华中科技大学计算机科学与技术学院

《机器学习》课堂二结课报告

专　　业：计算机科学与技术

班　　级：校交 1902

学　　号：U201912633

姓　　名：张睿

成　　绩：

指导教师：黄宏

完成日期： 2021 年 11 月 27 日

目录

新闻数据情感分析

一、 实验题目：新闻数据情感分析

二、实验要求

2.1 题目背景

2020 年初， 新型冠状病毒所导致的肺炎疫情袭来. 全世界的人们都关注着这场疫情， 作为第一个爆发疫情的国家， 中国在疫情治理方面取得了卓效的成就， 无论是武汉疫情的处理， 还是在疫情进入常态化的管控措施， 都是做到了尽善尽美. 在平静生活的背后有着医务人员和政府官员的倾情奉献， 我们才能取得这么完美的成就. 因为我国取得了光彩照人的成就， 越来越多的国外媒体也开始关注我们中国的疫情管控措施， 不少人对我们的管控措施提出了赞扬，称赞“中国疫情取得了卓然的成效”， 也有媒体或是立场问题或是意识形态的问题， 对我国的疫情进行了批评， 批评道“这是对人权的亵渎”. 对于新闻的立场的研究显得尤为重要， 如果我们能够分析外国报道对于我国疫情的态度， 借此我们可以侦测国际舆情并对我国打好“疫情舆论战”做好准备.

2.2 数据集

数据集分为训练集和测试集两部分。

训练集：train.json 文件，每一条记录有 content 和 label 两个字段，label 只有 0 和 1 两个值，即为二分类问题

测试集：test.json 文件，每一条记录有一个字段，为 content 字段。

2.3 任务描述

预测测试集的 label，并将结果复制到 educoder 的编辑器中，格式为一行一个记录的标签，如下：

```
1. 1
2. 0
3. 1
4. 1
5. ...
6. ...
```

2.4 评测标准

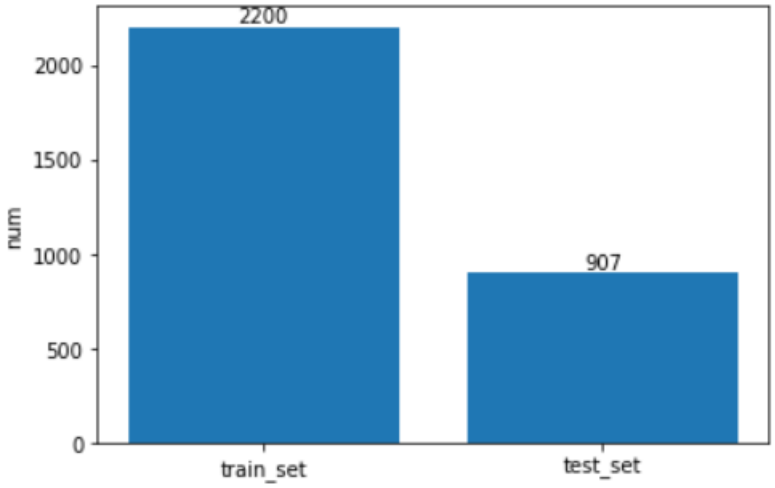
根据测试集结果的 micro-f1 和 macro-f1 进行评价。

三、数据处理

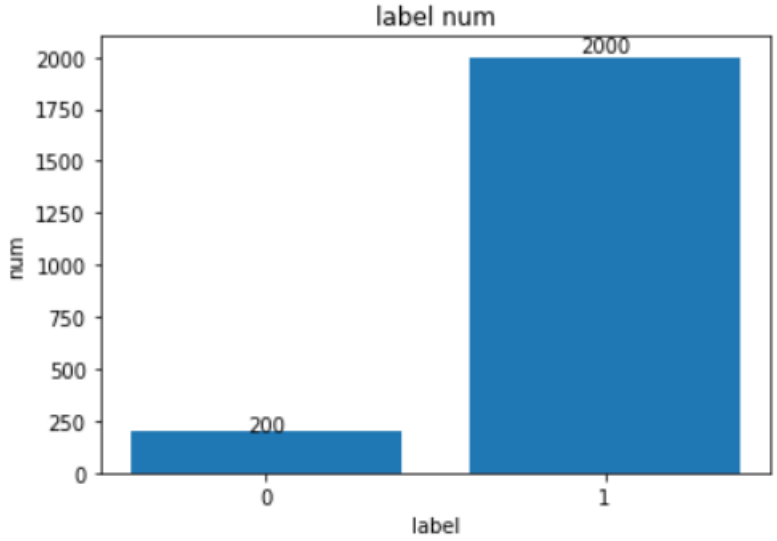
这部分内容的写作顺序按照实验顺序记录，即包括从拿到数据集到进行训练之间的一系列操作。

