期中项目

幽默文本检测

雷至祺

2021.10.21

目录

- 项目任务
- 项目知识梳理
- 项目要求

目录

幽默文本检测：数据竞赛

- 子任务1：幽默文本分类（二分类）
  - 给定文本，检测其是否幽默
  - 使用Kaggle平台，Kaggle竞赛邀请链接：https://www.kaggle.com/t/2d1345980fc04ba8b1114663e55f33ee

- 子任务2：幽默文本评分（回归）
  - 给定幽默文本，对其幽默程度进行打分
  - 使用Kaggle平台，Kaggle竞赛邀请链接：https://www.kaggle.com/t/75610f44a3e647acbada5375312b531c

幽默文本检测：项目任务

- 训练集介绍：train.csv
  - id – 整数，文本的唯一id
  - text – 字符串，文本内容
  - is_humor – 0/1，分类标签，0表示不幽默，1表示幽默
  - humor_rating – 范围[0,4]的实数，回归标签，幽默文本的幽默程度评分

	A	B	C	D
1	id	text	is_humor	humor_rati
2	1	TENNESSEE: We're the best state. Nobody even comes close. *Elevenn	1	2.42
3	2	A man inserted an advertisement in the classifieds "Wife Wanted". The	1	2.5
4	3	How many men does it take to open a can of beer? None. It should be	1	1.95
5	4	Told my mom I hit 1200 Twitter followers. She pointed out how my br	1	2.11
6	5	Roses are dead. Love is fake. Weddings are basically funerals with cake	1	2.78
7	6	'Trabajo,' the Spanish word for work, comes from the Latin term 'trepa	0	
8	7	I enrolled on some skill training and extra curricula activities that adde	0	
9	8	ME: I'm such an original. Truly one of a kind. ALSO ME: [holding a glas	1	1.79
10	9	Men who ejaculated 21 times or more a month had a lower risk of prc	0	
11	10	I got REALLY angry today and it wasn't about nothing, but you're goir	0	

子任务1：幽默文本分类

- 二分类
- 评测指标：正确率
- 测试集介绍：test_classification.csv
  - id - 整数，文本的唯一id（8001-9000）
  - text - 字符串，文本内容
- 测试集标签（is_humor）不公布
- 需要对 is_humor 进行预测，输出文件格式见 sampleSubmission_classification.csv

	A	B
1	id	text
2	8001	What's the difference between a Bernie Sanders supp
3	8002	Vodka, whisky, tequila. I'm calling the shots.
4	8003	French people don't masturbate They Jacque off
5	8004	A lot of Suicide bombers are Muslims - I don't blame
6	8005	What happens when you fingerbang a gypsy on her

	A	B
1	id	is_humor
2	8001	1
3	8002	1
4	8003	1
5	8004	1
6	8005	1

子任务2：幽默文本评分

- 回归
- 评测指标：均方根误差

$$RMSE(X,h) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(h(x^{(i)}) - y^{(i)})^2}$$

- 测试集介绍：test_regression.csv
  - id - 整数，幽默文本的唯一id（9001-10000，不连续）
  - text -字符串，文本内容

	A	B
1	id	text
2	9001	Finding out your ex got fat is like fi
3	9003	A girl runs up to her mother with a
4	9004	gotta wonder if baseball still would
5	9005	When you're dreading getting in th

- 测试集标签（humor_rating）不公布
- 需要对 humor_rating 进行预测，输出文件格式见 sampleSubmission_regression.csv

	A	B
1	id	humor_rati
2	9001	2.5
3	9003	2.5
4	9004	2.5
5	9005	2.5

https://blog.csdn.net/capecape/article/details/78623897

组队

- 在Kaggle的两个Task上完成组队操作
  - 每人注册自己的账号，加入竞赛，提交sampleSubmission，然后Merge
  - 详细说明：https://www.kaggle.com/docs/competitions#forming-a-team

- Kaggle队伍名称：队号-学号1-学号2
  - 例如：01-19881234-19885678
  - 例如：02-19993456
  - 队号已在Q群公布

- 建议提前作出合理分工

目录

- 项目任务
- **项目知识梳理**
  - **文本数据处理进阶**
  - 分类与回归算法
- 项目要求

文本数据预处理

- 脏数据清洗
- 分词（tokenize）
- 停用词（stop word）去除
  - 如‘i’, ‘me’, ‘my’, ‘there’, ‘when’, ‘where’等等
- 词干化（stemming）
- 低频词去除
- …

