深圳大学研究生课程论文

题目 Semester-end test paper 成绩

专业 计算机科学与技术 课程名称、代码 MachineLearning 2703004

年 级 研一 姓 名 杨 树

学 号 1800271003 时 间 2019 年 1 月 1 日星期二

任课教师 王熙照

Semester-end test paper

一、训练数据的统计值以及可视化

1. 训练数据集“Training-set.csv”的每个样本共有五个属性值，分别是“id”“a”“b”“c”“t”这五个。“id”代表编号对模型的训练没有帮助。属性“t”代表目标值，共有 0 和 1 这两类，其中标签为 0 的样本有 3476 个，标签为 1 的样本有 3524 个，共计 7000 个训练样本数据。
2. 如表 1 所示，展示的是训练数据的各种统计值。从表中数据可以初步推断，标签为 0 的样本的三个属性值均在[-12, 12]区间内，且数据的三个属性值都集中在 0 附近（从三个属性的第 25 百分位数大于-4 和第 75 百分位数小于 4 推测得到）；标签为 1 的样本的三个属性值均在[-22, 22]区间内，且数据的三个属性值都均匀分布在区间内（从三个属性的第 25 百分位数在-10 附近和第 75 百分位数在 10 附近推测得到）。

表 1　训练集数据的各种统计值

	标签为 0 的样本			**标签为 1 的样本**		
属性	**a**	**b**	**c**	**a**	**b**	**c**
均值	0.086	- 0.133	- 0.038	-0.223	0.007	0.365
方差	4.852	4.736	4.751	12.308	12.521	12.258
最小值	-11.592	-11.743	-11.828	-21.985	-21.994	-21.971
第 25 百分位数	-3.733	-3.815	-3.719	-10.659	-10.332	-9.942
中位数	-0.008	-0.143	-0.065	-0.385	-0.055	0.550
第 75 百分位数	3.802	3.455	3.544	10.310	10.249	10.554
最大值	11.941	11.838	11.567	21.999	21.988	21.989

3. 可将三种属性值对应三维空间中的坐标，用两种颜色来标记标签为 0 和标签为 1 的样本。如图 1 所示，左上的横坐标“a”纵坐标“b”，右上的横坐标“a”，纵坐标“c”，左下的横坐标“b”纵坐标“c”，右下的图为沿垂直于属性“c”的坐标轴的横截面，蓝色点是标签为 0 的样本，绿色点是标签为 1 的样本。可以看出，标签为 0 的样本分布在一个球心为坐标原点、半径为 12 的球体中，标签为 1 的样本分布在一个中间挖空的立方体中。
4. 为了更清楚的观察数据的分布情况，对训练数据集的每个样本计算其到坐标原点的 Euclidean 距离，并进行相关的统计，如表 2 所示。可以看到球心为坐标(0, 0, 0)、半径为 11 的球面可以近似对两个标签的数据进行划分。

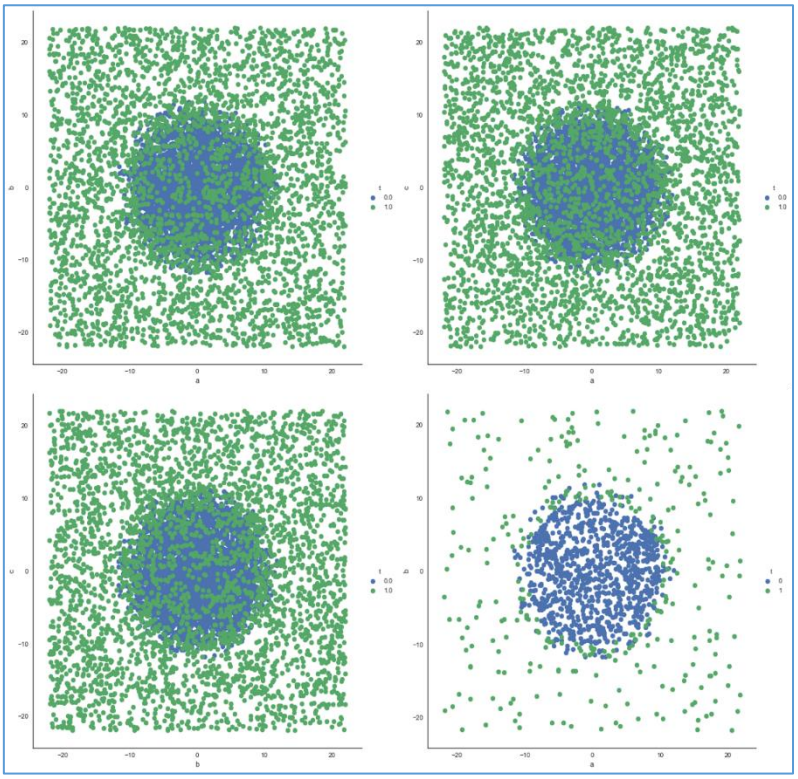

图 1　训练集数据在空间中的分布

表 2　训练集数据到坐标原点的 Euclidean 距离统计

标签	均值	方差	最小值	第 25 百分位数	中位数	第 75 百分位数	最大值
0	7.988	2.178	0.687	6.594	8.343	9.513	11.995
1	20.540	6.057	10.008	15.884	21.112	25.109	36.506

5.　下表 3 展示的是测试集数据的各种统计值。测试集共 1000 个样本。

表 3　测试集数据的各种统计值

	测试集样本		
属性	a	b	c
均值	0.086	- 0.133	- 0.038
方差	4.852	4.736	4.751
最小值	-11.592	-11.743	-11.828
第 25 百分位数	-3.733	-3.815	-3.719
中位数	-0.008	-0.143	-0.065
第 75 百分位数	3.802	3.455	3.544
最大值	11.941	11.838	11.567

