

# KNN Algorithm



[www.educba.com](http://www.educba.com)

Алгоритм К ближайших соседей относится к категории контролируемого обучения и используется для классификации (чаще всего) и регрессии. Это универсальный алгоритм, который также используется для вменения пропущенных значений и повторной выборки наборов данных. Как следует из названия (К ближайших соседей), он рассматривает К ближайших соседей (точек данных) для прогнозирования класса или непрерывного значения для новой точки данных.

Обучение алгоритма:

1. Обучение на основе экземпляров: здесь мы не изучаем веса из обучающих данных для прогнозирования вывода (как в алгоритмах на основе моделей), а используем целые обучающие экземпляры для прогнозирования вывода для невидимых данных.
2. Ленивое обучение: модель не изучается с использованием обучающих данных заранее, и процесс обучения откладывается до момента, когда запрашивается прогнозирование для нового экземпляра.
3. Непараметрический: в KNN нет предопределенной формы функции отображения.

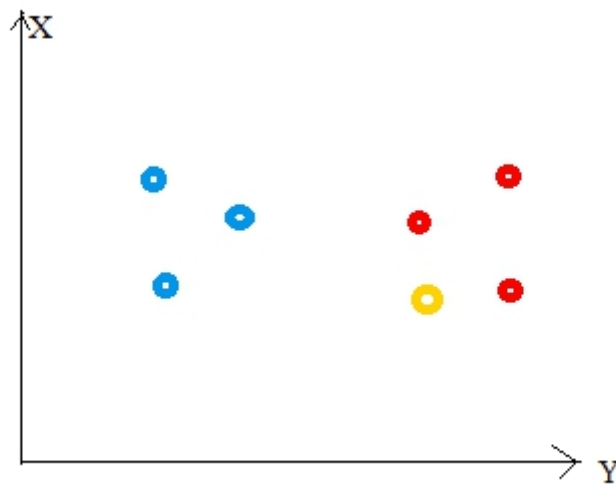
# Как работает KNN?

## 1. Принцип:

Рассмотрим следующий рисунок. Допустим, мы нанесли точки данных из нашего обучающего набора на двумерное пространство признаков. Как показано, у нас всего 6 точек данных (3 красных и 3 синих). Красные точки данных принадлежат «классу 1», а синие точки данных принадлежат «классу 2». А желтая точка данных в пространстве признаков представляет собой новую точку, для которой должен быть предсказан класс. Очевидно, мы говорим, что он принадлежит к «классу 1» (красные точки).

Почему?

Потому что его ближайшие соседи принадлежат к этому классу!



Да, это принцип К ближайших соседей. Здесь ближайшие соседи — это те точки данных, которые имеют минимальное расстояние в пространстве признаков от нашей новой точки данных. И К — количество таких точек данных, которые мы учитываем в нашей реализации алгоритма. Следовательно, метрика расстояния и значение К являются двумя важными факторами при использовании алгоритма KNN. Евклидово расстояние — самая популярная метрика расстояния. Вы также можете использовать расстояние Хэмминга, расстояние Манхэттена, расстояние Минковского в соответствии с вашими потребностями. Для прогнозирования класса/непрерывного значения для новой точки данных учитываются все точки данных в обучающем наборе данных. Находит ближайших соседей (точки данных) новой точки данных из пространства объектов и их меток классов или непрерывных значений.

Потом:

Для классификации: метка класса, назначенная большинству  $K$  ближайших соседей из обучающего набора данных, считается прогнозируемым классом для новой точки данных.

Для регрессии: среднее или медиана непрерывных значений, присвоенных  $K$  ближайшим соседям из обучающего набора данных, является прогнозируемым непрерывным значением для нашей новой точки данных.

