非参数方法—K-nearest neighbors（KNN）

王媛媛 经济系硕士 15320171151909

2019 年 4 月 13 日

0.1 非参数方法与参数方法

0.1.1 参数方法

当选择一个目标模型，利用样本数据去估计模型中的系数，得到估计后的模型，然后这个模型去预测所需要的值，这就是参数方法的基本思路。最常用的参数方法是把目标函数假设为线性模型，线性的假设大大简化了后续的计算过程，也使得系数容易估计得出。对数据进行分类的方法中常见的参数方法有 Logistic 回归，Logistic 回归分析法也具有参数方法的特点，即具有待定的系数需要估计，与线性模型不同的是，它的因变量不需要是连续的，可以是二分类变量甚至多种分类的，所以 Logistic 回归可以用来做数据的选择分类。并且引用 Logistic 回归主要是因为当因变量为二分类变量时，如果直接使用线性回归的模型，会造成方程两边取值区间不同和普遍的非直线关系。因为因变量为二分类变量，某个概率作为方程的因变量估计值取值范围为 0-1，但是，方程右边取值范围是无穷大或者无穷小，所以才引入了 Logistic 回归。

参数方法的优点是：很容易理解和解释结果，对数据的处理比较的快速，并且只要估计出了模型，每次计算就不需要再用全部的数据，有了模型可以很快的得出预测值。当然，参数方法也有其局限性：由于实际的总体分布是不可观的，所以我们永远都不知道实际分布应该是什么，可能要比我们假设的模型要复杂的多，所以当模型选择错误时，就会产生很大的错误估计，所以当背后的总体分布过于复杂时，参数方法的准确率就不够了。

0.1.2 非参数方法

实际中很多总体的分布不是已知的，非参数方法不需要假设模型，可以用来处理任意分布的数据。非参数方法的概率密度估计主要是先从样本中估计概率密度函数 $P(x|\omega_j)$，然后直接估计出后验概率 $P(\omega_j|x)$。

一个向量 x 落在区域 R 中的概率为：$P = \int_R p(x')dx'$，因此可以通过估计 P 来估计概率密度函数 p。

假设 n 个样本 $x_1, \cdots, x_n$，都是根据概率密度函数 $p(x)$ 独立分布抽取而来，k 个样本落在区域 R 的概率服从二项式定理：$P_k = (\frac{n}{k})P^k(1P)^{nk}$。k 的期望为 nP，比值 $\frac{n}{k}$ 就是对 P 的一个很好的估计，如果 $p(x)$ 是连续的，区域 R 很小，那么可以推出 $p(x) \approx \frac{k/n}{V}$。为了估计 x 的概率密度函数，$V_n$ 为区域 R_n 的体积，k_n 为落在 R_n 中的样本个数，$p_n(x)$ 表示对 $p(x)$ 的第 n 次估计：$p_n(x) = \frac{k_n/n}{V_n}$。

$p_n(x)$ 收敛到 $p(x)$ 的条件：[1]

