自然语言处理

Project 1 Task 2

院系：人工智能学院
姓名：王蔚昕
学号：211300042

2024 年 4 月 24 日

目录

1 子问题 1 **3**

1.1 难点 3

2 子问题 2 **3**

2.1 实验一 3

2.2 实验二 5

1 子问题 1

1.1 难点

1.1.1 言语与性格的相关度

首先必须承认的是，言语一定与性格相关，但是在这里要说的事，言语与性格能有多相关？因为我认为语言对于性格来讲并非一个强有力的证明因素，他可能只是一个更容易收集到的因素，而且与性格有关联。所以我提出的猜想是，仅仅凭借言语可能不足以完全确定一个人的性格。一个人格可能有代表性行为，代表性处理事情的偏向，但是语言收到影响的东西太多了，同样的性格，在不同的环境中成长表现出来的言语可能是不同的，甚至截然相反。因为人们对于语言的学习大多数来自于身边的人，说话习惯也可能和身边的某个人或某些人有相似之处。所以我决定验证不同性格人之间言语的相似度。

1.1.2 分类方法不合理

想到这一点的原因是，在 KNN 的实验中，因为我采用了将性格分为四类分别预测的方式，最后发现哪怕 KNN 这种相对原始的方法对于每一种人格预测的分辨率也能达到 60% 以上，如果使用神经网络进行这种程度的二分类问题，想必效果会出乎意料的好。所以我认为，这种过于细致的分类可能本身之间也不存在明显地界限，结合我经历过的 MBTI 性格测试题目以及最终得出结论的方法来看，这种方法只是给每个题目的选择在不同维度打分，综合计算各个维度的分数而已，所以四个维度有明显清晰地界限的也仅仅是他们对应的两个性格而已，直接进行 16 种性格的分类似乎是不合理的，于是我决定对于这一点展开实验。

2 子问题 2

2.1 实验一

2.1.1 实验方法

为了调查不同性格之间的话语相似度，我们预计统计两种完全对立的性格中使用的单词数目，计算比例，然后用内积来判断这两种性格语言的相似

度。

2.1.2 实验结果

实验结果表明，哪怕是完全相反的两种人格他们的语言之间的相似度也高达 98%。

2.1.3 实验结果分析

通过以上实验结果，可以发现的是，不同人格之间说的话或许并没有显著差距，或者说这个差距很小，这一点可能会降低神经网络分类器的性能。这一点更加证明了 KNN 在实验中取得效果并不好的原因，因为 KNN 的度量函数采用普通的二范数来度量，而已经证明了话语之间的不同并不完全来自性格，可能更多来源于个人，所以这种表面的学习方法不能很好理解话语深处可能的表达不同，导致 KNN 效果远远低于神经网络。神经网络可能在语言理解上超过了 KNN，但是，但是由于语言和性格之间关联并非十分强烈，所以哪怕是有能力理解语义的神经网络也不能完美区分性格。

2.1.4 一些思考

出现这个结果也不难预料，因为总有一些高频词汇，不论什么人说话总要用到，这样的词汇往往参考价值不大，结合上一个报告中的思考，面临一个境地，数量稀少的词汇到底应该怎么认定？这些词汇是只有这类人才会说的还是说这些词汇只是这几个人会说的？如果是前者无疑可以对模型起到很好地作用，但是后者反而可能对预测造成干扰。当然，这是基于词频统计的传统学习方法来讲，深度学习的问题可能要涉及语义理解了。但是语义就一定可以表达一个人的性格吗？一个极端的例子，如果所有人围绕 1+1 等于几这个问题来讨论，那么神经网络找到的语义一定是大差不差的，截然不同的语义可能是某些调皮的性格的人，也可能的的确确是所谓的“九漏鱼”。所以，我还是认为，仅仅用文本来判断区分如此细致的 MBTI 性格划分是不太合理的。

2.2 实验二

2.2.1 实验方法

为了探究不同分类标准的效果，我分别实验了从四个维度建立分类器，记录分类效果，对比与直接进行 16 分类的实验结果的效果区别。（每个分类器用 100 个训练样本和测试样本）

2.2.2 实验结果

实验结果如下表

类别	准确率
E&I	0.75
N&S	0.86
T&F	0.62
J&P	0.62
total	0.247938

2.2.3 实验结果分析

这个实验表明，猜想是正确的，可以看到的是不同性格的分类准确度相差极大，最高和最低有 20% 左右的差距，所以可以初步判定，有些性格对于文本内容的影响是更大的，有些性格则不然。而且，这样四个二分类器最终综合的准确率达到了 0.247938，远远高于用 100 个样本的直接得出 16 分类器的准确率，而且，随着更多样本的输入这个准确率可能还会提高，因为 100 个样本相对于神经网络来讲还是太少太少了，提升的空间一定存在。所以我猜测直接使用 MBTI 进行性格分类未必是一个很好地方法，因为 MBTI 分类和文本的相关性并不算高，至少其中某些方面是这样的。

2.2.4 一些思考

1. 首先在实验中我犯了一个很简单的缓存错误，预实验内容无关，这个错误曾经严重影响了实验的效率。

2. 其次，发现了一个在 tsak-1 中出现的一个问题，已经在 github 中完成代码的修正。

3. 在进行完实验后，我确信神经网络分类效果一般的原因是样本问题，因为一个简单的二分类问题神经网络没有理由不如 KNN，但是需要大量样本也说明了这两个性格标签的差距似乎也不是很大，于是我回顾了 task-1 中 KNN 的分类结果，我发现 KNN 中 I，E 性格的分类准确度是遥遥领先与其他相对人格的分类器的，所以我认为，并不是所有人格都在言语上有重大区别，或者说不同方面的性格在言语上的差别不同，比如 IE 的区别可能会导致在语言上有更大的差距。所以证明了一点，用文本来分类 MBTI 性格的确不是很科学，因为某些方面的分类对于文本的影响可能很小。

   那么什么样的性格与文本的关联度大？根据社会经验，理性和感性的性格的文本会呈现出很大的区别，理性的人可能更注重分析，感性的人可能更偏向于表达感受。理想主义者和现实主义者也会有较大不同，前者更愿意相信未来的可能性，后者更乐于结合可行性现实的分析等等。比如，J 和 P 判断和感知可能在行为上体现更好，所以在实验中通过文本判断正确率并不高。

   根据现在的 MBTI 性格测试的情况也可以看出，通过文本分析并非一个好手段，大多数测试都是由情景选择来得到评分而决定的，所以对于文本判断性格，可能更适合判断可以在文本上立竿见影判断出来的类型。

4. 除此之外，另外一个让我很在意的点是，在同样数量上用多个分类器反而比用一个更好？我思考后认为原因可能有：部分性格之间有交叉性，训练不到位。

   对于前者，比如 SN 和 TF，直觉理应和情感更接近，实感和思考往往也是绑定的，这可能会造成一些隐藏的粘连，导致 16 分类的分类器 ST 更容易出现在一起，NF 更容易出现在一起，蕴含了这样一个隐藏的规律，导致实际上分类器并没有理解 SN，TF 这两组的区别，因为 MBTI 中这两组其实的内涵是不太一样的，现在流行的性格判断方式做选择赋分制度就可以很好地完成这个区分，因为这个区分由专家完成，那么分类器夹杂了这种粘连可能会对综合的判断造成一定的影响。

   训练不到位是因为数据量不够大，因为 BERT 训练太过缓慢，大量数据在消费级显卡上难以维持，可能对于最终结论有一定影响。