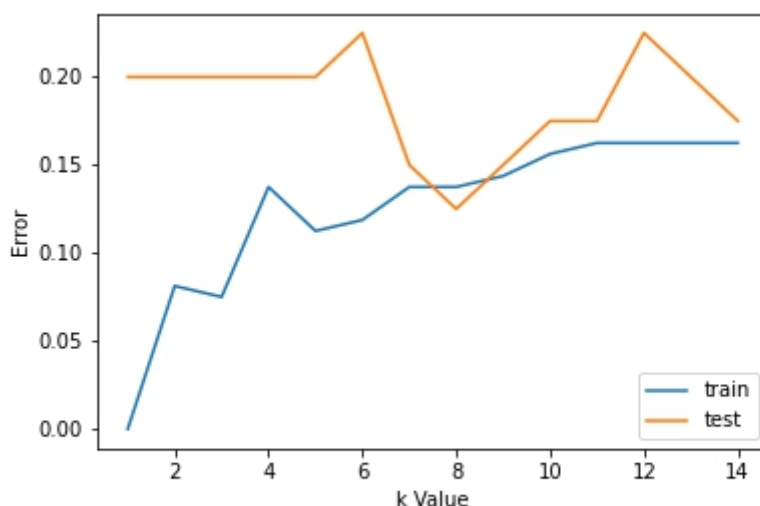
## 2. Представление модели

Здесь мы не изучаем веса и не сохраняем их, вместо этого весь обучающий набор данных хранится в памяти. Следовательно, модельное представление для KNN — это весь обучающий набор данных.

### Как выбрать значение для $K$ ?

$K$  является важным параметром в алгоритме KNN. Некоторые рекомендации по выбору значения  $K$ :

1. Использование кривых ошибок. На рисунке ниже показаны кривые ошибок для различных значений  $K$  для обучающих и тестовых данных.



### Choosing a value for $K$

При низких значениях  $K$  происходит переобучение данных/высокая дисперсия. Поэтому ошибка теста высока, а ошибка на тренировке низка. При  $K=1$  в данных тренировка ошибка всегда равна нулю, потому что ближайшей соседней точкой является сама эта точка. Следовательно, хотя ошибка обучения мала, ошибка теста высока при более низких значениях  $K$ . Это называется переобучением. По мере увеличения значения  $K$  ошибка теста уменьшается.

Но после определенного значения  $K$  вводится смещение/недообучение, и ошибка теста становится высокой. Таким образом, мы можем сказать, что изначально ошибка тестовых данных высока (из-за дисперсии), затем она

снижается и стабилизируется, а при дальнейшем увеличении значения  $K$  снова увеличивается (из-за смещения). Значение  $K$ , когда ошибка теста стабилизируется и является низкой, считается оптимальным значением для  $K$ . Из приведенной выше кривой ошибки мы можем выбрать  $K = 8$  для нашей реализации алгоритма KNN.

2. Кроме того, знание предметной области очень полезно при выборе значения  $K$ .

3. Значение  $K$  должно быть нечетным при рассмотрении бинарной (двухклассовой) классификации.

## **Необходимая подготовка данных:**

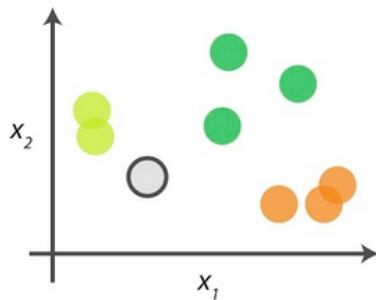
1. Масштабирование данных. Чтобы найти точку данных в многомерном пространстве признаков, было бы полезно, если бы все признаки были в одном масштабе. Следовательно, поможет нормализация или стандартизация данных.

2. Уменьшение размерности: KNN может работать плохо, если функций слишком много. Следовательно, могут быть реализованы методы уменьшения размерности, такие как выбор признаков и анализ основных компонентов.

2. Обработка отсутствующих значений: если из  $M$  функций отсутствуют данные одной функции для конкретного примера в обучающем наборе, мы не можем найти или рассчитать расстояние от этой точки. Поэтому необходимо удалить эту строку или вменение.

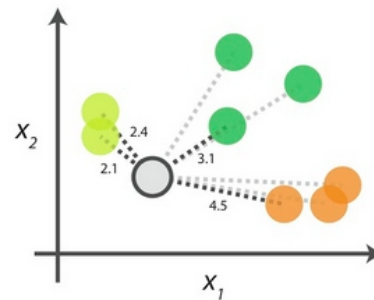
# kNN Algorithm

## 0. Look at the data











Say you want to classify the grey point into a class. Here, there are three potential classes - lime green, green and orange.

## 1. Calculate distances









Start by calculating the distances between the grey point and all other points.

## 2. Find neighbours

Point Distance		
 ... 	2.1	→ 1st NN
 ... 	2.4	→ 2nd NN
 ... 	3.1	→ 3rd NN
 ... 	4.5	→ 4th NN

Next, find the nearest neighbours by ranking points by increasing distance. The nearest neighbours (NNs) of the grey point are the ones closest in dataspace.

## 3. Vote on labels

Class	# of votes	
	2	→ Class  wins the vote! Point  is therefore predicted to be of class  .
	1	
	1	

Vote on the predicted class labels based on the classes of the  $k$  nearest neighbours. Here, the labels were predicted based on the  $k=3$  nearest neighbours.