基于机器学习的辩论激烈程度分类 设计报告

小组成员及分工

*** (2017******): 负责第一部分。
RainEggplant (2017******): 负责第二部分。
** (2016******): 负责第三部分。

工作开展及研究情况

一、基于给定图像特征的逻辑回归分类

原理

逻辑回归是用于处理因变量为分类变量的回归问题，本小题属于二分类问题，即可能的结果只有两种。该分类算法的核心部分为最小化损失函数来逼近最优解，具体算法如下：

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} Cost\left(h_\theta(x^{(t)}), y^{(t)}\right)$$

$$Cost(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$

$$h_\theta(x)) = g(\theta^T x) \qquad g(z) = \frac{1}{1 + e^{-z}}$$

我们的目标就是通过逐步减小损失函数 $J(\theta)$ 来得到最优的参数 θ。

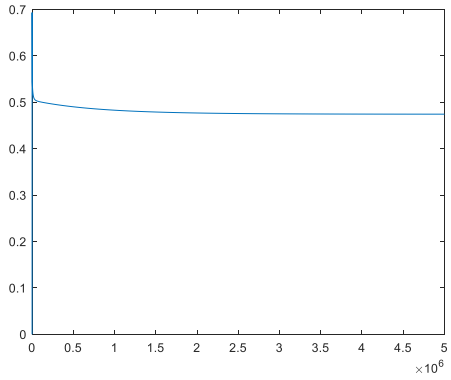
优化方法

本题中我分别采用了梯度下降法和牛顿法来逼近最优解，这两种算法的主要区别在于前者只利用了泰勒展开里的一阶项，而后者则是利用了一阶项和二阶项。

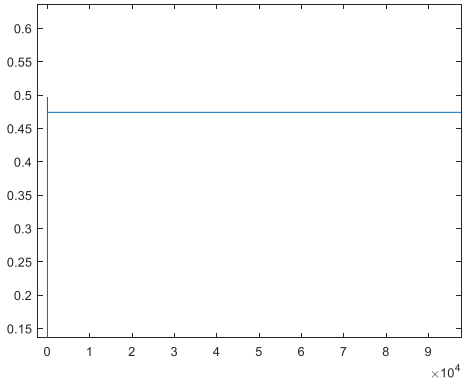
训练结果及分析：

训练集大小为200组数据，其中100组为“激烈”，另外100组为不激烈。

使用梯度下降法训练，训练步数为5000000步，步长为0.001，训练得到的参数保存在LR_first_order.mat中。将训练结果应用到训练集上得到的正确率为79.5%，另外得到训练过程中的目标损失值随着训练次数的关系函数如下：



使用牛顿法训练，训练步数为100000步，训练得到的参数保存在LR_second_order.mat中。将训练结果应用到训练集上得到的正确率为79%，另外得到训练过程中的目标损失值随着训练次数的关系函数如下：



比较两种优化方法后发现，显然牛顿法趋近最优解的速度比梯度下降法快很多，这也与原理分析相符（二阶下降较一阶更快），然后在算法复杂度方面，牛顿法显然要比梯度下降法大得多，也是因为在牛顿法中使用到了大量的矩阵运算，包括矩阵求逆、矩阵相乘等等，而梯度下降法中仅仅只是数值运算，这也导致了运行相同的时间，梯度下降法的执行步数要多得多，这也导致两种优化方法在运行相同时间的情况下，得到的结果正确率相差无几。

最终，在测试集上运行的结果保存在*A*.mat 中。

二、基于音频的特征提取和支持向量机分类

原理

支持向量机 (Support Vector Machine, SVM) 是一类按监督学习方式对数据进行二元分类的广义线性分类器，其决策边界是对学习样本求解的最大边距超平面。除了进行线性分类之外，SVM 还可以使用核方法有效地进行非线性分类。

在本题中，我们的目标是提取到合适的音频特征，调整 SVM 的参数，对 SVM 进行训练，使其能够对辩论的激烈程度进行分类。

特征提取方法

MFCC (Mel-Frequency Cepstral Coefficients) 是语音特征参数的提取方法之一，因其独特的基于倒谱的提取方式，更加的符合人类的听觉原理，因而也是最为普遍、最有效的语音特征提取算法。

我们采用 librosa 库来计算音频信号的 MFCC。注意计算 MFCC 时一般需要对信号进行预加重，即让信号通过一个高通滤波器，来增强信号中的高频部分。

但是，计算得到的结果是一个形状为 (n_mfcc, 帧数) 的大矩阵，无法直接用于 SVM 分类。因此我们还需要进一步从 MFCC 中提取信息。

经过思考与试验，我们选择了如下特征：

- MFCC 的均值：体现音频信号大体上的听觉响度特征
- MFCC 的方差：体现音频信号响度的分布
- MFCC 的偏度（三阶矩）：体现音频信号响度分布的不对称性
- MFCC 的一阶差分的均方值：体现音频信号响度变化的快慢程度