3.1 数据集观察

首先利用观察一下数据集的大小。



可以看到我们一共有 3107 条数据，其中训练集有 2200 条，而测试集有 907 条，数据集比较小，如果用较为复杂的模型，很有可能出现模型过拟合的情况。接着我们再观察一下训练集中 label 的分布情况。



可以看到，在训练集中，label 的分布极不均匀，label==0：label==1 为 1:10，在模型选择和设计上应该重复考虑其不平衡性。

通过观察数据集，我们得出以下结论：

1. 数据集比较小
2. 数据集标签分布很不均匀
3. 数据集中的文本语言不一致

3.2 翻译

通过观察数据集，我们知道数据集比较小且文本语言不一致，这启示我们可能需要对文本进行翻译，因为如果不翻译而是把每种语言单独考虑，由于数据集过小很可能出现某一种语言并没有足够的数据进行训练，甚至很有可能会出现某一种语言仅有一种标签，可以预想这将对结果产生极大的不利影响，对于这种问题，有两种解决的思路。

1. 设计可以支持无监督训练的模型，即尽管标签不存在也可以进行预测。
2. 将所有文本翻译成一种语言，统一到一个空间中。

对于第一种方案，有以下几个问题：

1. 就算支持无监督，但仍旧没有解决数据量过小的问题，无法使用标签传播等一系列方法，导致结果不够精确。
2. 无监督和有监督混在一起，设计难度较大，编程较难。

综合考虑，最终选择使用翻译解决文本语言不一致的问题，但这又会导致新的问题，即文本翻译的好坏也成了影响结果的因素之一，在尝试过多种文本翻译方法后，选择使用 Baidu 翻译的 API 接口对文本进行翻译。

Tips: 由于基础版账号限制，http 请求频率需要为每秒一个请求，并且由于网络不稳定，需要 cache 住已翻译的部分文本，并且在途中要经常 sleep 一下。

翻译结果对比：

翻译前：

```
La cifra de afectados se dispara:más 4.500 enfermos y 106 muertos. El número de infectados por elcoronavirus de Wuhanno deja de crecer y cada vez lo hace más rápido. Según los últimos datos ofrecidos por la Comisión Nacional de Sanidad de China, en las últimas 24 horas han fallecido 25 personas y el total de víctimas mortales alcanza los 106 en todo el país.El número de casos confirmados, por su parte, ya alcanza este martes los 4.515, el 59% más que los 2.835 ofrecidos este lunes. Entre los enfermos, 515 se encuentran en estado grave. El número de casos sospechosos roza los 7.000, mientras que se mantiene en observación a casi 45.000 personas.Para Santiago Moreno, jefe de servicio de enfermedades infecciosas del Hospital Ramón y Cajal de Madrid, estas cifras son "muy significativas" y demuestran que "el brote aún se encuentra en sus primeras fases". "La curva ascendente aún es muy marcada, lo que significa que falta mucho todavía para que se alcance el pico", sostiene Moreno, informa Oriol GÜell.Hong Kong, donde se han confirmado ocho casos, limitará drásticamente la admisión de visitantes que procedan de China a partir de la medianoche del jueves. En una rueda de prensa, la jefa del Gobierno hongkonés, Carrie Lam, cubierta con una máscara, ha anunciado que el territorio autónomo paralizará el servicio de tren de alta velocidad y los de transbordador que comunica el territorio autónomo con la China continental. También se reducirán a la mitad los vuelos entre Hong Kong y el resto de China, y se recortarán drásticamente los servicios de autobús.Pekín ha accedido, por su parte, a dejar de emitir permisos individuales de ciudadanos chinos para visitar Hong Kong. En noviembre, el último mes con datos disponibles, la excolonia británica recibió 1,92 millones de visitantes procedentes de China continental.Las medidas de Hong Kong llegan después de que un grupo de médicos de las principales universidades del territorio pidieran medidas más duras para la entrada de visitantes procedentes de China.El anuncio se ha producido tras los primeros casos de contagio de personas que no han estado en China. Después de que Vietnam confirmara el viernes pasado el primer caso, este martes lo han hecho Alemania y Japón, con un caso respectivamente. En el caso alemán se trata de un hombre que asistió a un curso de formación que impartió una empleada llegada de Shanghái. En el nipón, de un conductor de autobús de 60
```

