CENTRAL UNIVERSITY OF FINANCE AND ECONOMICS

中央财经大学

数学与统计建模案例

附加题

小组编号：29

指导老师：杨欣欣

统计与数学学院

2020 年 10 月 29 日

目录

1 题 **2**

1.1 随机梯度助推法 (GBM) 2

1.2 支持向量机 (SVM) 2

1.3 岭回归 (Ridge Regression) 3

2 题 **4**

参考文献 **4**

1 题

请列举一个数据科学中含有调整参数的模型或者数据处理方法。并说明调整参数在该模型中的作用。

1.1 随机梯度助推法 (GBM)

随机梯度助推法中我们需要设定最大树深度 (Max Tree Depth)

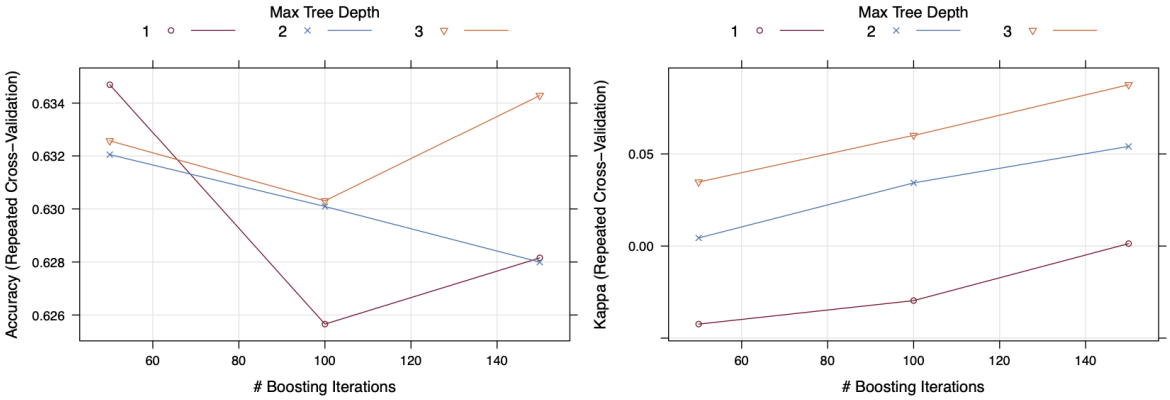

图 1: 调优参数不同取值下的准确率和 Kappa 指标变化（摘自小组之前的论文）

一般情况下，树的深度越深，叶节点个数越多，树的复杂度越高。当树的深度无穷时，理论上可以用大数定律证明训练误差与测试误差是收敛一致的，但树的深度过高，计算速度过慢，且此时会有过拟合现象。[1]

1.2 支持向量机 (SVM)

当使用 SVM 支持向量机时我们需要调节正则化参数 C，C 表示模型对误差的惩罚系数。

sigma	C	Accuracy	Kappa	AccuracySD	KappaSD
0.059	0.25	0.668	0.000	0.010	0.000
0.059	0.50	0.662	-0.010	0.021	0.038
0.059	1.00	0.644	0.025	0.053	0.123
0.059	2.00	0.644	0.087	0.069	0.142
0.059	4.00	0.640	0.100	0.071	0.162

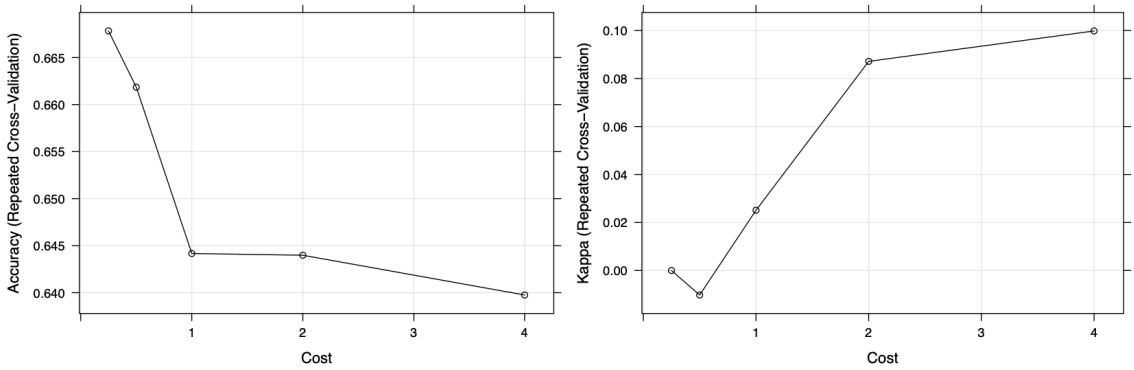

图 2: 调优参数不同取值下的准确率和 Kappa 指标变化（摘自小组之前的论文）

C 越大，模型越容易过拟合；C 越小，模型越容易欠拟合。

1.3 岭回归 (Ridge Regression)

岭回归中我们需要根据不同的 K 参数判断回归估计的优良性。

当设计矩阵存在多重共线性情况时，X'X 可能是奇异矩阵，此时求得的最小二乘回归系数不稳定；但如果将 X'X 加上正常数矩阵 KI，则 X'X + KI 的奇异性就会比 X'X 有所改善，此时求得的回归估计值比最小二乘估计稳定。

当 K=0 时，退化为普通最小二乘估计，当 K→∞ 时，回归系数趋于 0。由于 K 的选择是任意的，岭回归分析时一个重要的问题就是 K 取多少合适。

由于岭回归是有偏估计，K 值不宜太大；而且一般来说我们希望能尽量保留信息，即尽量能让 K 小些。因此可以观察在不同 K 的取值时方程的变动情况，然后取使得方程基本稳定的最小 K 值。

2 题

请说明使用交叉验证法的原因。

交叉验证是验证模型准确性的一种常见方法，将样本数据分为训练数据和测试数据，用训练数据来进行模型的训练，再用测试数据去测试模型。[2]

1. 交叉验证用于评估模型的预测性能，尤其是训练好的模型在新数据上的表现，可以在一定程度上减小过拟合现象的发生。
2. 使用交叉验证在训练集外进行预测并验证得到结果，可以使最后得到的结论有说服力。
3. 使用多折交叉验证可以从有限的数据中获取尽可能多的有效信息。[3]

参考文献

[1] 吴宇翀. 企业员工离职预测及模型比较 [EB/OL]. https://wuyuchong.com/projects/resign.

[2] 百度文库. 交叉验证法 [EB/OL]. https://wenku.baidu.com/view/e8f4bcd6f01dc281e53af0b0.html.

[3] CSDN. 为什么要用交叉验证 [EB/OL]. https://blog.csdn.net/aliceyangxi1987/article/details/73532651?utm_medium=distribute.pc_relevant.none-task-blog-title-1&spm=1001.2101.3001.4242.