# 最佳参数实验报告

一 实验目的：（可附王老师给出的最初文档）

查找最优参数

1. Decision Tree：申小虎已经完成。我们会使用“both”参数。假定使用“both”方法找到的最优深度为“O”。
2. Gradient Boosting – 分类

**考虑的因素：**

**因素I：**深度（5个维度）“O-2”,“O-1”, “O”, “O+1”, “O+2” （O是申小虎用“both”方法找到的最优深度）

**因素II：**学习率（4个维度）0.2, 0.4, 0.6, and 0.8

**因素III：**弱学习器的数目（3个维度）100, 150, 200

**（注：对不同因素的组合使用不同的种子重复两轮）**

除了 “Depth”, “Learn Rate”和“N\_Estimator”，保持其他选项如下：

* + 1. Loss: “deviance”, 损失函数与Logistic Regression的损失函数一致，使用概率输出。
    2. Criterion: “friedman\_mse”, 采用的切分策略，对应的最小平方误差的近似，加入了Friedman的一些改进，默认是gini
    3. Min\_sample\_split: （节点最小分割的样本数，表示当前树节点还可以被进一步切割的含有的最少样本数）

1. 训练数据集样本数N大于1,000,000,000: min\_sample\_split = **400**
2. 训练数据集样本数N大于 1,000,000 且小于1,000,000,000: min\_sample\_split = **200**
3. 训练数据集样本数N小于 1,000,000且大于10,000: min\_sample\_split = **100**
4. 训练数据集样本数N小于10,000 : min\_sample\_split = **50**
   * 1. Min\_sample\_leaf:（叶节点所需的最小样本数）
5. 训练数据集样本数N大于 1,000,000,000: min\_sample\_leaf = **200**
6. 训练数据集样本数N大于1,000,000 且小于1,000,000,000: min\_sample\_leaf = **100**
7. 训练数据集样本数N小于1,000,000 且大于10,000: min\_sample\_split = **50**
8. 训练数据集样本数N小于10,000 : min\_sample\_split = **25**
   * 1. Min\_weight\_fraction\_leaf: 叶节点最小权重总值为0
     2. Subsample: 子采样率1.0
     3. Max \_features: 分裂时考虑的最大特征数None
     4. Max\_leaf\_nodes: 最大叶节点数 None
     5. **Min\_impurity\_decrease :** 用来进行早停止的参数，判断树是否进一步分支，原先是比较不纯度是否仍高于某一阈值，0.19后是判断不纯度的降低是否超过某一阈值，设置为**10-8**
     6. Init: None。计算初始基学习器的预测，需要具备fit和predict方法，若未设置则默认为loss.init\_estimator
     7. Verbose: 0 (Default)。建模完成后对输出的打印方式，默认0不输出任何结果
     8. Warm\_start: False (Default)。使用它我们就可以用一个建好的模型来训练额外的决策树，默认关闭
     9. **Random State:** **Use Different Seed for each Repetition。**作为每次产生随机数的随机种子，任一个随机样本都有可能导致过拟合，可以用不同的随机样本建模来减少过拟合的可能。每次重复性试验采用不同的随机种子。
     10. Presort: “auto”。 决定是否对数据进行预排序，可以使得树分裂地更快。

* Data Splitting: 70% of Training and 30% of Testing。数据分割：70%作训练集，30%作验证集。
* 分析测量（不需要使用所有的测量）：

1. C-statistics。ROC曲线下面积, AUC
2. Lift in top ten, twenty and thirty percent。前10%，20%，30%的提升度
3. Catch rate for the top ten, twenty, and thirty percent。前10%，20%，30%的捕获率
4. Sum of Squared of Errors。SSE，误差平方和
5. Gradient Boosting – 回归

