好未来 AI camp 学习报告（语音）

概述:
前两周，我参加了好未来的 AI camp 学习，通过两周的抽取特征及调参（炼丹）后，获得两次优秀作业与camp优秀学员。现在我将分享我这两周的工作。
第一周。学习语音处理的一些基本概念和语音计算常用的一些技术，然后通过opensmile工具抽取音频的特征，最后使用GDBT+ridge regression 的混合模型对音频的P值预测，通过10KFlod 交叉验证，最终的模型效果为0.747(pearsonr score)。
第二周。学习神经网络基础知识，了解语音情感识别中常用的 attention mechanism 和 RNN 模型。最终通过librosa抽取音频特征，搭建了一个CNN+LSTM+attention 的网络模型，该模型最终效果为0.73(pearsonr score),再将该模型与上周模型计算结果做混合，最终得到的结果为0.756(pcc)。
因为缺少算力，我构建的模型都未调参，有兴趣的同学可以在我模型的基础上调参，预计调参效果能达到0.76-0.78。
目录:

第一周机器学习方法预测语音情感

论文笔记

语音情感识别研究进展综述（韩文静，阮华斌等，2014）

语音情感识别系统组成：语音信号采集，情感特征提取，情感识别
语音情感识别研究内容可分为：情感空间描述，情感识别数据收集，语音情感识别特征，语音情感识别模型，语音情感识别应用

1. **语音情感模型**
   离散型：
   美国心理学家Ekman 提出的6大基础情感(big six)
   连续型：

二维的激活度-效价空间理论(arousal-valence space)、三维的激励-评估-控制空间理论 (valence-activation-dominance space)

2. **语音数据集**
   CASIA 汉语情感语料库
   系列汉语情感数据库
3. **语音特征**
   用于语音情感识别的声学特征大致可归纳为**韵律学特征、基于谱的相关特征和音质特征**这 3 种类型.这些特征常常以帧为单位进行提取,却以全局特征统计值的形式参与情感的识别.全局统计的单位一般是听觉上独立的语句或者单词,常用的统计指标有极值、极值范围、方差等.

- 韵律学特征：时长(duration)、基频(pitch)、能量(energy)等
- 频域特征：线性谱特征：:LPC(linear predictor coefficient),倒谱特正:MFCC(mel-frequency cepstral coefficient)
- 语言学特征:共振峰频率及其带宽(format frequency and bandwidth)、频率微扰和振幅微扰(jitter and shimmer)[50]、声门参数(glottal parameter)等
  融合特征
  给予i-vector的特征

4. **常用识别模型**
   离散型：KNN,DC,SVM,HMM,GMM,MLP,ANN
   连续型：SVR,ANN
5. **应用**
   驾驶情感监控（疲劳驾驶），在在线课程情感监控，患者情感监控诊断和治疗辅助

基于PAD三位情绪模型的情感语音转换于识别(论文)

1. **情感语音研究**
   离散型：初级情感(primary emotion)和次级情感(secondary emotion)。Emkna 提出的六种基本情感恐惧、愤怒、悲伤、高兴、厌恶和惊奇。
   连续型：
   Schloberg 提出的愉快—不愉快、注意—拒绝和激活水平三个维度Plutchik提出的具有强度、相似性和两极性三维模型
   Russell提出的愉快度和强度模型
   Mehrabian提出的PAD模型
   - Pleasure-Displeasure:
     愉悦度。表示情绪状态的正、负性。
     Arousal-Nonarousal:
     激活度。表示情绪生理激活水平和警觉性。
     Dominance-Submissiveness:
     优势度。表示情绪对他人和外界环境的控制
     力和影响力。
2. **语音情感信号特征**
   韵律特征提取的参数:
   时长，语速，停顿，
   基频，能量（振幅），共振峰
   根据语音声学、人耳感知提取的特征参数:
   线性预测系数LPC,LPC倒谱系数LPCC,mel频率倒谱系数

