KNN 实验报告

赵泽华
（重庆大学软件学院，重庆，401331）

KNN 算法简介

k 近邻法（k-nearest neighbor，k-NN）是一种基本分类与回归方法。K 近邻法的输入为实例的特征向量，对应于特征空间的点；输出为实例的类别，可以取多个类别。

算法一：K 近邻法

输入：训练数据集 $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$，其中，$x_i \in X \subseteq R^n$ 为实例的特征向量，$y_i \in Y=\{c_1,c_2,\cdots,c_K\}$ 为实例的类别，$i=1,2,\cdots,N$；实例特征向量 x；

输出：实例 x 所属的类 y

（1）根据给定的距离度量，在训练集 T 中找出与 x 最邻近的 k 个点，涵盖这 k 个点的 x 邻域记做 $N_k(x)$

（2）在 $N_k(x)$ 中根据分类决策规则（如多数表决）决定 x 的类别 y：

$$y=\arg\max_{c_j}\sum_{x_i\in N_k(x)} I(y_i=c_j), i=1,2,\cdots,N; j=1,2,\cdots,K$$

其中，I 为指示函数，即当 $y_i=c_j$ 时 I 为 1，否则 I 为 0

K 近邻法有三个基本要素：k 值的选择、距离度量以及分类决策规则。

（1）k 值的选择

① k 值过小：整体模型变得复杂，容易发生过拟合，结果易受近邻点的影响。如果近邻点是噪声点，则会对分类结果造成干扰。

② k 值过大：整体模型变得简单，但较远点也会对结果产生影响

（2）距离度量

常见的距离度量公式如：

① L_p 距离： $L_p(x_i,x_j)=(\sum_{l=1}^{n}|x_i^{(l)}-x_j^{(l)}|^p)^{\frac{1}{p}}$

② 欧式距离： $L_2(x_i,x_j)=(\sum_{l=1}^{n}|x_i^{(l)}-x_j^{(l)}|^2)^{\frac{1}{2}}$

③ 曼哈顿距离： $L_1(x_i,x_j)=\sum_{l=1}^{n}|x_i^{(l)}-x_j^{(l)}|$

（3）分类决策规则

通常采用“多数表决”作为分类决策规则。多数表决规则等价于经验风险最小化

数据集介绍

本次实验选用 Iris Plants Database 作为实验数据集。该数据集中共有 150 个样本，每个样本包含四个属性：sepal length、sepal width、petal length、petal width，类别包括三种：Iris Setosa、Iris Versicolour 和 Iris Virginica。每个样本的四种属性均有值，无缺省情况。三种类别各有 50 个样本。具体数据统计如下表：

	MIN	MAX	MEAN
SEPAL LENGTH（CM）	4.3	7.9	5.84
SEPAL WIDTH（CM）	2.0	4.4	3.05
PETAL LENGTH（CM）	1.0	6.9	3.76
PETAL WIDTH（CM）	0.1	2.5	1.20

表一：数据集信息表

数据集来源：http://archive.ics.uci.edu/ml/datasets/Iris

实验设置

本次实验选用数据集中全部样本进行测试，通过五折交叉验证的方式进行实验，即 20%测试集和 80%训练集。

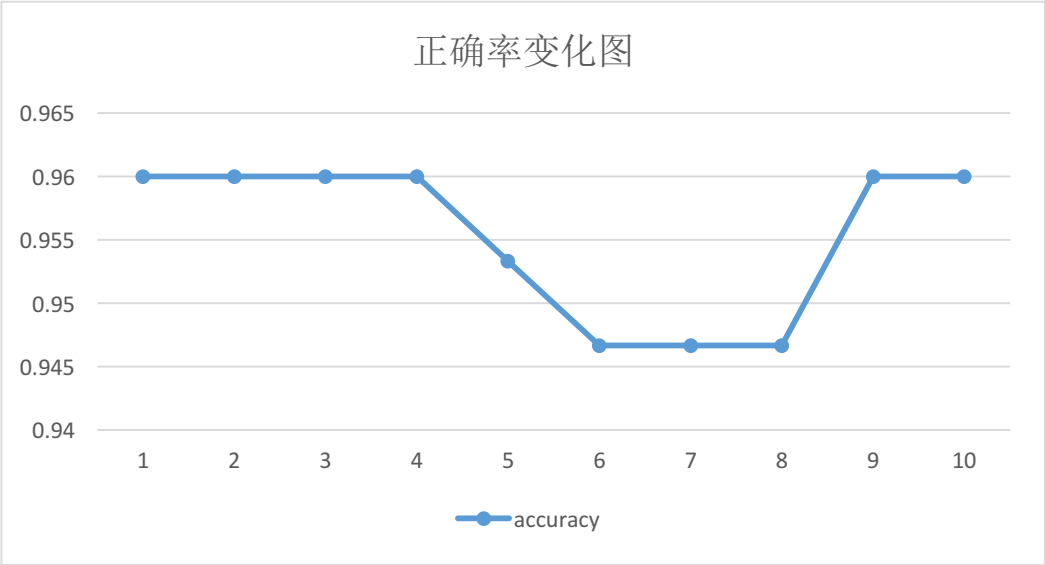
考虑到多数表决会出现两个类别数目相同的情况，为了减小分类误差，本次实验中分类决策规则选用平均距离最小化来进行划分。首先计算 k 个近邻与待分类样本的距离，然后计算每个类别到待分类样本的平均距离，平均距离最小的类别即为待分类样本的类别。

本实验中度量距离采用欧式距离。选用正确率（accuracy）作为评价指标。

实验环境为 python2.7

实验结果

分别对 k 取 1~10 进行测试，实验结果如下图所示：



图一：正确率变化图

正确率在 K 值取较小值时较为稳定，说明待分类样本和其最近邻样本同属一个类别概率较高，之后随着 K 值增大，新加入的样本点可能会起到一定程度的干扰，所以正确率略有下降。但是随着 K 值的再次增加，正确率会再次上升并重新趋于稳定。