# K-近邻算法

1. 农近邻算法（kNN）：

采用测量不同特征值之间的距离方法进行分类。

优点：精度高、对异常值不敏感、无数据输入假定。

缺点：计算复杂度高、空间复杂度高。

适用数据范围：数值型和标称型。

工作原理：

存在一个样本数据集合（训练样本集），并且样本集中每个数据都存在标签，即样本集中每一数据与所属分类的对应关系。

输入没有标签的新数据后，将新数据的每个特征与样本集中数据对应的特征进行比较，利用算法提取样本中特征最相似数据（最近邻）的分类标签。

选择样本数据集中前A个最相似的数据（A<=20）

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。

k- 近邻算法的一般流程：

(1) 收集数据：可以使用任何方法。

(2) 准备数据：距离计算所需要的数值，最好是结构化的数据格式。

(3) 分析数据：可以使用任何方法。

(4) 训练算法：此步骤不适用于农近邻算法。

(5) 测试算法：计算错误率。

(6) 使用算法：首先需要输入样本数据和结构化的输出结果，然后运行 4- 近邻算法判定输 入数据分别属于哪个分类，最后应用对计算出的分类执行后续的处理。

伪代码：

对未知类别属性的数据集中的每个点依次执行以下操作：

(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个点出现频率最高的类别作为当前点的预测分类。