语音特征参数提取:
采样-->预加重处理-->静音处理-->分帧与加窗-->特征参数计算

3. **特征工程**
   传统语音信号中提取特征主要使用短时分析技术:短时能量，短时过零率，短时平均幅度，短时自相关系数，短时平均幅度差函数，短时频谱，短时功率谱等。短时分析将音流分为一段一段来处理，其中每一段为一帧(10~30ms),帧移:0~1/2帧长,常用窗口:矩形，hamming，hann
   较为新颖特征:使用Hibert-huang的特征提取
4. **语音情感识别模型**
   SVR

opensmile操作抽取语音特征

1. opensmile操作
   在提取语音特征我是通过配置文件结合批处理操作使用opensmile
2. 语音特征
   IS09_emotion: 2009年国际语音情感挑战赛特征集合，包含对LLD(low-level descriptor)应用统计函数得到的384个特征,主要包含短时能量，MFCC，短时过零率，及时域与频域信息的统计量
   IS10_paraling: 2010年国际语音挑战赛特征集合，总共包含1582个特征，其中34个低级描述符（LLDs）及其逐差相关系数(delta cofficients),68个LLD使用21个函数生成的1428个特征，此外使用19个函数通过4个基频LLD(pitch-based LLD)及其逐差相关系数。
   IS11_speaker_state:
   IS12_speaker_trait:
   IS13_ComParE：
   emobase: 包含以下特征的LLD:强度(Intensity),响度(Loudness), 12 MFCC, Pitch (F0),发声概率(Probability of voicing), F0 envelope, 8 LSF (Line Spectral Frequencies), 过零率(Zero-Crossing Rate),在对这些特征计算统计特征(算术平均，线性拟合，逐差相关系数,标准差等)
   emobase2010: 根据opensmile的使用文档，此特征集合与 IS10_paraling基本相同一。
   emo_large：

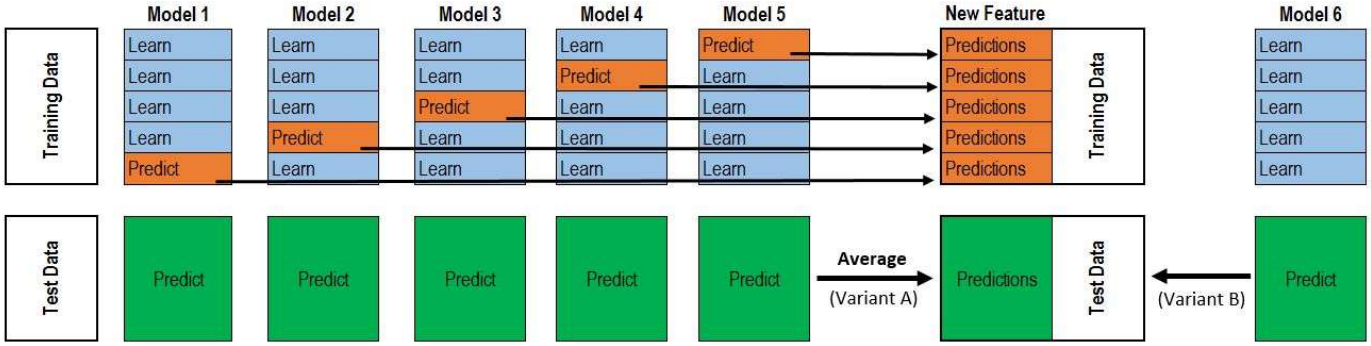
本实验在做特征工程时选用了上面所有特征，每个特征单独使用LightGBM训练并预测，最后通过岭回归(Ridge Regression)将单独的预测结果拟合得出最终结果

所用模型与模型评价

1. 模型:
   提升树(LightGBM):
   梯度提升树的优化模型，通过直方图算法优化了分割点搜索的速度，并对传统的并行算法优化。
   岭回归(Ridge Regression)
   通过添加L2范数惩罚项的最小二乘回归，它需要最小化的损失函数为:

$$\min_w ||X_w - y||_2^2 + \alpha||w||_2^2$$

   混合模型(stacking,blend):
   模型构建如下图



在本次实验中，我对模型修改使得混合模型不是通过不同的基础模型在同一数据集上的不同预测结果最为下一级模型的输入，而是通过使用相同的基础模型(LightGBM)在不同的数据集上训练作为下一级模型的输入。

2. 模型评价：

可解释方差得分(explained variance score):

$$explained\_variance(y, \hat{y}) = 1 - \frac{Var\{y - \hat{y}\}}{Var\{y\}}$$

评价绝对误差(Mean absolute error)

$$MAE(y, \hat{y}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} |y_i - \hat{y}_i|$$

均方差(Mean squared error)

$$MAE(y, \hat{y}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} (y_i - \hat{y}_i)^2$$

R2 决定系数(R² score)

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=0}^{n_{samples}-1}(y_i - \hat{y}_i)^2}{\sum_{i=0}^{n_{samples}-1}(y_i - \bar{y}_i)^2}$$