翻译后：

```
the affected figure is triggered  plus 4 500 patients and 106 dead  the number of infected by wuhanno elcoronavirus stops growing and every time it makes it faster  according to the latest data offered by the chinese national health commission  25 people have died in the last 24 hours and the total number of fatalities reaches 106 throughout the country  the number of cases confirmed  on the other hand already reaches this tuesday 4 515  59  more than the 2 835 offered this monday  among the sick  515 are in a serious state  the number of cases suspected brushes the 7 000  while it is maintained in observation to almost 45 000 people  for santiago moreno  head of service of infectious diseases of ram n and cajal de madrid hospital  these figures are  very significant  and demonstrate that  the bud is still in its first phases     the ascending curve is still very marked  which means that it is still missing so that the peak is reached  says moreno  reports oriol g ell hong kong  where eight cases have been confirmed  will drastically limit the admission of visitors that come from china from midnight on thursday  at a press conference  the hongkonian government chief  carrie lam  covered with a mask  has announced that the autonomous territory will paralyze the high speed train and ferry service that communicates the autonomous territory with mainland china  flights between hong kong and the rest of china will also be reduced  and bus services will also be shredled pek n has accessed  on the other hand  to stop issuing individual permissions of chinese citizens to visit hong kong  in november  the last month with data available  the british excolonia received 1 92 million visitors from china continental  hong kong measures arrive after a group of doctors from the main universities of the territory requested higher measures for the entry of visitors from china the announcement has occurred after the first cases of contagion of people who have not been in china  after vietnam confirmed last friday the first case  germany and japan have been made on friday  with a case respectively  in the german case  it is a man who attended a training course that gave an arrival employee of shanghai  in the japanese  a 60 year old bus driver in the tourist city of nara in a
```

3.3 文本清洗

文本清洗主要包括以下几个方面：

1. 去除标点符号等非文本内容
2. 分词
3. 去除停用词
4. 词干还原

要进行文本清洗的原因十分简单明了，过多杂乱的符号和不同类型的词语会导致数据集中的噪声过大，并且标点符号等非文本表示，不利于表达和模型训练。

最开始，选择采用手动进行相关操作，但是后来发现在相关库种已经打包好相关函数，于是采用相关库种自带的函数进行文本清洗。

文本清洗结果：

```
the ascend curv is still veri mark which mean that it is still miss so that the peak is reach say moreno report oriol g ell hong kong where eight case have been confirm will drastic limit the admiss of visitor that come from china from midnight on thursday at a press confer the hongkonian govern chief carri lam cover with a mask ha announc that the autonom territori will paralyz the high speed train and ferri servic that commun the autonom territori with mainland china flight between hong kong and the rest of china will also be reduc and bu servic will also be shredl pek n ha access on the other hand to stop issu individu permiss of chines citizen to visit hong kong in novemb the last month with data avail the british excolonia receiv 1 92 million visitor from china continent hong kong measur arriv after a group of doctor from the main univers of the territori request higher measur for the entri of visitor from china the announc ha occur after the first case of contagion of peopl who have not been in china after vietnam confirm last friday the first case germani and japan have been made on friday with a case respect in the german case it is a man who attend a train cours that gave an arriv employe of shanghai in the japanes a 60 year old bu driver in the tourist citi of nara in a meet in beij with the chines foreign minist wang yi member of the work group in charg of coordin the respons to the epidem the gener director of the world health organ tedro adhanom ghebreyesu ha express it support for the measur that china ha taken to tackl the progress of the diseas as report by the chines ministri
```

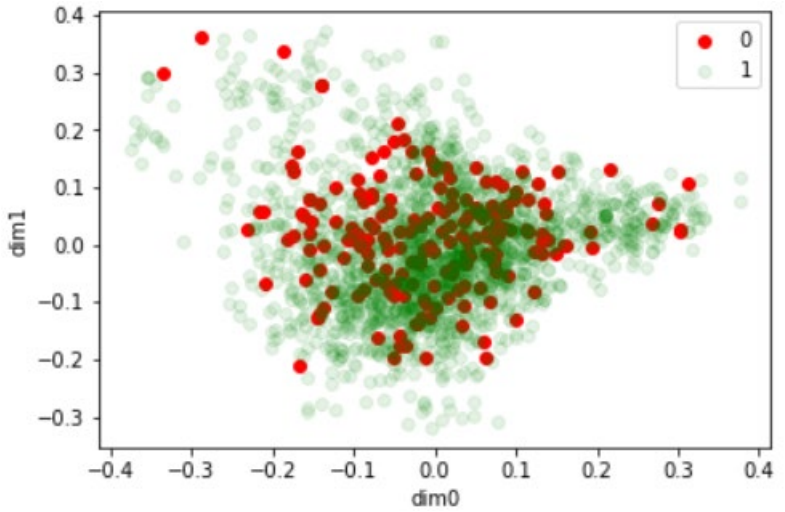
可以发现有一些不准确，主要是词干还原的问题，词干还原不够智能，把一些以 s 或者 e 等结尾的词错判为非词干，比如 has 被还原成 ha，显然这是不对的。于是选择不使用词干还原。

