# 机器学习大作业

21751191 林琦宏

## 实验目的

预测客户是否订阅定期存款

## 数据描述

采用bank-additional-full数据，有20个特征，41188条数据。

特征包括：

1. 年龄（数值型）

2. 工作（类别型）：12个类别，admin.，blue-collar， entrepreneur， housemaid， management， retired， self-employed， services， student， technician， unemployed， unknown

3. 婚姻（类别型）：4个类别，divorced， married， single， unknown

4. 学历（类别型）：8个类别， basic.4y， basic.6y， basic.9y， high.school， illiterate， professional.course， university.degree，unknown

5. 有无信用卡（类别型）：3个类别，no，yes，unknown

6. 有无房贷（类别型）：3个类别，no，yes，unknown

7. 个人贷款（类别型）：3个类别，no，yes，unknown

8. 通话方式（类别型）：2个类别，手机，电话

9. 最后一个通话的月份（类别型）：12个月，

10. 最后一个通话的星期（类别型）：4个类别，

11. 最后的通话持续时间（数值型）

12. 在此活动期间和这个客户端进行的接触次数（数值型）

13. 客户最后一次与之前的活动联系的天数（数值型）（数字：999意味着客户之前没有联系过）

14. 在此活动之前和该客户进行的接触次数（数值型）

15. 先前市场营销活动的结果（类别型）：3个类别失败，不存在，成功

16. 比率：就业变化率-季度指标（数值型）

17. 消费者价格指数-月度指标（数值型）

18. 消费者信心指数-月度指标（数值型）

19. 3个月利率-每日指标（数值型）

20. 就业人数：雇员人数-季度指标（数值型）

21.y列，客户是否订阅定期存款，yes or no。

## 预处理方法

* 将类别数据one-hot处理
* 将y列的yes/no 转化为1/0。

转化完后数据共有64列。

## 模型建立

采用K最近邻，逻辑回归，随机森林以及GBDT这四种算法。

逻辑回归采用L2正则化。

随机森林采用50棵树。

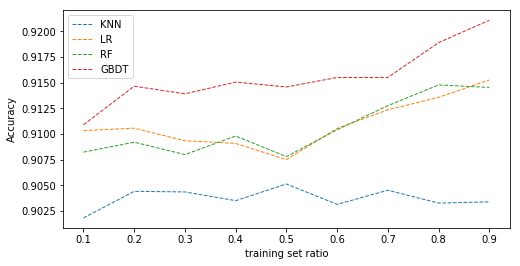
GBDT最大迭代次数100。

其他采用默认值。

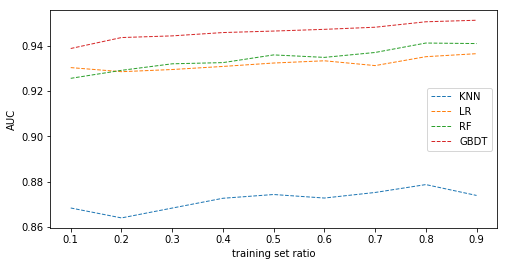
## 模型评估

按不同比例将data分成训练集与测试集，比例有0.1，,02, 0.3，...,0.9，模型评估采用 Accuracy与AUC值。

## 结果可视化



我们可以看到，GBDT准确率最高，LR,RF,GBDT准确率随训练集的扩大而上升，KNN一直保持平稳且准确率较低。



我们可以看到，GBDT AUC值最高，四个模型AUC值都随训练集的扩大而上升，KNN AUC值远远低于其他三者。