



Метод К-ближайших соседей (K nearest neighbors)



Метод К-ближайших соседей

- KNN (k-nearest neighbors) – один из простейших алгоритмов машинного обучения.
- Обзор раздела:
 - Теория KNN
 - Пример кода для KNN
 - Упражнения по KNN
 - Решения упражнений по KNN



Метод К-ближайших соседей

- Хотя метод KNN и можно применить для задач регрессии, производительность и эффективность будут ниже, чем у других алгоритмов. Поэтому мы решили не показывать примеры для регрессии.
- Но если Вы захотите это сделать, то можете просто подставить в качестве модели KNNRegressor с помощью Scikit-Learn.



Метод К-ближайших соседей

- Возможно Вы также слышали об алгоритме К-средних – K-means.
- Несмотря на похожее название, это совсем другой алгоритм! Не перепутайте эти два алгоритма, у них очень похожие названия.



Метод K-ближайших соседей

Теория



Метод К-ближайших соседей

- KNN (k-nearest neighbors) – один из простейших алгоритмов машинного обучения.
- Для каждой точки этот алгоритм присваивает значение, основываясь на расстоянии между старыми данными и новыми данными
- Посмотрим на примере...



Метод К-ближайших соседей

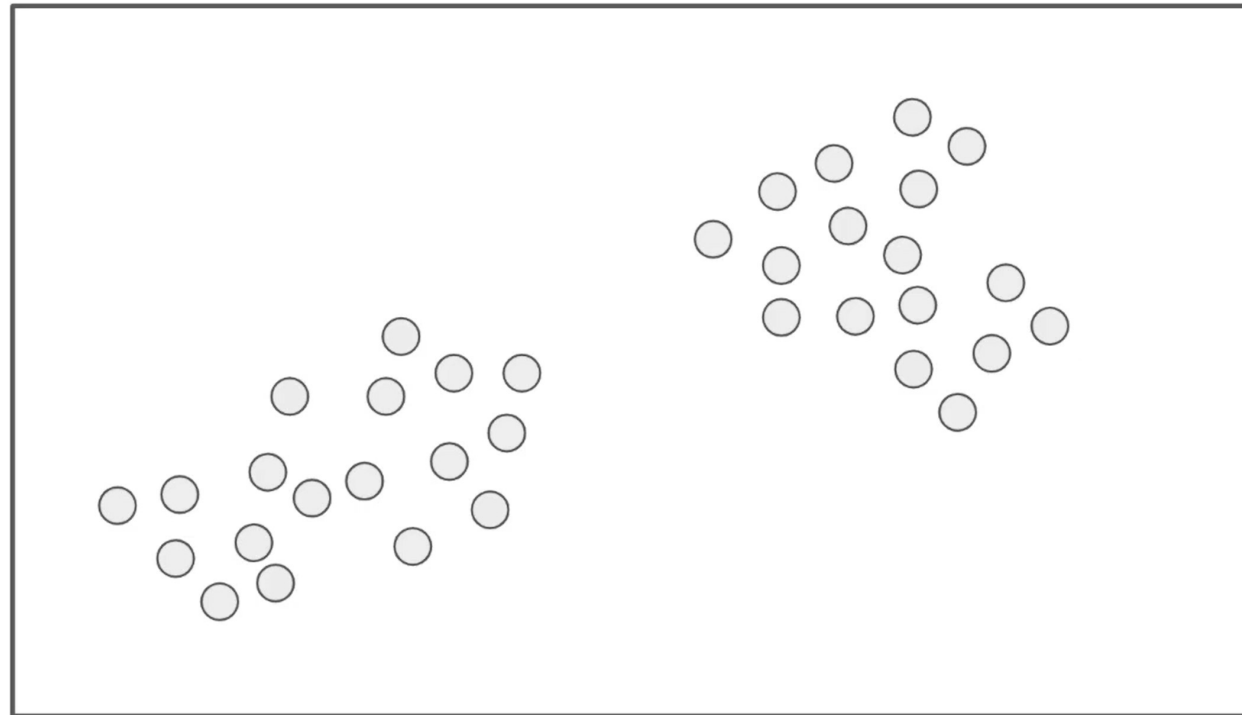
- В птицеводстве есть задача определять пол цыплят.
- Предположим, что у нас есть набор данных с ростом и весом цыплят.
- Можем ли мы написать алгоритм, определяющий пол цыплёнка на основе его роста и веса?



Метод К-ближайших соседей

- Набор данных с ростом и весом цыплят.

Рост



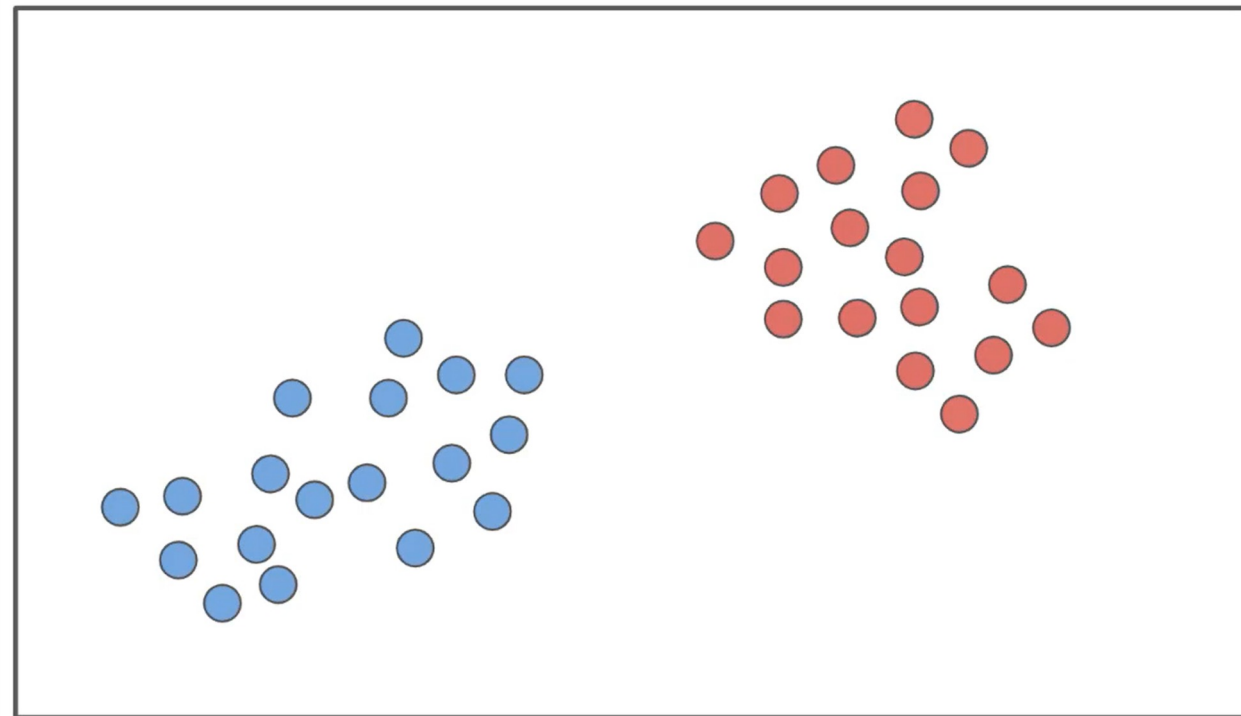
Вес



Метод К-ближайших соседей

- Набор данных с ростом и весом цыплят.