3.4 文本向量化

在本次实验中，选择以文本为粒度进行向量化，之所以选择这个以文本为粒度，是因为我们要对文本进行分类，这样更为直观，容易解释，当然也可以采用 word2vec 等词向量模型进行向量化，这里只叙述文本向量化的做法。
在实验中，选用了三种方法对文本进行向量化：

1. 词袋模型
2. TF-IDF 模型
3. Text2vec 模型

最终选取了 TF-IDF 作为向量化方法，结果参见结果说明部分。在向量化的过程中，需要对向量进行降维，以便降低噪声的影响，并且有利于发现潜在的关系，降维方法为 PCA，在此不过多叙述。
向量化结果：



对向量结果进行 2 维可视化分析，可以发现两种类型的数据混在了一起，选用线性分类器，有一定可能会导致效果较差，当然这也不是绝对的，毕竟降到了 2 维，数据如此分布也是可以理解的，有可能在别的维度两者可以被分开。

至此，已经准备好了所有数据，可以进行模型训练。

四、算法模型

根据之前的分析，有以下观察和结论：
1. 数据集较小---不适合使用较大规模的网络，并且大规模网络调参较为困难
2. 标签分布很不平衡
3. 有可能线性模型分类较差

所以选择使用较为简单的分类模型，并且要对标签分布有一定平衡能力。

模型选择了朴素贝叶斯模型、逻辑回归模型和 svm，其中 svm 采用高斯核函数作为这三个中的非线性分类器，并且选择加上 attention 机制对三者进行了一个集成。接下来将简要叙述四个模型的原理和算法。

4.1 朴素贝叶斯

根据贝叶斯公式

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)}$$

可以将后验概率的求解转化为关于先验概率的函数，在本例中，y 即为 label，x 为文本向量，由于$P(y)$可以直接由数据集中的 label 的概率分布得到，所以重点在于$P(x|y)$的求取，对此进行一个较为强大的假设。

$$P(\boldsymbol{x}|y) = \prod_{i=1}^{n} P(x_i|y)$$

即x_i的每个维度对于 y 来说都是独立同分布的，这个假设其实并不是十分正确的，因为其忽略了词语与词语之间的语义关系，但是这样求出来的结果还是比较令人满意的。
接着，便为对$P(x_i|y)$进行建模，由于这里的向量是由词频模型的出来的结果，所以这里采用多项式模型对其进行建模。

$$P(\mathbf{x} \mid m, y = c) = \frac{m!}{x_1! \cdot x_2! \cdot \cdots \cdot x_d!} \prod_{\alpha=1}^{d} (\theta_{\alpha c})^{x_\alpha},$$

我们要使得 $P(\boldsymbol{x}|y = c)$ 最大，即做极大似然估计：

$$P_\alpha = \frac{\sum_{i=1}^{n} I(y_i = c) x_{i\alpha}}{\sum_{i=1}^{n} I(y_i = c) m_i}$$

上式中 mi 代表第 i 篇文章所有的单词数，yi 代表第 i 篇文章的标签，x_{ia}表示第 i 篇文章，第 α 个单词的词频.$I(yi = c) = (yi == c)?1 : 0$，就是判断这个标签 yi 是

不是 c。注意，由于 sklearn 的多项式朴素贝叶斯继承的是贝叶斯基类，其可以接受连续的特征。

4.2 逻辑回归

和朴素贝叶斯不同，逻辑回归直接对后验概率进行建模

$$P(y|x_i) = \frac{1}{1 + e^{-y(\mathbf{w}^T x_i + b)}}$$

这使其拥有更加强的适应能力，因为其不再事先假设$P(x|y)$具有某种分布，事实上，逻辑回归本质即为高斯分布的朴素贝叶斯。并且，在 sklearn 中，逻辑回归的函数中可以对每类标签所对应的数据自动计算其权重大小，以平衡由于标签分布不均带来的精度损失，刚好和我们的数据集的情况相吻合。

4.3 SVM

本次实验中，根据分析，有可能数据非线性可分，于是采用 SVM 加上非线性核函数来当作一个非线性分类器，观察其效果，SVM 的原理如下：
SVM 首先就是通过坐标的伸缩变换使得离超平面最近的一个点的距离为 1, 即 $min|xi \cdot w+b| = 1$, 放缩后的超平面称为规范超平面, 对于规范超平面, 任何一个点到它的距离为

$$\frac{|w \cdot x + b|}{||w||^2}$$

对于离之最近的一个点, 已知其距离为 1, 那么问题就转变成求$||w||^2$ 的最小值了, 但是这个最小值需要满足松弛条件.即问题转化为一个优化问题

$$\begin{cases} \min \frac{1}{2}||w||^2 \\ s.t. y_i|\mathbf{w} \cdot \mathbf{x} + \mathbf{b}| \geq 1 \end{cases}$$

