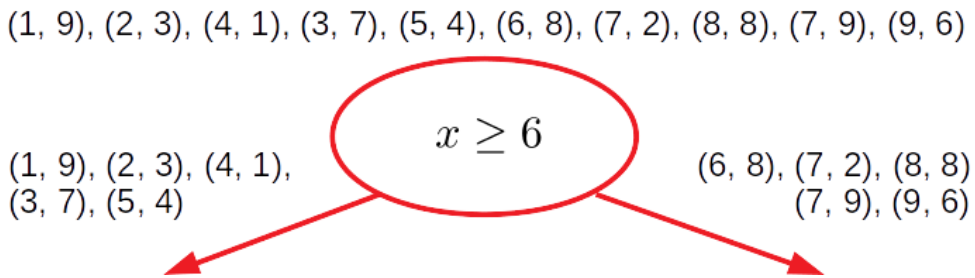
KD-tree - сбалансированное бинарное дерево поиска, в узлах которого находится решающее правило вида $x_{(d)} \geq a$.

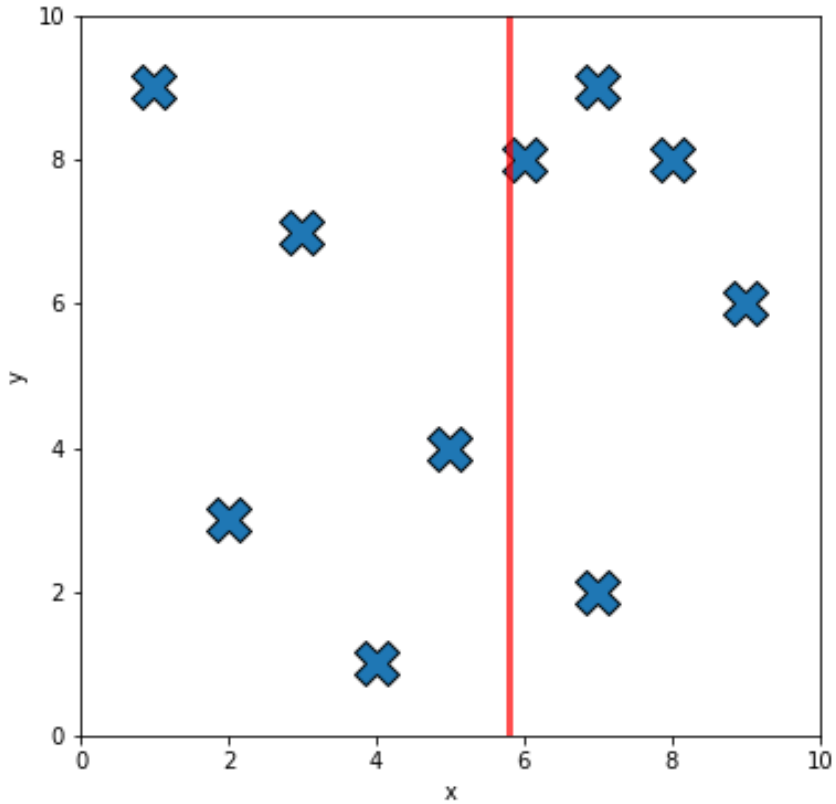
Построение происходит следующим образом:

1. Выбрать случайный признак
2. Найти медиану в данных по этому признаку
3. Разделить данные на две равных части
4. Если размер части больше чем *leafMax*, рекурсивно перейти для нее к пункту 1

