**LAB01-Regression 补充报告**

# 1 实验回顾及存在的问题

在LAB01的小组实验中，我们进行了如下工作：

* 在数据提供的8个特征的基础上，查阅相关资料和公式，构造影响混凝土强度的组合特征2个；
* 对共计10个特征，使用主成分分析PCA，获得前6个主成分，利用这6个主成分展开后续实验；
* 绘制了各个主成分之间的分布图；
* 使用三个回归模型进行回归，分别是线性回归、支持向量机回归和高斯过程回归；
* 进行5折交叉验证，使用RMSE和作为评价指标，得到实验结果；
* 对实验结果进行假设检验分析；
* 对结果最好的模型（高斯过程回归模型）绘制了预测结果和实际结果之间的分布图以及残差的分布图。

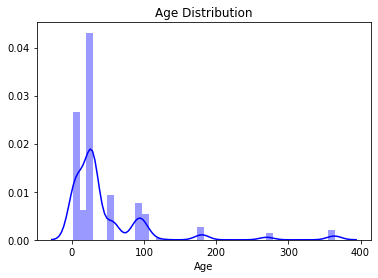
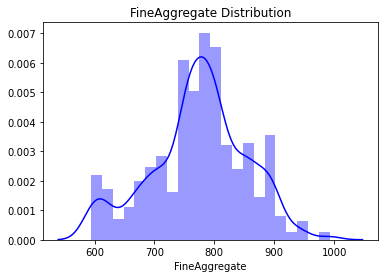
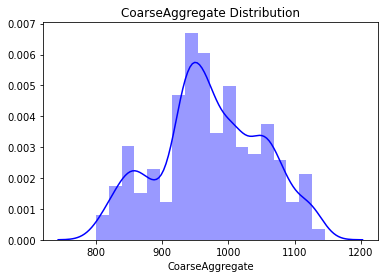
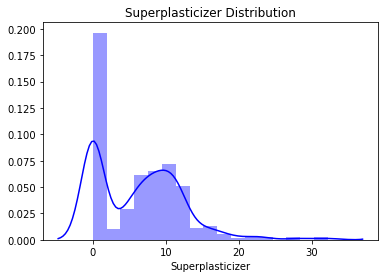
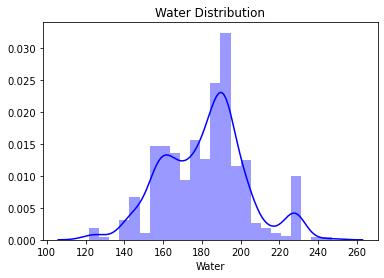
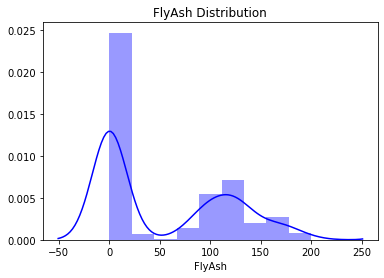
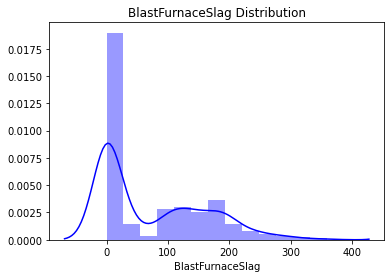
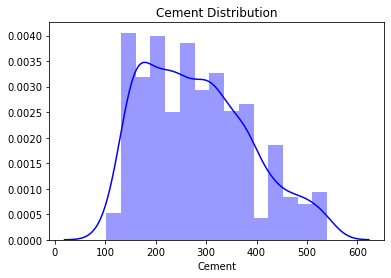
实验的完成度一般，主要存在以下几个问题：

* 缺少原始数据分布图，没有分析原始数据之间的相关度；
* 使用模型较少，仅仅考虑了三个回归的模型，且原实验该部分使用 matlab 编程，没有使用计算机系更常使用的 python 来编程；
* 在答辩时，被询问到有关KKT条件的内容，对此缺少了解。

针对以上问题，我做了本次的补充报告。

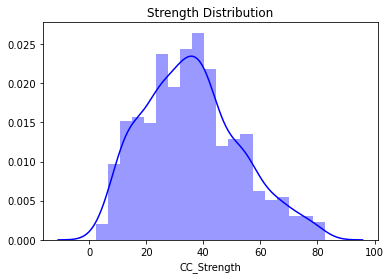
# 2 原始数据分布及数据相关度

**（1）原始的8个特征数据分布图如下：**



可以看到Cement, Water, CoarseAggregate, FineAggregate这4个特征的分布近似服从于正态分布；而其他剩余四个特征的分布也具有相似性，它们都有两个峰，并且左侧的峰高于右侧的峰。

