贝叶斯判决脑电波分类实验报告

李浩然

2017210914

1. 实验目的

给出二十个人的脑电波采样数据以及对应的睡眠阶段。利用前19个人的脑电波-睡眠阶段作为训练数据。根据贝叶斯判决预测第二十个人的脑电波采样数据对应的睡眠阶段。将预测结果和真是结果作比较，并进行分析。

1. 数据描述
   1. 脑电波数据

实验给出的数据是对一个人脑电波的200Hz的时域采样数据，每个采样点的值在±500之间。数据以一个1\*M的矩阵形式保存在.mat文件中并命名为data。M为每个人对应的采样点个数。

* 1. 睡眠阶段数据

实验给出的睡眠阶段数据包括-1~5七类，仅要求对1~3类做分类。

* 1. 数据对应关系

睡眠阶段的数据和脑电波采样数据不是一一对应，而是一对一千的形式对应，即每一千个脑电波采样对应一个睡眠阶段。

1. 实验环境

本次实验在Windows 10系统中借助Octave 4.2.1软件完成。

处于编程方便考虑，本次实验将“数据”和“标签”文件夹重命名为“data”和“label”。

1. 实验过程
   1. 读入数据
      1. 文件读入

脑电波采样文件保存成.mat文件，可以用load函数很方便读入，再将data变量用assignin重命名即可。

睡眠阶段数据由于是纯文本的txt文件，需要简单的文件读入操作，即用fopen得到文件句柄、feof判断文件结束、fgetl读入一个睡眠阶段标签。其他需要注意的就是文件第一行是注释信息，需要丢弃。

* + 1. 数据过滤

读入的采样数据和标签包含非1~3类睡眠阶段的数据，需要进行过滤。具体做法为：利用idx = find(label > 0 & label < 4)命令得到需要处理的数据的标号，再如下所示将对应标号的采样数据和标签提取出来。

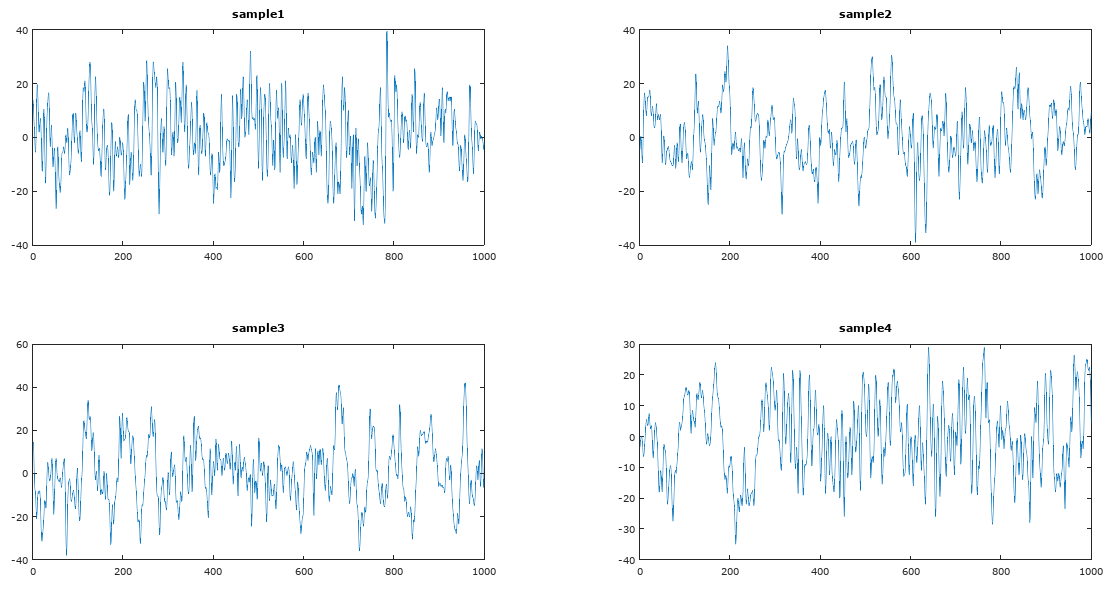
filered\_sample = sample(indices(:),:);

filered\_label = label(indices(:));