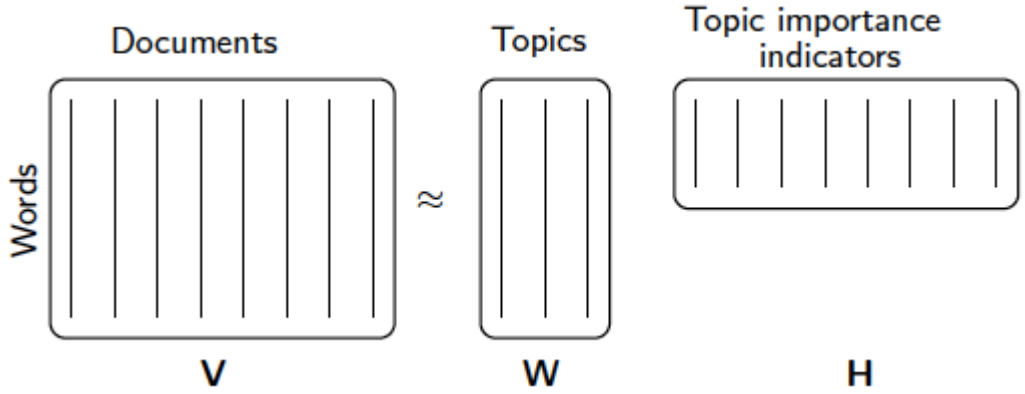
降维

- 数据维度大 → 模型的复杂度高 → 训练时间长，计算量大
  - 文档数量为N，词表长度为V，则文档集的矩阵表示是N*V维的。
  - V可能很大，故需要降维
- 维数灾难：数据稀疏化[1]
  - 高维空间下距离度量失效[2]
- 降维方法
  - 通用方法：PCA (Principal Component Analysis)、LDA (Linear Discriminant Analysis)
  - 单词表示：词嵌入，如Word2Vec、Glove
  - **文档表示**：主题模型、文档嵌入

1. https://www.zhihu.com/question/27836140
2. https://stats.stackexchange.com/questions/99171/why-is-euclidean-distance-not-a-good-metric-in-high-dimensions

文本降维：主题模型

- Latent Semantic Indexing (LSI)
  - 是奇异值分解（SVD）在文本数据（单词-文档矩阵）中的应用
  - https://www.cnblogs.com/pinard/p/6805861.html
- Non-negative Matrix Factorization (NMF)
  - https://www.cnblogs.com/pinard/p/6812011.html
  - https://blog.csdn.net/winycg/article/details/83005881



- Latent Dirichlet Allocation (LDA)
  - https://www.cnblogs.com/pinard/p/6831308.html （了解即可）

文本表示：总结

- 单词表示
  - One-hot向量
  - 降维
    - 词嵌入：Word2vec、Glove等（如何得到文档表示？）
- 文档表示
  - Bag-of-Words向量与TF-IDF向量
  - 降维
    - 主题模型：LSI、NMF、LDA等
    - 文档嵌入：Doc2vec等

允许使用的高级库和工具

- sklearn.decomposition[1]
  - TruncatedSVD
  - NMF
  - LatentDirichletAllocation
  - PCA
- sklearn.discriminant_analysis[2]
  - LinearDiscriminantAnalysis
- gensim[3]
  - models.word2vec
  - models.doc2vec
- Glove[4]

1. https://scikit-learn.org/stable/modules/classes.html#module-sklearn.decomposition
2. https://scikit-learn.org/stable/modules/classes.html#module-sklearn.discriminant_analysis
3. https://radimrehurek.com/gensim/apiref.html
4. https://nlp.stanford.edu/projects/glove/

目录

- 项目任务
- **项目知识梳理**
  - 文本数据处理进阶
  - **分类与回归算法**
- 项目要求

学过的基本算法

分类

- *k*-NN
- Decision Tree
- Logistic Regression
- PLA

回归

- *k*-NN
- Decision Tree
- Linear Regression

如何针对本次任务，在基本算法的基础上进行改进与优化？

可尝试的基本算法

- 贝叶斯学派
  - Naïve Bayes
  - …
- 统计机器学习
  - 分类：SVM / Kernel SVM
  - 回归：SVR / Kernel SVR
  - …
- …