Рост



Пол

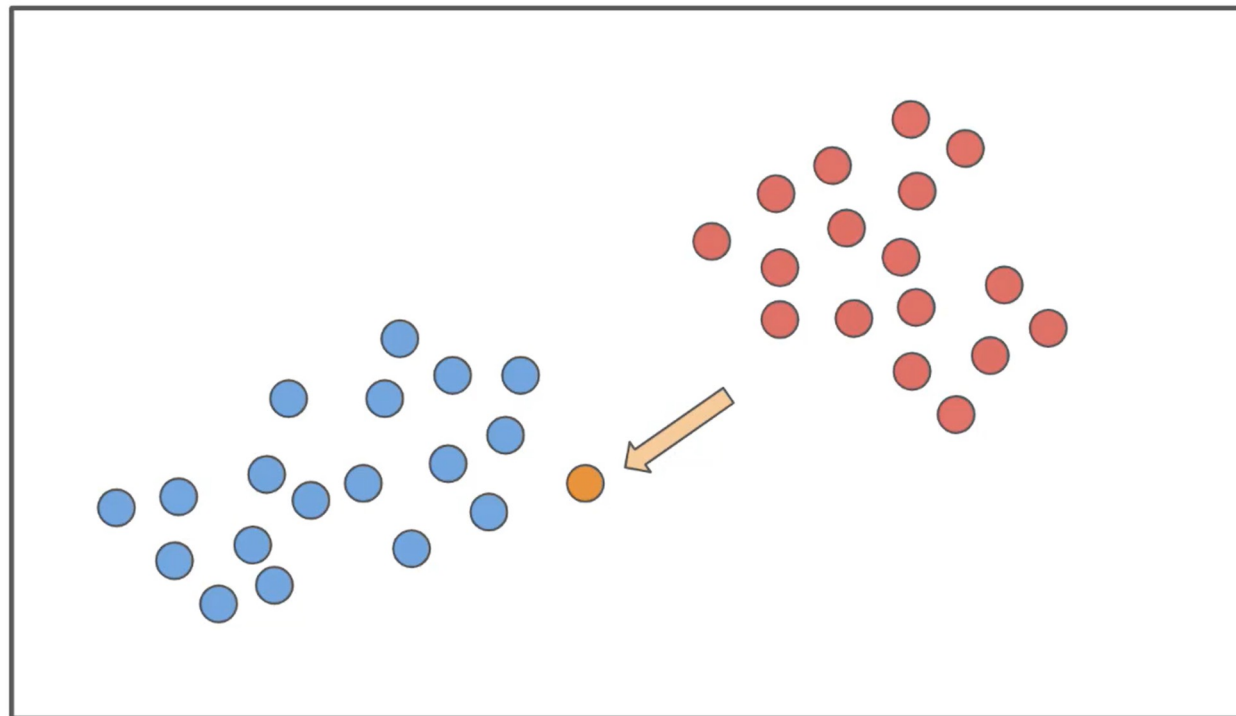
● Мужской
● Женский



Метод К-ближайших соседей

- Как определить пол нового цыплёнка?

