



社会计算课程实验

任务讲解

— 目录 —

1. 题目概述
2. 基本思路
3. 注意事项

题目概述

五道题目中任选一题

- 电影评论情感分类 (☆ ☆)
- 虚假新闻识别 (☆ ☆ ☆)
- 中文垃圾短信识别 (☆)
- 中文幽默类型识别 (☆ ☆)
- 相似疾病问句匹配 (☆ ☆ ☆)

题目一：电影评论情感分类

要求：

给出电影评论，识别出这条评论是积极的还是消极的，积极用标签1表示，消极用0表示。

关键词：

分类任务、英文、长文本

id	sentiment	review
"5814_8"	1	xxxxxxx
"7759_3"	0	xxxxxxx
.....

题目二： 虚假新闻识别

要求：

给出一段新闻，判断是否为虚假新闻，1表示是，0表示不是。

关键词：

分类任务、英文、有扩展数据

Id	数据编号
Url	新闻网页地址
Title	新闻标题
Tweet_ids	分享了该新闻的推特id，可根据接口获取对应推特相关数据（点赞数、转发数等）

题目三：中文垃圾短信识别

要求：

给出短信文本内容，判断这条短信是否为辣鸡短信，是则标记为1，不是标记为0。

关键词：

分类任务、中文、短文本

0 所有的装修细节让我真是喜欢我家
1 豪美装饰xx年最新活动，工程总款返还xx%，地址：建华大街与中山路交叉口东行xxx米钻石广场A座xxx室。电话xxxxxxxxxxx
0 投资机会仍然集中在微观结构在此轮调整中出现明显改善的品种
0 港版原封小6金色16G仅需4199
1 亲，金汕教育春季班从x月x号起陆续开班啦！报名热线xxxxxxxx，或者直接回复需要补习的年级科目，我们会尽快跟您联系的。
0 徐州市区x条主干道违停将罚款xxx元记x分
0 重新审视、展览、出版中国与西方交流文化的始端
1 家长您好：旗帜数学本着提高学生成绩的宗旨，新学期开课啦。招生电话：xxxxxxxxxxx xxxxxxxxxxxx地址：五完小西十
0 昆山爆炸——铝粉燃爆的军用级威力

题目四：中文幽默类型识别

要求：

给出笑话文本内容，判断该笑话的类型，共三类，分别用1， 2， 3表示。

关键词：

分类任务、中文、隐含特征

尼采去面试，面试官问：“你叫什么？” “尼采。” “猜你个姥姥啊！下一个！” 1
有一天，我给电视拆了。 我爸对我说：你若安好，便是晴天，你若安不好，老子打死你。 2
甲：我妻子常提起她以前的丈夫，真气人。 乙：你真幸运，我妻子常提起她将来的丈夫。 3

题目五：相似疾病问句匹配

要求：

给出两个疑问句，判断他们是否相同（或相近），相同用1进行标记，不相同用0标记。

关键词：

分类任务变体、中文、相似语义

问句1:糖尿病吃什么?

问句2:糖尿病的食谱?

label:1

问句1:乙肝小三阳的危害?

问句2:乙肝大三阳的危害?

label:0

基本思路



数据处理 (1/2)

"\Those who envision themselves as amateur \"critics\" look only to criticize everything they can. Others rate a movie on more important bases,like being entertained, which is why most people never agree with the \"critics\".

The movie delivers the goods with lots of blood and gore as beheading

Train Data

text	label
今天特别开心	1
我有点难过	0
.....	