挑战性算法

- Ensemble Learning
- Feedforward Neural Network

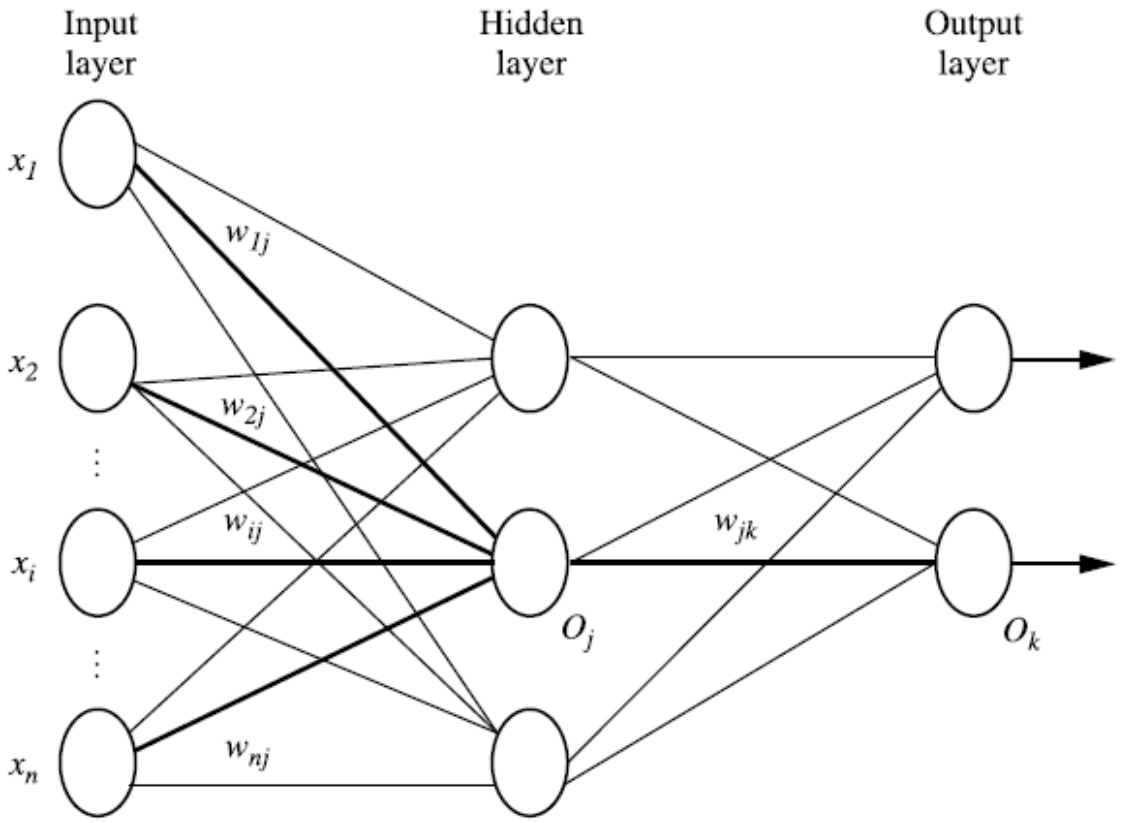
Ensemble Learning

- 在各种数据竞赛中，Random Forest、GBDT、XGBoost、LightGBM等经典的集成学习模型或工具库，常常能取得优秀的表现

- 然而，本次项目中大家如果要使用，则需要在不调用高级库的前提下，手动实现集成学习方法
  - Bagging & Random Forest
  - Boosting & Adaboost
  - …

Feedforward Neural Network

- 不得调用高级库
  - 例如：若使用Python，可以并建议基于numpy/scipy等实现，但不得调用tensorflow、pytorch或sklearn等。
- 可以考虑如下尝试：
  - 不同的激活函数
    - Sigmoid
    - Tanh
    - ReLU
    - LeakyReLU
  - 不同的层数/层内神经元数
  - L2正则化
  - Mini-batch梯度下降及变种
  - …



目录

项目评分

- Leaderboard（40%）
  - 按队伍算分，即同一个队伍所有成员相同
- 验收（20%）
  - 以队伍为单位，但按个人算分
- 小组项目报告（40%）
  - 每个队伍共同完成一份报告

项目时间节点与说明

- **11月10日23:00：Kaggle结果提交截止时间**
  - 在此之前，每个队伍每天（UTC时间0点，即北京时间早上8点为界）可在Kaggle上提交5次测试集预测结果。Kaggle将计算其中一部分测试集（约30%）上的指标分值，并以该分值参与”Public Leaderboard”排名。
  - 截止后，Kaggle将对各个队伍的结果（截止前自选5个，否则Kaggle自动选取Public分值最高的5个）在另一部分测试集（约70%）上计算指标分值，并以该分值得到最终的”Private Leaderboard”排名。
  - 我们将综合Public Leaderboard与Private Leaderboard的分值与排名情况（更重视Private Leaderboard，因此要小心过拟合），得到Leaderboard部分的实验分数。
- 11月11日实验课：验收
- 11月17日23:00：超算习堂代码和报告提交截止时间

