中山大学数据科学与计算机学院

移动信息工程专业-人工智能

本科生项目报告

(2017-2018学年秋季学期)

课程名称：**Artificial Intelligence**

教学班级	专业方向	组号	组名	学号	姓名
1501	移动（互联网）	04	神经源	15352005	蔡景韬

一、 Project最终结果展示

1. 最终结果

类型	二元	多元	回归
评判标准值	0.905484	0.628413	58.77263
排名	34	16	23

2. 组内分工

二元	多元	回归
曾华浪	蔡景韬	蔡泽杰

3. 个人工作

12.21--12.24	12.25--12.26	12.27--12.31	1.1--1.4
数据处理——Matlab调用函数分析（放弃） 数据处理——C++库挂载计算词向量（放弃） 数据处理—使用已有的词向量	基于词向量的KNN算法 基于词向量的LR算法	数据处理——卡方检测筛词 基于卡方检测的KNN算法 基于卡方检验的LR算法	归纳所做工作，准备Pre汇报结果 收尾工作

二、 工作流程

1. 算法简介

1. **KNN**
   - 在k-NN分类中，一个对象的分类是由其邻居的“多数表决”确定的，k个最近邻居（k为正整数，通常较小）中最常见的分类决定了赋予该对象的类别
   - 影响KNN分类准确性的要素主要有以下几个
     1. K值的选择
     2. 分类决策规则：多数表决
     3. 距离度量
     4. 文本分析的数据挖掘方式：OneHot、TF-IDF
2. **LR**
   - 逻辑回归算法是单层神经网络的一种，也属于线性分类器的一种，它是通过梯度下降对权重矩阵进行训练，用以表示各列属性对结果的贡献值，最终使用权重矩阵对输入数据矩阵进行加权求和，通过激活函数即可得到预测结果
   - 逻辑回归擅长处理两元分类，得出两元的概率。
   - 在本次多元分类任务中，使用三个逻辑回归分类器进行分类：是否为LOW，是否为MID，是否为HIG，最终取最大概率的标签作为预测结果
3. **词向量**
   - 将word映射到一个新的空间中，并以多维的连续实数向量进行表示叫做词向量
   - 自从21世纪以来，人们逐渐从原始的词向量稀疏表示法过渡到现在的低维空间中的密集（词向量）表示。
     - 用稀疏表示法在解决实际问题时经常会遇到维数灾难，并且语义信息无法表示，无法揭示word之间的潜在联系。
     - 而采用词向量表示法，不但解决了维数灾难问题，并且挖掘了word之间的关联属性，从而提高了向量语义上的准确度。
   - 词向量在多元的使用目的，主要是用于降维
4. **卡方检验**
   - 卡方检验是用途非常广的一种假设检验方法，通过计算卡方值，得到统计样本的实际观测值与理论推断值之间的偏离程度，从而判断假设是否正确[1]
   - 在本次多元分类任务中[2]

1. 我们先假设，词“no”与标签“HIG”不相关。
2. 简单统计四格表

特征选择	属于HIG	不属于HIG	合计
含“no”的文本	A	B	A+B
不含“no”的文本	C	D	C+D
合计	A+C	B+D	A+B+C+D=N

3. 根据原假设（不相关），“HIG”类别中包含“no”的文档比例应与所有文档中包含“no”的文档比例相同，则四格表中A的理论值应为$E_{11} = (A + C) * \frac{A+B}{N}$，根据卡方值计算公式，差值为$D_{11} = \frac{(A-E_{11})^2}{E_{11}}$
4. 同理，可以分别求出B、C、D的差值D_{12}、D_{21}、D_{22}，最终计算化简得出，词“no”与标签“HIG”的卡方值公式为

$$\begin{aligned}\chi^2(no, HIG) &= D_{11} + D_{12} + D_{21} + D_{22} \\ &= \frac{N(AD - BC)^2}{(A + C)(A + B)(B + D)(C + D)}\end{aligned}$$

5. 查χ^2 -- P 表格[3]

