2016 年春人工智能导论文本分类编程实验

◆ 实验题目

使用 SVM 及朴素贝叶斯两种方法对给定的文本数据集进行文本分类。要求编写代码实现朴素贝叶斯方法（方法主体要求自己实现，可以参考各类开源实现），SVM 方法可以自己实现或使用现有工具包实现。

◆ 实验描述

文本分类是信息检索应用中的经典问题，而 SVM 及朴素贝叶斯也是两种经典的分类方法，本实验的目标是让同学们了解课上所学知识如何在实际中得到应用，同时对不同的方法的效果及优缺点有所对比。

◆ 实验数据

实验数据由独立的文本文件组成，每个文本文件代表一篇文档，按不同主题共分为 4 类分别放在不同的文件夹下，四种分类分别为：c1_atheism（无神论），c2_sci.crypt（洞穴），c3_talk.politics.guns（政治），c4_comp.sys.mac.hardware（计算机）。

◆ 实验要求

对上述给定数据自己设计方法构建训练集和测试集，使用 SVM 及朴素贝叶斯两种方法进行文本分类，评价实验效果，针对实验进行分析，提交源代码及实验报告。最终评分将根据实验设计、实验实现及实验分析的撰写情况给出。

◆ 提示

1. 算法实现：
   SVM 推荐 libSVM(http://www.csie.ntu.edu.tw/~cjlin/libsvm/)。
2. 构建训练集和测试集：
   采用 5 折（10 折）交叉验证的方法。

◆ 注

1. 实验并不以最终的分类效果作为主要评价标准（对于本次实验的数据集，一般来说只要适当的进行参数调整最终的准确率都可以达到 90%），而是以每位同学的工作量和报告撰写情况来评价，所以请务必在报告中体现自己的工作量以便助教核对（例如自己实现 SVM 而非使用工具包等等）。
2. 报告中的实验结果请整理成表格等形式列出，请不要将程序输出直接粘贴到报告中（或截图）。
3. 千万不要抄袭，助教存有之前几届所有作业的数据，一旦被判定为抄袭，将可能对课程通过产生影响！