Python Machine Learning

讲师：陈博

Python机器学习

- 隆重推出scikit-learn机器学习库
- Scikit-Learn是基于python的机器学习模块
- Scikit-Learn中的机器学习模型非常丰富，包括SVM，决策树，GBDT，KNN等等，可以根据问题的类型选择合适的模型
- Scikit-Learn的安装需要numpy，scipy，matplotlib等模块
- 首先安装VCForPython27.msi->wheel->numpy
- 其次安装scipy
- 然后安装scikit_learn

微博聚类

- 数据集(微博数据)
- 算法使用(scikit-learn中的kmeans)
- 期望结果(相似微博聚到同一类)
- 额外支持模块(jieba中文分词库)

原始数据

```
1  3793992720744105    #九阳有礼 无需多滤#陷入被窝温柔乡，起床靠毅力？九阳免滤豆浆机C668SG耀世首发！智能预约免过滤
2  3793993084926422    #谢谢你陪我走过2014#好吧，这一年马上就要过去了，同样这一年有欢笑，有泪水，但更多的还是幸福。
3  3793993291060111    跨年啦。小伙伴们，新年快乐～[笑哈哈][笑哈哈][笑哈哈]@美的电饭煲官方微博 @美的生活电器 @九阳
4  3793993588106975    我的胆有0.9斤，我想要3.1斤重的铁釜，有份量才够胆量！九阳Alva0716
5  3793995102741635    《太上青玄慈悲太乙救苦天尊寶懺》 - 起讚　元始運元　神運元神 &#57359;化太一尊　九陽天上布
6  3793995370610238    #九阳有礼 无需多滤#新年交好运！有了九阳，让生活免滤无忧！@誰能許诺給我一世柔情 @索心进 @错爱
7  3793995484592300    #谢谢你陪我走过2014#2014年将至，希望能中一个好东西来送给我的家人。@九阳 @枫叶红了112
8  3793995781905340    免过滤，更顺滑，#九阳有礼 无需多滤# 更多营养更安心！@princess佳妮昂 @木凝眉 @单纯会让人受伤
9  3793996277455995    #谢谢你陪我走过2014#2014年将至，希望能中一个好东西来送给我的家人。@九阳 @枫叶红了112
10 3793996323668014    #谢谢你陪我走过2014#2014年将至，希望能中一个好东西来送给我的家人。@九阳 @枫叶红了112
11 3793996390629648    #谢谢你陪我走过2014#2014年将至，希望能中一个好东西来送给我的家人。@九阳 @枫叶红了112
12 3793997111551610    #谢谢你陪我走过2014#2014年将至，希望能中一个好东西来送给我的家人。@九阳 @田雨贝儿
13 3793997170105703    #谢谢你陪我走过2014#2014年将至，希望能中一个好东西来送给我的家人。@九阳 @田雨贝儿
14 3793997229018013    #谢谢你陪我走过2014#2014年将至，希望能中一个好东西来送给我的家人。@九阳 @田雨贝儿
15 3793997380280014    #九阳有礼 无需多滤#陷入被窝温柔乡，起床靠毅力？九阳免滤豆浆机C668SG耀世首发！智能预约免过滤
16 3793997920780506    #九阳有礼 无需多滤#给力活动，亲们快来支持哦~@茹卫满同 @贾昼研牡 @笨笨de有来有去
17 3793999376436708    【二〇一五·美意延祥年】奉上一张清代乾隆年间的九阳消寒图。此图以缂丝加刺绣制成，其背景为缂丝
18 3793999473378969    #掌阅iReader阅读分享#无聊度日是可耻的，看看这本《九阳帝尊剑棕著》吧！让你的生活丰富起来！@掌
19 3793999640732305    新年快乐[太开心][太开心][太开心][太开心]@九阳 @好孕妈妈杂志 @宝贝第一babyfirst官微 @澳贝婴幼
20 3794000605348165    #九阳有礼 无需多滤#新年交好运！有了九阳，让生活免滤无忧！@曾澄鸿 @狄曲淑 @舒素贡素
```