**考虑的因素：**

**因素I：**深度（5个维度）“O-2”,“O-1”, “O”, “O+1”, “O+2” （O是申小虎用“both”方法找到的最优深度）

**因素II：**学习率（4个维度）0.2, 0.4, 0.6, and 0.8

**因素III：**弱学习器的数目（3个维度）100, 150, 200

除了 “Depth”, “Learn Rate”和“N\_Estimator”，保持其他选项如下：

1. Loss: “ls”, 损失函数采用最小平方
2. Criterion: “friedman\_mse”。 采用的切分策略，对应的最小平方误差的近似，加入了Friedman的一些改进
3. Min\_sample\_split: （节点最小分割的样本数，表示当前树节点还可以被进一步切割的含有的最少样本数）
   1. 训练数据集样本数N大于1,000,000,000: min\_sample\_split = **400**
   2. 训练数据集样本数N大于 1,000,000 且小于1,000,000,000: min\_sample\_split = **200**
   3. 训练数据集样本数N小于 1,000,000且大于10,000: min\_sample\_split = **100**
   4. 训练数据集样本数N小于10,000 : min\_sample\_split = **50**
4. Min\_sample\_leaf: （叶节点所需的最小样本数）
   1. 训练数据集样本数N大于 1,000,000,000: min\_sample\_leaf = **200**
   2. 训练数据集样本数N大于1,000,000 且小于1,000,000,000: min\_sample\_leaf = **100**
   3. 训练数据集样本数N小于1,000,000 且大于10,000: min\_sample\_split = **50**
   4. 训练数据集样本数N小于10,000 : min\_sample\_split = **25**
5. Min\_weight\_fraction\_leaf: 叶节点最小权重总值为0
6. Subsample: 子采样率1
7. Max \_features: 分裂时考虑的最大特征数None
8. Max\_leaf\_nodes: 最大叶节点数 None
9. **Min\_impurity\_Decrease:** 用来进行早停止的参数，判断树是否进一步分支，原先是比较不纯度是否仍高于某一阈值，0.19后是判断不纯度的降低是否超过某一阈值，设置为**10-8**
10. alpha: Not used since it is not available in “ls” 。不使用这个参数，因为在“ls”里它不可用。
11. Init: No, 计算初始基学习器的预测，设置为不需要预先估计。
12. Verbose: 0 (Default), 建模完成后对输出的打印方式，默认0不输出任何结果
13. Warm\_start: False (Default), 使用它我们就可以用一个建好的模型来训练额外的决策树，默认关闭，每步拟合一棵新树
14. Random\_state: Since “init” option is not used, this option is not available.因为没有使用“init”，此参数不可用。
15. Presort: “auto”, the default。决定是否对数据进行预排序，可以使得树分裂地更快。

Data Splitting: 70% of Training and 30% of Testing。数据分割：70%作训练集，30%作验证集。

分析测量：

1. Mse，均方误差
2. Ase，绝对误差
3. AIC，赤池信息准则
4. BIC，贝叶斯信息准则
5. Random Forest – 分类

**考虑的因素：**

**因素I：**深度（5个维度）“O-2”,“O-1”, “O”, “O+1”, “O+2” （O是申小虎用“both”方法找到的最优深度）

**因素II：**子树的数目（3个维度）100, 150, 200

**（注：对不同因素的组合使用不同的种子重复两轮）**

* 除了 “MAX\_Depth”和“N\_Estimator”，保持其他选项如下：
  + 1. Criterion: Gini (Default)。默认为gini，与树相关
    2. Max \_features: None, 随机森林允许单个决策树使用特征的最大数量。设置为考虑所有的特征。
    3. Min\_Sample\_Split:（分裂所需的最小样本数）