项目时间节点与说明

- 11月10日23:00：Kaggle结果提交截止时间
- **11月11日实验课：验收**
  - 以队伍为单位，对其中每个成员进行验收
  - 考察每个成员的工作量及在队伍内的贡献等
- 11月17日23:00：超算习堂代码和报告提交截止时间

项目时间节点与说明

- 11月10日23:00：Kaggle结果提交截止时间
- 11月11日实验课：验收
- **11月17日23:00：超算习堂代码和报告提交截止时间**
  - 提交一个压缩包，命名为：队号-学号1_姓名拼音1-学号2_姓名拼音2.zip
    - 或：队号-学号1_姓名拼音1
  - 例如：01-19881234_wangxiaoming-19885678_lixiaohong.zip。其中包含：
    - 01-19881234_wangxiaoming-19885678_lixiaohong.pdf：项目报告
    - /result：分别存放分类任务和回归任务在Private Leaderboard上最好的实验结果
      - 命名为submission_classification.csv和submission_regression.csv
    - /code：存放实验代码，并附上readme
      - readme内要讲清楚如何运行代码来复现出两个任务最优的Private结果
  - 两位成员均需在超算习堂上提交一样的压缩包
    - 作为确认机制，如有特殊情况请在提交的压缩包内附上文字说明

项目时间节点与说明

- 11月10日23:00：Kaggle结果提交截止时间
- 11月11日实验课：验收
- **11月17日23:00：超算习堂代码和报告提交截止时间**

报告项	说明	分值
概述	对本队伍完成的工作进行介绍性的简要概述	
实验原理	对于所尝试方法中使用到的主要模型与算法，介绍其原理	20
方法	有条理地介绍所使用的一些主要方法。应包含整套方法的流程图，并对流程图中的各个模块详细展开介绍，并贴上少量关键代码	40 +5
实验结果与分析	展示并对比分析所有尝试的实验结果（除了指标外，还应从其它角度分析，例如运行速度、过拟合、输出结果的稳定性等）	40 +5
总结	对本队伍完成的工作内容与成果进行总结	
参考资料	参考的资料，如文献、博客、网上资源等，注意引用规范	
分工	总结成员各自做了什么工作，包括代码实现和实验报告的具体分工	

一些Leaderboard的建议

- 很多时候，简单的模型调出一个好的参数，跑出来的结果不比那些看起来高大上的算法的结果差。
- 觉得自己调参不好，提交各种版本，总会有表现好的时候，不要浪费每天5次的机会。但最后选取版本时，要小心过拟合。
- 因此，比起在Kaggle上每天查看5个版本的表现，要学会自己线下通过验证集来判断自己模型的表现，这样不仅更有效率，更方便，更自由，也有利于避免过拟合。
- 记得控制和保存版本和结果（Kaggle提交版本时记录参数和设置）。

整体工作量建议

- 如果验收与报告部分想拿到较高分数，每个队伍应做到1~2条：
  - 实现前馈神经网络及其BP更新算法
  - 实现1种集成学习方法
  - 学习并实现1~2种理论课上没学过的基本算法
  - 不调库实现1~2种文本降维表示方法
  - 深度钻研并对任一算法做出了较好的有亮点的创新与优化
    - 鼓励，可以是学过的基本算法，也可以是没学过的基本/挑战性算法
- 上述内容在写报告时，最好进行全方位深入对比分析
- 并未实现但是调研过的相关方法，也鼓励写上自己的理解及未能实现的原因

其它注意事项与要求

- 请不要作弊或者上网搜原竞赛及数据集，每个数据集会有一个state-of-the-art基准线，如果你超过了目前世界上最好的效果，在验收时会让你详细解释模型如何设计，并要求提供完整数据和代码。与其搜索答案胆战心惊故意搞错部分答案骗取Leaderboard排名，不如认真实现与训练模型。

- 队伍内积极展开思路代码等全方位的讨论合作和分享，但队伍间应互相拒绝关于思路与代码等的分享和抄袭，否则两队一律0分。

- 请仔细阅读Kaggle竞赛页面中的说明。

期待看到大家的方法

Q & A