案例流程

- 一行行读入原始微博
- 读的同时进行分词并存入语料库
- 使用sklearn包中feature_extraction的方法计算出每条微博每个词中的tf-idf值
- 将计算出的微博向量矩阵带入到算法中去聚类
- 将聚类结果和原始微博数据进行整合存入一个结果文件

结果文件

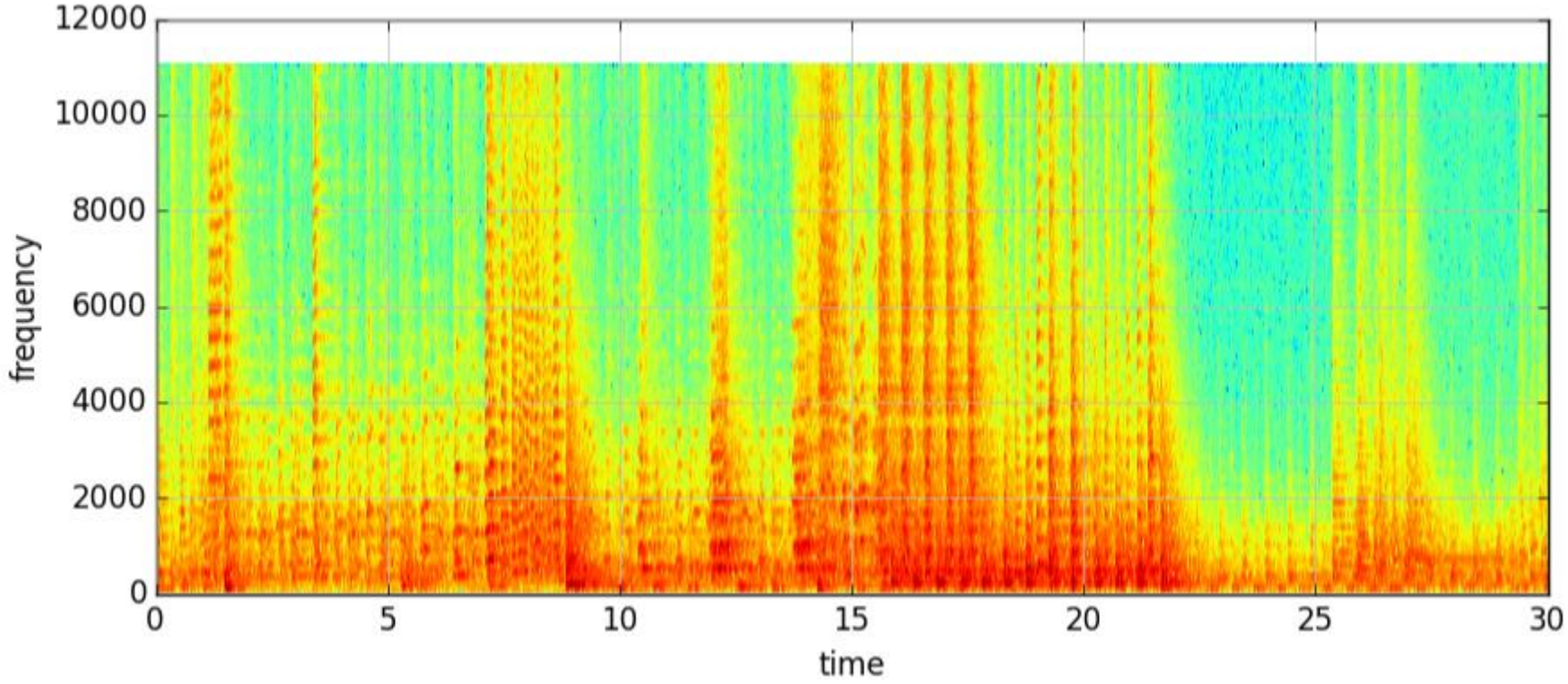
```
3       # 谢谢 你 陪 我 走过 2014 # 2014 年 将 至 ， 希望 能 中 一个 好 东西 来 送给 我 的 家人 。 @ 九阳
9       免 过滤 ， 更 顺滑 ， # 九阳 有 礼
3       # 谢谢 你 陪 我 走过 2014 # 2014 年 将 至 ， 希望 能 中 一个 好 东西 来 送给 我 的 家人 。 @ 九阳
3       # 谢谢 你 陪 我 走过 2014 # 2014 年 将 至 ， 希望 能 中 一个 好 东西 来 送给 我 的 家人 。 @ 九阳
3       # 谢谢 你 陪 我 走过 2014 # 2014 年 将 至 ， 希望 能 中 一个 好 东西 来 送给 我 的 家人 。 @ 九阳
3       # 谢谢 你 陪 我 走过 2014 # 2014 年 将 至 ， 希望 能 中 一个 好 东西 来 送给 我 的 家人 。 @ 九阳
3       # 谢谢 你 陪 我 走过 2014 # 2014 年 将 至 ， 希望 能 中 一个 好 东西 来 送给 我 的 家人 。 @ 九阳
3       # 谢谢 你 陪 我 走过 2014 # 2014 年 将 至 ， 希望 能 中 一个 好 东西 来 送给 我 的 家人 。 @ 九阳