1. 训练数据集样本数N大于1,000,000,000: min\_sample\_split = 400
2. 训练数据集样本数N大于 1,000,000 且小于1,000,000,000: min\_sample\_split = 200
3. 训练数据集样本数N小于 1,000,000且大于10,000: min\_sample\_split = 100
4. 训练数据集样本数N小于10,000 : min\_sample\_split = 50
   * 1. Min\_sample\_leaf: （叶节点最小样本数）
5. 训练数据集样本数N大于 1,000,000,000: min\_sample\_leaf = **200**
6. 训练数据集样本数N大于1,000,000 且小于1,000,000,000: min\_sample\_leaf = **100**
7. 训练数据集样本数N小于1,000,000 且大于10,000: min\_sample\_split = **50**
8. 训练数据集样本数N小于10,000 : min\_sample\_split = **25**
   * 1. Min\_weight\_fraction\_leaf: 0 , 叶节点最小样本权重总值，设置为0没有权重
     2. Max\_leaf\_nodes: None, i.e., considering all leaf nodes for continuing split，最大叶节点数设置为None，考虑到所有的叶节点连续分裂
     3. Min\_impurity\_Decrease: use default , i.e., 10-7  ，用来进行早停止的参数，判断树是否进一步分支，原先是比较不纯度是否仍高于某一阈值，0.19后是判断不纯度的降低是否超过某一阈值，设置为默认值10-7
     4. Bootstrap: True (Default), i.e., use bootstrapping sample。是否bootstrap对样本抽样，默认为是，子模型间不要强相关
     5. oob\_score: False (Default), i.e., not use out-of-bag-sample to estimate the error。是否计算[袋外得分](http://scikit-learn.org/stable/auto_examples/ensemble/plot_ensemble_oob.html)，默认为否，即不用袋外样本估计误差
     6. N\_jobs: -1 (Default), i.e., use all available cores。并行化参数，-1表示使用所有的CPU核心
     7. Random\_state: None, (Default), i.e., random number generator will be used for each new tree. 作为每次产生随机数的随机种子，任一个随机样本都有可能导致过拟合，可以用不同的随机样本建模来减少过拟合的可能。默认为None，随机数生成器用来生成每棵新树。
     8. Verbose: 0 (Default), i.e., not used, 建模完成后对输出的打印方式，默认0不输出任何结果
     9. Warm\_start: False (Default), i.e., fit a whole new tree at each step, 使用它我们就可以用一个建好的模型来训练额外的决策树，默认关闭，每步拟合一棵新树
     10. Class\_weight: None (Default), i.e., do not use any weightsMin\_sample\_split: 类别的权值，不使用任何的权值

* Data Splitting: 70% of Training and 30% of Testing。数据分割：70%作训练集，30%作验证集。
* 分析测量（不需要使用所有的测量）:

1. C-statistics。ROC曲线下面积, AUC
2. Lift in top ten, twenty and thirty percent。前10%，20%，30%的提升度
3. Catch rate for the top ten, twenty, and thirty percent。前10%，20%，30%的捕获率
4. Sum of Squared of Errors。SSE，误差平方和
5. Random Forest – 回归

**考虑的因素：**

**因素I：**深度（5个维度）“O-2”,“O-1”, “O”, “O+1”, “O+2” （O是申小虎用“both”方法找到的最优深度）

**因素II：**子树的数目（3个维度）100, 150, 200

**（注：对不同因素的组合使用不同的种子重复两轮）**

* 除了 “MAX\_Depth”和“N\_Estimator”，保持其他选项如下：
  + 1. Criterion: “mse”, i.e., use the criterion to minimize the L2 norm。判断节点是否继续分裂采用的计算方法，采用均方误差，也就是使用此准则最小化L2正则化范数
    2. Max\_features: None, , 随机森林允许单个决策树使用特征的最大数量。设置为考虑所有的特征。
    3. Min\_Sample\_Split:（分裂所需的最小样本数）

