k-邻近算法

实现原理

讲师：李宁

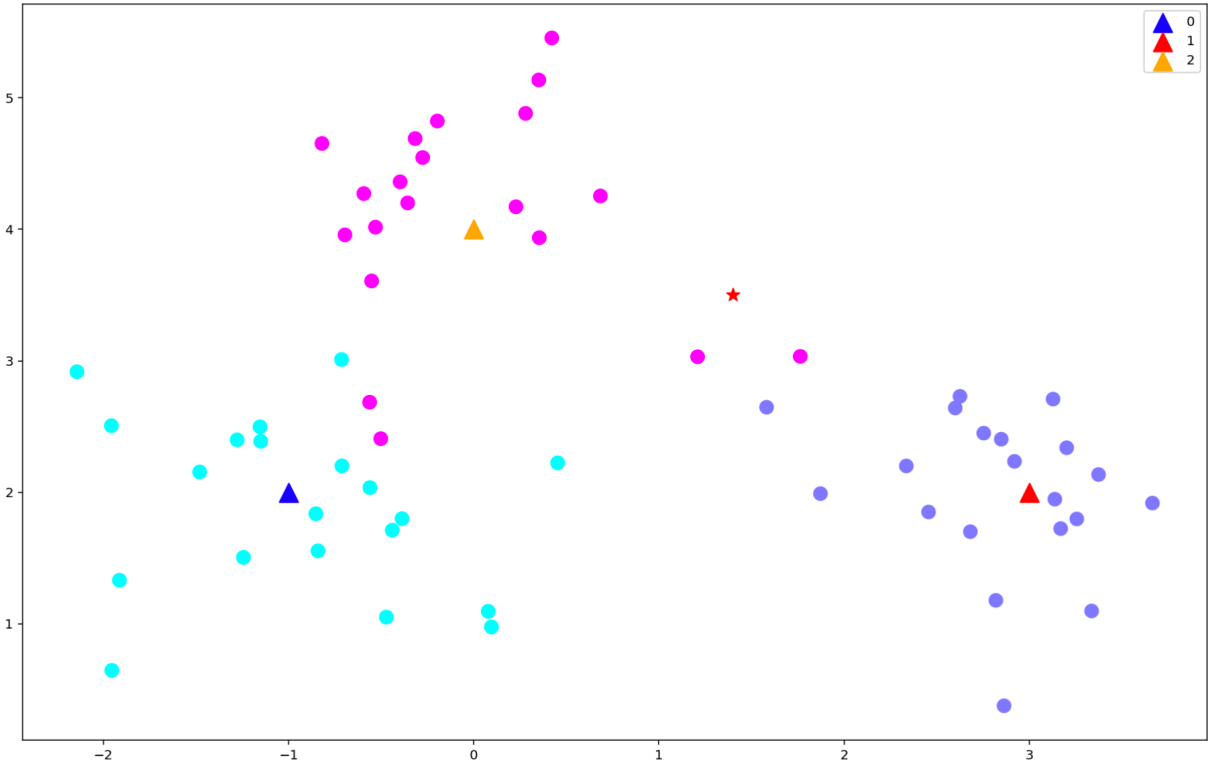
K-邻近算法

K-邻近：KNN(K-Nearest Neighbor）

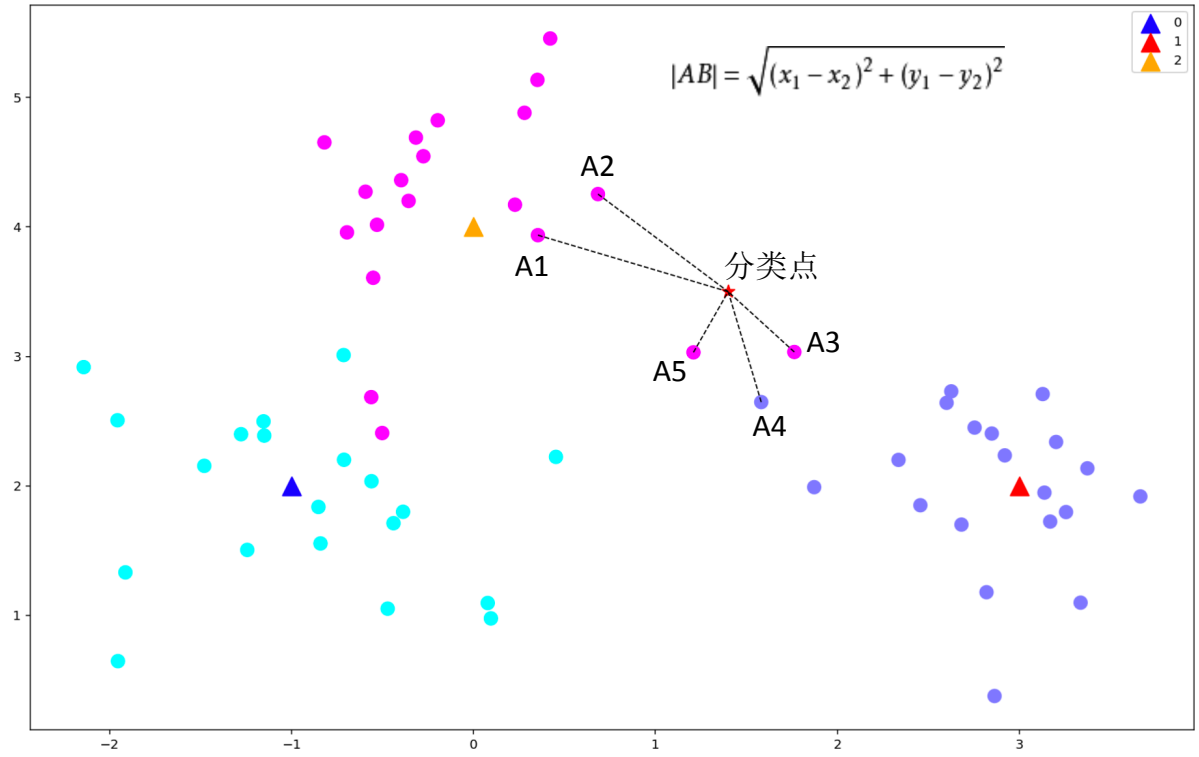
一种有监督的机器学习算法，可以用于分类问题，也可以用于回归（预测）问题。

使用k-邻近算法，需要提供一组做了标记的数据集进行训练。

实现原理



距离待分类点最近的k个点



K = 5

[A1 A2 A3 A4 A5]

类别1：A4
类别2：A1 A2 A3 A5

类别1一共投了1票
类别2一共投了4票

类别2获胜，所以分类点属于类别2

算法描述

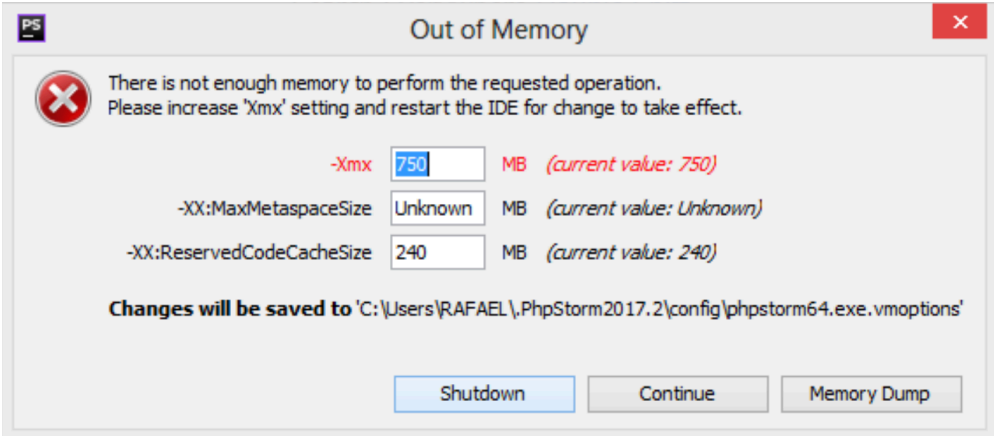
1. 遍历训练集中的所有样本，计算训练集中每个样本与测试集中的每个样本的距离，并将距离保存在一个数组Distance中。如果训练集的样本数是n，测试集的样本数是m，那么需要计算n × m个距离。
2. 对Distance数组进行排序，取距离最近的k个点，记为X_knn。
3. 在X_knn中统计每个类别的个数，例如，类别1在X_knn中有c1个样本，类别2在X_knn中有c2个样本，类别3在X_knnz中有c3个样本。这里假设只有3个类别。
4. 比较c1、c2和c3，哪个最大，待分类的样本就属于哪一个类别，例如，c2最大，那么待分类样本就属于类别2。

算法的优点

准确性高，对异常值和噪声有较高的容忍度。

算法的缺点

计算量较大，对内存的需求也比较大。时间复杂度是n × m。因此，在做实验时，频繁执行k-邻近算法，可能会造成内存不足的现象。



PyCharm

算法的参数

只有一个参数：k

参数的设置需要根据数据来决定。K值越大，模型的偏差越大，对噪声数据越不敏感，当k值过大时，可能会造成欠拟合；k值越小，模型的方差就会越大，当k值太小，就会造成模型过拟合。

算法的变种

变种1： 增加邻居的权重，例如，距离越近，权重越高

变种2： 使用一定半径内的点取代距离最近的k个点

“极客起源”技术公众号

“极客题库”小程序

“欧瑞科技”官方公众号