

第五讲 分类算法 作业

724602210635 黄奕中

分类

序号	分类器	准确率	设置的参数
01	支持向量机 SVM	1.0	c=1.0; kernal="linear"
02	支持向量机 SVM	1.0	c=10.0; kernal="linear"
03	支持向量机 SVM	1.0	c=0.1; kernal="linear"
04	支持向量机 SVM	1.0	c=1.0; kernal="rbf"
05	支持向量机 SVM	1.0	c=1.0; kernal="poly"; degree=2
06	支持向量机 SVM	1.0	c=1.0; kernal="poly"; degree=3
07	支持向量机 SVM	1.0	c=1.0; kernal="poly"; degree=4
08	支持向量机 SVM	0.5	c=1.0; kernal="sigmoid"
09	K近邻 KNN	0.5	default
10	K近邻 KNN	1.0	n_neighbors=3
11	K近邻 KNN	0.0	n_neighbors=7
12	K近邻 KNN	1.0	n_neighbors=3; metric="manhattan"
13	K近邻 KNN	0.5	n_neighbors=3; metric="chebyshev"
14	决策树 Decision Tree	0.0	default
15	决策树 Decision Tree	0.46895	splitter="random"; repeated for 100000 times and averaged
16	决策树 Decision Tree	0.0	criterion="entropy"
17	决策树 Decision Tree	0.0	max_depth=100
18	决策树 Decision Tree	0.0	max_depth=10
19	决策树 Decision Tree	0.5	min_samples_split=6
20	决策树 Decision Tree	0.5	min_samples_leaf=2
21	决策树 Decision Tree	1.0	min_samples_leaf=8
22	决策树 Decision Tree	0.5	min_samples_leaf=20
23	决策树 Decision Tree	1.0	max_features="log2"
24	决策树 Decision Tree	0.0	max_features="sqrt"

分析

- SVM : 在各参数下均表现良好。
- KNN : 结果随参数改变而变化较大，且存在一个较为奇怪的问题：
 - 一般来说， K 越大，分类结果越准确，但在这组数据中， $K = 3$ 反而优于 $K = 5$ 或 $K = 7$ ，这可能是因为数据集过小，无法完全反应算法真实水平。
- DT : 结果随参数改变而变化较大，且普遍较差。
 - 将 splitter 改为 "random" 后，结果如其名字几乎完全随机，多次实验取均值后发现甚至不如完全随机（ $0.46895 < 0.5$ ）。
 - 考虑到 DT 在实际生产生活中的广泛使用，其结果较差应该也是由于数据集过小。