1       # 九阳 有 礼
1       # 九阳 有 礼
9       【 二 〇 一五 · 美意 延祥 年 】 奉 上 一张 清代 乾隆年间 的 九阳 消寒 图 。 此 图 以 缂丝 加 刺绣 制成 ， 其
6       # 掌阅 iReader 阅读 分享 # 无聊 度日 是 可耻 的 ， 看看 这本 《 九阳 帝尊剑 棕 著 》 吧 ！ 让 你 的 生活 丰
2       新年快乐 [ 太 开心 ] [ 太 开心 ] [ 太 开心 ] [ 太 开心 ] @ 九阳
1       # 九阳 有 礼
9       # 谢谢 你 陪 我 走过 2014 # 马上 要 和 2014 告别 了 ， 这 一年 七七 又 长大 一岁 ， 今年 带 着 七七 走过 了
2       # 谢谢 你 陪 我 走过 2014 # 好 吧 ， 这 一年 马上 就要 过去 了 ， 同样 这 一年 有 欢笑 ， 有 泪水 ， 但 更
0       # 谢谢 你 陪 我 走过 2014 # 谢谢 我 的 好 朋友 们 ， 有 你们 的 陪伴 我 不 孤单
9       # 谢谢 你 陪 我 走过 2014 # [ 馋嘴 ]
9       2014 年 的 最后 一秒 买 了 瓷宝 煲 ， 199 还 送 料理 棒 ， 希望 萌 哒 哒 的 瓷宝煲 和 可爱 的 料理 棒 能 给
1       # 九阳 有 礼
0       # 谢谢 你 陪 我 走过 2014 # 感谢 你 陪伴 我 走过 2014 年 ， 谢谢 你 ！ 我 的 好 朋友 ！
1       # 九阳 有 礼
0       # 谢谢 你 陪 我 走过 2014 #
```