Pearson相关性系数(PCC,Pearson correlation coefficient)

$$\rho_{X,Y} = \frac{cov(X, Y)}{\sigma_X \sigma_Y}$$

模型(代码)整体脉络

1. 使用上面所述的语音特征集合的配置文件通过opensmile抽取特征
2. 混合模型的第一级模型：对每组特征使用 LightGBM 模型拟合，并通过10-FLod记录每次拟合效果
3. 混合模型的第二级模型：将 LightGBM 的拟合结果再通过岭回归拟合得出最终结果,并通过10-FLod记录拟合效果

实验结果与评估

首先通过各个语音特征集合的opensmile配置文件(IS09_emotion,IS10_paraling,IS11_speaker_state,IS12_speaker_trait,IS13_ComParE,emobase,emobase2010,emo_large)抽取语音特征，再使用 LightGBM 拟合，最终使用岭回归拟合上面的预测结果，效果如下(预测效果都是使用5-KFlod的交叉验证):

	method	explained_variance_score	mae	mse	r2_score	pearsonr_score
0	IS09_emotion-lightgbm	0.440435	0.570866	0.557346	0.436991	0.682871
1	IS10_paraling-lightgbm	0.524198	0.521948	0.471702	0.520788	0.725706
2	IS11_speaker_state-lightgbm	0.469320	0.554042	0.525173	0.465982	0.707846
3	IS12_speaker_trait-lightgbm	0.498083	0.536783	0.498760	0.494411	0.719321
4	IS13_ComParE-lightgbm	0.520890	0.526819	0.476636	0.517492	0.725607
5	emobase-lightgbm	0.490612	0.542987	0.506003	0.484586	0.706525
6	emobase2010-lightgbm	0.516928	0.526999	0.478623	0.513516	0.724016
7	emo_large-lightgbm	0.523099	0.524991	0.473867	0.520473	0.727198
8	blend-ridge	0.556291	0.497225	0.440209	0.551923	0.747586

第二周构建神经网络预测语音情感

论文概括

论文：ADVANCED LSTM: A STUDY ABOUT BETTER TIME DEPENDENCY MODELING IN EMOTION RECOGNITION

该论文提出一种改进的LSTM模型，相比于传统的LSTM，该改进模型不仅只接受上次网络单元的状态变量h,还接受更远的状态变量h.

论文：AUTOMATIC SPEECH EMOTION RECOGNITION USING RECURRENT NEURAL NETWORKS WITH LOCAL ATTENTION

该论文提出一种基于attention mechanism 的 local attention 模型，具体实现是通过RNN的输出计算一组权重并将权重与RNN输出矩阵相乘作为新的输出

论文：Multimodal emotion recognition base on deep neural network

改论文提出一种机基于视频与音频的多模块识别。通抽取音频 低频(low-level)信息将其压缩为1维后用CNN与RNN处理，而对于视频信息则是通过CNN与RNN处理抽取面部变化分析情绪。

神经网络模型与所使用的音频特征

所使用的特征：

使用librosa抽取音频特征

- chroma_cens
- mfcc
- spectral_centroid

- spectral_contrast
- spectral_rolloff
- zero_crossing_rate
- melspectrogram

模型:

CNN+LSTM+Attention Mechanism,首先构建一组CNN网络抽取时序信息，然后通过LSTM模型处理CNN抽取出的信息，使用 Attention Mechanism 通过LSTM的输出计算权重并将权重与LSTM输出矩阵相乘作为结果，再通过若干层全连接神经网络得出最终输出结果。模型代码与结构如下:

```
x_input = Input(shape=(shape[-2],shape[-1]),name='input')
x_mid = Conv1D(filters=32,kernel_size = window_length,activation='relu')(x_input)
x_mid = MaxPooling1D(2)(x_mid)
x_mid = Conv1D(32,kernel_size=window_length,activation='relu')(x_mid)
x_mid = LSTM(512,dropout=0.1, recurrent_dropout=0.2,activation='relu',return_sequences=True)(x_mid)
attation = Dense(1)(x_mid)
attation = Reshape((1,x_mid._keras_shape[1]))(attation)
attation = Dense(x_mid._keras_shape[1],activation='softmax')(attation)
output = dot([attation,x_mid],axes=(2,1))
output = Flatten()(output)
output = Dense(512)(output)
output = Dense(512)(output)
output = Dense(512)(output)
output = Dense(512)(output)
output = Dense(128)(output)
output = Dense(1)(output)
model = Model(inputs=[x_input],outputs=output)

model.compile(optimizer=Adam(lr=5e-4),loss=['mse'],metrics=[pearson_r])
```