Рост



Пол

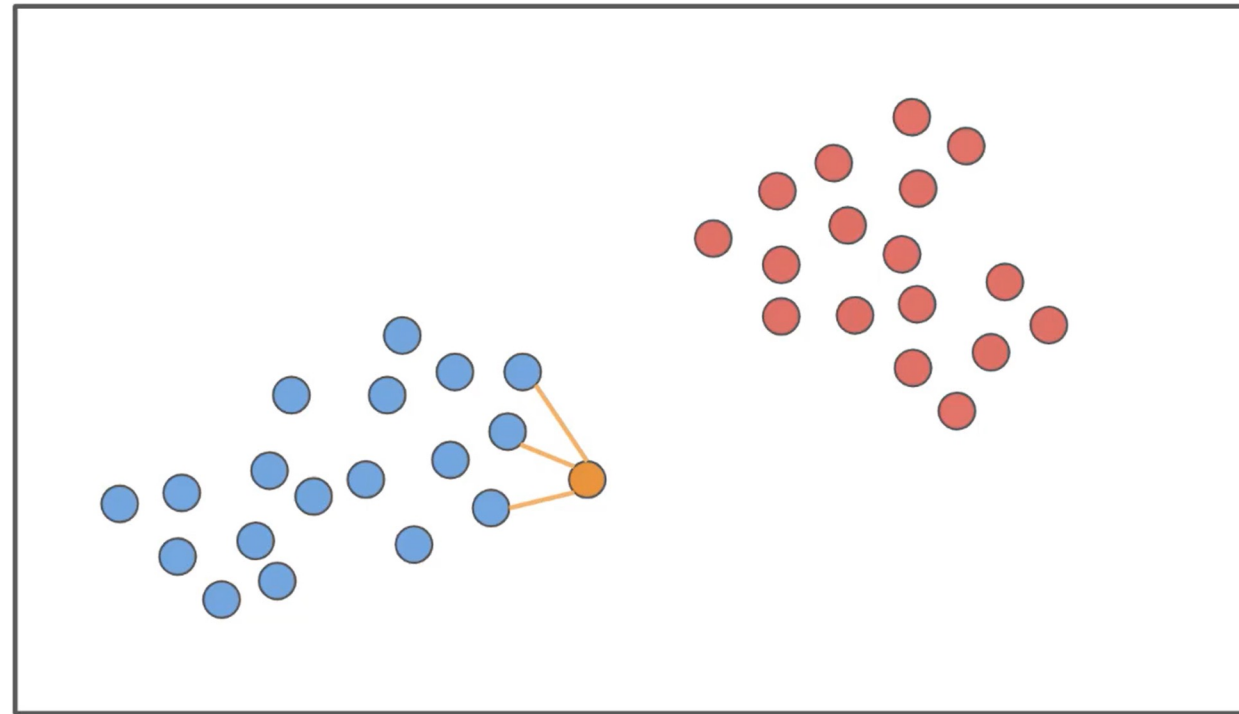
- Мужской
- Женский



Метод К-ближайших соседей

- По расстоянию до ближайших соседей

Рост



Пол

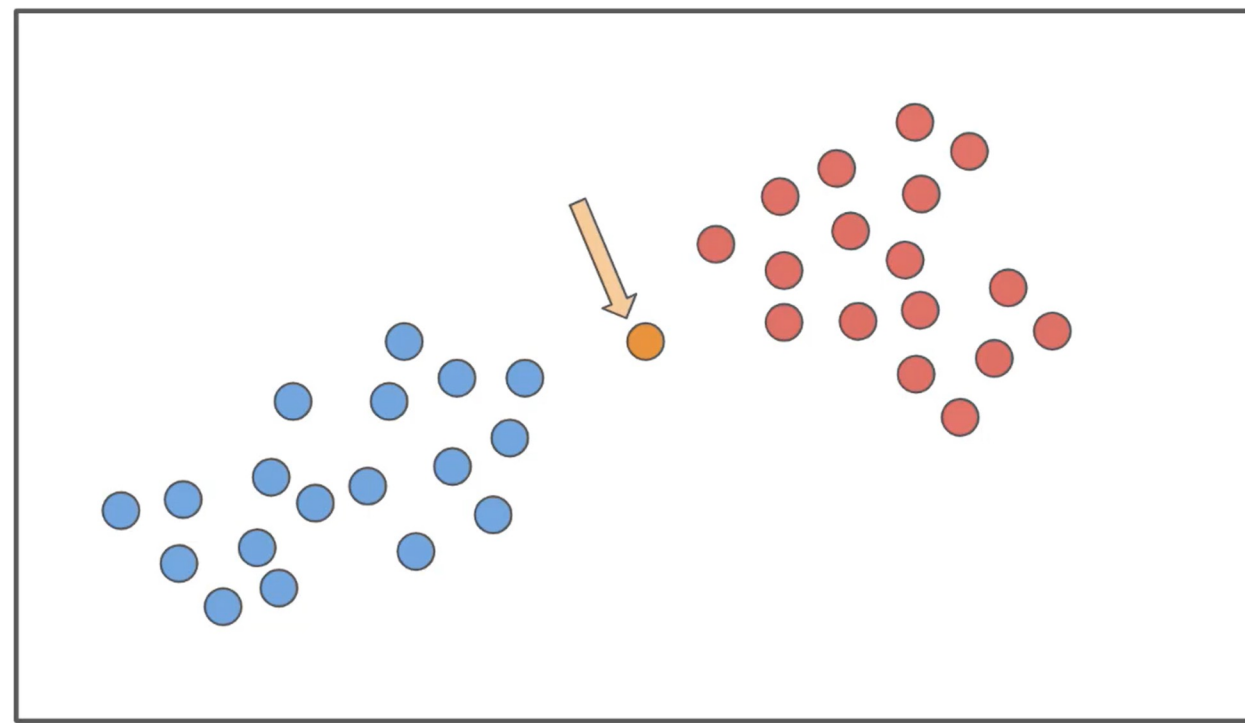
● Мужской
● Женский



Метод К-ближайших соседей

- А как насчёт менее очевидного варианта?

Рост



Пол

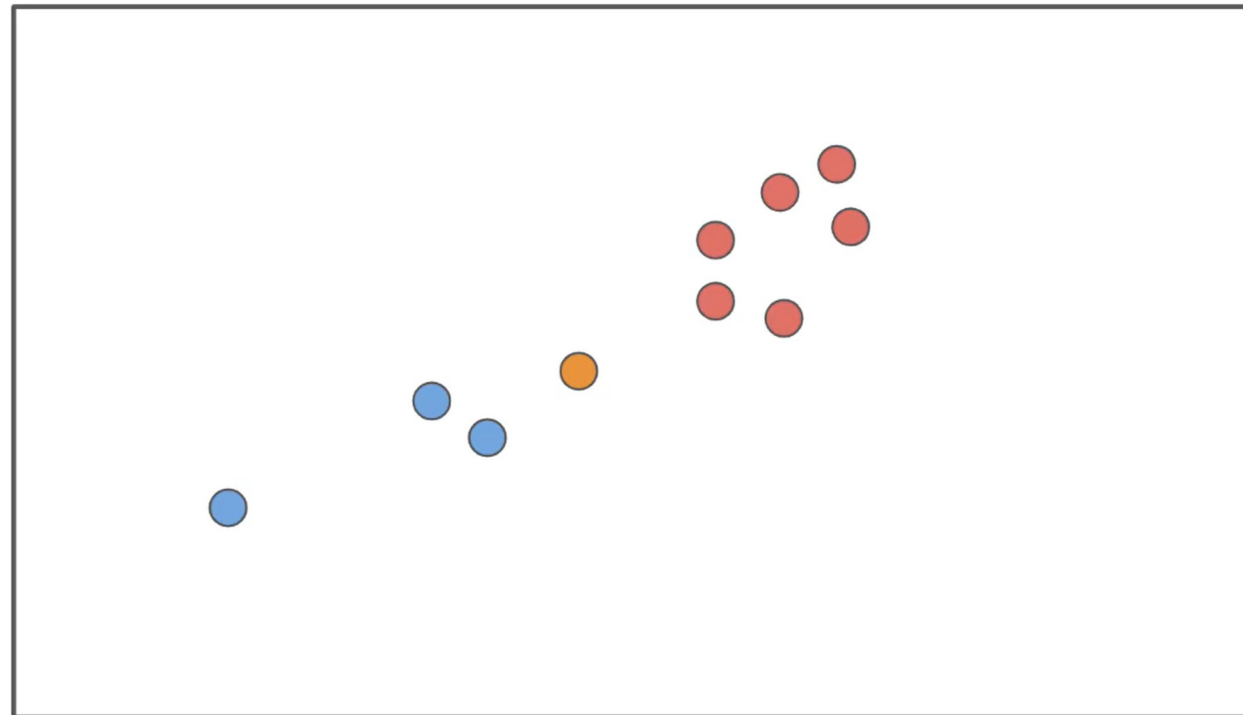
● Мужской
● Женский



Метод К-ближайших соседей

- Представим себе следующую ситуацию:

Рост



Пол

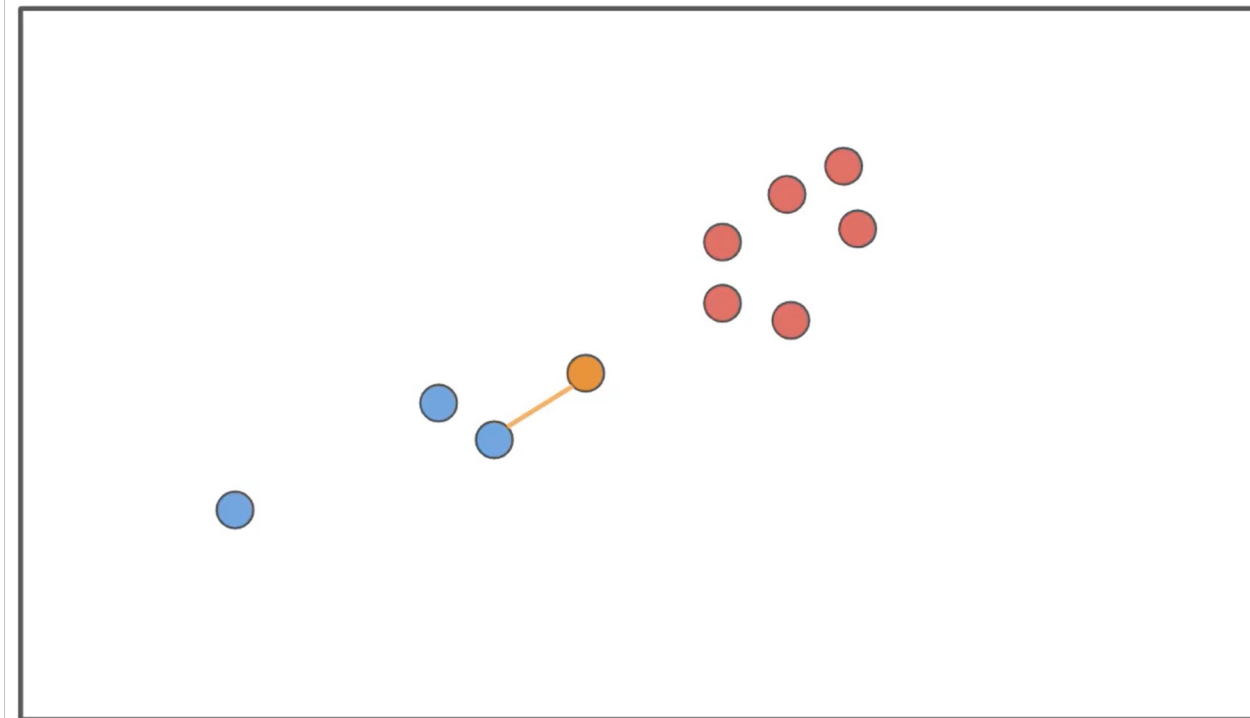
● Мужской
● Женский



Метод К-ближайших соседей

- $K=1$

Рост



Пол

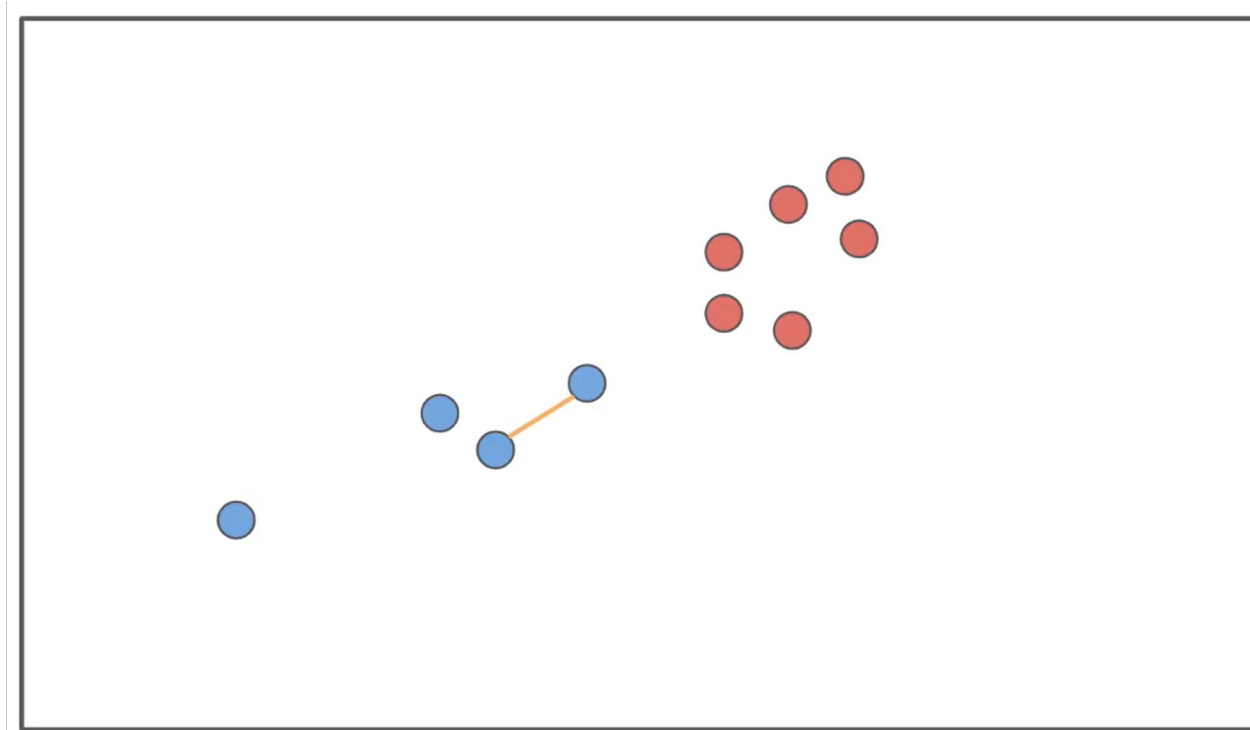
- Мужской
- Женский



Метод К-ближайших соседей

- $K=1$

Рост



Пол

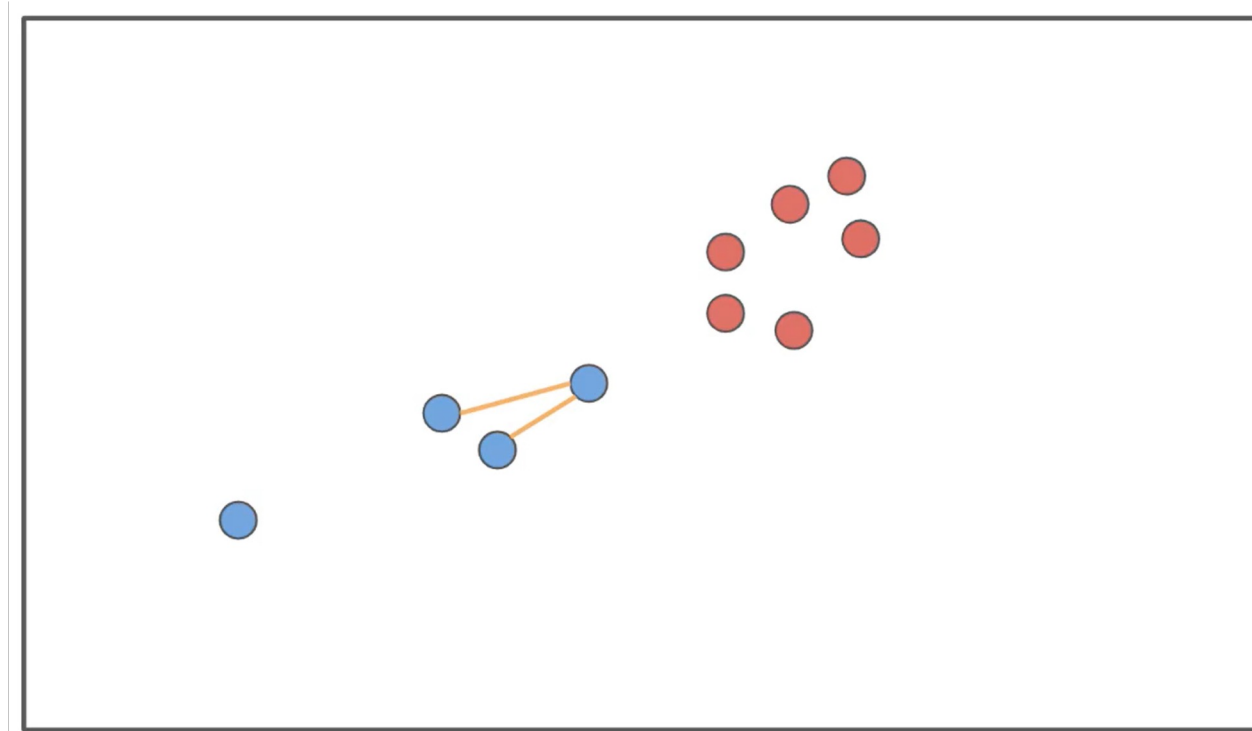
- Мужской
- Женский



Метод К-ближайших соседей

- $K=2$

Рост



Пол

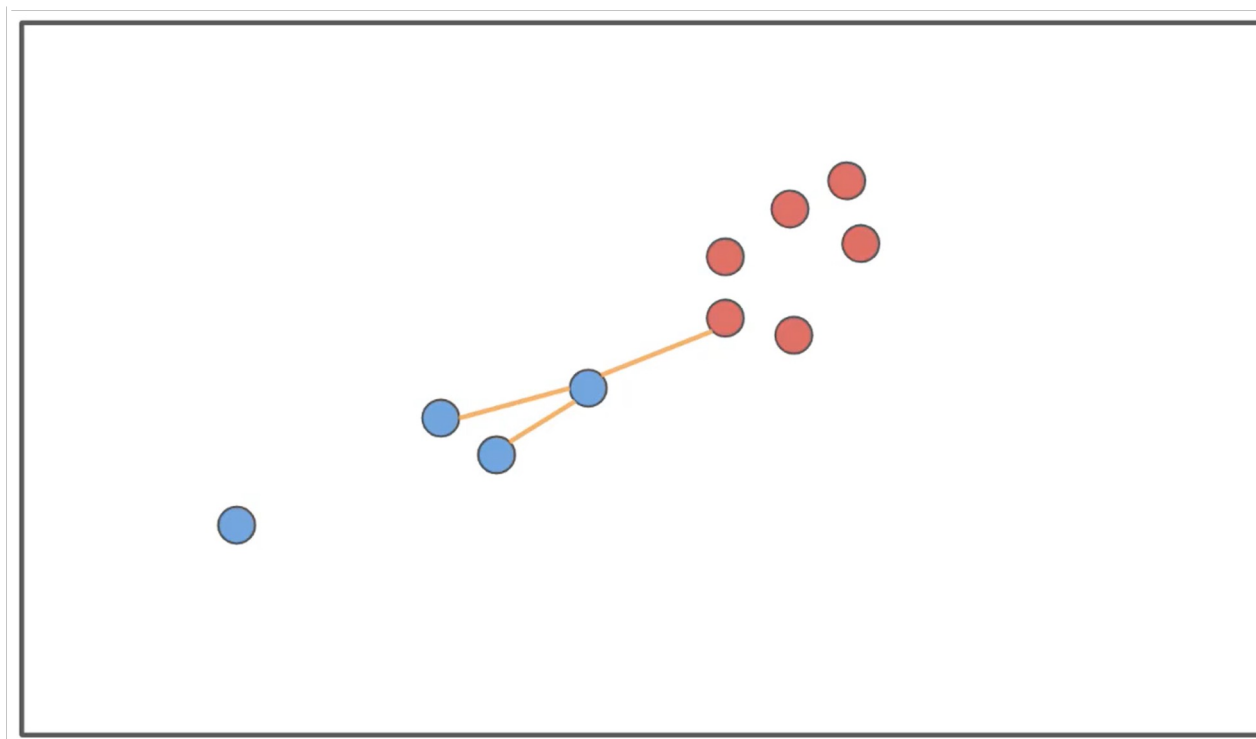
- Мужской
- Женский



Метод К-ближайших соседей

- $K=3$

Рост



Пол

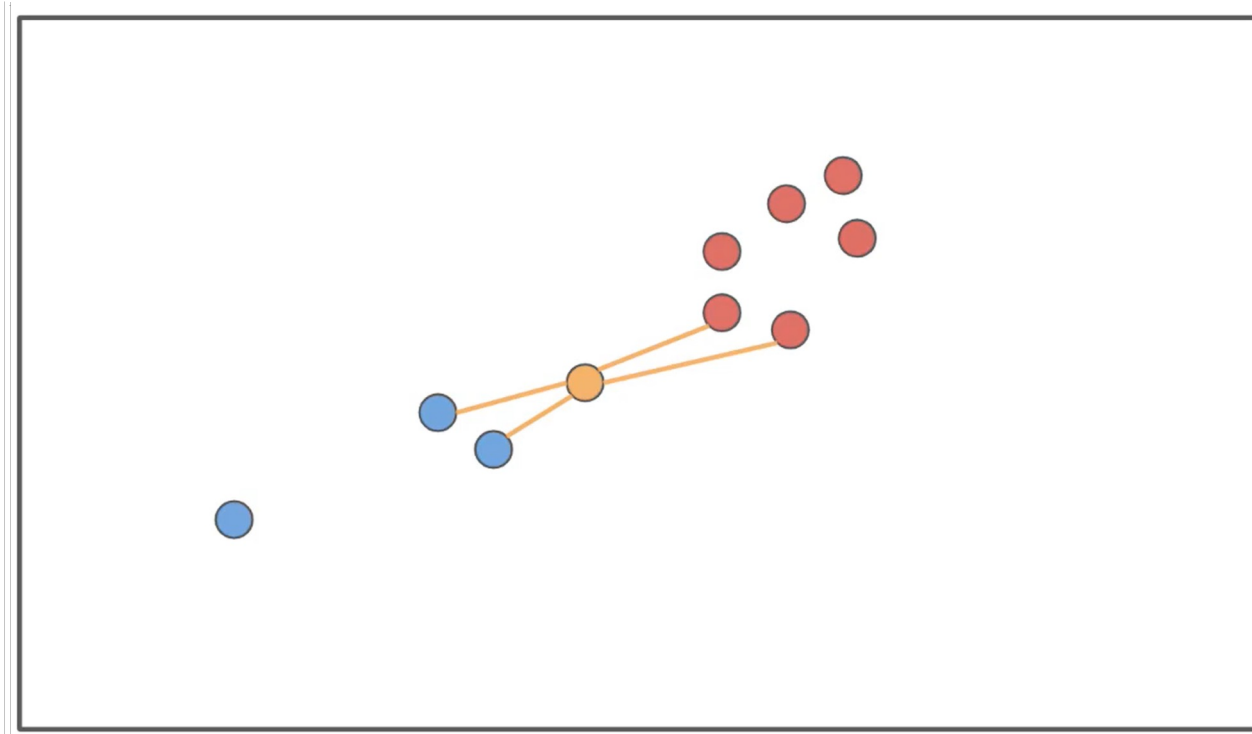
- Мужской
- Женский



Метод К-ближайших соседей

- $K=4$ – равное количество (tie) !

Рост





Метод К-ближайших соседей

- Варианты:

- Всегда выбирать нечётные значения K
- В случае равенства, уменьшить K на 1
- Случайно выбрать тот или иной вариант
- Выбрать ближайшую точку



Метод К-ближайших соседей

- Как с такими ситуациями работает Scikit-Learn?
 - Предупреждение: если для алгоритмов KNN выяснится, что две соседних точки, сосед $K+1$ и сосед K , имеют одинаковые расстояния, но разные классы, то результат будет зависеть от сортировки обучающих данных.



