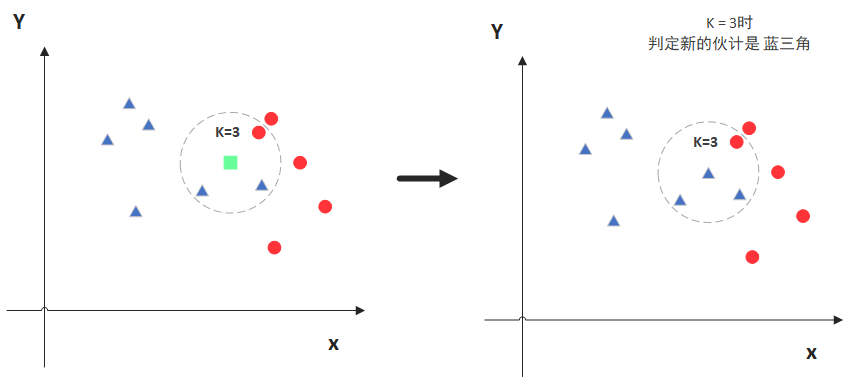
# [深入浅出KNN算法（一） KNN算法原理](https://www.cnblogs.com/listenfwind/p/10311496.html)

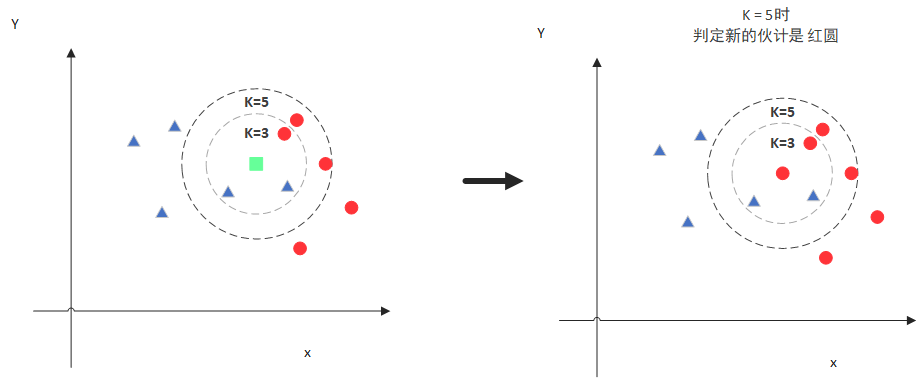
**一.KNN算法概述**

KNN可以说是最简单的分类算法之一，同时，它也是最常用的分类算法之一，注意KNN算法是有监督学习中的分类算法，它看起来和另一个机器学习算法Kmeans有点像（Kmeans是无监督学习算法），但却是有本质区别的。那么什么是KNN算法呢，接下来我们就来介绍介绍吧。

**二.KNN算法介绍**

KNN的全称是K Nearest Neighbors，意思是K个最近的邻居，从这个名字我们就能看出一些KNN算法的蛛丝马迹了。K个最近邻居，毫无疑问，K的取值肯定是至关重要的。那么最近的邻居又是怎么回事呢？**其实啊，KNN的原理就是当预测一个新的值x的时候，根据它距离最近的K个点是什么类别来判断x属于哪个类别**。听起来有点绕，还是看看图吧。  


图中绿色的点就是我们要预测的那个点，假设K=3。那么KNN算法就会找到与它距离最近的三个点（这里用圆圈把它圈起来了），看看哪种类别多一些，比如这个例子中是蓝色三角形多一些，新来的绿色点就归类到蓝三角了。



但是，**当K=5的时候，判定就变成不一样了**。这次变成红圆多一些，所以新来的绿点被归类成红圆。从这个例子中，我们就能看得出K的取值是很重要的。

明白了大概原理后，我们就来说一说细节的东西吧，主要有两个，**K值的选取**和**点距离的计算**。

**2.1距离计算**

要度量空间中点距离的话，有好几种度量方式，比如常见的曼哈顿距离计算，欧式距离计算等等。不过通常KNN算法中使用的是欧式距离，这里只是简单说一下，拿二维平面为例，，二维空间两个点的欧式距离计算公式如下：