Layer (type)	Output Shape	Param #	Connected to
input (InputLayer)	(None, 313, 170)	0	
conv1d_1 (Conv1D)	(None, 294, 32)	108832	input[0][0]
max_pooling1d_1 (MaxPooling1D)	(None, 147, 32)	0	conv1d_1[0][0]
conv1d_2 (Conv1D)	(None, 128, 32)	20512	max_pooling1d_1[0][0]
lstm_1 (LSTM)	(None, 128, 512)	1116160	conv1d_2[0][0]
dense_1 (Dense)	(None, 128, 1)	513	lstm_1[0][0]
reshape_1 (Reshape)	(None, 1, 128)	0	dense_1[0][0]
dense_2 (Dense)	(None, 1, 128)	16512	reshape_1[0][0]
dot_1 (Dot)	(None, 1, 512)	0	dense_2[0][0] lstm_1[0][0]
flatten_1 (Flatten)	(None, 512)	0	dot_1[0][0]
dense_3 (Dense)	(None, 512)	262656	flatten_1[0][0]
dense_4 (Dense)	(None, 512)	262656	dense_3[0][0]
dense_5 (Dense)	(None, 512)	262656	dense_4[0][0]
dense_6 (Dense)	(None, 512)	262656	dense_5[0][0]
dense_7 (Dense)	(None, 128)	65664	dense_6[0][0]
dense_8 (Dense)	(None, 1)	129	dense_7[0][0]

Total params: 2,378,946
Trainable params: 2,378,946
Non-trainable params: 0

结果与分析：

因为没有计算资源所以基本没有做超参数调节，模型的效果不理想。神经网络模型10-KFold交叉验证的结果如下：

	explained_variance_score	mae	mse	r2_score	pearsonr_score
0	0.527569	0.495816	0.425169	0.522505	0.726343
1	0.474277	0.499342	0.448130	0.436024	0.688691
2	0.487600	0.551819	0.554921	0.487584	0.698292
3	0.407286	0.671261	0.751298	0.378726	0.793954
4	0.421979	0.523098	0.472775	0.415848	0.679902
5	0.492131	0.550054	0.534562	0.491082	0.712520
6	0.472808	0.540471	0.500103	0.465861	0.687853
7	0.602861	0.497938	0.439866	0.598729	0.782873
8	0.642362	0.486932	0.428794	0.618192	0.803376
9	0.615350	0.487246	0.414849	0.613681	0.797926

平均PCC值为0.737，如下：

```
explained_variance_score    0.514422
mae                         0.530398
mse                         0.497047
r2_score                    0.502823
pearsonr_score              0.737173
```

将神经网络的结果与上周的计算结果一同使用岭回归混合训练，10-KFold训练后，平均结果如下：

	explained_variance_score	mae	mse	r2_score	pearsonr_score
0	0.564271	0.491913	0.434691	0.559234	0.756951

可改进的部分

主要缺少算力，加大算力我还可以做更多尝试：

- 可通过librosa提取更多特征
- 构建更复杂的网络结果
- 尝试更多的超参数组合

参考文献：

http://www.audeering.com/research/opensmile

http://scikit-learn.org/stable/modules/model_evaluation.html#regression-metrics

https://lightgbm.readthedocs.io

https://librosa.github.io/librosa/

Cormen T H, Leiserson C E, Rivest R L, et al. Introduction to algorithms[M]. MIT press, 2009.

Data classification: algorithms and applications[M]. CRC Press, 2014.

王炳锡. 实用语音识别基础[M]. 国防工业出版社, 2005.

语音情感识别进展 赵双全

韩文静, 李海峰, 阮华斌, 等. 语音情感识别研究进展综述[J]. 软件学报, 2014.

周慧. 基于 PAD 三维情绪模型的情感语音转换与识别[D]. 兰州: 西北师范大学, 2009

张雪英, 张婷, 孙颖, 等. 情感语音数据库优化及 PAD 情感模型量化标注[J]. 太原理工大学学报, 2017, 48(3): 469-474.

Ruder S. An overview of gradient descent optimization algorithms[J]. arXiv preprint arXiv:1609.04747, 2016.

Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks[C] 2010.
He, Kaiming, et al. "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification." Proceedings of the IEEE international conference on computer vision. 2015.
Mirsamadi S, Barsoum E, Zhang C. Automatic speech emotion recognition using recurrent neural networks with local attention[C]//Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on. IEEE, 2017: 2227-2231.
Tao F, Liu G. Advanced LSTM: A study about better time dependency modeling in emotion recognition[J]. arXiv preprint arXiv:1710.10197, 2017.