$$lim_{n\to\infty}V_n = 0$$

$$lim_{n\to\infty}k_n = \infty$$

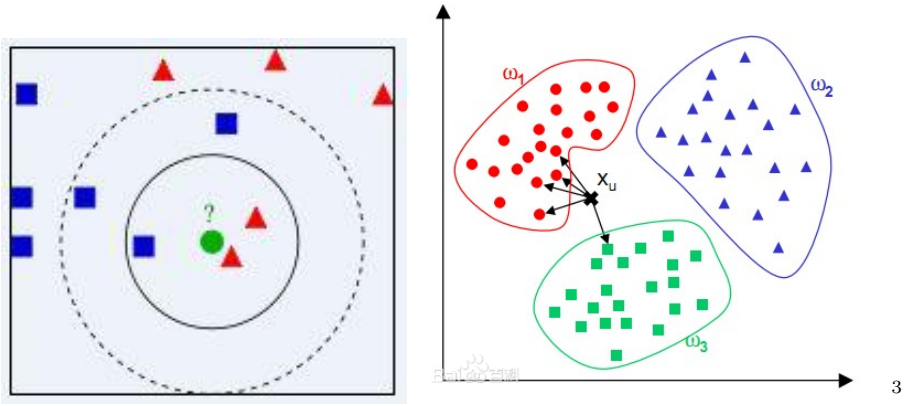
$$lim_{n\to\infty}\frac{k_n}{n} = 0$$

[1]参考自 https://blog.csdn.net/u013413471/article/details/79123405

0.2 K 最近邻居分类算法（KNN）

所谓 K 最近邻，就是 k 个最近的邻居的意思，说的是每个样本都可以用它最接近的 k 个邻居来代表。Cover 和 Hart 在 1968 年提出了最初的邻近算法。KNN 是一种分类算法，它属于基于实例的学习（instance-based learning），属于懒惰学习（lazy learning）。即 KNN 没有显式的学习过程，也就是说没有训练阶段，数据集事先已有了分类和特征值，待收到新样本后直接进行处理。与急切学习（eager learning）相对应。

KNN 的基本思路是：如果一个样本在空间内的 k 个邻居的样本中大多数属于某一个类别，那么这个样本也被认为是这个类别的。因为我们可以认为大多数分布都是连续的，所以邻居可以告诉我们要预测的信息。如图[2]，如果我们想要确定图中绿点是属于什么颜色的，要做的就是选出距离目标最近的 k 个点，看这 k 个点中大多数是红色还是蓝色。当 k=3 时，最近的 3 个点红色最多，所以当 k=3 时，预测目标点为红色；当 k=5 时，我们可以发现最近的 5 个点中蓝色为 3 个，比红色的多，所以我们预测目标点为蓝色。



[3]

由上面的例子也可以看出，k 的取值非常的重要。如果 k 的值比较小的话，其实意味着整体的模型比较的复杂，但是如果一旦有错误的点，偏差就会很大，容易出现过度的拟合，就像是线性回归模型里面拟合了所有的点一样，如果有噪声存在，就会偏离背后真正的总体分布。如果 k 的值比较大的话，误差就会增咋，这时即使与目标较远的数据点也会对目标产生影响，预测就容易发生错误。总之，k 越小的话意味着复杂度越高，k 越大意味着模型越平滑，但是容易有偏差。

KNN 算法有其优点，当然也存在缺陷。优点在于这种分类算法简单容易实现，便于理解，不需要估计模型参数，这对于非很复杂、不清楚真实分布的模型是非常适用的。并且 KNN 分类方法非常适合多分类的问题，即对象是具有很多个类别标签的，这种情况下 KNN 分类方法表现较好。但是 KNN 分类方法也有很多问题存在，如果样本集是随机分布，同类的分布较为分散，那么 KNN 的表现就不是很好，预测没有那么准确。另外如果样本不平衡时，例如一个类别的样本容量很大，其他类别的样本容量较小时，有可能导致在预测新的样本时，周围 k 个邻居中大样本容量的样本占多数，可能带来预测偏差。这种分类算法还有一个重要的缺点是计算量太大，如果样本数据集很大的时候，这种算法需要存储所有的数据就会耗用很大的空间，并且在输入新样本后要计算它与其他所有数据点的距离，非常耗时。但是针对这些缺陷，也出现了一些改进的方法。例如针对计算效率低下，可以先对样本进行简化，删除对分类结果影响较小的属性，；针对分类的效果不好，可以采用权值的方法，即和目标样本小的距离小的

[2]参考自 https://www.cnblogs.com/jyroy/p/9427977.html
[3]该图片来自 https://baike.baidu.com/item/%E9%82%BB%E8%BF%91%E7%AE%97%E6%B3%95/1151153?fr=aladdin

权值大等。

总之，没有哪一种统计方法适用于一切分布，我们要根据实际情况需要来选择最合适的统计方法，正确方法的选择，是统计分析成功的前提！