对这个优化问题的求解可以采用对偶问题求解，即为

$$g(\alpha) = \operatorname{argmin} \sum_{i=1}^{m} \alpha_i - \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (\mathbf{x_i} \cdot \mathbf{x_j})$$

详细证明略。
在数据线性不可分时, 无法找到参数(w, b) 使得 $w \cdot xi$ 与 yi 的符号一致, 也就是说, 对任意超平面 $w \cdot x + b = 0$, 存在 xi, 使得 $yi[w \cdot xi] < 1$. 为此, 我们引入松弛变量 $\zeta i \geq 0$, 使得 $yi[w \cdot xi] < 1 - \zeta i$.并且还可以加入核函数，将数据投影到高维空

间中，使得其线性可分，在加入核函数后，优化问题即变为：

$$\begin{cases} \min \frac{1}{2}||w||^2 + C\sum_{i=0}^{m}\zeta_i^2 \\ s.t. y_i K(|\mathbf{w}\cdot\mathbf{x}+\mathbf{b}|) \geq \mathbf{1}-\zeta_\mathbf{i} \end{cases}$$

4.4 模型集成

在之前多种模型的预测后，可以将多个模型的结果统一起来，来达到的更好的效果，于是问题转化为如何对结果进行集成，这里借鉴深度学习的思想，对每个模型施加权重，利用结果计算 mean square loss。

$$L = w_1 L_{NB} + w_2 L_{LogReg} + w_3 L_{svc} \qquad s.t.\ \Sigma_i w_i = 1$$

再利用梯度下降，对权重进行更新。

$$w_1 = w_1 - \eta\frac{\partial L}{\partial w_1}$$

$$w_2 = w_2 - \eta\frac{\partial L}{\partial w_2}$$

$$w_3 = w_3 - \eta\frac{\partial L}{\partial w_3}$$

最后将权值作用在预测结果上，以 0.5 作为阈值，将标签规约到 0 和 1。

$$y = w_1 y_1 + w_2 y_2 + w_3 y_3$$

$$result = \begin{cases} 0 & y \leq 0.5 \\ 1 & y > 0.5 \end{cases}$$

五、实验与结果分析

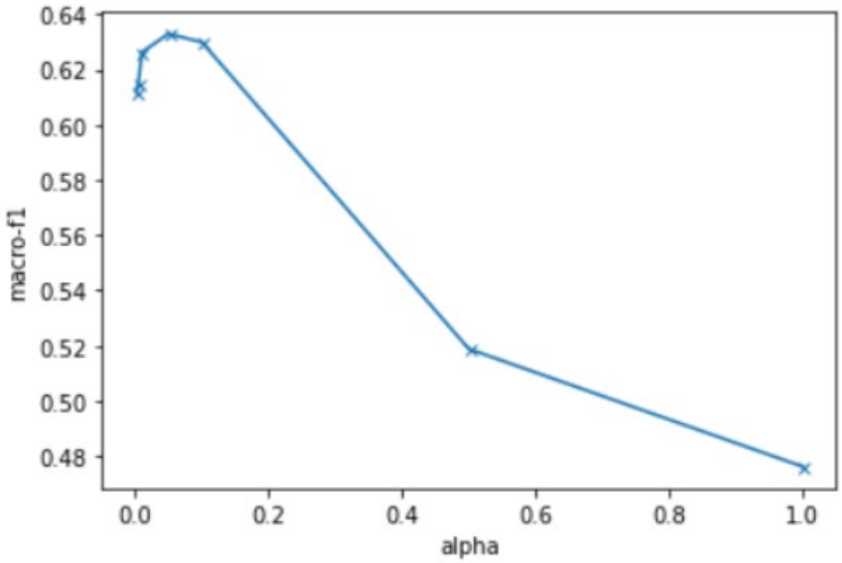
5.1 朴素贝叶斯实验

1. 实验设置

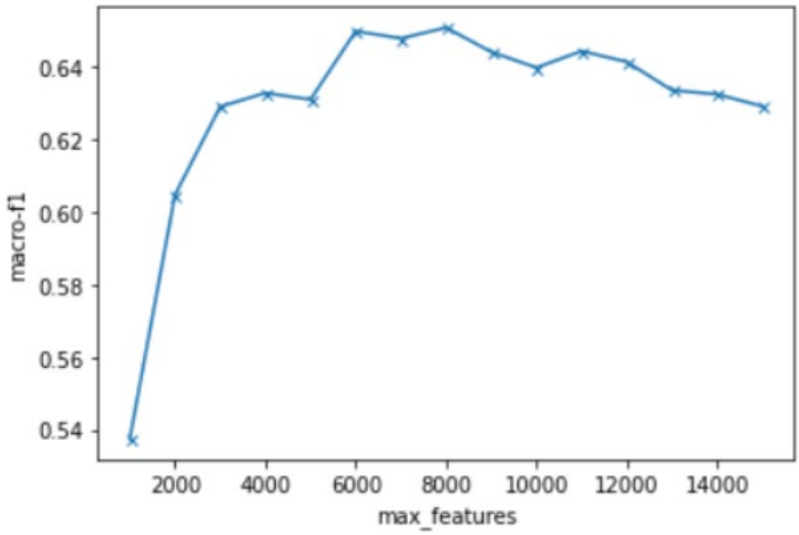
利用管道和网格搜索，对各项参数进行评估，并利用交叉验证对结果进行评估并记录。在实验中，K-fold 的值为 5，即每次取 25%作为测试集进行验证。

