Lab 5: Datasets for Regression and Classification

1. Objective

Know how to write a machine learning/artificial intelligence report using larger datasets.

2. Datasets

给定一篇新闻的 58 个属性（特征），比如新闻标题中词的个数（n_tokens_title）等等，预测这篇新闻在社交网络中被分享的次数（shares）。被分享的次数越多，表示这篇新闻越“热门/受欢迎”。如下数据文件供回归实验采用：

Datar_all.csv（全部数据，每行代表一篇新闻，每篇新闻都给出了 shares 的真实值）；

Datar_train.csv（训练数据，27751 篇新闻，每篇新闻都给出了 shares 的真实值）；

Datar_test.csv（测试数据，11893 篇新闻，每篇新闻的 shares 的真实值需要预测）。

输入的属性（特征）：数据文件中的第 1-58 列。说明如下：

n_tokens_title	新闻标题中词的个数
n_tokens_content	新闻正文中词的个数
n_unique_tokens	新闻正文中不重复的词占总词数的比例
n_non_stop_words	新闻正文中非停用词占总词数的比例
n_non_stop_unique_tokens	新闻正文中不重复的非停用词占总词数的比例
num_hrefs	新闻正文中超链接的个数
num_self_hrefs	新闻正文中指向 Mashable 其它新闻的超链接数
num_imgs	新闻正文中图片的张数
num_videos	新闻正文中视频的个数
average_token_length	新闻正文中词的平均长度
num_keywords	元数据中关键词的个数
data_channel_is_lifestyle	新闻是否放置在“Lifestyle”频道下？
data_channel_is_entertainment	新闻是否放置在“Entertainment”频道下？
data_channel_is_bus	新闻是否放置在“Business”频道下？
data_channel_is_socmed	新闻是否放置在“Social Media”频道下？
data_channel_is_tech	新闻是否放置在“Tech”频道下？
data_channel_is_world	新闻是否放置在“World”频道下？
kw_min_min	差的关键词最少分享数
kw_max_min	差的关键词最多分享数
kw_avg_min	差的关键词平均分享数
kw_min_max	好的关键词最少分享数
kw_max_max	好的关键词最多分享数
kw_avg_max	好的关键词平均分享数
kw_min_avg	一般关键词最少分享数
kw_max_avg	一般关键词最多分享数
kw_avg_avg	一般关键词平均分享数
self_reference_min_shares	参考新闻中的最少分享数

self_reference_max_shares	参考新闻中的最多分享数
self_reference_avg_sharess	参考新闻中的平均分享数
weekday_is_monday	新闻是否在周一发布的？
weekday_is_tuesday	新闻是否在周二发布的？
weekday_is_wednesday	新闻是否在周三发布的？
weekday_is_thursday	新闻是否在周四发布的？
weekday_is_friday	新闻是否在周五发布的？
weekday_is_saturday	新闻是否在周六发布的？
weekday_is_sunday	新闻是否在周日发布的？
is_weekend	新闻是否在周末发布的？
LDA_00	与 LDA（一种主题模型）主题 0 的相似度
LDA_01	与 LDA 主题 1 的相似度
LDA_02	与 LDA 主题 2 的相似度
LDA_03	与 LDA 主题 3 的相似度
LDA_04	与 LDA 主题 4 的相似度
global_subjectivity	新闻正文的主观性
global_sentiment_polarity	新闻正文的情感强度
global_rate_positive_words	新闻正文中正面词占全部词的比例
global_rate_negative_words	新闻正文中负面词占全部词的比例
rate_positive_words	新闻正文中正面词占非中性词的比例
rate_negative_words	新闻正文中负面词占非中性词的比例
avg_positive_polarity	新闻正文中正面词的平均情感强度
min_positive_polarity	新闻正文中正面词的最小情感强度
max_positive_polarity	新闻正文中正面词的最大情感强度
avg_negative_polarity	新闻正文中负面词的平均情感强度
min_negative_polarity	新闻正文中负面词的最小情感强度
max_negative_polarity	新闻正文中负面词的最大情感强度
title_subjectivity	新闻标题的主观性
title_sentiment_polarity	新闻标题的情感强度
abs_title_subjectivity	绝对主观性水平
abs_title_sentiment_polarity	绝对情感强度水平

