超级码力数据挖掘工程师笔试题

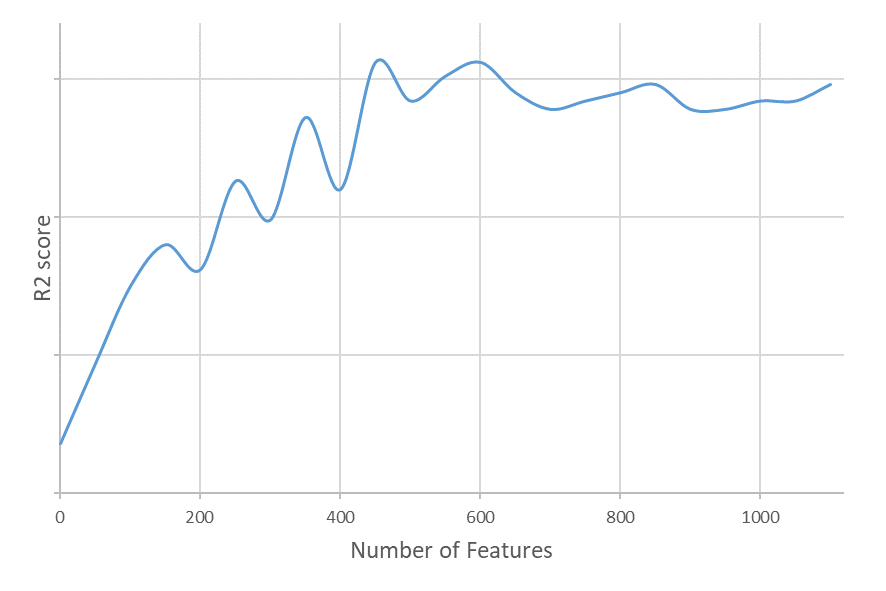
**题目：**

使用提供的数据集建立Regression模型，数据含有1257条样本，300个特征，1个目标变量‘Label’。因为特征数量过多，需要你使用RFE**特征选择**的方法对数据集进行处理，同时选用你认为合适的Regression算法训练，并基于三折交叉验证画出在特征选择过程中R2 score变化过程，以及regression算法的三维调参图（选用两个你认为最重要的参数变量来调）。

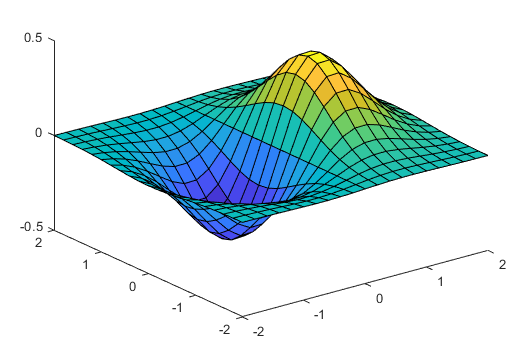
**要求：**

1. 简述解题流程并提交代码（最终R2 score越高越好）；
2. 画出特征选择过程的R2 score变化图；
3. 画出所选算法在最终特征集上的调参图（如计算较慢，可通过调整步长及网格尺寸减少计算量）；

示例图：R2 score随着特征选择过程中特征数量的变化而变化



示例图：R2 score随着算法参数变化的调参图



1.

代码见code.py

第一步，数据读取、数据清洗：

清除特殊字符‘-’、‘+’等；

删除缺失值较多的行(thres=285)和列(thres=1200）；

以列均值填充剩余缺失值。

第二步，特征值提取：

设定变量集：解释变量集feature和被解释变量集label；

回归方式：尝试sklearn.linear\_model的Ridge,Lasso,ElasticNet（最终选择岭回归 Ridge）；

递归特征消除：借助sklearn.feature\_selection下的REFCV进行特征选择，并选用三折交叉检验和R2 score；绘制特征选择过程的R2 score变化图。

提取特征值，并删选出新的feature集。

第三步，调参：

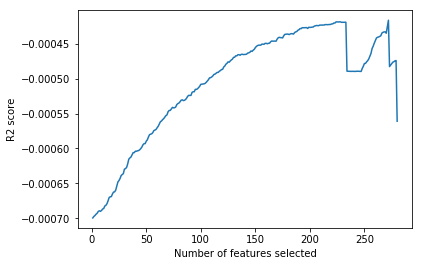
回归方式：岭回归Ridge；

调参方式：借助sklearn.model\_selection下的GridSearchCV进行，并选用三折交叉 检验和R2 score。

调参图：提取参数集和R2 score集，绘制调参过程的R2 score变化图。

图画的有问题，没找到问题所在

1. 特征选择过程的R2 score变化图



1. 调参过程的R2 score变化图