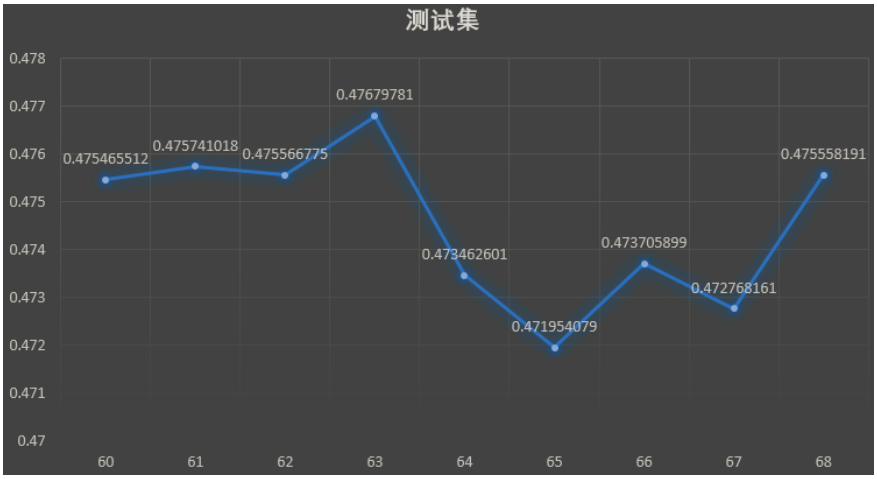
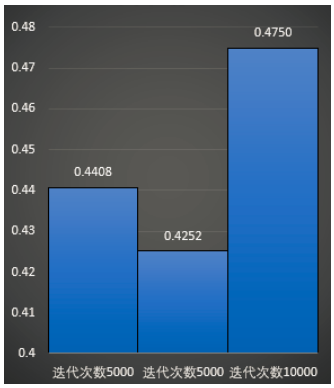
   在查表之前应知自由度：按χ^2检验的自由度v=(行数-1)(列数-1)，则四格表的自由度v=1，查x表可得

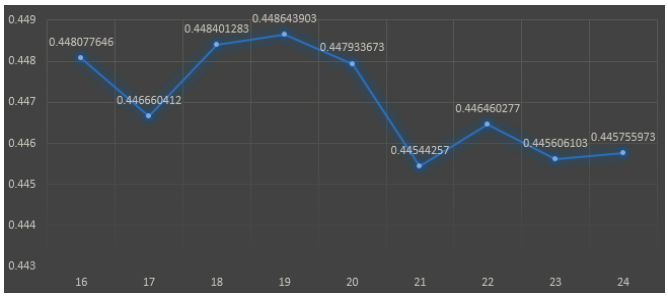
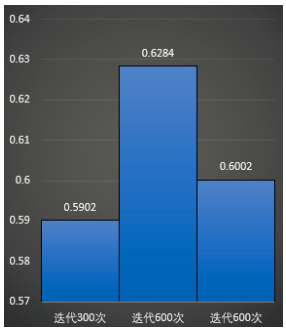
v	P			V	P		
	0.05	0.01	0.001		0.05	0.01	0.001
1	3.84	6.63	10.83	16	26.30	32.00	39.25
2	5.99	9.21	13.81	17	27.59	33.14	40.79
3	7.81	11.34	16.27	18	28.87	34.18	42.31
4	9.49	13.28	18.47	19	30.14	36.19	43.82

   当卡方值大于10.83，则假设成立的概率仅为0.001=0.1%
6. 最终我们分别对所有词做三次（三种标签）卡方值计算，取卡方值前K个词即可。（卡方值越大，说明假设不成立概率越大，即词与标签的相关性越大）

数据处理——词向量	**数据处理——卡方检验**
词向量能很好的把握词的相关性、词与词之间的距离，所以当使用距离指向性的算法**如KNN时，往往能表现出良好的效果**，但是在使用像LR这种神经网络类型的算法时，**会由于文本词向量相加求平均而损失较多信息从而失真训练**，最终的结果不尽如人意	卡方检验是一种简单的数据统计，关系挖掘的方法，它能较为准确的得出假设的成立概率，从而筛选出相关性大的词。从多元分类任务的应用上来看，它表现良好，但也**说明在90000+维的原数据集稀疏One-Hot表示矩阵中有96%的数据属于冗余数据**

2. 调参过程

词向量+KNN	词向量+LR
	 第一个为固定学习率 第二三个都为动态学习率

卡方检验（选取卡方值大于100的词） +LR迭代4400次	卡方检验（三个标签各取前300个词） +KNN	卡方检验（三个标签各取前300个词） +LR
准确率： 0.593071359676		 第一二个为固定学习率 第三个为动态学习率