Dev Data

text	label
昨天特别开心	1
他很愧疚	0

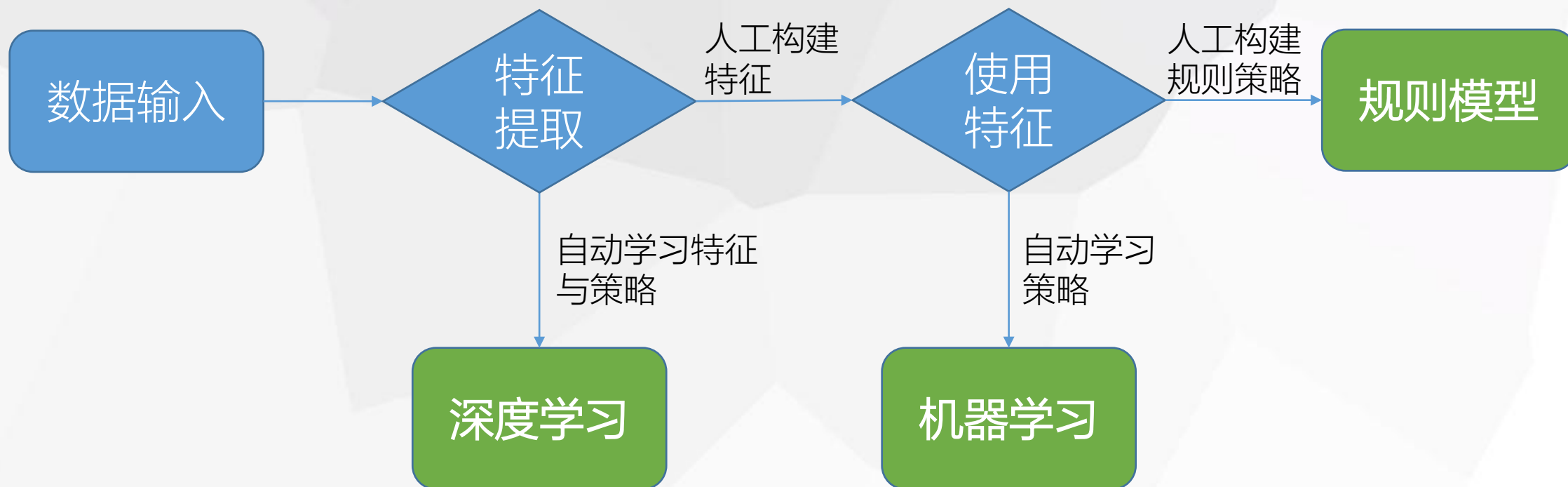
Test Data

text	label
他昨天很开心	
我难过	

数据处理 (2/2)

Train Data	Dev data(Valid Data)	Test Data
训练集	开发集 (验证集)	测试集
输入数据, 标准答案	输入数据, 标准答案	输入数据
训练模型	模型的制作者判断模型的效果	模型的审核者评估模型
60%-70%	15%-20%	20%

建立与训练模型



1. 基于规则

text	label
今天特别开心	1
我有点难过	-1
昨天特别开心	1
他很愧疚	-1
他昨天很不开心?	0
我特别难过	1
这家店菜好吃, 但是服务不好	

If "开心" in sentence and "不" not in sentence :
return 1

If "难过" in sentence or "愧疚" in sentence :
return 0

情感词典:

积极词、消极词、否定词、副词
(知网、台湾大学、大连理工)

思路简单, 但规则
复杂, 泛化性弱

2. 机器学习

0 所有的装修细节让我真是喜欢我家
1 豪美装饰xx年最新活动，工程总款返还xx%，地址：建华大街与中山路交叉口东行xxx米钻石
0 投资机会仍然集中在微观结构在此轮调整中出现明显改善的品种
0 港版原封小6金色16G仅需4199
1 亲，金汕教育春季班从x月x号起陆续开班啦！报名热线xxxxxxxx，或者直接
0 徐州市区x条主干道违停将罚款xxx元记x分
0 重新审视、展览、出版中国与西方交流文化的始端
1 家长您好：旗帜数学本着提高学生成绩的宗旨，新学期开课啦。招生电话：
0 昆山爆炸——铝粉燃爆的军用级威力

怎么使用呢？

	发布时间	文本长度	是否有特征词	是否中英语言混杂	是否有感叹号	结果
句子1	2:00	100	1	是	1	1
句子2	18:00	30	0	否	0	0
.....	

特征工程

2. 机器学习

不同于结构化数据，非结构化的文本结构，有没有能更优雅的、专属于文本的特征构造方法呢？

	发布时间	文本长度	是否有特征词	是否中英语言混杂	是否有感叹号	结果
句子1	3	1	1	1	1	1
句子2	4	0	0	0	0	0
.....	

依赖特征构造质量，需要标注数据，进行训练

方法一：

$$y = \max(\begin{aligned} &P(\text{结果} = 1) | p(\text{文本长度} = 1) \text{ and } p(\text{感叹号} = 1), \\ &P(\text{结果} = 0) | p(\text{文本长度} = 1) \text{ and } p(\text{感叹号} = 1) \end{aligned})$$

方法二：

$$y = W_1X_1 + W_2X_2 + \cdots + W_nX_n$$

结果评价

准确率(P): $TP / (TP + FP)$

召回率(R): $TP / (TP + FN)$

F1 score(F): $2 * (P * R) / (P + R)$

	真实为正	真实为负
预测为正	TP	FP
预测为负	FN	TN

例:

真实序列: 0 0 1 1 0 0

预测序列: 0 1 1 1 1 0



$$P = 2 / 4 = 0.5$$

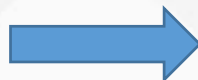
$$R = 2 / 2 = 1$$

$$F1 = 2 / 3$$

例:

真实序列: 0 0 1 1 1 0


预测序列: 0 0 1 0 0 0



$$P = 1 / 1 = 1$$

$$R = 1 / 3$$

$$F1 = 1 / 2$$



注意事项

提交方式:

将过程文档(.docx/.pdf)、结果文件(.txt)与源代码文件(.rar), 发送至 932974672@qq.com , 邮件主题与各文件命名方式均为“题目编号-学号-社会计算-姓名”, 例如“1-xxx-社会计算-大佬”。

结果格式:

请参考各题目对应文件夹下的readme文件。

截止时间:

11月30日24点



—谢谢—

