**机器学习（进阶）纳米学位**

**文档归类 开题报告**

吕奇峰  
December 28st, 2017

**开题报告**

**项目背景**

自然语言处理（NLP）是人工智能极为重要的一部分。

是计算机科学，人工智能，语言学关注计算机和人类（自然）语言之间的相互作用的领域。因此，自然语言处理是与人机交互的领域有关的。在自然语言处理面临很多挑战，包括自然语言理解，因此，自然语言处理涉及人机交互的智能。在NLP诸多挑战涉及自然语言理解，即计算机源于人为或自然语言输入的意思，和其他涉及到自然语言生成。  
 机器学习下的自然语言处理不同于一般的语言处理算法。在数据量不足，计算机算力也不大的时候，人们普遍大规模的使用规则模型去理解语意。将注意力主要集中在规则上，不同的规则将会输出不同的结果。而机器学习则改变着这种形态。机器学习通过海量的数据，淡化人为的规则，将更多的注意力放在了数据本身。通过巨大的算力，进行巨量的运算而自然生成自然语言的模型。这种做法为自然语言处理带来了跨越性的进展，突破了传统的瓶颈。

一些最早使用的算法，如决策树，产生硬的if-then规则类似于手写的规则，是再普通的系统体系。然而，越来越多的研究集中于统计模型，这使得基于附加实数值的权重，每个输入要素柔软，概率的决策。此类模型具有能够表达许多不同的可能的答案，而不是只有一个相对的确定性，产生更可靠的结果时，这种模型被包括作为较大系统的一个组成部分的优点。

自然语言处理研究逐渐从词汇语义进一步的到叙事的理解。然而人类水平的自然语言处理，是一个人工智能的极限问题。它是相当让人工智能和人一样聪明。这是自然语言处理的未来，因此密切结合人工智能发展。

**问题描述**

自然语言处理有个非常复杂的问题解决过程。首先是一个难点是分词，只是分词现在已经有非常多的库可以使用，也已经有相对比较成熟的算法，得到比较好的成功率。分词对于不同语言的含义是不一样的，比如中文，正向思维是把一个句子分成短语，短语分成词。而逆向思维则是字组成词，词组成短语，短语组成句子。而英文则不太一样，因为的最小单位是字母，字母组成词，词组成习惯短语，然后再组成句子。

分词之后就是词、语句以及文章的表达。以英文为例，最常见的词语表述方式比如”cat“、”dog“，这些都是利用字母表示意思。统计语言处理里面，比较容易利用字母来描述概率模型，比如ngram模型 ，计算两个单词或者多个单词同时出现的概率，但是这些符号难以直接表示词与词之间的关联，也难以直接作为机器学习模型输入向量。对句子或者文章的表示，可以采用词袋子模型，即将段落或文章表示成一组单词，例如两个句子：”She loves cats.“、”He loves cats too.“ 我们可以构建一个词频字典：{"She": 1, "He": 1, "loves": 2 "cats": 2, "too": 1}。根据这个字典, 我们能将上述两句话重新表达为下述两个向量: [1, 0, 1, 1, 0]和[0, 1, 1, 1, 1]，每1维代表对应单词的频率。

​近年来，借助深度学习概念和性能强劲的硬件平台，Geofrey Hinton, Tomas Mikolov, Richard Socher等学者深入开展了针对词向量的研究，将自然语言处理推向了新的高度。以词向量为基础，可以方便引入机器学习模型对文本进行分类、情感分析、预测、自动翻译等。最简单的词向量就是独热编码(one-hot encoder)，比如有三个单词“man"、”husband“、”dog“，将之分别表示为[0,0,1]，[0,1,0]，[1,0,0]，这些词向量可以作为机器学习模型的输入数值向量，但是它们依然难以表达关联性，而且当词库单词量庞大时，独热编码的维度也会十分巨大，给计算和存储带来不少问题。Mikolov、Socher等人提出了Word2Vec、GloVec等词向量模型，能够比较好的解决这个问题，即用维数较少的向量表达词以及词之间的关联性。关于这些词向量模型的具体原理，可以阅读他们所发表的论文，主要是英文，中文网站上也出现了不少精彩的翻译和解读，可以参考某些关于自然语言处理的中文博客。

​本项目目的就是利用上述自然语言处理技术结合所学机器学习知识对文档进行准确分类。

**输入数据**

分类文本数据使用经典的20类新闻包，里面大约有20000条新闻，比较均衡地分成了20类，是比较常用的文本数据之一。

​此外，词向量的训练也需要大量数据，如果20类新闻数据样本量不足以训练出较好的词向量模型，则增加Mikolov使用的text8数据包进行训练。

**解决方法**

探索文本表示的方式

使用词袋子模型来表示每篇文档，常见的一种思路是首先将文本进行分词，也就是将一个文本文件分成单词的集合，建立词典，每篇文档表示成特征词的频率向量或者加权词频TF-IDF向量，这样可以得到熟悉的特征表。接下来，就可以方便利用机器学习分类模型进行训练。

利用Word2Vec方式即词向量模型表示每篇文档，这里面包含两部分主要工作：

利用文本数据对词向量进行训练，将每个单词表示成向量形式。词向量训练后需要进行简单评测，比如检验一些单词之间相似性是否符合逻辑，下图是我在text8数据上训练的词向量模型，其中"school"、“university"、”college“三个意义比较类似的词基本上聚在一起。词向量示意图

探讨怎样用文档中每个词的向量来表达整个文档。学有余力的同学还可以尝试Dov2Vec的模型来直接训练表示每篇文档。

分别在词袋子、词向量表达基础上采用你认为适当的模型对文本分类，优化模型并分析其稳健性。

**基准模型**

文本表示模型在前面已经谈到，这里不作赘述。 注意文本预处理方式可能对最终结果有影响，对于某些语言比如英文，是否要考虑区分单词大小写、是否要对同一词不同形式（如单复数）进行统一、是否要保留标点符号？这些讨论都应该反映在报告中。

​下列分类模型可供参考：

决策树模型

支持矢量机(SVM)模型

朴素贝叶斯模型

神经网络模型

**评估指标**

*(approx. 1-2 paragraphs)*

In this section, propose at least one evaluation metric that can be used to quantify the performance of both the benchmark model and the solution model. The evaluation metric(s) you propose should be appropriate given the context of the data, the problem statement, and the intended solution. Describe how the evaluation metric(s) are derived and provide an example of their mathematical representations (if applicable). Complex evaluation metrics should be clearly defined and quantifiable (can be expressed in mathematical or logical terms).

**设计大纲**

*(approx. 1 page)*

In this final section, summarize a theoretical workflow for approaching a solution given the problem. Provide thorough discussion for what strategies you may consider employing, what analysis of the data might be required before being used, or which algorithms will be considered for your implementation. The workflow and discussion that you provide should align with the qualities of the previous sections. Additionally, you are encouraged to include small visualizations, pseudocode, or diagrams to aid in describing the project design, but it is not required. The discussion should clearly outline your intended workflow of the capstone project.

**Before submitting your proposal, ask yourself. . .**

* Does the proposal you have written follow a well-organized structure similar to that of the project template?
* Is each section (particularly **Solution Statement** and **Project Design**) written in a clear, concise and specific fashion? Are there any ambiguous terms or phrases that need clarification?
* Would the intended audience of your project be able to understand your proposal?
* Have you properly proofread your proposal to assure there are minimal grammatical and spelling mistakes?
* Are all the resources used for this project correctly cited and referenced?