2. 参数实验

在朴素贝叶斯的实验中，参数主要由 alpha 和 max_features，即平滑系数和向量维度。首先看平滑系数的参数实验。

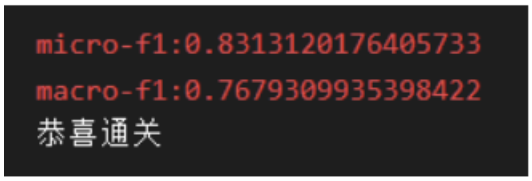


可以看到，随着 alpha 的上升，结果呈现先上升而后迅速下降的趋势，说明平滑不是越大越好，这个主要是由于本身文本向量的值较小，若是平滑系数过大，则平滑主导了结果，必然结果不好。
接着，对维度大小实验。



可以看到，总体来说，随着维度增大，macro-f1 值呈现出先迅速增长，再而后缓慢下降的趋势，这也比较好理解，当维度过低的时候，特征维度较少，难以反映文章的特征，而当维度太大，会导致噪声增加，最终导致结果降低。

3. 最终结果

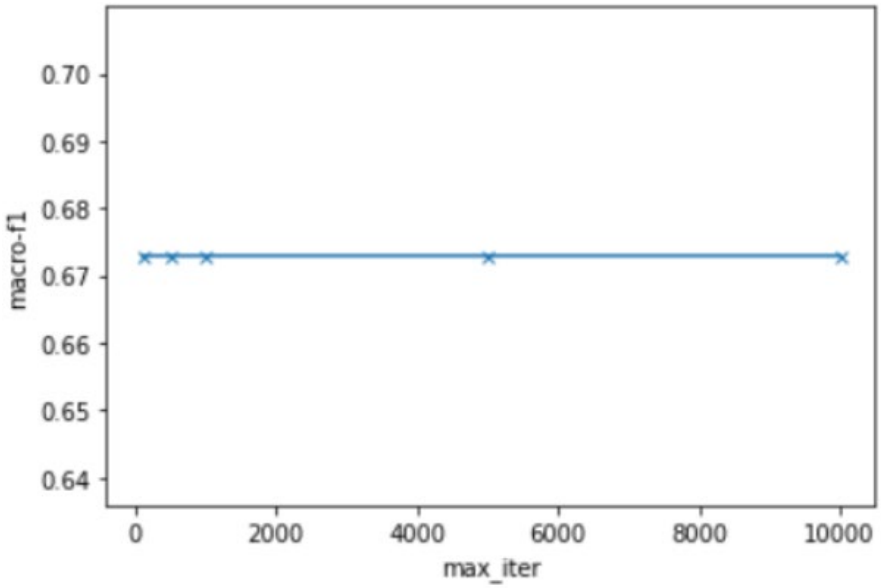


可以发现，朴素贝叶斯的结果非常之好，根据朴素贝叶斯的原理，朴素贝叶斯是一个线性分类器，这是否在一定程度上说明了数据集是线性可分的？

5.2 逻辑回归实验

1. 收敛情况

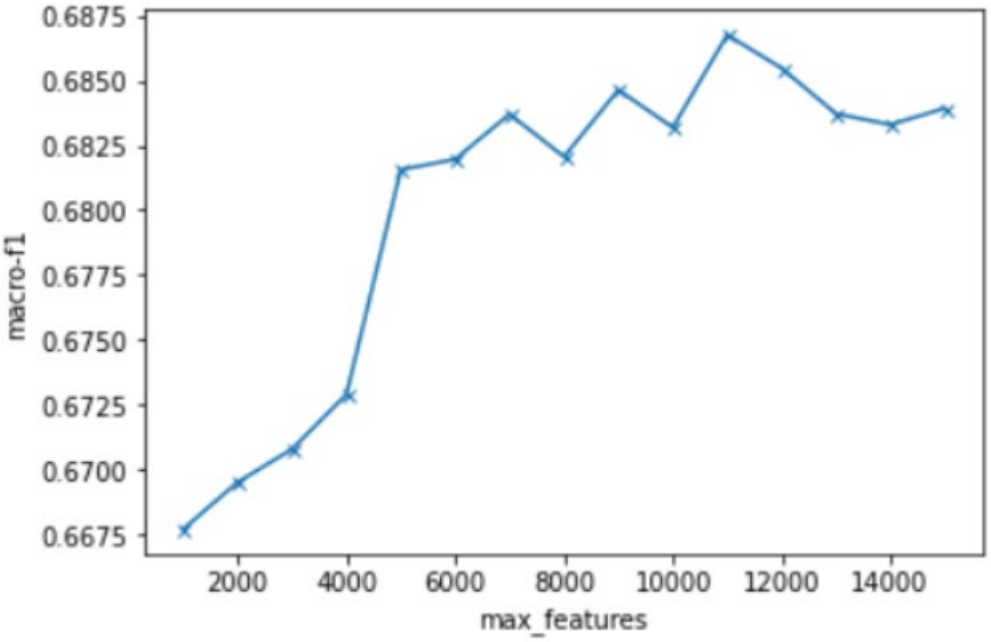
首先看一下逻辑回归的收敛情况



可以发现，逻辑回归很快就收敛了，并且一直没有过拟合的现象出现，推测逻辑回归在一定程度上具有比较强的泛化能力。