音乐分类

- 数据集(音乐数据)
- 算法使用(scikit-learn中的logistic regression)
- 期望结果(输入一首歌,可以对输入的歌曲进行分类)
- 额外支持模块(安装dateutil->six->pyparsing->pytz->matplotlib)

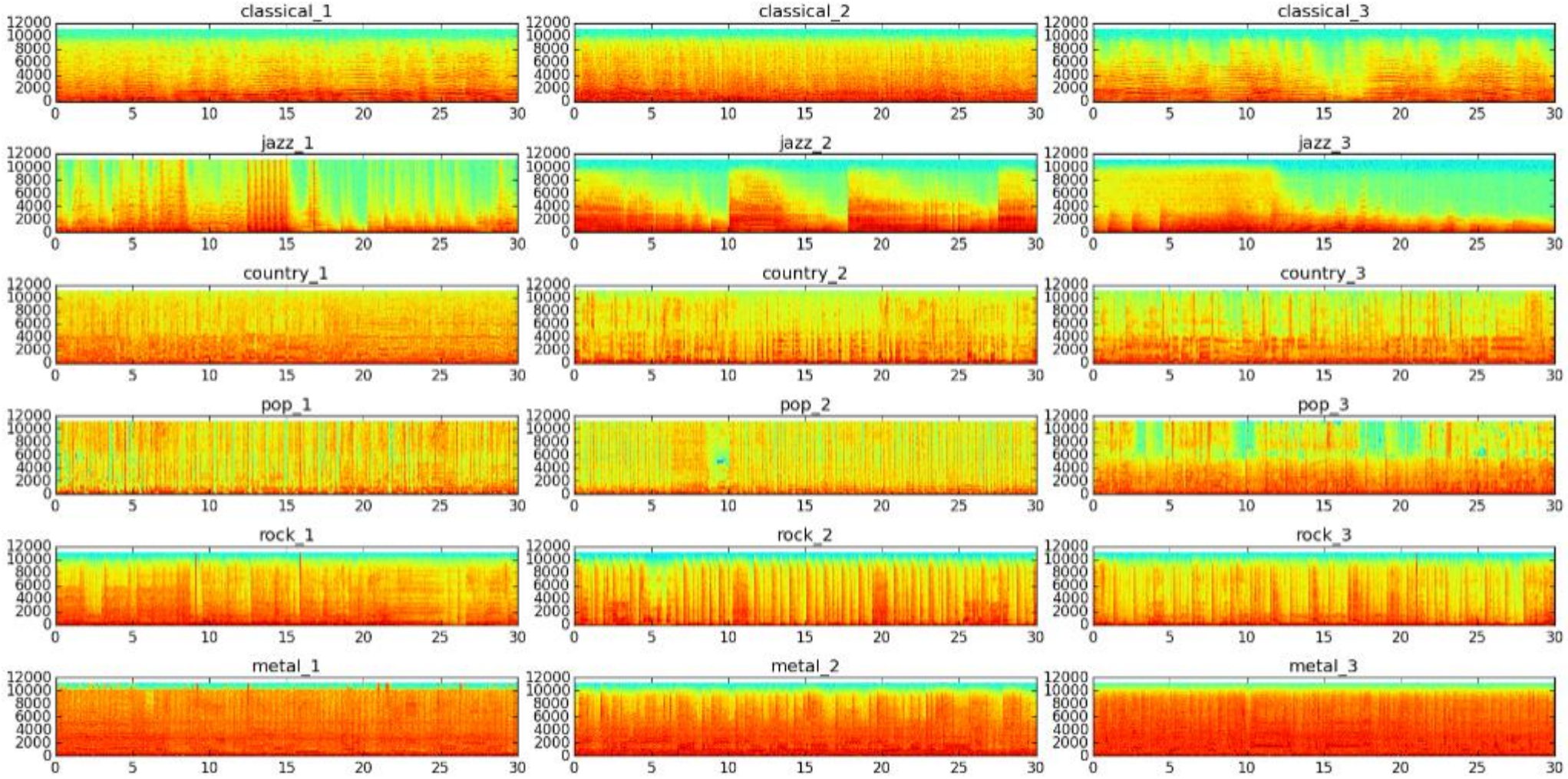
