# 机器学习纳米学位

## 文档归类

### 背景

随着网络的普及以及惊人的发展速度，伴随而来的是越来越多数据的积累，包括文本、图像、声音等。文本数据所占用的网络资源少，网络资源中大部分的数据是以文本的形式出现的，对这些文本的研究成为自然语言处理（NLP)的热门。自然语言处理经历好几个阶段，从最开始的基于规则的方法论，到基于统计的学习的方法，再到后来的深度学习。文本分类是NLP中的应用之一，目前主流的方法有基于TF-IDf特征提取再结合SVM/贝叶斯等模型进行训练，或者基于词嵌入的深度学习方法都达到了不错的效果。本毕业设计将集中研究这些方法解决文本分类问题。

### 问题描述

文档归类在本项目中属于一个有监督的分类问题。需要先将文字信息转化为计算机能够识别的特征，由机器学习不同分类的特性，然后输入新的文章时能够自动归类。

### 相关数据集

训练集和测试集为20类新闻包 (http://www.qwone.com/~jason/20Newsgroups/)，一共有近2万篇的文章被均匀的分为20类，以及用来训练词向量的维基百科的部分英文语料 (http://mattmahoney.net/dc/text8.zip)。可以在官方网站上直接下载,数据的加载和预处理(大小写转换，词干提取等等)可以直接使用sklearn的fetch_20newsgroups。预处理完成之后抽取文章的特征并向量化表示(TF-IDF/词袋模型等)然后使用svm/贝叶斯等模型进行分类并评估结果。在进行具体的学习任务前，首先需要对数据集进行分析。每个分类的常用词,每篇文章大概有多少可用的词特征，这些词的分布规律等。梳理清楚这些数据对后面选择模型以及分析调优模型都有巨大的作用。

### 解决方案

总共分为四个部分：

1. 文本的预处理。
2. 将文本转化为可量化的特征(TF-IDF)。
3. 训练SVM/贝叶斯/神经网络等模型进行训练。
4. 结合验证集和测试集对模型进行调优并输出最终的模型。

### 基准模型

文档分类已经有很多成熟的解决方案，包括kaggle上也有很多类似的比赛 (https://www.kaggle.com/c/cmsc-5622-document-classification)，正确率能达到80%以上。

### 验证方法

在训练好模型后，使用测试集来进行验证，正确率的计算方法为算法预测正确的分类数除以总预测数。

### 方案设计

在进行机器学习任务之前，首先要熟悉数据，仔细浏览数据了解数据。在了解数据之后对数据进行预处理，比如除去异常的字符、大小写转化、去停止词等等。并且将数据按照8:1:1的比例分为训练集、验证集、测试集。完成数据的预处理之后使用sklearn或者gensim统计单词的TF-IDF的值，TF-IDF值越大则代表一个词的越有代表意义。统计完成TF-IDF值后将文档使用词袋模型表示，一篇文章就能使用向量来表示，每一个维度的值就是词的TF-IDF值。 将文章使用词袋模型表示后使用SVM/贝叶斯等模型进行训练，并且在验证集上进行测试。针对不合理的结果反向分析原因调优模型，直到在验证集上得到不多的准确率，然后再在测试集上进行最终的验证，确认模型的效果，如果模型的效果还是不合理继续反向分析原因直到准确率合理。 在时间富裕的情况下将考虑使用深度学习的方法进行文档的分类，包括使用维基百科的语料训练词向量，并且使用词向量表示文章，使用深度学习的方法进行文档的分类，验证的流程和之前一致。