KNN算法

KNN(K-nearest neighbors)核心思想

最近邻算法：以全部训练样本作为代表点计算未知样本与所有训练样本的距离，并以最近邻的类别作为决策未知样本类别的唯一依据。缺点是对于噪声数据过于敏感。

KNN：选择未知样本一定范围内确定个数的K个样本，该K个样本大多数属于某一类型，则未知样本判定为该类型。KNN是一种基于实例的学习，或者说是局部近似和将所有计算推迟到**分类之后**的懒惰学习(lazy learning)，没有显示的训练过程。

> KNN可以用于分类和回归

1. 在KNN分类中，输出是一个分类族群。**一个对象的分类是由其邻居的多数表决确定的**(即选择这k个样本中出现最多的类别作为结果)
2. 在KNN回归中，输出是该对象的属性值。该值是K个最近邻居的平均值

KNN算法的执行与分析

KNN的执行过程

1. 初始化距离为最大值
2. 计算未知样本(未知标签)和每个训练样本(已知标签)的距离$dist$
3. 得到目前K个最邻近样本中的最大距离$maxdist$
4. 如果$dist < maxdist$，则将该训练样本作为K最近邻样本
5. 重复操作步骤2、3、4知道未知样本和所有训练样本的距离都算完
6. 统计K个最近邻样本中每个类别出现的次数
7. 选择出现频率最大的类别作为未知样本的类别

KNN的核心要素：距离度量

> L_p距离(闵可夫斯基距离)：$L_p(x_i,x_j)=[\sum_{l=1}^{n}|x_i^{(l)}-x_j^{(l)}|^p]^{\frac{1}{p}}$
>
> 1. $p=1$时，该距离称为曼哈顿距离(L_1距离)。$L_1(x_i,x_j)=\sum_{l=1}^{n}|x_i^{(l)}-x_j^{(l)}|$
> 2. $p=2$时，该距离称为欧氏距离(L_2距离)。$L_2(x_i,x_j)=[\sum_{l=1}^{n}|x_i^{(l)}-x_j^{(l)}|^2]^{\frac{1}{2}}$
> 3. $p\to\infty$时，该距离称为切比雪夫距离。$L=max(|x_{1k}-x_{2k}|)$

KNN参数的选择

K值的选取往往取决于数据。一般情况下，在分类时较大的K值可以减小噪声的影响，但会使得类别的界限变得模糊，减少了学习的估计误差但是学习的近似误差增大；较小的K值会使模型容易发生过拟合，减少了模型学习的近似误差但是增大了估计误差。

e.g. 在二分类问题中，选取K为奇数有助于避免两个分类平票的情况，选取最佳经验K值的方法是自助法。

KNN的优缺点

KNN的本质是**对特征空间的划分**，适于数值型/标称型(只在有限目标集中取值)数据。

优点：

1. 精度高
2. 对异常值不敏感
3. 无数据输入假定

缺点：

1. 计算复杂度高
2. 空间复杂度高
3. 只考虑近邻不同类别的样本数量而忽略了距离