

朴素贝叶斯分类实验

机器学习概论-第一次实验

2025-春



目标

- 实现一个朴素贝叶斯分类器并在真实数据集上进行评测
- 主要内容：
 - 如何**实现**一个机器学习算法并将其应用到真实数据集上？
 - 如何**评估**机器学习模型的性能？
 - 如何**分析**实验结果？
- 注意：上述三个部分都很重要，应体现在代码/报告中



朴素贝叶斯分类器

- 假设： $P(y|x_1, \dots, x_n) \propto P(y) \prod_{i=1}^n P(x_i|y)$
- 训练阶段：
 - 估计 $P(y)$ 和 $P(x_i|y)$
- 测试阶段：
 - 输出 $\hat{y} = \operatorname{argmax}_y P(y) \prod_{i=1}^n P(x_i|y)$



任务和数据

- 任务：确定一封电子邮件是否是垃圾邮件
- 数据集：英文电子邮件数据集

<https://plg.uwaterloo.ca/~gvcormac/treccorpus06/>

- 文件格式：
 - `./data/`: 每个文件是一封电子邮件，包括正文和元信息 (共37,822封)
 - `./label/index`: 每行包括标签 (spam/ham) 和对应电子邮件的相对路径
 - 可能会有一些噪声数据 (例如不同的编码)



模型评估

- 在训练集上训练分类器并在测试集上测试其性能

- 使用五折交叉验证

- 至少汇报平均准确率(Accuracy):

- $Accuracy = \frac{\text{正确分类样本数}}{\text{测试样本数}}$

- 鼓励使用其他相关的评价指标 (例如 precision, recall or F1)



实验分析

- 实验中遇到或者探究了什么问题？
- 如何分析和解决问题？
 - 如何设计进一步的实验？
 - 如何调整算法？
- 解决方案是否有效？
 - 是否提升了模型的效果？
- 最终解释方案有效的原因(或者无效的原因)



问题 1: 训练集大小对性能的影响

- 训练集大小对分类器的性能有着怎样的影响？
- 建议方案：
 - 采样 5%, 50% 和 100% 的训练集数据进行训练，观察性能变化



问题 2: 零概率问题

- 假设在训练集中没有样本出现 $x_i = k, y = c$
- 那么 $\hat{P}(y = c | x_1, \dots, x_i = k, \dots, x_n) = 0$
- 这个问题对性能有什么影响？它在什么情况下可能发生？
- 一种可能的方案：
 - 平滑: $\hat{P}(x_i = k | y = c) = \frac{\#\{y=c, x_i=k\} + \alpha}{\#\{y=c\} + M\alpha}$
 - M 是样本标签类别的数量



问题 3: 特征设计

- 除了词袋模型之外，还可以使用什么特征？
- 提示：
 - Received from ...
 - Time
 - Priority/Mailer



评分要求

- 实现一个朴素贝叶斯分类器 (30%)
- 解决上述三个问题：
 - 问题 1 (30%)
 - 问题 2 & 3 ($2 \times 20\% = 40\%$)
- 注意:评分并非基于模型性能,而是基于你如何实现算法、如何评估模型以及如何分析实验结果



作业提交格式

一个包含以下内容的压缩文件：

- **Source Code**

- 包含必要的注释
- 确保助教能够理解并运行代码复现结果 (注意 **设置随机种子**)

- **README**

- 一个文本文件 (utf8编码)：姓名、学号、联系方式以及代码运行指南

- **Report**

- 一个PDF文件：实验设计/结果/分析/讨论
- 请不要直接复制代码

- **请不要上传数据集**



截止时间以及其他信息

- 截止时间: **2025.03.23 Sunday 23:59**
 - 上传压缩文件到网络学堂, 文件名为 姓名_学号
 - 迟交作业会有扣分
 - 迟交 \leq 一周: 0.8; 迟交 $>$ 一周: 0.6
 - 如果有特殊情况, 请提前告知助教
 - 抄袭代码和报告是不被允许的。学校提供了作业查重系统, 被确认为抄袭的作业会严重扣分
- 如有任何问题, 请联系助教
 - 王亦凡, 何祉瑜, 江楚萌
 - {yf-wang21, hezy22, jcm24}@mails.tsinghua.edu.cn



参 考 文 献

- http://scikit-learn.org/stable/modules/naive_bayes.html

(基础理论以及平滑技术)

- 仅供参考，核心算法应当自己实现

