实验概述

本次实验实现(或调用)以下方法，并探究不同参数对实验结果的影响。

集成学习方法：Bagging，AdaboostM1

分类法：SVM，DTree，KNN(K-NN)，NB(Naive Bayes)

归一化方法：None(不使用)，l1-Norm，Standardizing

迭代次数

实验结果

一、探究Ensemble对分类结果的影响：

Ensemble	Classifier	Normalize	Iteration	RMSE
Bagging	DTree	None	25	**0.76445**
AdaBoostM1	DTree	None	25	**0.78769**
Bagging	SVM	None	25	**0.78492**
AdaBoostM1	SVM	None	25	**0.72778**

由上述结果可知，AdaBoostM1对SVM的修正效果较DTree好。

二、探究Classifier对分类结果的影响：

Ensemble	Classifier	Normalize	Iteration	RMSE
None	**SVM**	None	0	**0.80404**
None	**DTree**	None	0	**0.81896**
None	**KNN**	None	0	**0.87713**
None	**NB**	None	0	**1.08812**

从上述结果可知，以上分类算法在未经集成学习的情况下的优劣程度大致为：SVM > DTree > KNN > NB

三、探究Normalize(归一化方法)对分类结果的影响：

Ensemble	Classifier	Normalize	Iteration	RMSE
Bagging	SVM	**None**	25	**0.78492**
Bagging	SVM	**l1-norm**	25	**0.79585**
Bagging	SVM	**Standardizing**	25	**0.78419**

由上述结果可知，对SVM而言，归一化方法的优劣程度大致为：Standardizing >= None > l1-norm

四、探究Iteration(迭代次数)对分类结果的影响：

Ensemble	Classifier	Normalize	Iteration	RMSE
Bagging	DTree	None	**25**	**0.76445**
Bagging	DTree	None	**100**	**0.76130**
Bagging	DTree	None	**400**	**0.76225**

由上述结果可知，迭代次数并不是越多越好，过多的迭代可能导致"过度学习"从而降低准确率。

五、最终结果：

以上探究中RMSE相对较小的组合有：

Ensemble	Classifier	Normalize	Iteration	RMSE
Bagging	DTree	None	100	**0.76130**
Bagging	KNN	None	25	**0.65549**
Bagging	SVM	None	25	**0.78492**

我决定用以下组合进行深入测试，探索最佳实验效果，结果如下：

#	Ensemble	Classifier	Normalize	Iteration	RMSE
0	Bagging	**DTree**	None	**400**	**0.76225**
1	Bagging	**KNN**	None	100	**0.60519**
2	Bagging	**SVM**	None	100	**0.68449**

由上述结果可知，增加迭代次数无助于DTree精度的提高，而KNN与SVM相比之下，KNN精度较高，故选取第一组作为最终的实验结果。

实验总结

本次实验的因变量是RMSE，而自变量相当多，导致在探索单一变量的影响时比较麻烦，另外程序运行也花费不少时间。下面针对不同的自变量进行分析：

1. 特征词提取：人工选取比较有代表性的词，可能不够全面，导致实验精度不高；
2. 训练集大小：选择合适大小的训练集比较困难，训练集过小则精度不高，过大则增加程序运行时间；
3. 集成学习方法+分类法+迭代次数+归一化方法：个人认为，其实在本次实验中用控制变量法并不能准确说明孰优孰劣，因为这四个变量并非独立，但是若考虑各种组合情况工作量又过于庞大。

总之，由于没有明确的思路以及对原理理解的不到位，导致对本次实验的探究工作做得很不理想，不够深入。在今后的实验中，应该基于对原理的熟练把握之上先做好明确的规划，不能想当然地盲目进行实验！

另外，在实验中发现一个问题：本地测试的结果优于线上测试，二者基本成平方关系。猜想可能与Kaggle的RMSE计算方法有些出入。