Метод К-ближайших соседей

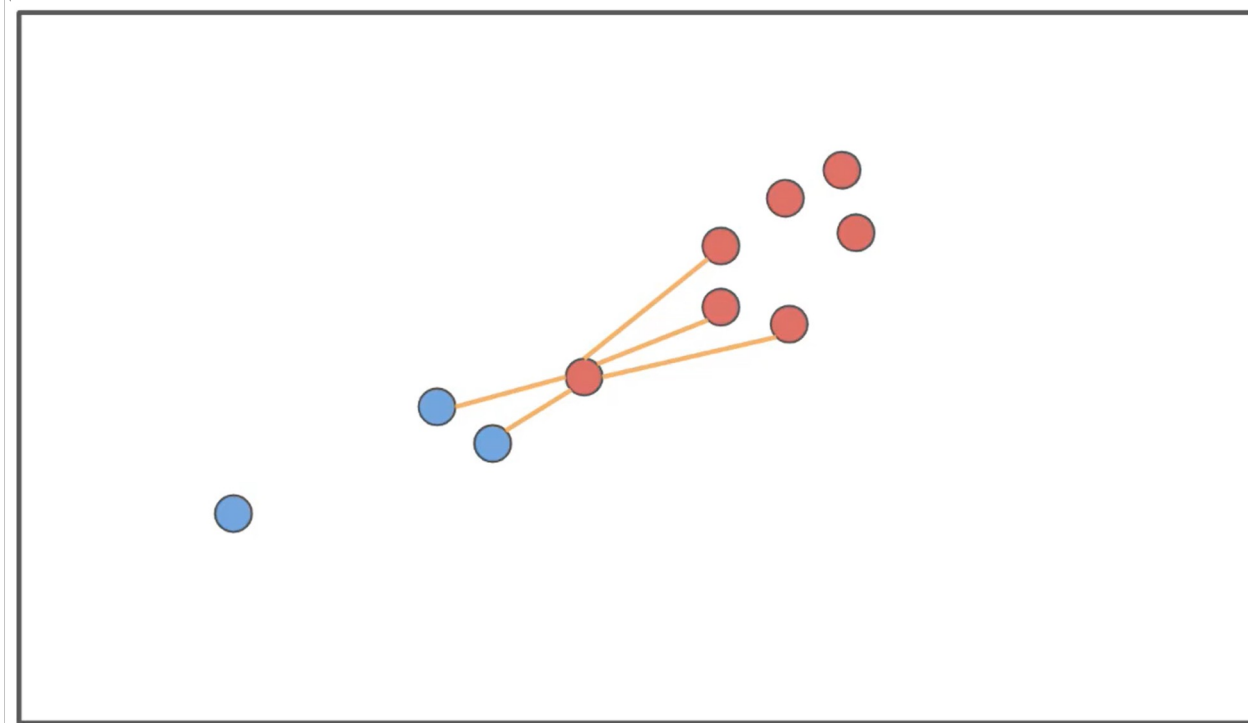
- Как с такими ситуациями работает Scikit-Learn?
 - В случае равенства (ties) будет выбран класс, идущий первым в множестве соседей.
 - Результаты отсортированы по расстоянию, так что будет выбран класс ближайшей точки.



Метод К-ближайших соседей

- $K=5$ – мы переключаемся на другой класс.

Рост



Пол

● Мужской
● Женский



Метод К-ближайших соседей

- Как выбрать число К?
 - Хотим минимизировать $\text{error} = 1 - \text{accuracy}$



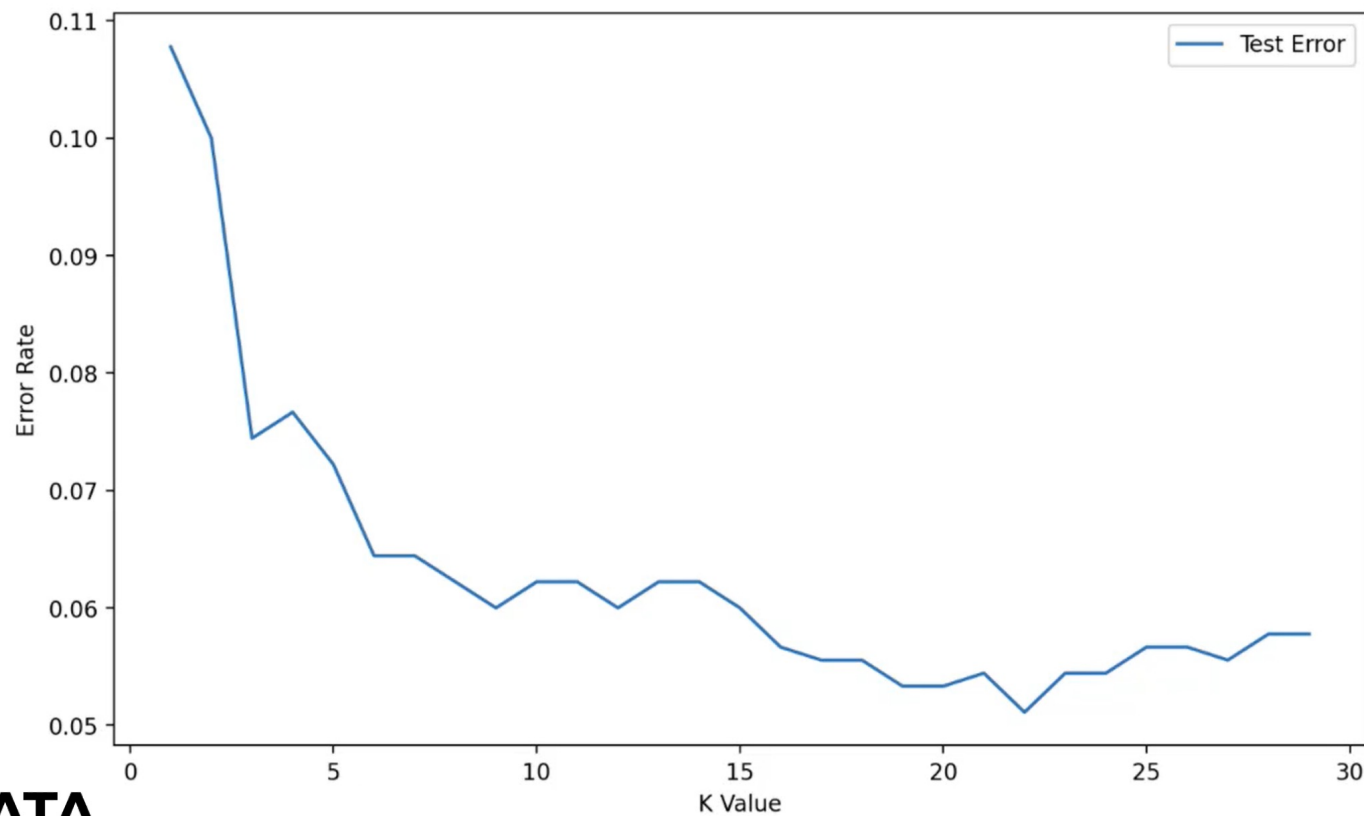
Метод К-ближайших соседей

- Как выбрать число К?
 - Хотим минимизировать $\text{error} = 1 - \text{accuracy}$
- Два метода:
 - Метод локтя (elbow method)
 - Кросс-валидация - перебор различных значений К по сетке (grid search), чтобы найти значение К с наименьшими ошибками