二维空间欧式距离

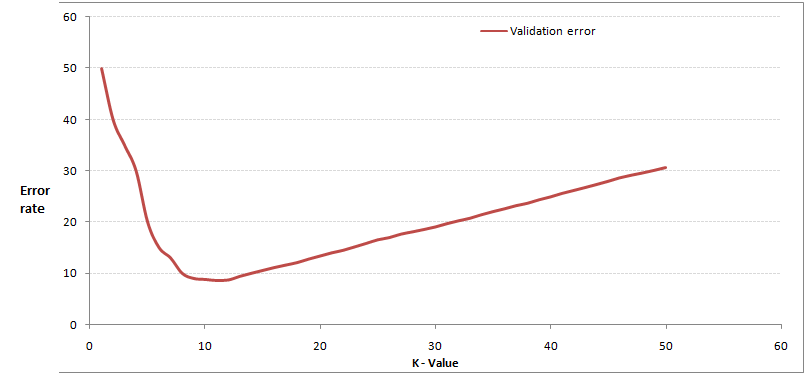
这个高中应该就有接触到的了，其实就是计算（x1,y1）和（x2,y2）的距离。拓展到多维空间，则公式变成这样：

多维空间欧式距离

这样我们就明白了如何计算距离，KNN算法最简单粗暴的就是将预测点与所有点距离进行计算，然后保存并排序，选出前面K个值看看哪些类别比较多。但其实也可以通过一些数据结构来辅助，比如最大堆，这里就不多做介绍，有兴趣可以百度最大堆相关数据结构的知识。

**2.2 K值选择**

通过上面那张图我们知道K的取值比较重要，那么该如何确定K取多少值好呢？答案是通过交叉验证（将样本数据按照一定比例，拆分出训练用的数据和验证用的数据，比如6：4拆分出部分训练数据和验证数据），从选取一个较小的K值开始，不断增加K的值，然后计算验证集合的方差，最终找到一个比较合适的K值。

通过交叉验证计算方差后你大致会得到下面这样的图：  


这个图其实很好理解，当你增大k的时候，一般错误率会先降低，因为有周围更多的样本可以借鉴了，分类效果会变好。但注意，和K-means不一样，当K值更大的时候，错误率会更高。这也很好理解，比如说你一共就35个样本，当你K增大到30的时候，KNN基本上就没意义了。

所以选择K点的时候可以选择一个较大的临界K点，当它继续增大或减小的时候，错误率都会上升，比如图中的K=10。具体如何得出K最佳值的代码，下一节的代码实例中会介绍。

**三.KNN特点**

KNN是一种**非参的**，**惰性**的算法模型。什么是非参，什么是惰性呢？

**非参**的意思并不是说这个算法不需要参数，而是意味着这个模型不会对数据做出任何的假设，与之相对的是线性回归（我们总会假设线性回归是一条直线）。也就是说KNN建立的模型结构是根据数据来决定的，这也比较符合现实的情况，毕竟在现实中的情况往往与理论上的假设是不相符的。