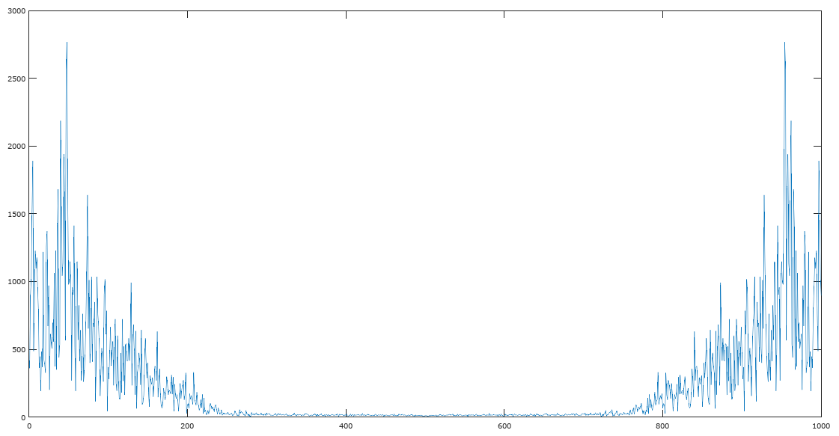
上述读入数据部分的功能在“load\_data.m”脚本中实现。

* 1. FFT转换

脑电波采样的数据是时域上的数据，所以各个1000点采样数据之间很可能存在不同程度的时间偏移。直接将每组的1000个采样点作为特征很有可能受时间偏移的影响，各个类之间难以曲分。如下图所示，sample1~4均属于第三类睡眠阶段，sample1中0~800这个区间的采样值和sample4中200~1000区间的分布类似、sample2中200时间附近的峰和sample3中700时间点附近的峰类似，但因为时间偏移，这个特征就难以被学习到。所以考虑将时域变换到频率以消除该时间偏移的影响。



脑电波采样数据是离散的，而且以1000个点为一组。所以考虑利用DTFT做变换。Octave提供了fft(sample, 1000, 2)函数，对M\*N规模的sample矩阵每一行的1000个点分别做DTFT变换，得到M\*N规模的频域分布矩阵fft\_sample，由于DTFT求出来的是复数表示的频谱分布，所以需要再用abs函数求各个频率的幅值。下图是其中一行的频谱分布。



根据DTFT的共轭特性，所有的频谱分布都是轴对称的，所以在之后的运算中可以只对其中一般进行分析，所以做fft\_sample = fft\_sample(:,1:500)的处理。

以上FFT变换由”preprocess.m”脚本实现。

* 1. PCA变换

即使利用DTFT的共轭特性将维度降低了一半，但500维对于后续分析仍然太多，而且在上面所示的频谱分布中，高频分量的幅值非常小。所以考虑用PCA进行降维。Octave提供了[COEFF, SCORE, latent] = princomp(fft\_sample)函数，以fft\_sample低维作为特征做PCA，得到的COEFF的每一列保存成分降序排列的基，SCORE保存原矩阵在COEFF基上的系数，latent表征各个维度上包含原矩阵成分的多少。通过cumsum(latent) ./ sum(latent)得到累计成分比重可以作为选取几个主成分的依据。

如下图所示，为了保留大部分信息，选取0.9累计成分对应的15个主成分。又有princomp函数默认对数据做均值规整化操作，所以经过PCA降维的数据集可以表示为：pca\_sample = (fft\_sample – mean(fft\_sample)) \* COEFF(:,1:15)。

pca\_vec自定义函数根据输入指定的保留信息占比求出对应的变换矩阵vec，后续的pca变换可以简单地用(fft\_sample – mean(fft\_sample)) \* vec操作实现。