2. 向量维度

接着分析向量维度对逻辑回归的影响



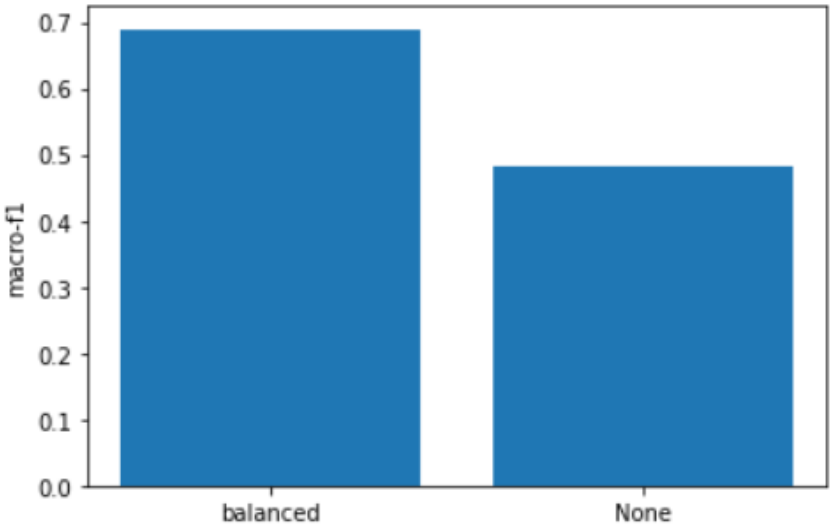
可以发现，其趋势和朴素贝叶斯很像，这说明之前的解释较大可能成立，另一方

面，其和朴素贝叶斯不同的是，朴素贝叶斯在大约 8000 左右的维度后，效果开始缓慢下降，而逻辑回归则有继续上升的趋势，推测有以下两个原因：

1. 逻辑回归并不对数据分布进行假设而是直接对后验概率进行建模，需要更多的维度。
2. 逻辑回归更能够抵抗或筛选噪声，从而在噪声较多的情况下，也能发现其中蕴藏的信息。

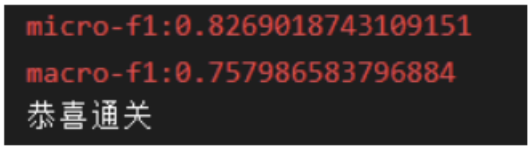
3. balanced 影响

接着，看一下“balanced”的影响



可以发现，增加“balanced”之后，结果有非常显著的提高，这说明，标签几不均衡对结果有非常大的影响，这也验证了之前的猜测。

4.最后结果：



可以发现，逻辑回归的效果似乎并不如朴素贝叶斯，这就比较耐人寻味了，按理来说，逻辑回归的结果应该是好于朴素贝叶斯的，但是在我们的训练集上，其实逻辑回归的效果是要好于朴素贝叶斯的，但是在测试集上却并不是这样，很自然的想到这是过拟合，但是之前才分析过逻辑回归应该比朴素贝叶斯的泛化能力更强，推测原因如下：