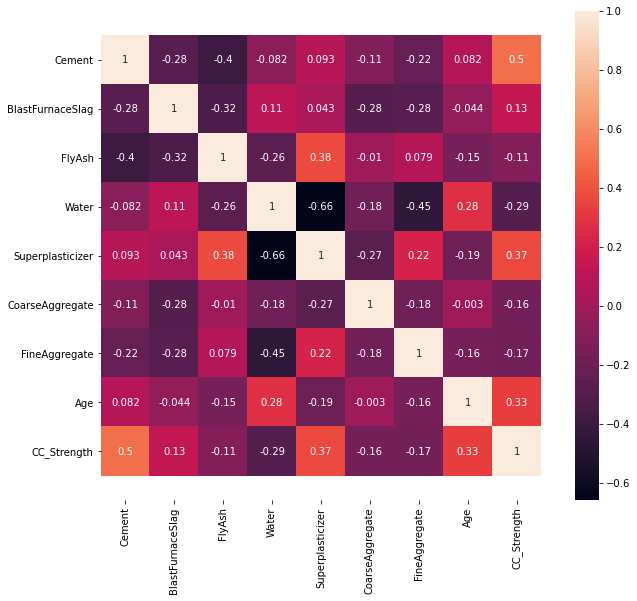
**（2）原始的水泥强度数据分布入下：**



水泥强度的数据分布也近似服从正态分布。从分布样式来说，同样服从近似正态分布的Cement, Water, CoarseAggregate, FineAggregate四个特征似乎会对水泥强度起到更加决定性的影响。

**（3）数据相关性**

数据相关性通过计算各个特征之间的相关系数得到，如下：



从其中可以看出，Cement、Superplasticizer 和 Age 三个特征对水泥强度的影响最大。

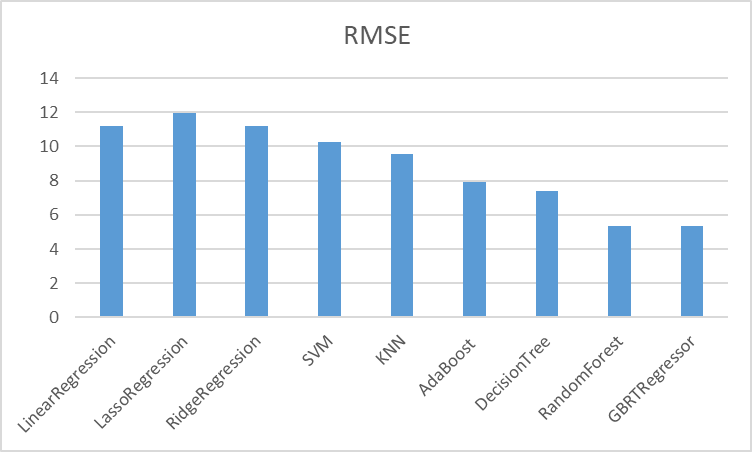
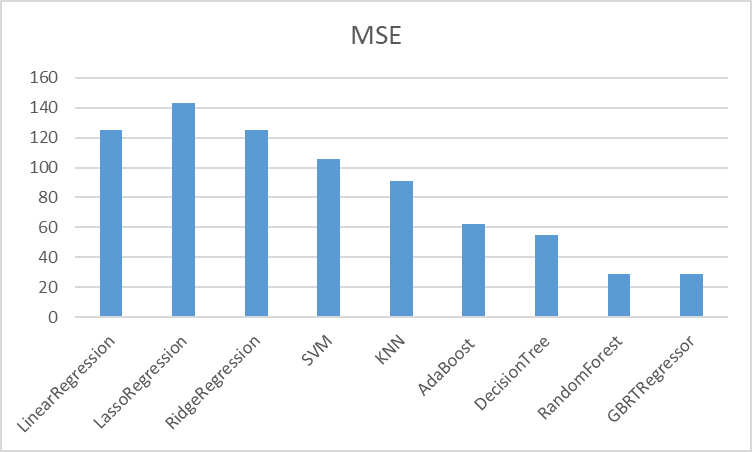
# 3 增加使用的回归模型