**惰性**又是什么意思呢？想想看，同样是分类算法，逻辑回归需要先对数据进行大量训练（tranning），最后才会得到一个算法模型。而KNN算法却不需要，它没有明确的训练数据的过程，或者说这个过程很快。

**KNN算法的优势和劣势**

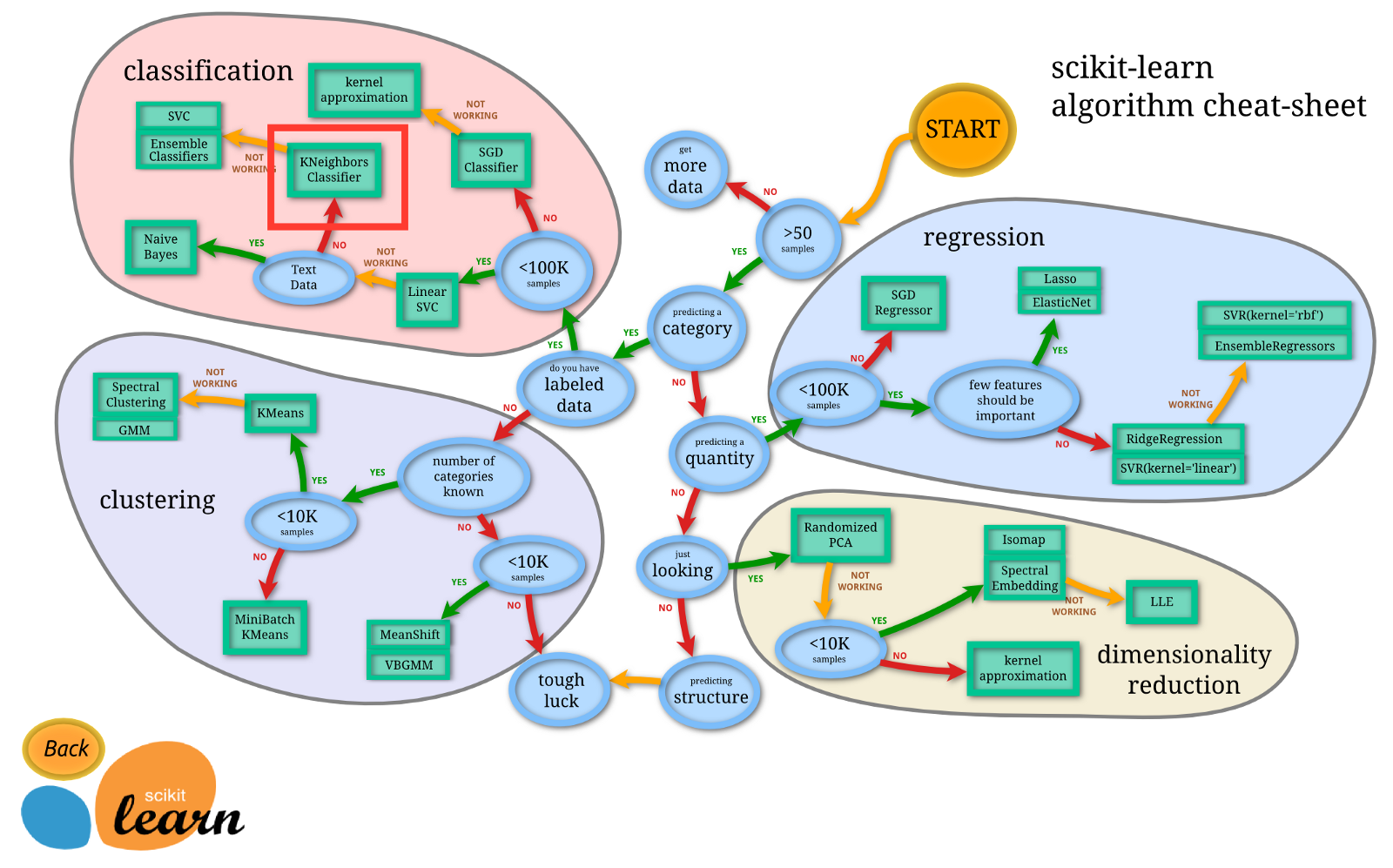
了解KNN算法的优势和劣势，可以帮助我们在选择学习算法的时候做出更加明智的决定。那我们就来看看KNN算法都有哪些优势以及其缺陷所在！

**KNN算法优点**

1. 简单易用，相比其他算法，KNN算是比较简洁明了的算法。即使没有很高的数学基础也能搞清楚它的原理。
2. 模型训练时间快，上面说到KNN算法是惰性的，这里也就不再过多讲述。
3. 预测效果好。
4. 对异常值不敏感

**KNN算法缺点**

1. 对内存要求较高，因为该算法存储了所有训练数据
2. 预测阶段可能很慢
3. 对不相关的功能和数据规模敏感

至于什么时候应该选择使用KNN算法，sklearn的这张图给了我们一个答案。  
  
简单得说，当需要使用分类算法，且数据比较大的时候就可以尝试使用KNN算法进行分类了。

OK，本次先对KNN算法做一个介绍，下一节解析sklearn的参数，以及K值选取。