预测的变量：shares（测试数据文件中的最后一列）。

如下数据供分类实验采用：

Datac_all.csv（全部数据，每行代表一篇新闻，每篇新闻都给出了 shares 的真实值）；
Datac_train.csv（训练数据，27751 篇新闻，每篇新闻都给出了 shares 的真实值）；
Datac_test.csv（测试数据，11893 篇新闻，每篇新闻的 shares 的真实值需要预测）。

预测的变量：shares（测试数据文件中的最后一列）。
其中，“1”代表 popular（即该篇新闻为“热门/受欢迎”的），“0”代表 unpopular。

3. Experiment

无论是选择第一种实验模式（代码实现类），还是第二种实验模式（工具或代码应用类）的同学，请基于上述数据集开展实验。每位同学可以选择回归或分类两种任务中的任意一种（或者两种）获得实验结果并撰写实验报告。

假如选择了回归任务，请采用 Datar_all.csv、Datar_train.csv、Datar_test.csv 开展实验，通过 Datar_train.csv 训练模型，然后预测 Datar_test.csv 中每篇新闻的 shares 值，接着将预测值粘贴到 Evaluater.xlsx 文件的第二列，将得到模型预测值与真实值的相关系数值（D3 单元格）。可以采用的回归模型如下：

（1）kNN 回归：k 值请固定取为 1。**注意不能**根据不同 k 值的 kNN 回归模型，看它们在测试数据集上的相关系数大小，然后取最大值。但是，如果同学提出了从训练数据分布中自动确定 k 值大小的方法，可以提交自动确定的 k 值得到的相关系数值。

（2）NB 回归：请参照实验二及 TA 的讲稿。

（3）其它自行实现的或者工具中包含的回归模型。

请将你目前能够得到的最好的一个相关系数值，比如 0.80，写入“学号.txt”文件，并上传到 FTP 的 Regression results 目录中。实验报告文件，即“学号.doc”或“学号.pdf”，请上传到 FTP 的 Regression reports 目录中。

假如选择了分类任务，请采用 Datac_all.csv、Datac_train.csv、Datac_test.csv 开展实验，通过 Datac_train.csv 训练模型，然后预测 Datac_test.csv 中每篇新闻的 shares 值，接着将预测值粘贴到 Evaluatec.xlsx 文件的第二列，将得到模型预测的准确率（D3 单元格）。可以采用的分类模型如下：

（1）kNN 分类：k 值请固定取为 1。**注意不能**根据不同 k 值的 kNN 分类模型，看它们在测试数据集上的准确率大小，然后取最大值。但是如果同学提出了从训练数据分布中自动确定 k 值大小的方法，可以提交自动确定的 k 值得到的准确率值。

（2）NB 分类：请参照课件 Lec 4_kNN and NB.pdf。

（3）决策树分类：请参照实验四及 TA 的讲稿。

（4）其它自行实现的或者工具中包含的回归模型。

请将你目前能够得到的最好的一个准确率，比如 0.80，写入“学号.txt”文件，并上传到 FTP 的 Classification results 目录中。实验报告文件，即“学号.doc”或“学号.pdf”，请上传到 FTP 的 Classification reports 目录中。

【注】实验模式选择的 deadline 为 11 月 15 日（本周日）晚上 10:00，请尚未选择的或需要修改的联系朱和胜。自 11 月 12 日起，同学们可以提交上述实验结果和实验报告文件至 FTP，我们会在 11 月 18 日（下周三）晚上 11:00 从 FTP 中剪切已提交的文件作为参考。后续将会每周三晚上 11:00 从 FTP 中剪切已提交的文件作为参考，一直延续到学期结束。