3. 数据集分析

- 二元数据集

  通过信息增益分析各列属性值

属性列	A	B	C	D	E	F
连续/离散	连续	离散	连续	连续	连续	连续
取值范围	1880~3860	1/2/3/4/5	-2~5455	-67~6970	-165~7000	29~7111
ID3	0.232261	0.150562	0.008189	0.0369893	0.0364817	0.0958766

属性列	G	H	I	J	K	L	M
连续/离散	连续	连续	连续	连续	离散	离散	连续
取值范围	3~6225	2~7082	0~6888	0~7139	0/1	0/3	2.000005~2.999996
ID3	0.010859	0.0391927	0.0342441	0.103483	0.00174809	0.00188828	0.99956

从信息增益可以看到，A、B、J列属性是比较重要的，所以在尝试对属性加权的时候，可以加大这三项的权重，而M属性取值过多，导致计算ID3时趋近于1，如果计算C4.5可以明显发现其增益率较低，所以在取属性可以不做考虑

- 三元数据集

1. 三元数据集包含**62522**行数据，其中不重复词有**90295**个，如果不降维是完全无法进行操作处理的。
2. 数据集中，各标签的数量比LOW：MID：HIG = 2：3：2。
3. 分析词单，可以发现大部分词是没有意义的，统计了一下，非单词的个数有**28204**个。
4. 使用现有的50维40万词容量的词向量集，对词单取了一个交集，最终只剩下了**49794**个词。现有的40万容量的词是常用的词，且容量巨大，所以我们有理由相信，筛掉的词是不常用的词或者说没有意义的词。
5. 使用卡方检验，根据χ^2 -- P 表格，当卡方值大于10.83时，对应的相关性才具有可观性，统计所有90000+个词进行统计计算，最终发现所有词中只有**3098**个词可以说与标签有相关性。

- 回归数据集

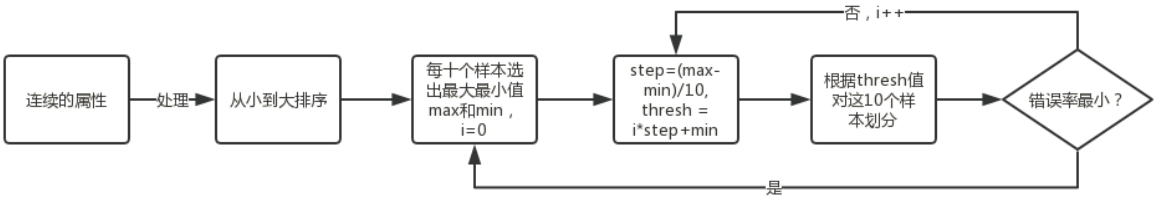
属性	特征
instant	只是样本的下标，连续型，不应参与模型的构建
dteday	具体日期，这个属性隐藏着较多信息，月份、星期这三个属性十分重要，且能从中挖掘出节假日、季节等属性，这些属性对于模型的构建起到了较大的帮助，例如，周末的单峰以及工作日的上下班导致的双峰；各个月份使用单车量不同；节假日的单车量单峰等
hr	小时，每日每小时的自行车使用量记录，同样是个重要的属性
weathersit	天气，有四个属性，根据天气的不同会导致自行车使用数量的增减，需要保留
temp	温度，和天气相类似，存在高温或低温导致自行车使用数量减少的情况，需要保留
atemp	与温度相关，与temp类似
hum	规范化温度，与temp类似
windspeed	风速，存在风速过大导致自行车使用数量减少的情况，需要保留
缺省值	在数据中存在某些数据是“？”，一般来说可以直接使用众数标签，如天气出现缺省，通过列表计算出1的属性值有11076个样本，属于众数标签，所以天气的缺省一律取1
噪音	在每个月以及每年都会有一行的统计数据，这个数据在训练的时候会造成很大的干扰，所以需要在代码中判断出来并删除掉