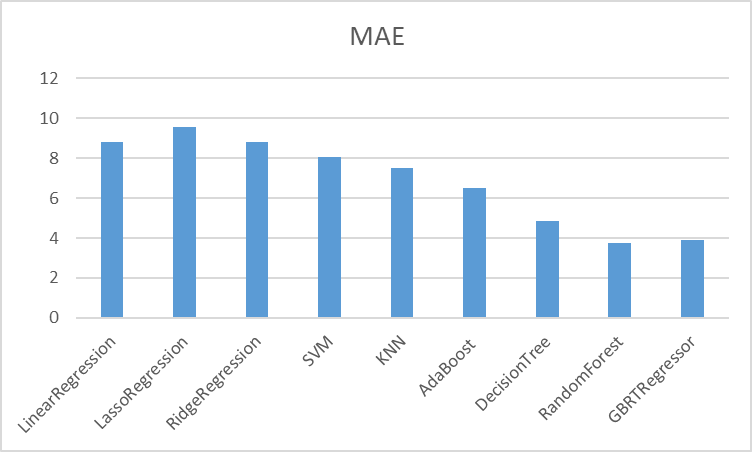
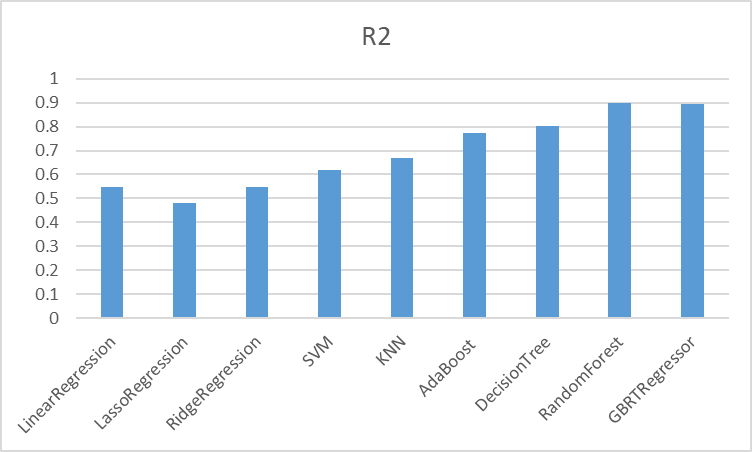
补充回归模型：Lasso回归，Ridge回归，KNN回归，决策树回归，随机森林回归，Adaboost回归，梯度提升树回归。

**（1）回归结果：**



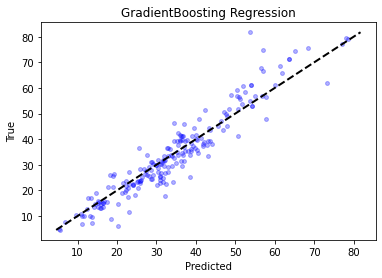
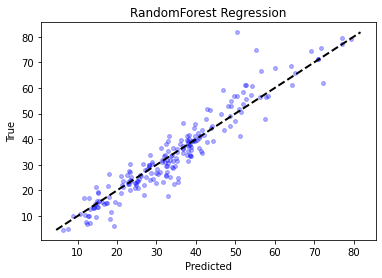
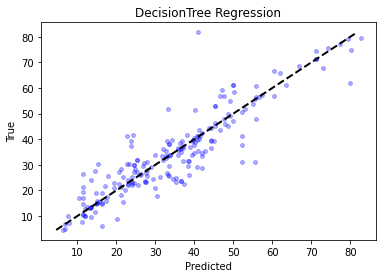
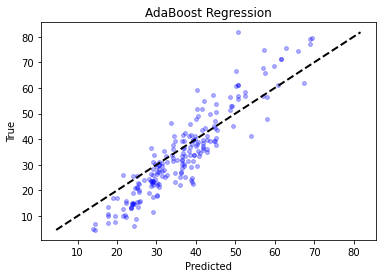
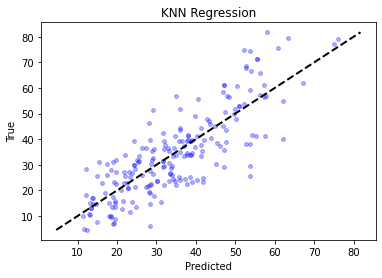
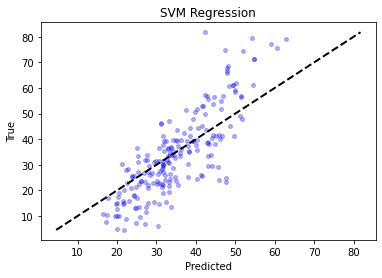
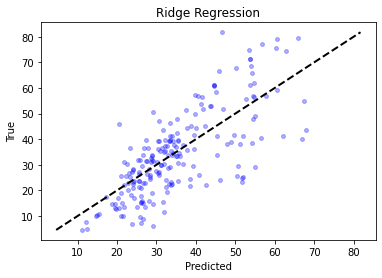
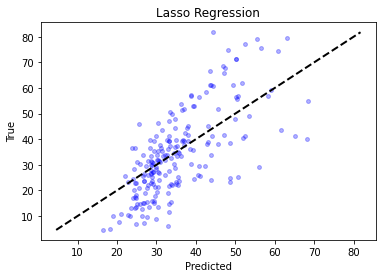
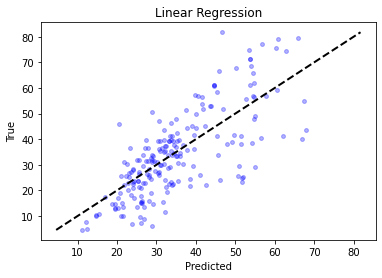
**（2）各评级指标的条形图：**

