kNN算法的核心思想是如果一个样本在特征空间中的k个最相邻的样本中的大多数属于某一个类别，则该样本也属于这个类别，并具有这个类别上样本的特性。该方法在确定分类决策上只依据最邻近的一个或者几个样本的类别来决定待分样本所属的类别。 kNN方法在类别决策时，只与极少量的相邻样本有关。由于kNN方法主要靠周围有限的邻近的样本，而不是靠判别类域的方法来确定所属类别的，因此对于类域的交叉或重叠较多的待分样本集来说，kNN方法较其他方法更为适合。

The core idea of the kNN algorithm is that if most of the k closest samples in the feature space belong to a certain category, the sample also belongs to that category and has the characteristics of the samples in that category. This method only determines the category of the sample to be divided based on the category of the nearest one or several samples in determining the classification decision. The kNN method is only relevant to a very small number of adjacent samples when making category decisions. Due to the fact that the kNN method mainly relies on limited neighboring samples around it, rather than relying on the method of discriminating class domains to determine the category it belongs to, the kNN method is more suitable for sample sets with more overlapping or overlapping class domains.

k近邻法(k-nearest neighbor, k-NN)是1967年由Cover T和Hart P提出的一种基本分类与回归方法。它的工作原理是：存在一个样本数据集合，也称作为训练样本集，并且样本集中每个数据都存在标签，即我们知道样本集中每一个数据与所属分类的对应关系。输入没有标签的新数据后，将新的数据的每个特征与样本集中数据对应的特征进行比较，然后算法提取样本最相似数据(最近邻)的分类标签。一般来说，我们只选择样本数据集中前k个最相似的数据，这就是k-近邻算法中k的出处，通常k是不大于20的整数。最后，选择k个最相似数据中出现次数最多的分类，作为新数据的分类。

The k-nearest neighbor (k-NN) method is a basic classification and regression method proposed by Cover T and Hart P in 1967. Its working principle is that there is a set of sample data, also known as a training sample set, and each data in the sample set has labels, that is, we know the corresponding relationship between each data in the sample set and its corresponding classification. After inputting unlabeled new data, compare each feature of the new data with the corresponding feature in the sample set, and then the algorithm extracts the classification label of the most similar data (nearest neighbor) in the sample. Generally speaking, we only select the first k most similar data in the sample dataset, which is the origin of k in the k-nearest neighbor algorithm. Usually, k is an integer not greater than 20. Finally, select the classification with the highest number of occurrences among the k most similar data as the classification for the new data.

距离的定义：只要满足非负、自反、三角不等式就可以称之为距离。主要有以下几种距离计算方法：马氏距离（闵可夫斯基距离）、欧氏距离、曼哈顿距离、切比雪夫距离、海明距离。

Definition of distance: As long as it satisfies non negative, reflexive, and triangular inequalities, it can be called distance. There are mainly several distance calculation methods: Markov distance (Minkowski distance), Euclidean distance, Manhattan distance, Chebyshev distance, and Hamming distance.

k-近邻法算法步骤如下：

1. 计算已知类别数据集中的点与当前点之间的距离；
2. 按照距离递增次序排序；
3. 选取与当前点距离最小的k个点；
4. 确定前k个点所在类别的出现频率；
5. 返回前k个点所出现频率最高的类别作为当前点的预测分类

The k-nearest neighbor algorithm steps are as follows:

1. Calculate the distance between points in a known category dataset and the current point;
2. Sort in ascending order of distance;
3. Select k points with the smallest distance from the current point;
4. Determine the frequency of occurrence of the first k points in the category;
5. Returns the category with the highest frequency of the first k points as the predicted classification for the current point.

k-近邻法算法优缺点分析

优点：

简单好用，容易理解，精度高，理论成熟，既可以用来做分类也可以用来做回归；

可用于数值型数据和离散型数据；

训练时间复杂度为O(n)；无数据输入假定；

对异常值不敏感

缺点

计算复杂性高；空间复杂性高；

样本不平衡问题（即有些类别的样本数量很多，而其它样本的数量很少）；

一般数值很大的时候不用这个，计算量太大。但是单个样本又不能太少，否则容易发生误分。

最大的缺点是无法给出数据的内在含义。

Analysis of the advantages and disadvantages of k-nearest neighbor algorithm

Advantages:

Simple and easy to use, easy to understand, high accuracy, mature theory, can be used for both classification and regression;

Can be used for both numerical and discrete data;

The training time complexity is O (n); Assumption of no data input;

Not sensitive to outliers

Shortcomings：

High computational complexity; High spatial complexity;

Sample imbalance problem (i.e. some categories have a large number of samples while others have a small number);

Usually, this is not used when the numerical value is large, as it requires too much computation. However, a single sample should not be too small, otherwise it is prone to misclassification.

The biggest drawback is the inability to provide the intrinsic meaning of the data.