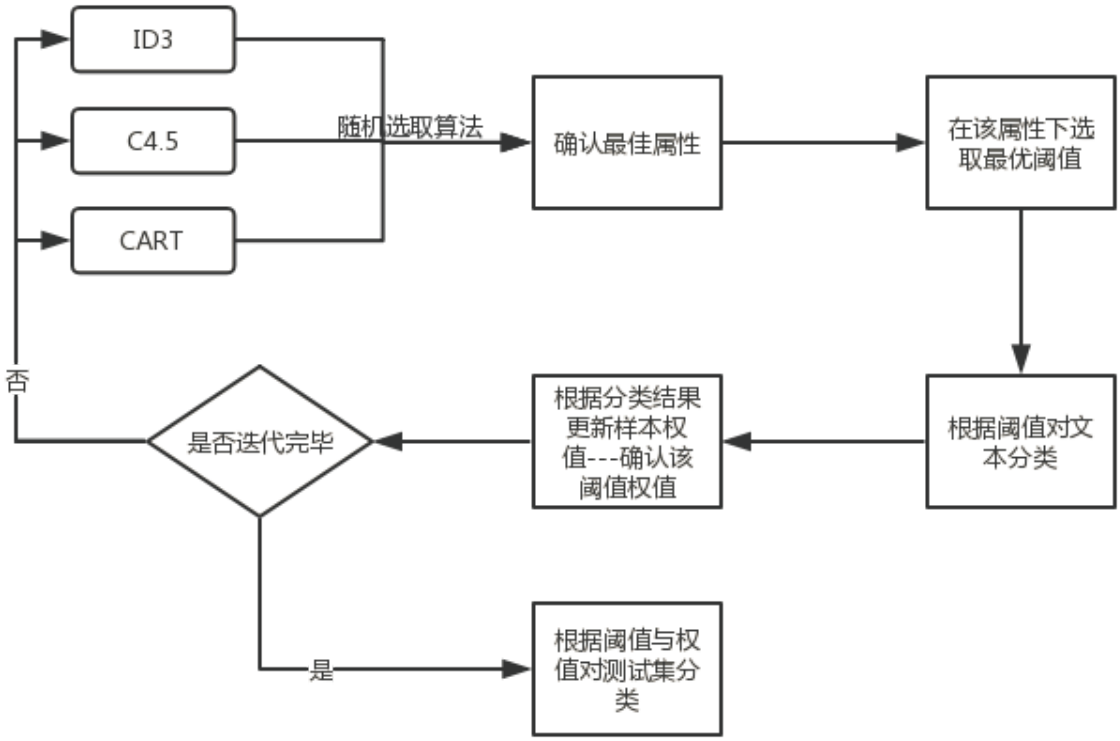
4. 集成学习方法(AdaBoost)

- 使用的是单层决策树+AdaBoost的方法，应用到二元分类上，效果F1=0.7063

1. 首先要将将连续值处理成离散值，处理完成之后再对根据不同标准选取最佳属性
2. 然后选取属性下的阈值作为分类器，更新样本权值，直到迭代完毕。

- 因为这里要根据错误率(样本权值)来选取最佳属性，所以对于ID3，C4.5，CART等值的计算也要有相应的改变，累加的不再是1而是样本权值。





三、 引用

见尾注

四、 课程总结

1. 本学期的实验课，从LAB1的数据处理，到LAB6的神经网络的实现，学到了很多，从完全没接触过人工智能，到现在已经能够初步看懂AlphaGo的对外技术介绍，确实是受益匪浅。
2. 人工智能课程，第一次的震撼便是距离的计算，通过OneHot或TF-IDF矩阵就能模拟数学过程得到训练集的距离，获得不同文本/数据之间的关系，从而得以发展出其他众多的算法，且能够较为基本的预测分类。这种对于文本/数据的挖掘是我从未想过的，刷新了三观。
3. 如果说NB、KNN只是通过概率或者纯粹使用距离玩点数学的把戏，那么从PLA开始的LR、神经网络就完全运用了计算机高强度计算的特点，模仿生物神经网络不断自适应纠错更新参数，从而得到一个预测较为准确的模型。以牛顿迭代的思想为基础使用的梯度下降法十分巧妙的只需要通过足够多的迭代次数就能逼近正确值，为纠错参数提供了更新公式。

4. 而决策树是另一个使用统计学领域知识实现的分类预测模型，其通过信息熵判断数据的不确定性，通过信息增益率（GINI系数）判断属性的重要性，从而形成树型的逐层决策，可解释性很强也很明了，十分类似人类的理性判断思维。
5. 总之，这次实验课通过亲自实现算法，熟悉了人工智能领域的常见算法，也学到了机器学习的基本思路，收获匪浅！

---

1. https://baike.baidu.com/item/%E5%8D%A1%E6%96%B9%E6%A3%80%E9%AA%8C/2591853?fr=aladdin↩

2. http://blog.sina.com.cn/s/blog_6622f5c30101datu.html↩

3. https://www.cnblogs.com/emanlee/archive/2008/10/25/1319569.html↩