音乐数据

- 分类型存在文件夹中
- 以先把一个wma文件读入python,然后绘制它的频谱图(spectrogram)来看看是什么样的jazz



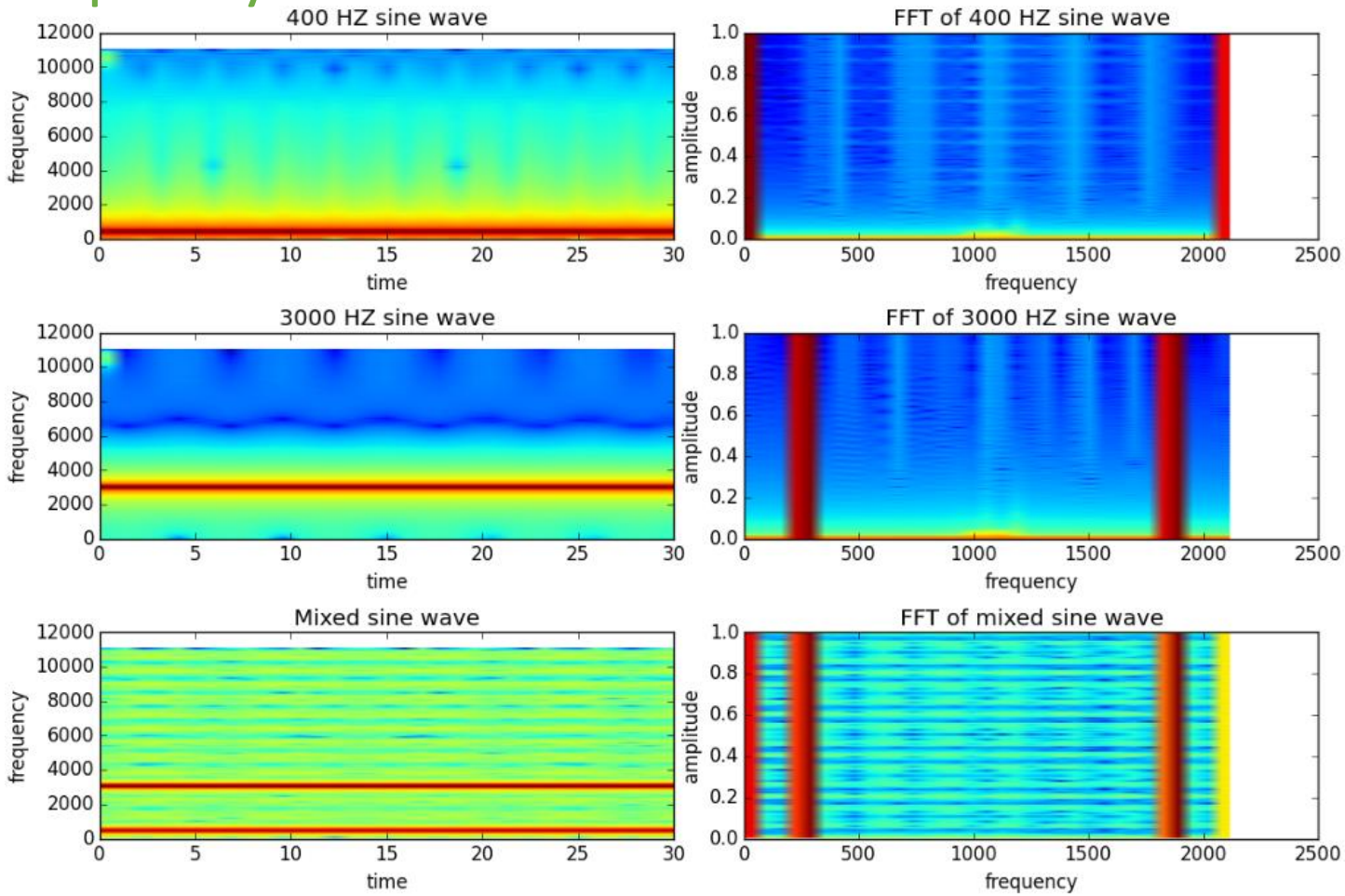
音乐数据

- 可以把每一种的音乐都抽一些出来打印频谱图以便比较,如下图:



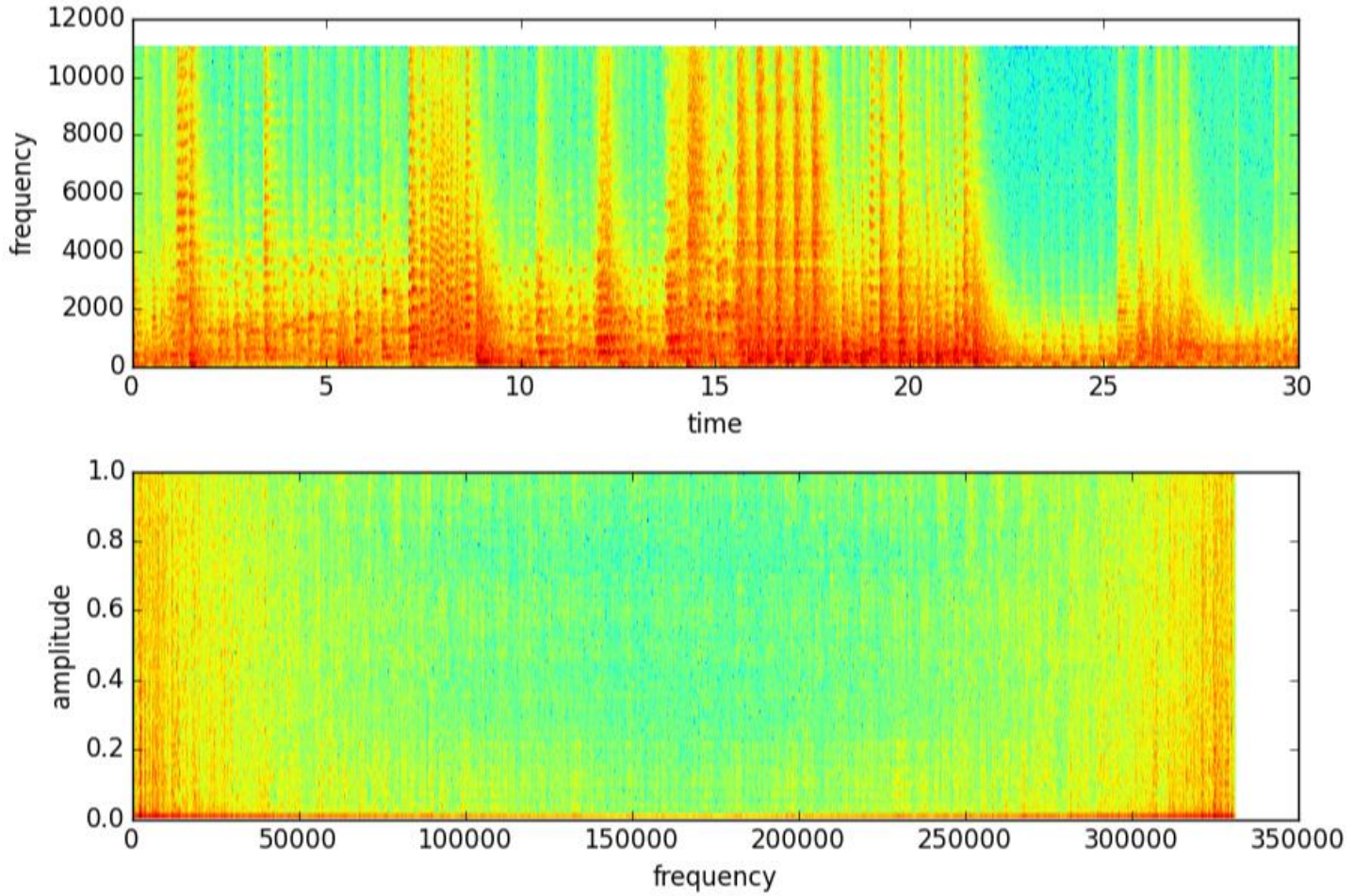
傅里叶变换

- 可以把time domain上的数据,例如一个音频,拆成一堆基准频率,然后投射到frequency domain上

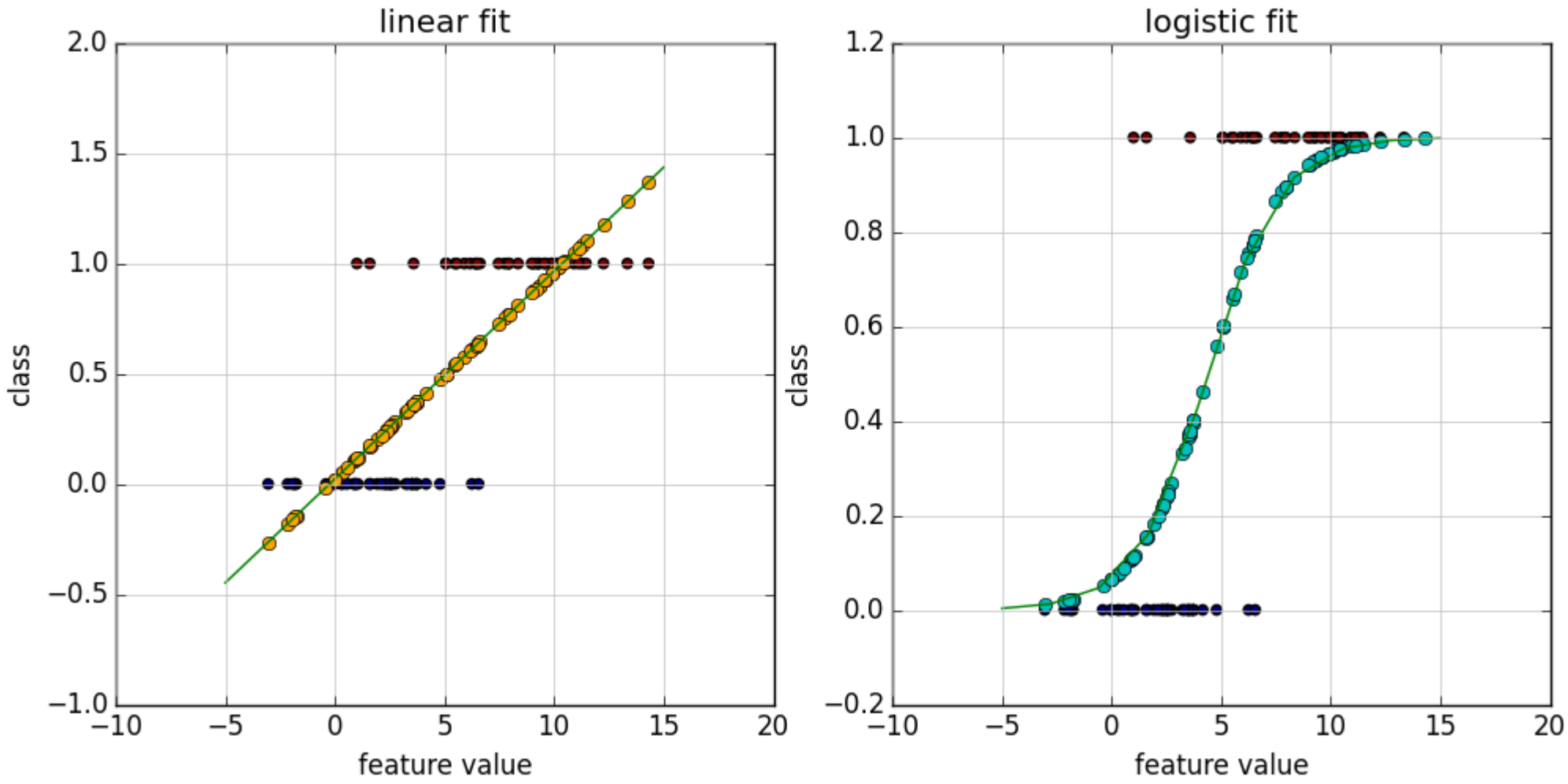


傅里叶变换

- 可以把time domain上的数据,例如一个音频,拆成一堆基准频率,然后投射到frequency domain上



逻辑回归


linear fit
class
feature value
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−10
−5
0
5
10
15
20
logistic fit
class
feature value
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−10
−5
0
5
10
15
20

案例流程

- *["classical", "jazz", "country", "pop", "rock", "metal"]*
- 通过傅里叶变换将以上6类里面所有原始wav格式音乐文件转换为特征,并取前1000个特征,存入文件以便后续训练使用
- 读入以上6类特征向量数据作为训练集
- 使用sklearn包中LogisticRegression的fit方法计算出分类模型
- 读入黑豹乐队歌曲”无地自容”并进行傅里叶变换同样取前1000维作为特征向量
- 调用模型的predict方法对音乐进行分类,结果分为rock即摇滚类

confusion matrix