基于LightGBM的抑郁倾向、焦虑倾向和压力倾向调查

Light GBM是基于决策树的一种改进机器学习算法。

算法参考：https://zhuanlan.zhihu.com/p/99069186

Google Health一直尝试将自己的业务范畴由生理问题扩展到更多的地方，尤其是心理学相关的尖端领域。但是由于缺乏硬件测量数据，多数情况下心理学的测试只能通过问卷完成。那么，在医学界至今还不能清晰解释一些神经科学的现象的情况下，我们可以尝试依靠机器学习去分类和度量用户的负面倾向程度。

泰勒显性焦虑量表，简称MAS，是J.A.泰勒于1953年为配合瞬眼条件反应的研究而编制的一种测定焦虑水平的量表，目的是为了研究焦虑对学习的动机或内驱力作用而编制。

问卷的第一部分是单选题：

Q1 I found myself getting upset by quite trivial things. Q2 I was aware of dryness of my mouth. Q3 I couldn't seem to experience any positive feeling at all. Q4 I experienced breathing difficulty (eg, excessively rapid breathing, breathlessness in the absence of physical exertion). Q5 I just couldn't seem to get going. Q6 I tended to over-react to situations. Q7 I had a feeling of shakiness (eg, legs going to give way). Q8 I found it difficult to relax. Q9 I found myself in situations that made me so anxious I was most relieved when they ended. Q10 I felt that I had nothing to look forward to. Q11 I found myself getting upset rather easily. Q12 I felt that I was using a lot of nervous energy. Q13 I felt sad and depressed. Q14 I found myself getting impatient when I was delayed in any way (eg, elevators, traffic lights, being kept waiting). Q15 I had a feeling of faintness. Q16 I felt that I had lost interest in just about everything. Q17 I felt I wasn't worth much as a person. Q18 I felt that I was rather touchy. Q19 I perspired noticeably (eg, hands sweaty) in the absence of high temperatures or physical exertion. Q20 I felt scared without any good reason. Q21 I felt that life wasn't worthwhile. Q22 I found it hard to wind down. Q23 I had difficulty in swallowing. Q24 I couldn't seem to get any enjoyment out of the things I did. Q25 I was aware of the action of my heart in the absence of physical exertion (eg, sense of heart rate increase, heart missing a beat). Q26 I felt down-hearted and blue. Q27 I found that I was very irritable. Q28 I felt I was close to panic. Q29 I found it hard to calm down after something upset me. Q30 I feared that I would be "thrown" by some trivial but unfamiliar task. Q31 I was unable to become enthusiastic about anything. Q32 I found it difficult to tolerate interruptions to what I was doing. Q33 I was in a state of nervous tension. Q34 I felt I was pretty worthless. Q35 I was intolerant of anything that kept me from getting on with what I was doing. Q36 I felt terrified. Q37 I could see nothing in the future to be hopeful about. Q38 I felt that life was meaningless. Q39 I found myself getting agitated. Q40 I was worried about situations in which I might panic and make a fool of myself. Q41 I experienced trembling (eg, in the hands). Q42 I found it difficult to work up the initiative to do things.

这些问题的选项值含义如下：

1 = Did not apply to me at all 2 = Applied to me to some degree, or some of the time 3 = Applied to me to a considerable degree, or a good part of the time 4 = Applied to me very much, or most of the time

在数据集中，这些问题的答案以Q1A、Q2A、QxA为列名进行存储。

同时，QxE也存储了用户回答问题所耗的时间，以毫秒为单位。

由于上述的42道题目是随机出题，所以引入了一个QxI列用来表示该题目在用户填写问卷时的顺序。

introelapse表明用户在介绍页面停留的时间，以秒为单位。

testelapse表明用户完成上述所有问题的时间，以秒为单位。

surveyelapse表明用户完成下一部分统计问卷的时间。

第二部分统计问卷的问题为"I see myself as: ____"，并且有以下十题：

TIPI1 Extraverted, enthusiastic. TIPI2 Critical, quarrelsome. TIPI3 Dependable, self-disciplined. TIPI4 Anxious, easily upset. TIPI5 Open to new experiences, complex. TIPI6 Reserved, quiet. TIPI7 Sympathetic, warm. TIPI8 Disorganized, careless. TIPI9 Calm, emotionally stable. TIPI10 Conventional, uncreative.

答案选项对应的含义为：

1 = Disagree strongly 2 = Disagree moderately 3 = Disagree a little 4 = Neither agree nor disagree 5 = Agree a little 6 = Agree moderately 7 = Agree strongly

该条目在数据集中以TIPIx存储。

第三部分是常识判断题，问卷列举一系列单词，考察用户是否理解下述单词的含义：

VCL1 boat VCL2 incoherent VCL3 pallid VCL4 robot VCL5 audible VCL6 cuivocal VCL7 paucity VCL8 epistemology VCL9 florted VCL10 decide VCL11 pastiche VCL12 verdid VCL13 abysmal VCL14 lucid VCL15 betray VCL16 funny

结果以VCLx存储，1代表是，0代表否。

第四部分是个人信息调查，包含：

education

"How much education have you completed?",

1=Less than high school, 2=High school, 3=University degree, 4=Graduate degree

urban

"What type of area did you live when you were a child?",

1=Rural (country side), 2=Suburban, 3=Urban (town, city)

gender

"What is your gender?",

1=Male, 2=Female, 3=Other

engnat

"Is English your native language?",

1=Yes, 2=No

age

"How many years old are you?"

hand

"What hand do you use to write with?",

1=Right, 2=Left, 3=Both

religion

"What is your religion?",

1=Agnostic, 2=Atheist, 3=Buddhist, 4=Christian (Catholic), 5=Christian (Mormon), 6=Christian (Protestant), 7=Christian (Other), 8=Hindu, 9=Jewish, 10=Muslim, 11=Sikh, 12=Other

orientation

"What is your sexual orientation?",

1=Heterosexual, 2=Bisexual, 3=Homosexual, 4=Asexual, 5=Other

race

"What is your race?",

10=Asian, 20=Arab, 30=Black, 40=Indigenous Australian, 50=Native American, 60=White, 70=Other

voted

"Have you voted in a national election in the past year?",

1=Yes, 2=No

married

"What is your marital status?",

1=Never married, 2=Currently married, 3=Previously married

familysize

"Including you, how many children did your mother have?"

major

"If you attended a university, what was your major (e.g. "psychology", "English", "civil engineering")?"

最后一部分则是自动生成的技术信息，包含：

country ISO country code of where the user connected from

screensize 1=device with small screen (phone, etc), 2=device with big screen (laptop, desktop, etc)

uniquenetworklocation 1=only one survey from user's specific network in dataset, 2=multiple surveys submitted from the network of this user (2 does not necessarily imply duplicate records for an individual, as it could be different students at a single school or different memebers of the same household; and even if 1 there still could be duplicate records from a single individual e.g. if they took it once on their wifi and once on their phone)

source how the user found the test, 1=from the front page of the site hosting the survey, 2=from google, 0=other or unknown

综合运用上述大量的数据，尝试使用LightGBM对DASS/data.csv构建一个预测器，输出为0到10的负面情绪水平。

由于数据种类极其复杂，请注意在开始训练之前需要花费大量时间对一些数据进行预处理，包括但不仅限于：空数据、重复数据、完成时间过短或者过长的无效数据、全部选项一致的恶意数据等。

强烈建议为了更好的准确度，先进行数据的清洗。