

模式识别实验四

实验人：叶平

实验内容：

P. 411, Prob. 7

1、研究“验证技术”未必会改善分类器的性能的情况。实验中分类器为“k-近邻分类器”，其中 k 是通过“验证技术”来设置。考虑一个二维的两类问题，其先验分布在范围 $0 \leq x_i \leq 1$ ($i=1, 2$) 内是均匀分布。

(a) 首先形成一个 20 个点的测试集 D_{test} ——10 个点属于 w_1 ，10 个点属于 w_2 ——并根据“均匀分布”的方式任意选出。

(b) 接下来产生 100 个点——每类 50 个模式。置 $\gamma=0.1$ ，将该集合划分成一个训练集 D_{train} (90 个点) 和一个验证集 D_{val} (10 个点)。

(c) 产生一个“k-近邻分类器”，其中 k 一直增加到验证误差的第一个极小值被找到。(限定 k 为奇数值，以避免出现不分胜负的情况。) 现利用测试集来确定该分类器的误差。

(d) 重复 (c)，但通过验证误差的第一个极大值来确定 k。

(e) 重复 (c) 和 (d) 5 次，注意所有 10 种情况下的测试误差。

(f) 讨论结论——尤其是，它们是如何的依赖于（或不依赖于）其数据是“均匀分布”的事实。

实验结果：

运行实验 4 文件夹下 Res.m 文件可以得到实验结果，如下图所示：

Command window			
第1次的最差的k=1	错误率为0.400000	最好的k=3	错误率为0.300000
第2次的最差的k=3	错误率为0.700000	最好的k=13	错误率为0.100000
第3次的最差的k=1	错误率为0.500000	最好的k=3	错误率为0.500000
第4次的最差的k=5	错误率为0.600000	最好的k=3	错误率为0.700000
第5次的最差的k=3	错误率为0.500000	最好的k=5	错误率为0.500000

由上图可以看出，无论是在训练集上表现的最好的 k 还是表现的最差的 k ，在验证集上的表现大体相当。错分率都在 0.5 上下波动。事实上，由于先验分布是在 $(0,1)$ 范围内的均匀分布，无论是否采用验证技术，分类的错分率应该都在 0.5 左右。通过画出 k 从 1 到 20 (k 为奇数)，KNN 分类器在测试集和验证集上的错分率的折线图(见下图)可以清楚的看出，错分率围绕 0.5 上下波动。