KD tree. Пример построения

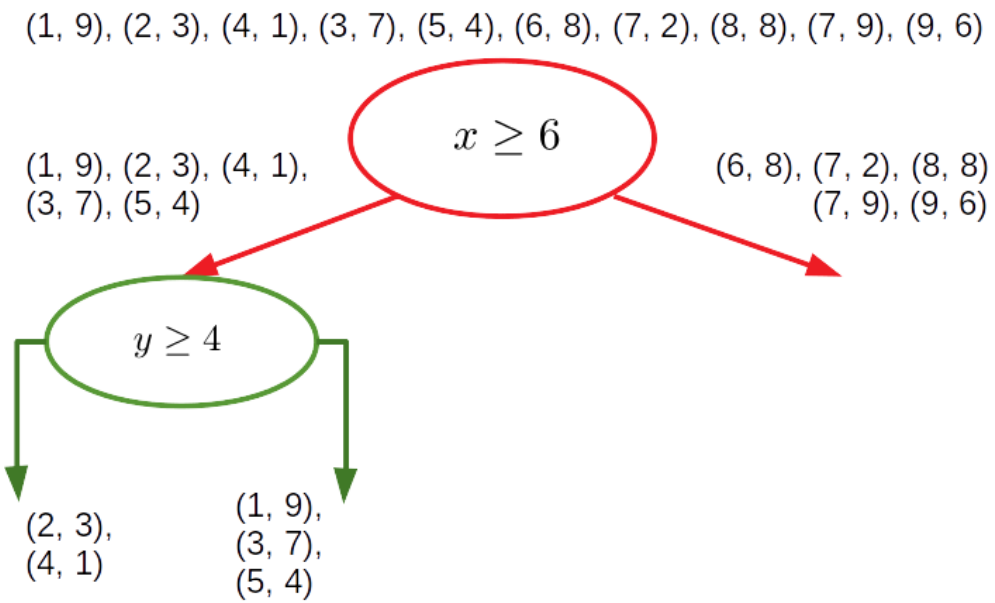
Выборка из 10 элементов. $leafMax = 3$

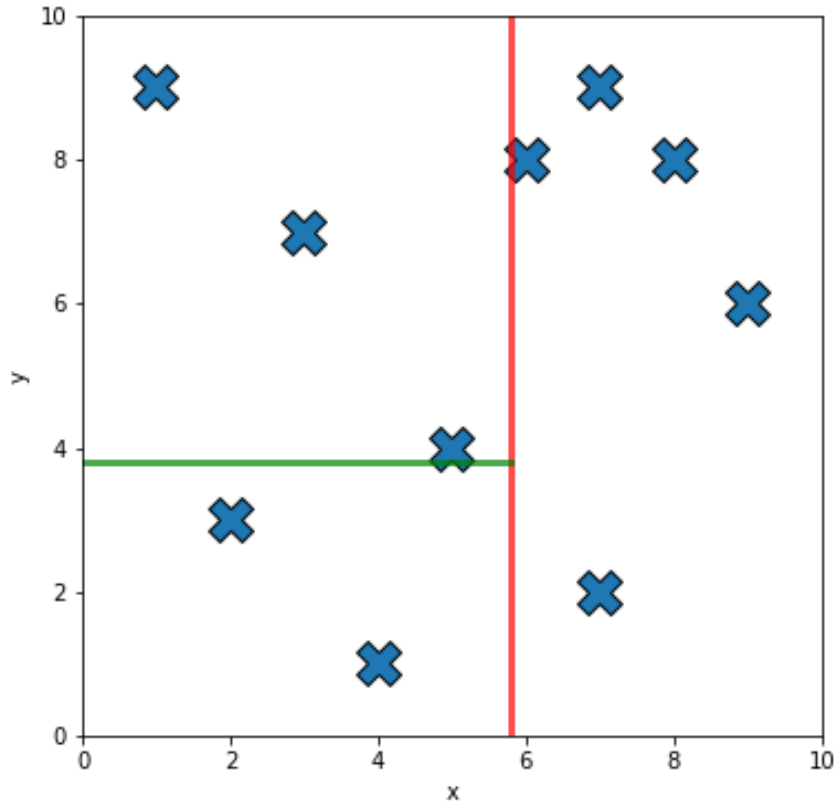




KD tree. Пример построения

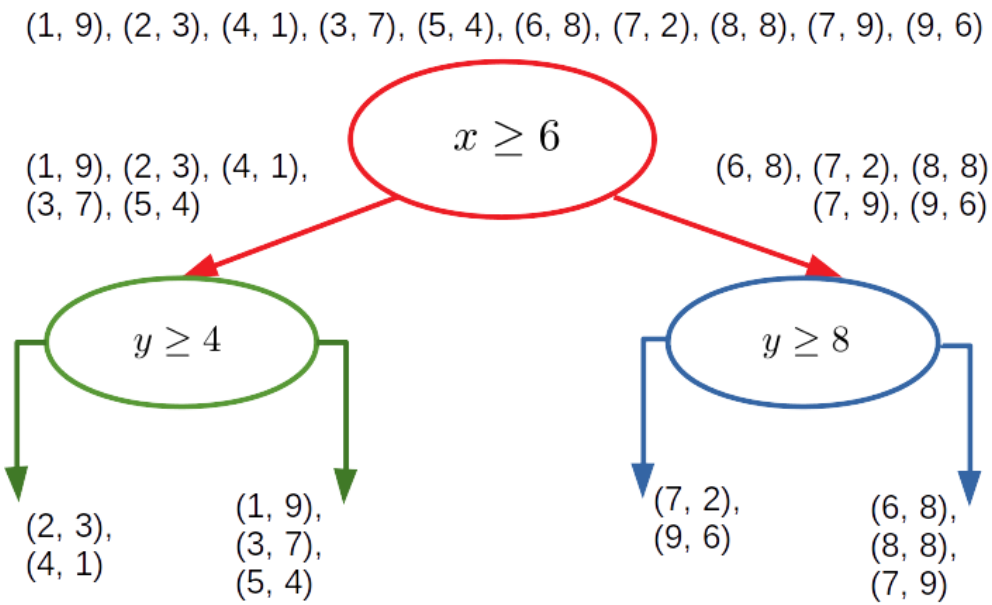
Выборка из 10 элементов. $leafMax = 3$

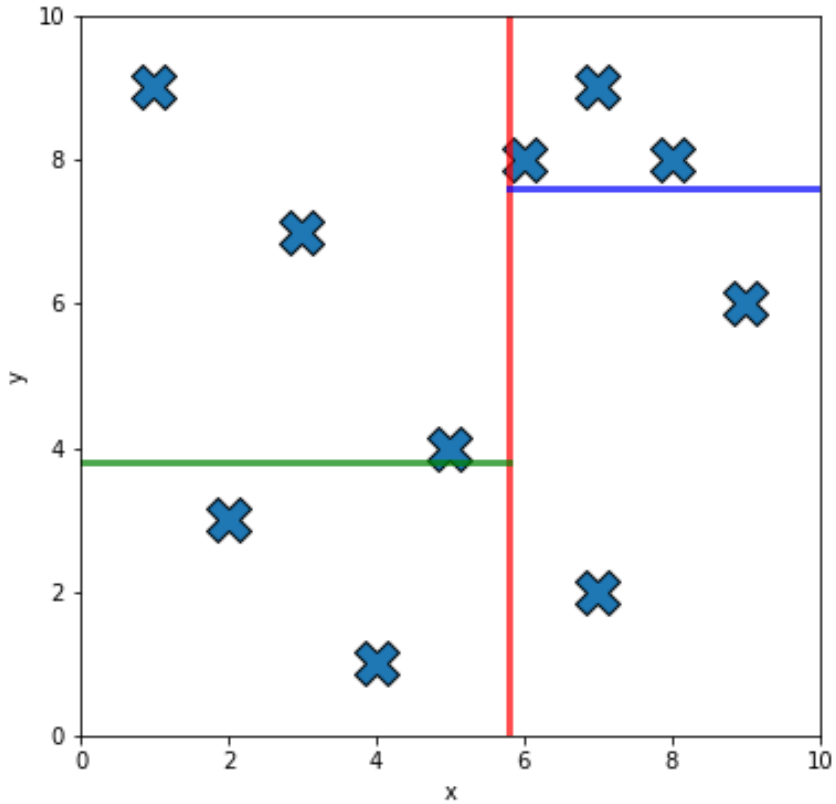




KD tree. Пример построения

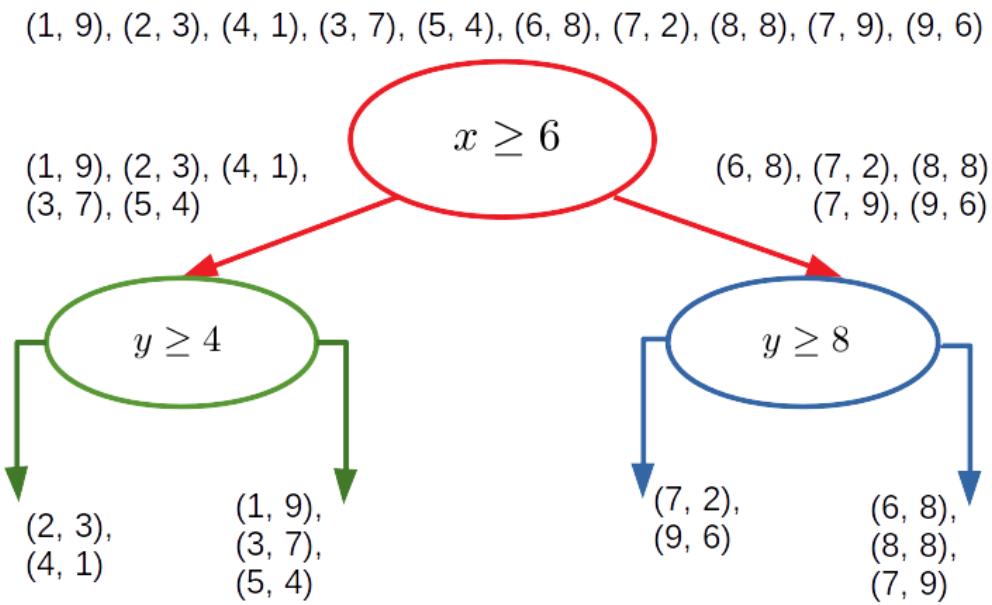
Выборка из 10 элементов. $leafMax = 3$