1. 训练数据集样本数N大于1,000,000,000: min\_sample\_split = 400
2. 训练数据集样本数N大于 1,000,000 且小于1,000,000,000: min\_sample\_split = 200
3. 训练数据集样本数N小于 1,000,000且大于10,000: min\_sample\_split = 100
4. 训练数据集样本数N小于10,000 : min\_sample\_split = 50
   * 1. Min\_sample\_leaf: （叶节点最小样本数）
5. 训练数据集样本数N大于 1,000,000,000: min\_sample\_leaf = **200**
6. 训练数据集样本数N大于1,000,000 且小于1,000,000,000: min\_sample\_leaf = **100**
7. 训练数据集样本数N小于1,000,000 且大于10,000: min\_sample\_split = **50**
8. 训练数据集样本数N小于10,000 : min\_sample\_split = **25**
   * 1. Max\_leaf\_nodes: None, i.e., considering all leaf nodes for continuing splits。最大叶节点数设置为None，考虑到所有的叶节点连续分裂
     2. Min\_impurity\_Decreased: 用来进行早停止的参数，判断树是否进一步分支，原先是比较不纯度是否仍高于某一阈值，0.19后是判断不纯度的降低是否超过某一阈值，设置为**10-8**
     3. Boostrap: True, i.e., to use bootstrap sample。是否bootstrap对样本抽样，默认为是，子模型间不要强相关
     4. Oob\_score: False, i.e., not to use out-of-bag data to estimate the error。是否计算[袋外得分](http://scikit-learn.org/stable/auto_examples/ensemble/plot_ensemble_oob.html)，默认为否，即不用袋外样本估计误差
     5. N\_jobs: -1, i.e., to use all available cores。并行化参数，-1表示使用所有的CPU核心
     6. Random\_state: None, (Default), i.e., random number generator will be used for each new tree.
     7. Verbose: 0 (Default), i.e., not used。作为每次产生随机数的随机种子，任一个随机样本都有可能导致过拟合，可以用不同的随机样本建模来减少过拟合的可能。默认为None，随机数生成器用来生成每棵新树。
     8. Warm\_start: False (Default), i.e., fit a whole new tree at each step, 使用它我们就可以用一个建好的模型来训练额外的决策树，默认关闭，每步拟合一棵新树

* Data Splitting: 70% of Training and 30% of Testing。数据分割：70%作训练集，30%作验证集。
* 分析测量（不需要使用所有的测量）:
  1. Mse，均方误差
  2. Ase，绝对误差
  3. AIC，赤池信息准则
  4. BIC，贝叶斯信息准则

