知识分析与处理——主动学习

实验数据

- 本次使用的数据集为先行回归数据集
- 直接使用之前处理好的词向量数据
- 采用逻辑回归为基准
- 输出结果在每个query对应的run.txt和runt.txt中，第一行为使用数据，之后为分类结果和准确率

实现功能

- 读取数据,使用基准模型得出分类结果和准确度
- 采用不同的主动学习策略，自动的选取需要加入训练集的数据

算法分析

首先读入词向量数据,划分出初始的训练集和测试集合
对于第一种主动学习方法：
    每一轮首先计算出测试数据的分类情况
    选取其中预测值最接近0.5的5个点，加入训练集
    重复一轮算法直至达到目标准确率
对于投票方法：
    首先训练三个不同的分类器
    每一轮首先计算出测试数据的分类情况
    选取对于三个分类器分歧最大的数据，加入训练集
    重复一轮算法直至达到目标准确率

实验结果

- 将测试结果与标注的结果做对比: 准确率 = 正确分类数/测试数据总数
- 实验结果：

方法一

query	1	2	3	4	5	6	7	8	9	10	均值
数据量	180	270	195	140	290	260	110	165	90	190	189
准确率	0.755	0.8016	0.691	0.771	0.6783	0.725	0.8266	0.81	0.766	0.80	0.763

方法二

query	1	2	3	4	5	6	7	8	9	10	均值
数据量	175	250	360	155	220	220	140	165	110	220	181
准确率	0.733	0.816	0.695	0.793	0.6866	0.6799	0.823	0.81	0.733	0.815	761