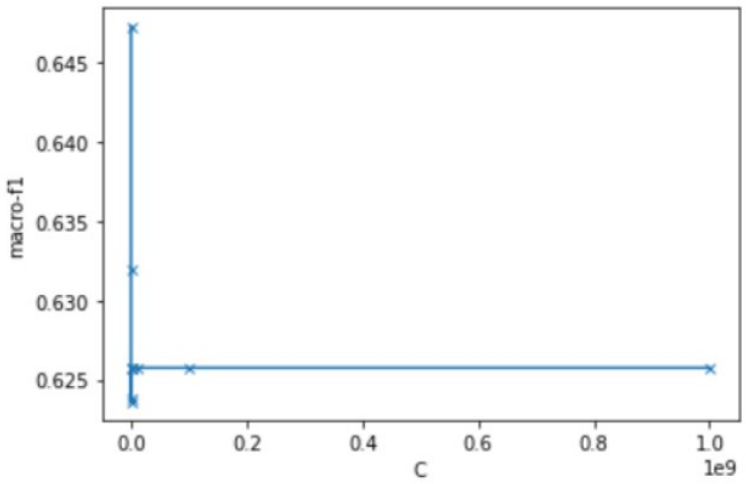
数据集较小，导致没有足够的数据为逻辑回归来拟合模型，而朴素贝叶斯默认了数据的模型，相当于确定了一个特征空间，而刚好这个空间和真正的空间比较吻合，或者说默认的数据分布和真实分布相近，使得结果较好。

5.3 SVM 实验

1. 参数

SVM 主要为惩罚的系数和核函数的权重两个参数，由于 sklearn 对惩罚系数已有自适应的机制，主要调的参为惩罚系数。

惩罚系数参数结果如下



可以发现，随着惩罚系数的增大，SVM 的效果匀速下降并平稳。

2. 最终结果

```
micro-f1:0.8257993384785006
macro-f1:0.7560433349675875
恭喜通关
```

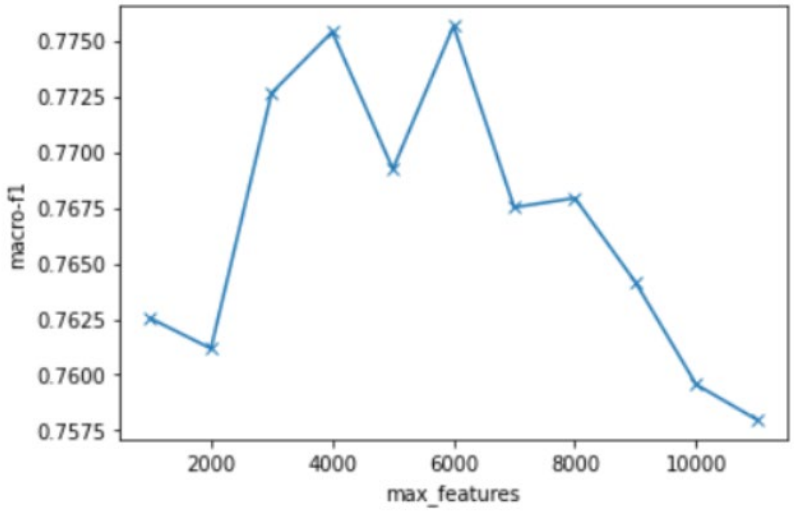
可以发现，非线性分类器效果并没有朴素贝叶斯这种线性分类器好，但是也有一定的效果。

5.4 集成实验

由于数据集中 label 为 0 的数据过少，经过实验，决定采用将三者的权重固定为 1，并进行投票，只要有一个模型认为其为 0，则该数据为 0。

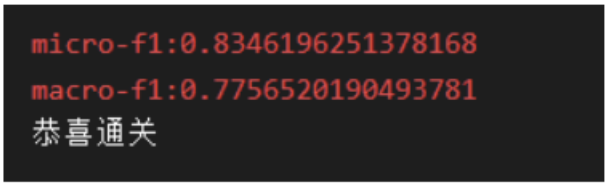
1. 参数实验

这里主要对特征维度进行实验。



可以发现，总的来说仍旧遵循着上文所说的标准，注意到在 5000 维的时候，效果出现一个波谷，经检查，并没有发现代码有错误，只好将其归因于偶然事件。

2.最终结果



结果确实有所改善，提高了大概一个百分点左右，这说明采用模型集成的方法确实能对结果带来一定的提升。

六、想法

总的来说，这次机器学习大作业还是令人受益匪浅的，但是由于时间的限制，而且到了考试周，也有一些做的不足的地方，算是有一些遗憾：

1. 没有尝试更加先进的模型，本次实验所用的模型已经是很多年前的产物了，本来还想试一下 text2vec 和 bert 等比较牛的模型，一方面确实是时间不够，另一方面也考虑到网络的复杂性和本次实验数据的数据量过小而且噪声过多。
2. 非常想自己去实现一个算法，或者在已有的算法上进行本质上的改进而并不是简单的调参。
3. 在实验中对结果做了很多推测，但其实并没有证明，不过确实也比较难去证明。

希望自己今后有时间，能继续深入，探索数据的奥秘。