Метод К-ближайших соседей

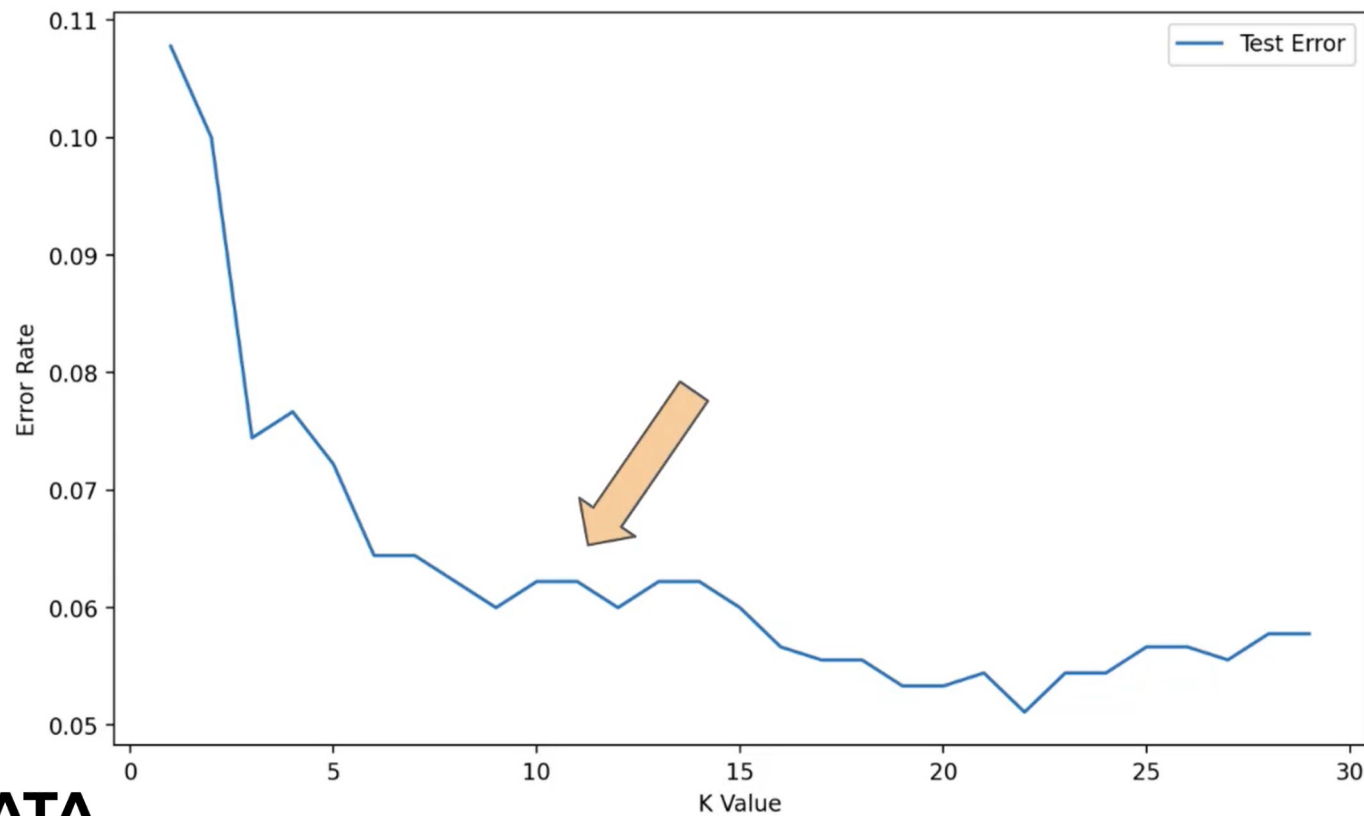
- Метод локтя (elbow method)





Метод К-ближайших соседей

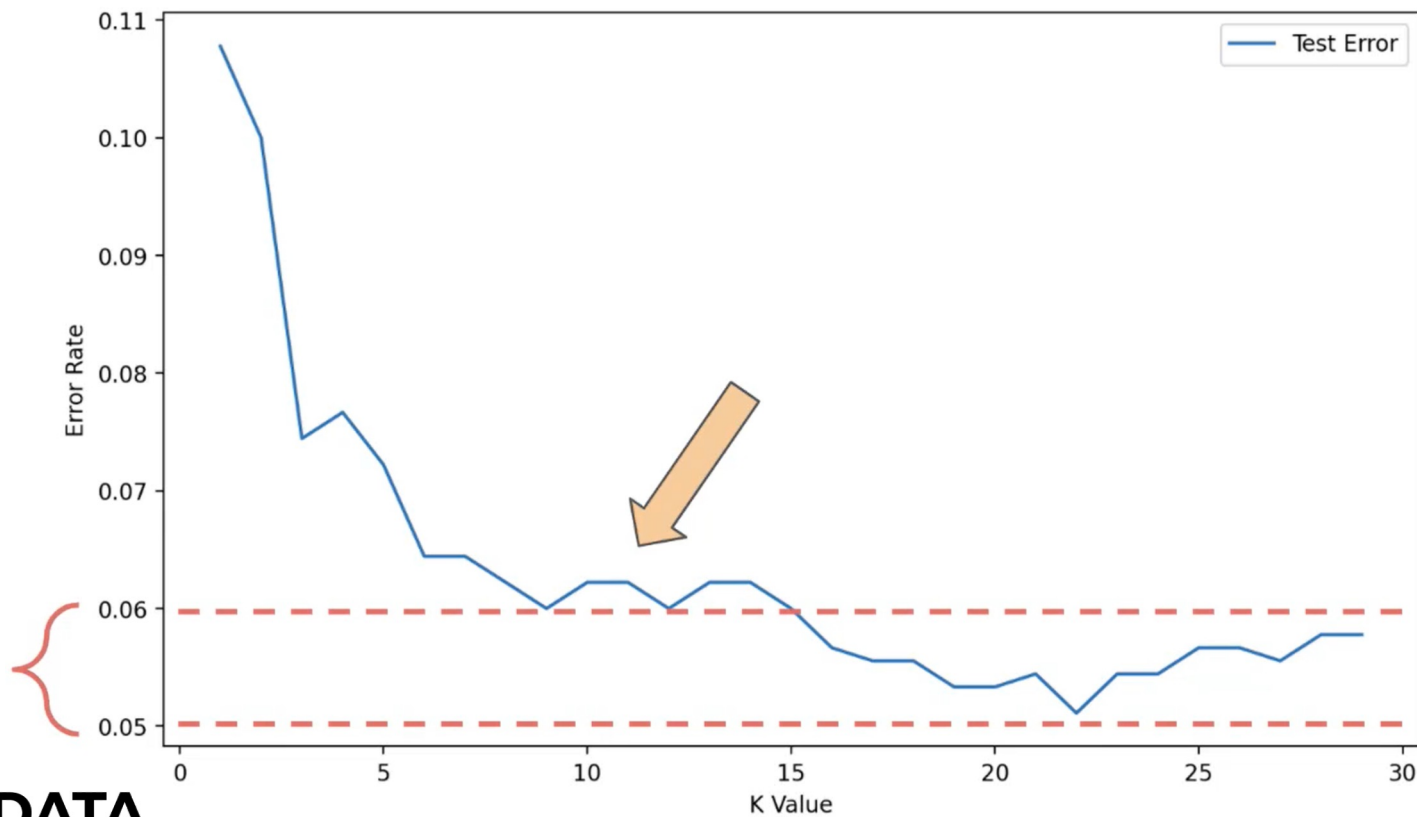
- Метод локтя (elbow method)





Метод К-ближайших соседей

- Метод локтя (elbow method)





Метод K-ближайших соседей

- Кросс-валидация ищет значение K , ориентируясь только на минимизацию ошибок
- Это может привести к более сложной модели (более высокому значению K)
- Учитывайте специфику задачи, чтобы понять, насколько приемлемо увеличение значения K



Метод К-ближайших соседей

- Алгоритм KNN
 - Выбираем значение K
 - Сортируем вектора признаков (в N -мерном пространстве) по метрике расстояния
 - Выбираем класс точек на основе K ближайших векторов признаков



Метод К-ближайших соседей

- Алгоритм KNN
 - Метрики расстояния – разные варианты
 - Расстояние Минковского
 - Евклидова метрика
 - Расстояние городских кварталов (Manhattan)
 - Расстояние Чебышёва



Метод К-ближайших соседей

- Алгоритм KNN
 - Масштабирование расстояний
 - Различные признаки могут иметь очень разный масштаб!