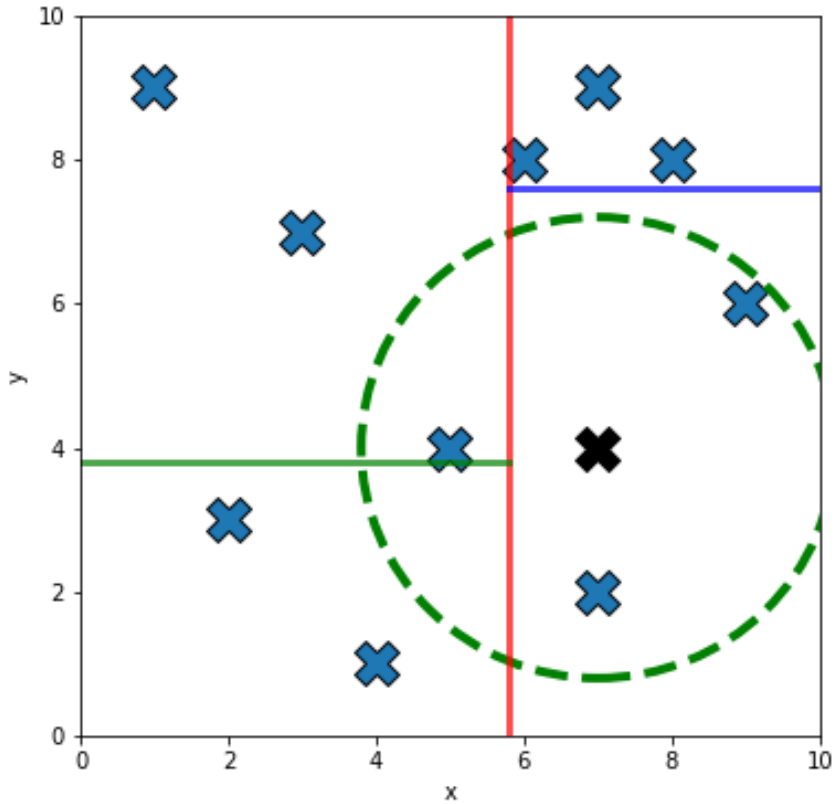




KD tree. Пример построения

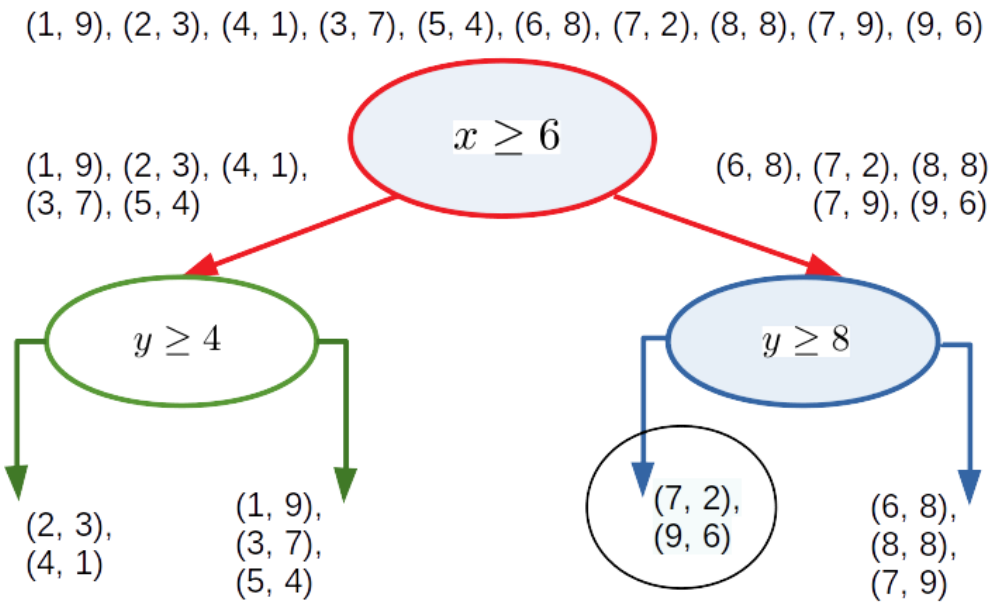
Выборка из 10 элементов. $leafMax = 3$

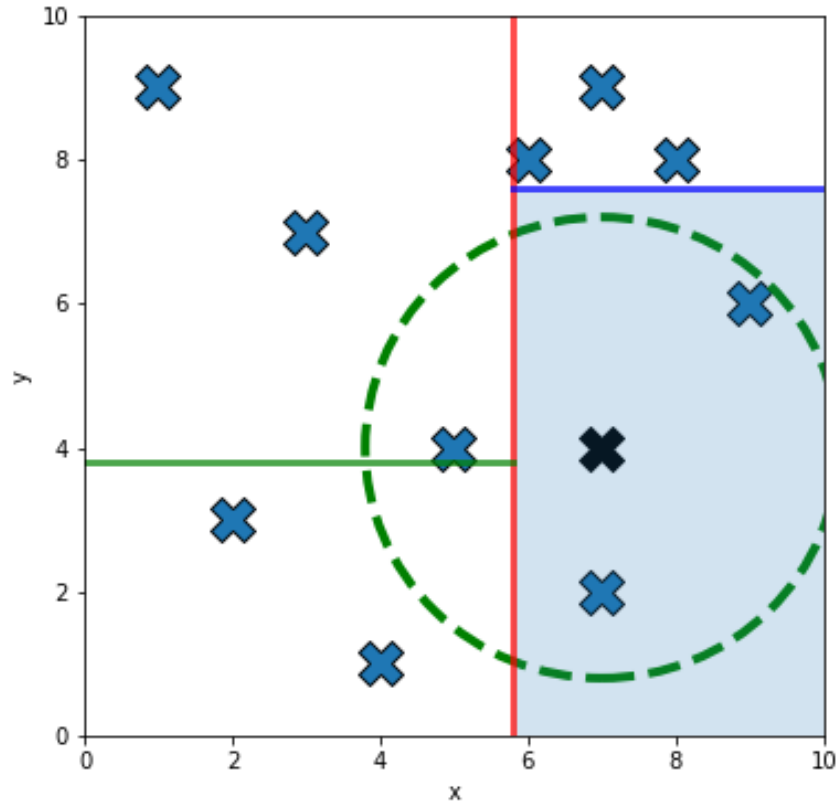




KD tree. Пример построения

Выборка из 10 элементов. $leafMax = 3$





Заметим, что листов не может быть больше, чем $\frac{N}{leafMax}$. А значит, максимальная глубина не превышает $log_2 N$.

Итоги

- Метрические методы классификации, такие как KNN, применимы при создании рекомендательных систем, для принятия решений, произвольных задач классификации разного рода
- Решение такого классификатора всегда можно объяснить, показав случаи, похожие на данный
- Существует множество различных вариаций KNN, отличающихся набором гиперпараметров и весовыми функциями
- KNN можно значительно ускорить, использовав приближенные методы поиска ближайших соседей. Например, KD-tree

Список литературы

- К.В. Воронцов. Видеолекция «Метрические методы классификации» https://www.youtube.com/watch?v=G_y16XfSMkw
- Sklearn classifier comparison http://scikit-learn.org/stable/auto_examples/classification/plot_classifier_comparison.html
- http://www.machinelearning.ru/wiki/index.php?title=Метод_ближайшего_соседа