我们使用的 n_mfcc=16，故每条音频的特征为一 1×64 向量。

此时，虽然向量的维数减小了，但是其仍不能直接作为用于训练的特征。这是由于不同维度上的特征的尺度差异很大，不同特征对模型参数的影响程度差别很大。因此，我们还需要对特征进行**正规化处理**，这里采用的是 z-score 方法。

结果展示

经过试验，我们采用上述特征和 RBF 核对 SVM 进行训练。其交叉验证结果如下，具有 97% 的正确率：

```
[0.96 0.98 0.96 0.98]
Avg accuracy: 0.97
```

结果分析

在完成本题过程中，我们还尝试了使用其他特征与 SVM 参数。现举例对比、分析如下：

使用线性核

```
[0.9  0.96 0.96 0.96]
Avg accuracy: 0.945
```

准确度差于使用 RBF 核的情况。其原因应当是数据不是完全线性可分的，而使用 RBF 核能引入非线性克服该问题。

特征提取中 n_mfcc = 20

```
[0.94 0.96 0.96 0.98]
Avg accuracy: 0.96
```

准确度差于 n_mfcc=16 的情况。经试验，该值不能太小，否则提取的特征太过粗糙；也不能太大，否则受噪声影响较大。

使用上一节特征的子集

经测试，如果仅使用上一节特征的子集进行训练，准确度也会降低。这说明我们提取到的四组特征都是有效的。

三、基于图像和音频特征的神经网络分类

实现方法

我们利用神经网络分别对图像和音频特征进行训练，利用训练好的模型进行辩论激烈程度推断。

其中，图像特征使用给定的 feat.npy，即根据人体姿态检测和人脸表情分类结果得到的 13 维特征向量；音频特征使用我们在第二问中提取的 MFCC 后处理的 64 维特征向量。

神经网络框架使用 Pytorch，并利用了 Skorch 这一结合了大量 sklearn 方法的 Pytorch wrapper，进一步简化调参、训练、验证的过程。

训练数据集通过 Skorch 自带的 CVsplit 将输入的训练数据集的 80% 作为训练集，20% 作为验证集，进行

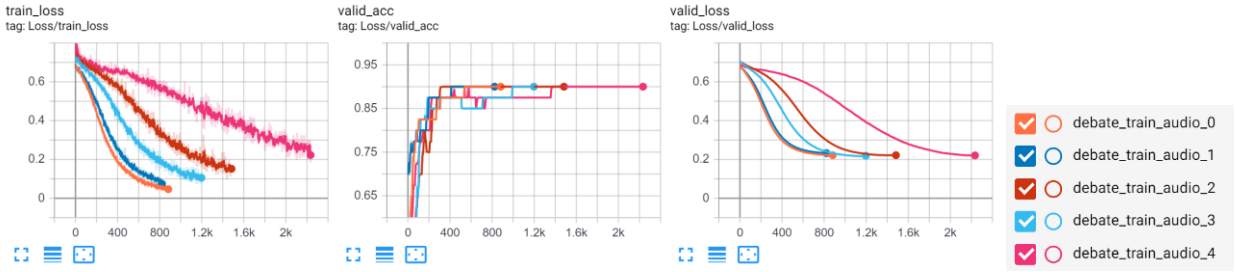
交叉验证，每一 epoch 输出 training_loss、valid_accur（验证集准确率）、valid_loss（验证集 loss），通过 early stopping 在 valid_loss 不再减小时及时停止训练，保存模型。

结果展示

音频特征训练

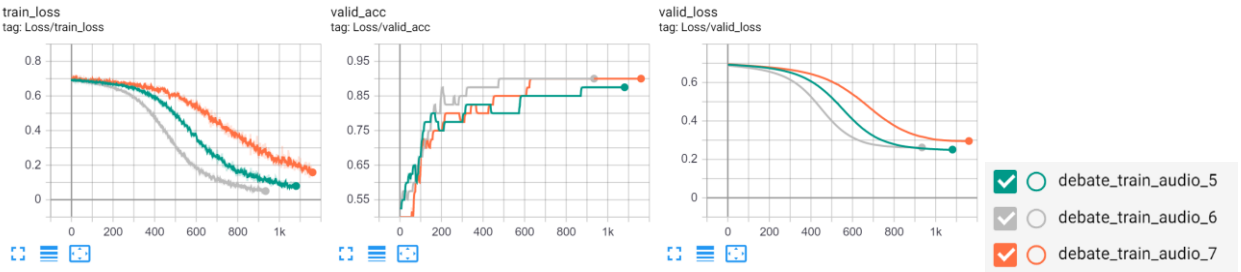
网络 1：全连接（64，128）->全连接（128，64）->全连接（64，1），ReLU 激活层，结合参数为 n 的 Dropout。

