浙江工业大学

文本分析与挖掘

上机实验

计算机科学与技术学院

实验六、基于多层感知机的文本分类

一、 实验目的

1. 使用多层感知机对文本数据进行分类，对比不同预处理和参数对分类结果的影响。
2. 熟悉多层感知机的分类过程。

二、 实验内容

把 20newsgroups（全部 20 个类）按 4：1 分成训练和测试集。再从训练集里面分出 10%作为验证集。

1.基于词袋表示为输入的多层感知机分类

a. 采用实验 3 中的预处理对训练集和测试集进行预处理并得到词袋表示。

b. 创建具有单层隐藏层（256 个节点）的多层感知机，设置激活函数为 ReLu，初始学习率为 0.001，轮数 epoch 为 50,用训练集对模型进行训练,画出学习曲线(训练、验证损失以及准确率)。观察曲线，讨论模型的学习情况（欠拟合、过拟合）

c. 得到测试集准确率，并对比实验 3 中基于朴素贝叶斯算法的结果。

d. 尝试改变实验设置（增加隐藏层节点个数、增加层数、初始学习率、优化器、增加 epoch 数目等）来提升测试集准确率并讨论结果。

2.基于词嵌入为输入的多层感知机分类

得到训练集和测试集中每个词的 100 维词嵌入向量，可以调用 gensim 或直接从 **http://nlp.stanford.edu/data/glove.6B.zip** 下载。

b.基于词向量得到的文档表示(直接对包含的词的向量平均或其他方法）作为输入，重复上面 1(b-d)的内容，对比后进行讨论。

3.（选做）参照课件代码和步骤，逐步实现 CBOW 和 Skip-gram 算法。