四个评价指标都指出，随机森林回归和梯度提升树回归的回归效果是最好的，比之前使用的一般线性回归和支持向量机回归要好很多。

**（3）预测值/真实值散点图：**



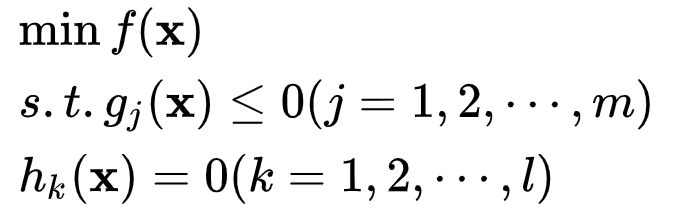
这个散点图的意义是：距离中线越近的点表明模型对这个数据的预测是越准确的。从散点图可以看出，随机森林回归和梯度提升树回归的散点更加贴近中线，在中线两侧的方差最小，回归效果最好；而其他模型相比之下散点就很分散。

# 4 关于KKT条件

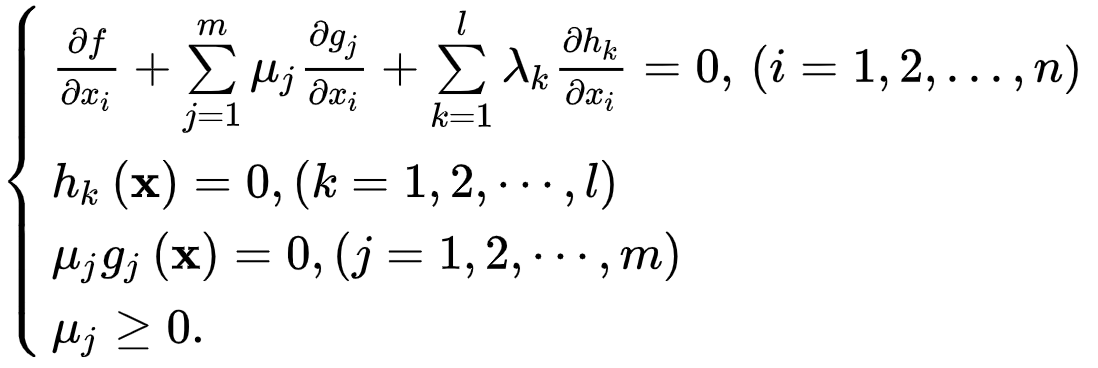
**我理解的KKT条件是这样的：**

**在不等式约束的优化问题中，变量满足KKT条件是变量为该优化问题最优解的必要条件。**详细说是：

对于具有等式和不等式约束的一般优化问题：



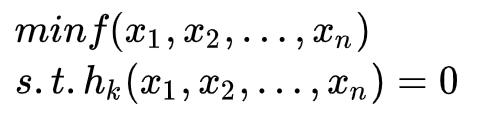
KKT条件给出了判断是否为最优解的必要条件，即：



**关于KKT条件的简单推导如下：**

**（1）等式约束的拉格朗日乘数法：**

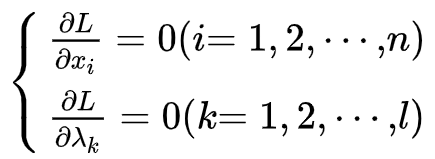
对于等式约束问题：



它要求极小化目标函数f，在等式约束的条件下，这时，我们使用高数中学习到的拉格朗日乘数法。

令

分别对变量和拉格朗日乘子求偏导，令偏导为0并联立，即：



该方程得到的解就是可能的极值点。

**（2）对不等式约束的处理**

在等式约束中加上形如下列的约束

等式约束问题就成为了不等式约束问题。

对于不等式约束的处理，思想是引入松弛因子从而转化为等价的等式约束问题，具体做法是：对于约束，我们分别引入松弛变量，得到.注意，这里直接加上平方项而非，是因为这个不等式的左边必须加上一个正数才能使不等式变为等式.若只加上，又会引入新的约束，这不符合我们的意愿。

由此可以得到拉格朗日函数

分别对求偏导，令其等于0并联立就得到KKT条件：

其中，不等式约束前的乘子是由几何性质导致的。

进一步处理，对于：若，，则有；若，，且.

合并两种情况，就有,所以KKT条件是：