训练编号	Dropout 比例 n	交叉验证集准确度	交叉验证集 loss
0	0.3	0.9	0.2216
1	0.5	0.9	0.2323
2	0.7	0.9	0.219
3	0.8	0.9	0.2219
4	0.9	0.9	0.2208



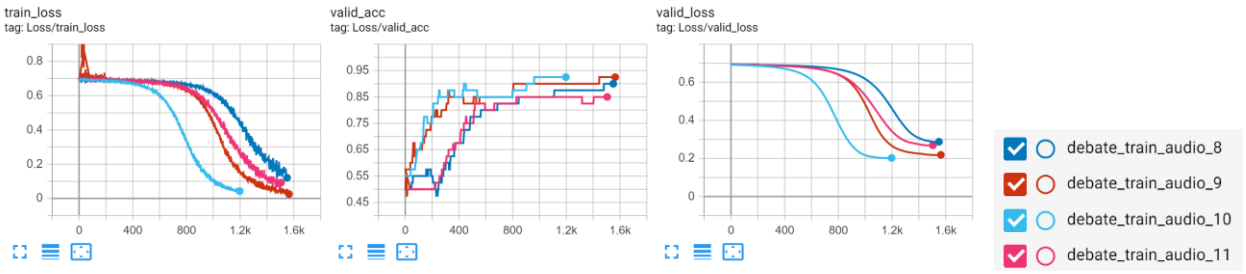
网络 2：全连接（64，128）->全连接（128，256）->全连接（256，64）->全连接（64，1），ReLU 激活层，结合参数为 n 的 Dropout。

训练编号	Dropout 比例 n	交叉验证集准确度	交叉验证集 loss
5	0.5	0.875	0.2511
6	0.3	0.9	0.2619
7	0.7	0.9	0.2956



网络 3：全连接（64，128）->全连接（128，256）->全连接（256，128）->全连接（128，64）->全连接（64，1），ReLU 激活层，结合参数为 n 的 Dropout。

训练编号	Dropout 比例 n	交叉验证集准确度	交叉验证集 loss
8	0.5	0.925	0.2027
9	0.3	0.85	0.2691
10	0.2	0.9	0.2878
11	0.4	0.925	0.2189



图像特征训练

网络 1：全连接（13，32）->全连接（32，1），ReLU 激活层

训练编号	交叉验证集准确度	交叉验证集 loss
0	0.875	0.4288

分析

显然，13 维图像特征的训练效果不如音频的 64 维特征，因此我们最后采用音频特征进行训练与调参，优化其在交叉验证集上的 loss，并得出其在测试集上的结果。

1. 我们在训练中使用 early stopping，就是因为交叉验证的 loss 随着训练迭代，先下降后上升。这种现象最可能由过拟合导致。Early stopping 可以提前阻止过拟合，其次，我们还应用了 Dropout 来随机丢弃网络参数，防止网络对输入的记忆，并通过更改 Dropout 控制变量，观察训练效果。对于简单网络，可以看到有无 Dropout 对网络 loss 有比较明显的作用；当层数增加时，Dropout 的有无甚至是其参数似乎对网络效果没有影响。
2. 由于训练集太小，本质上网络结构的更改抑或是各种参数的修改对结果的影响无法呈现明显的趋势，所以个人感觉调参并没有意义。任何一点参数的变化在如此小的训练集上都会产生很大的影响。
3. 最后选取网络 2：全连接（64，128）->全连接（128，256）->全连接（256，64）->全连接（64，1），ReLU 激活层，结合参数为 0.5 的 Dropout，得到的分类结果和我们对测试集的判断大多数吻合。
4. 复杂度分析：选取的该音频特征网络，在 batch size 为 160，且输入维数 64 时，网络参数量是 57.86k，运算量 FLOPs 为 18.37M。

问题与不足

1. 我们未尝试将图像与音频特征结合起来进行训练的方法，这样可能会有更好的结果，也可能有利于对抗过拟合。
2. 针对控制变量，只对网络结构和 Dropout 参数进行了改变，未尝试改变激活函数、优化方法等。

附：提交文件清单、依赖库信息与运行方法

见报告同目录下的 README.md