6. 下图 2 展示的是测试集数据在三维空间中的位置，左上的横坐标“a”纵坐标“b”，右上的横坐标“a”，纵坐标“c”，左下的横坐标“b”纵坐标“c”，右下的图为沿垂直于属性“c”的坐标轴的横截面。

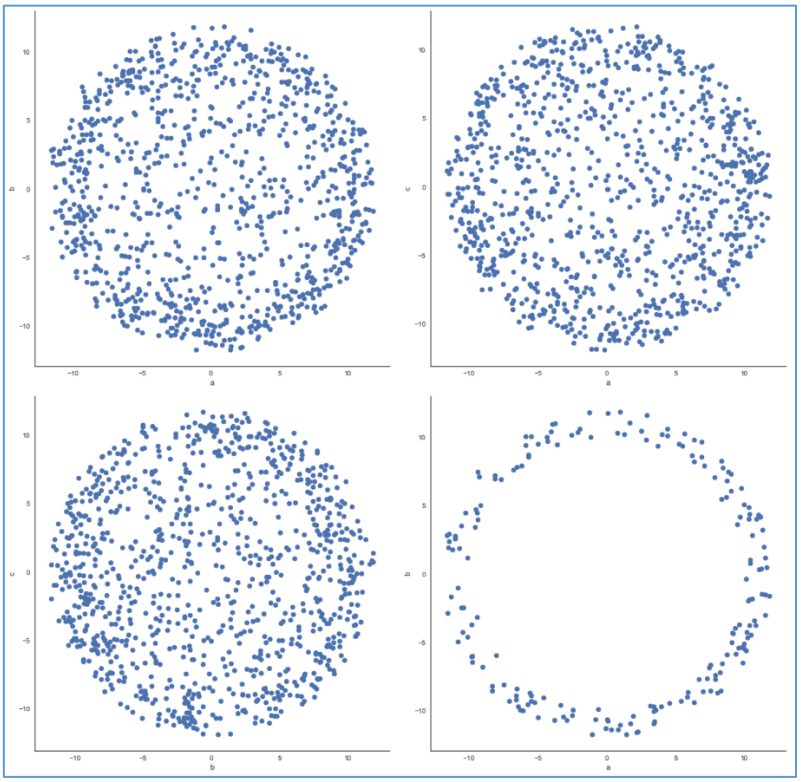

图 2　测试集数据在空间中的分布

7. 对测试集的每个样本计算其到坐标原点的 Euclidean 距离，并进行统计得到下表 4 中的数据。

表 4　测试集数据到坐标原点的 Euclidean 距离统计

	测试集样本到坐标原点的 Euclidean 距离统计
均值	11.062
方差	0.575
最小值	10.000
第 25 百分位数	10.550
中位数	11.091
第 75 百分位数	11.562
最大值	11.995

8. 从表 4 可以看出，测试集样本距离坐标原点最近为 10、最远为 12，主要分布在球心为坐标(0, 0, 0)、直径为 22 的球面附近。

二、模型的训练及测试

9. 通过前边的数据可视化和数据集统计数据的分析，我决定选用 K 近邻（K NearestNeighbor）和支持向量机（Support Vector Machine）这两种算法来建立模型。
10. 使用 KNN 不需要事先训练，可以直接利用训练集对测试集的数据进行预测。如下表 5 所示，该表展示的是随着 K 值的变化，测试准确率（Accuracy）的变化情况。此处 KNN 的距离函数使用的是 Euclidean 距离。

表 5　KNN 模型的测试准确率

K 值	1	2	3	4	5	6	7	8	9
准确率	77.2%	61.8%	62.6%	58.3%	58.9%	56.2%	56.3%	54.8%	56.3%

11. 第二种算法是支持向量机，在实验中采用了高斯核，使用了软间隔和正则化，下表 6 展示的是训练集和测试集的准确率随正则项系数 C 改变的变化情况。

表 6　SVM 模型的训练集和测试集的准确率变化情况

C 值	1	5	10	20	80	100	200	400	500
训练集准确率(%)	96.66	98.34	99.00	99.47	99.91	99.93	99.97	99.97	99.99
测试集准确率(%)	65.0	72.1	73.6	75.1	76.8	76.5	76.9	77.2	77.2

三、分析和总结

12. KNN 模型不需要预训练，测试集数据可以直接进行预测。SVM 是对训练集进行训练，使用训练好的模型去进行预测。当训练集收集到更多数据的时候，KNN 模型随时可以进行预测，而 SVM 则需要重新训练。从另外一个角度来说，SVM 可以从数据集提取到一些信息，而 KNN 没有这方面的优势。
13. 从前边的分析结果可以看出来，在训练集中，两类数据在球心为坐标原点且半径为 10 到 12 之间的球面附近相互交错，两类数据在这一区域附近过渡。对于这一区域，建立有效的边界对两类数据进行划分是比较困难的。而测试集数据也主要分布在这一区域附近，所以不管是 KNN 还是 SVM，模型的测试数据集的准确率均只有 77.2%，若要继续提升测试集准确率则比较困难。
14. 由于两类数据交界处需要不规则的边界才可以比较好的对数据进行划分，所以可以通过训练深度神经网络模型来获得不规则的边界，而如何设置神经网络的层数以及每层的神经元个数则又是一个新的问题。