Метод К-ближайших соседей

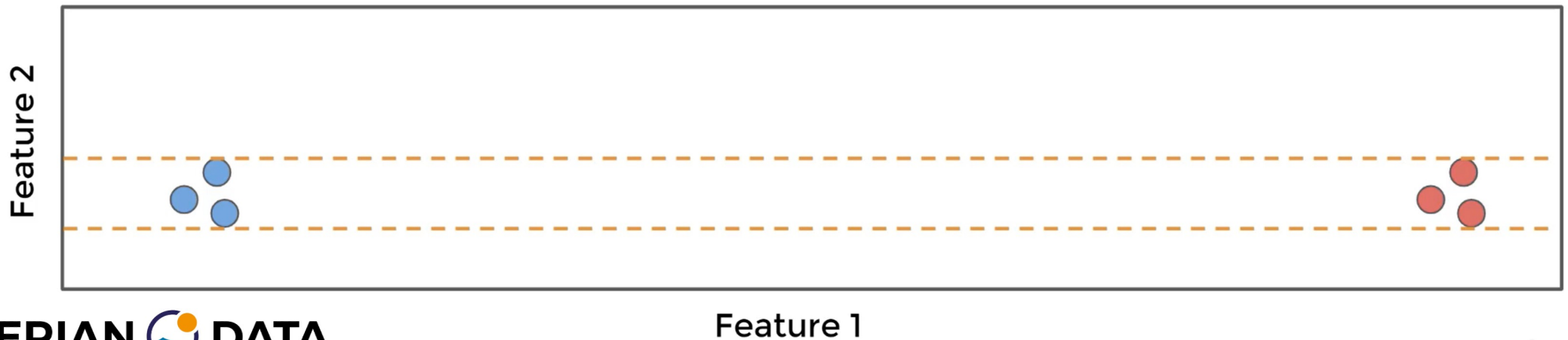
- Алгоритм KNN
 - Масштабирование расстояний
 - Различные признаки могут иметь очень разный масштаб!





Метод К-ближайших соседей

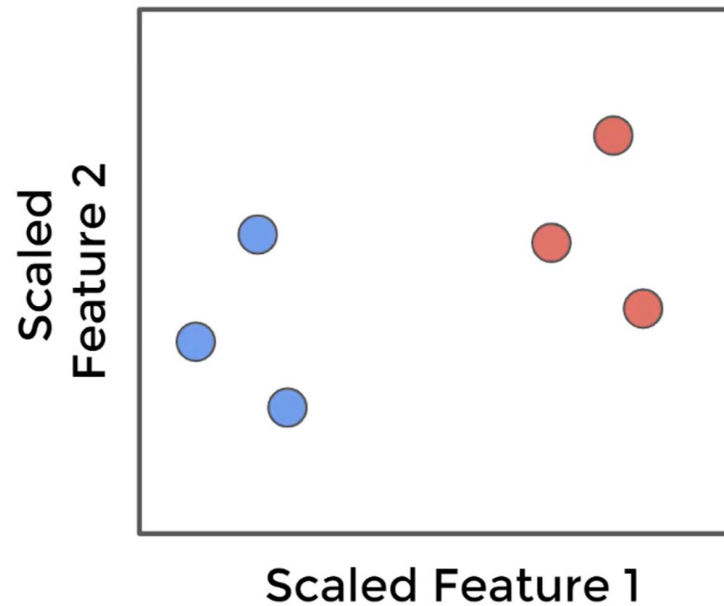
- Алгоритм KNN
 - Масштабирование расстояний
 - Различные признаки могут иметь очень разный масштаб!





Метод К-ближайших соседей

- Алгоритм KNN
 - Для KNN необходимо масштабировать признаки





Метод К-ближайших соседей

- Хотя алгоритм KNN и является очень простым, учитывайте необходимость:
 - Выбирать значение K
 - Масштабировать признаки
- Далее мы применим алгоритм KNN для задач классификации!



KNN для классификации

Часть 1: данные и модель



KNN для классификации

Часть 2: выбираем K



Метод К-ближайших соседей

- Объект Pipeline в Scikit-Learn может включать в себя несколько шагов, таких как масштабирование (scaler) и работа с моделью
- Удобство в том, что мы можем вызывать эти шаги для разных данных единой командой



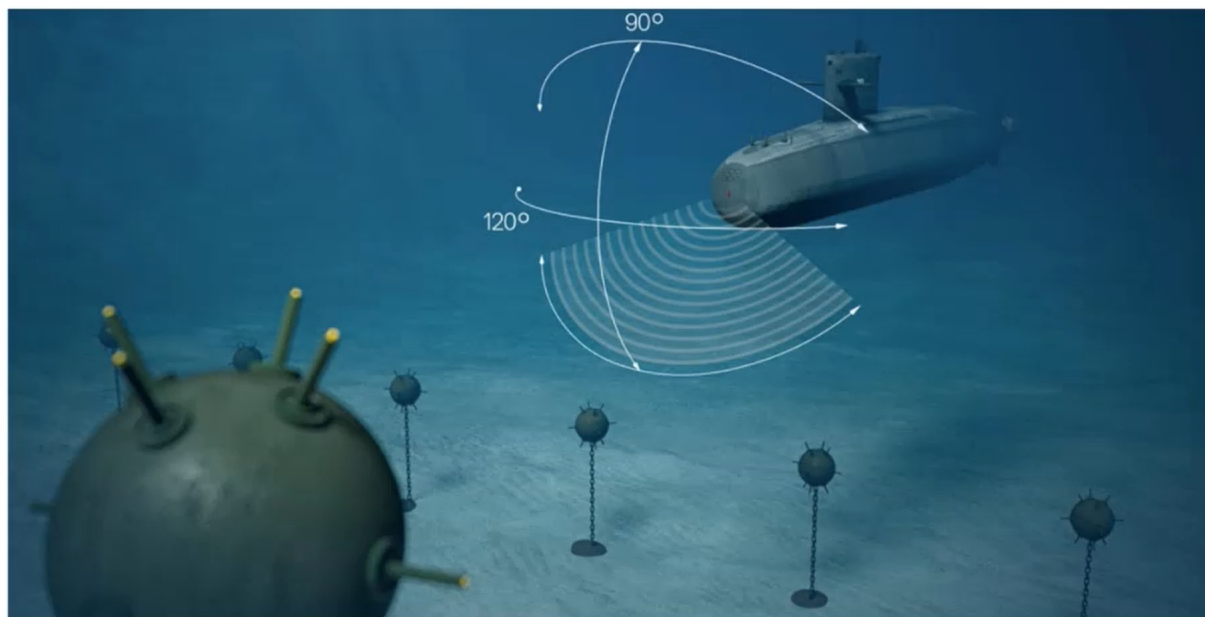
Упражнения по KNN

Обзор



Метод К-ближайших соседей

- Проверим Ваши навыки на реальных данных!
- Мы будем анализировать звуковые сигналы, чтобы отличить камни и подводные бомбы друг от друга.





Упражнения по KNN

Решения