选题报告

一、选题描述

有一些研究工作提出了词向量的预训练方法，即根据大量无标注的自然语言文本训练得到其中每个词语的词向量

使其能够较好的表达出每个词语的含义（例如相似词语的词向量也较为相近），并最终为下游任务提供较好的初始化参数。

现在给定一个小规模的词向量文件，其中每行都包含了 1 个英文单词的类别和它所对应的 50 维的词向量（类别与数字都以空格隔开），一共有 2 类（country，sports）40 个单词的词向量。期望构建一个【基于 50 维词向量】进行【单词类别预测】的模型。

需要回答/解决下述问题，数据见附件 word_embedding.txt:

1、说明单词类别预测是一个回归任务还是分类任务，请说明理由。

2、构建模型，使用【全部的 50 维特征对单词类别进行预测。】

每类前 16 条数据作为训练集，剩下数据作为测试集

最后需要给出模型在测试集上预测的效果（自己选择评估指标并说明原因）。

3、选择合适的降维方法，将特征【降维到适当的维度】，然后构建模型，使用降维后的特征对单词类别进行预测。

前 16 条数据作为训练集，剩下数据作为测试集，最后需要给出模型在测试集上预测的效果（自己选择评估标准并说明原因）。

4、思考该问题中能否使用 KNN，如果不能，请说明原因；如果可以请说明如何使用。、

二、其他问题

2.1 说明任务是一个回归任务还是分类任务，并说明理由。

是分类任务，给的训练集和测试集之中有两个分类 country 和 sport，且题目要求说要把未知数据分成两个类别。

回归问题的输出是连续的。

回归问题是结合原有的多重信息，拿 “预测波士顿房价”来说，结合房子“大小”、“地理位置”、“价格”等等因素，最后拟合出一条曲线。这条曲线是连续的，你给出任意一个输入都可以得到一个预测的输出。

分类问题的输出是离散的。

顾名思义，同样也是根据已有的信息进行整合，但最后输出的值是一大类一大类的，例如通过网络后判断是猫还是狗，这就是一个分类问题

而通过“单位阶跃函数”和“对数几率函数”来模拟的问题也是可以用 KNN 方法来解决的。（满足离散的问题）。当需要使用分类算法，且数据比较大的时候就可以尝试使用 KNN 算法进行分类了

2．2 降维方法描述：选择的什么降维方法，降到多少维，依据是什么。

降到 2 维度，但是此时应该调高 K 的值（不能再是 50 维下的 5 了），变成 2 维之后经验误差的敏感度增加，所以需要尽量让决策更加接近数据本身（可以提供判断的数据见笑了），即增加考虑的范围（从原来周围 5 个数据，变成现在 7 个数据）

2.3 该问题中能否使用 KNN，如果不能，请说明原因；如果可以请说明如何使用（只需说明即可）。

可以，是分类问题，且数据集较大，可以采用 KNN 算法。具体请参见代码。