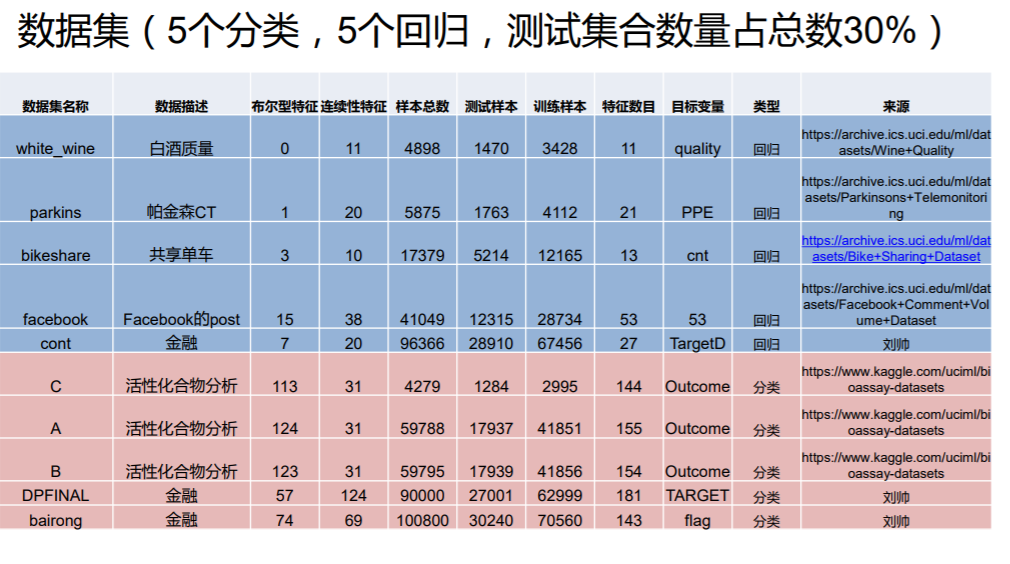
二 实验内容：

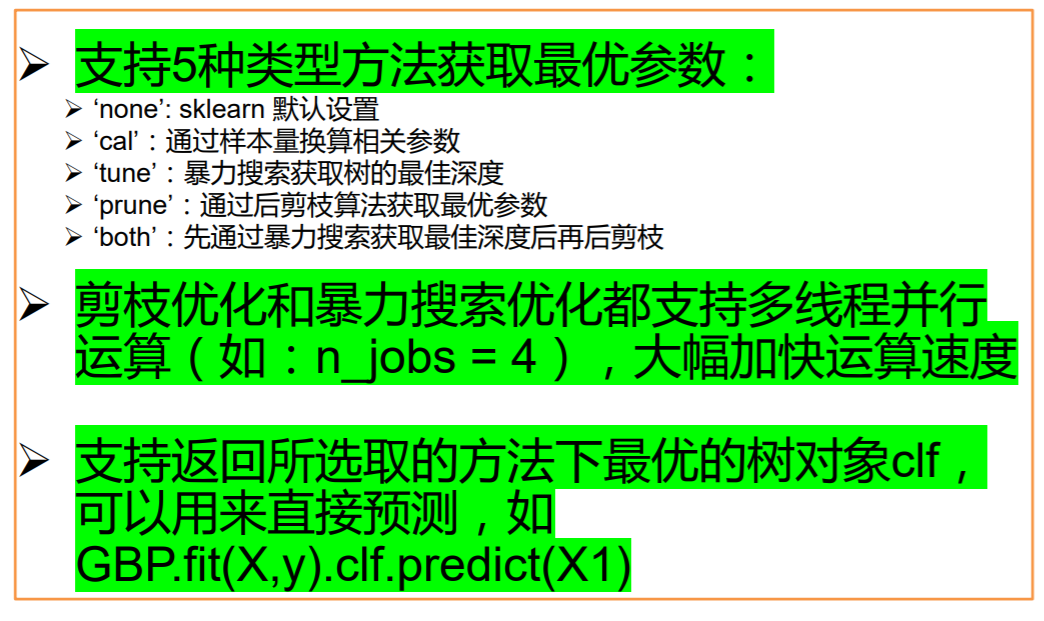
建最优树，剪枝等等

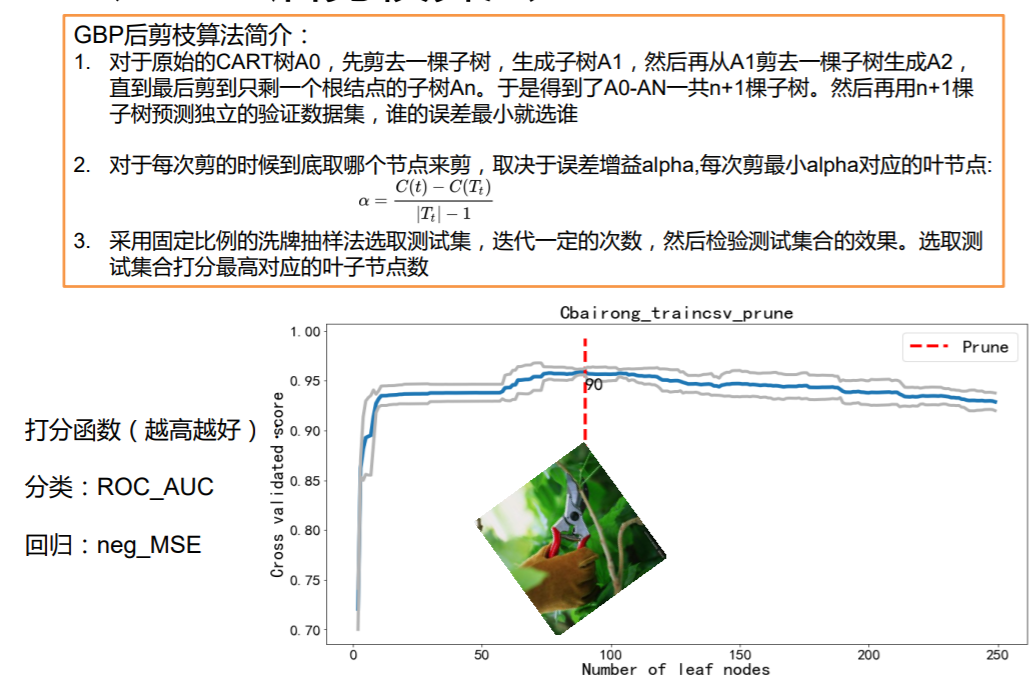


三 实验步骤：

1读数据







四 实验结果：