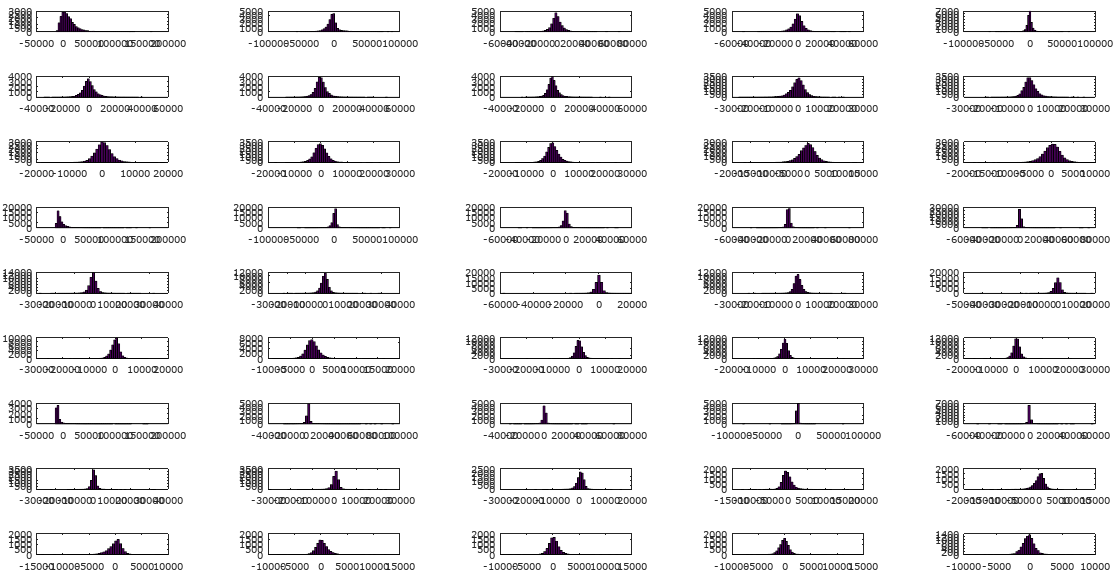
* 1. 贝叶斯判决
     1. 先验概率

对训练集根据睡眠阶段的标签分成三类，统计各类是数据集的大小，求得各类的先验概率分别为0.28407、0.60567和0.11026。

* + 1. 类条件概率

贝叶斯判决需要各类数据中特征的概率分布，所以对分好的三类数据集，需要估计PCA选出的15维特征分别的概率分布函数。

因为没有关于特征分布的先验知识，所以对所有15类做出统计直方图观察数值的分布形式。下图第1~3行列出了第一类睡眠阶段十五个特征的数值分布，4~6行、7~9行则分别对应第二、第三类的十五个特征的数值分布。



由上图可以看出，各个特征的数值分布都大致符合正太分布，所以考虑用正态函数作为数值分布的函数模型，再利用极大似然估计得到每个特征正态分布的参数。

上述求类条件概率的过程由build\_model函数实现，具体做法就是对训练集的每一列求均值、方差作为正态分布的参数。

* + 1. 贝叶斯判决

本次实验中假设每个特征相互独立，所以考虑对每个特征做贝叶斯判决（类先验概率乘类条件概率），得到各个特征对应类的概率。

这样，对于十五个特征，每个特征有三个类的后验概率，本次实验尝试两种方法预测某个测试样本属于哪个类：

1. 将十五个特征的后验概率做连乘，比较三个类的最终结果取大值。
2. 算出每个特征中概率大的类，综合15个特征投票得到预测的类。
3. 实验结果

利用4.4.3中描述的两种方法求得的准确率均为0.57431。观察预测结果：

|  |  |  |  |
| --- | --- | --- | --- |
| 预测类 真实类 | 1 | 2 | 3 |
| 1 | 0 | 0 | 0 |
| 2 | 1662 | 2736 | 366 |
| 3 | 0 | 0 | 0 |

1. 结果分析

由5中结果可以看出，贝叶斯判决的方法将所有的测试数据预测成了第二类，分析原因有以下两点：

1. 第二类数据占训练数据的大部分，所以先验概率比其他类别大很多，在做贝叶斯判决的时候乘上先验概率第二类的后验概率明显高于其他类。
2. 观察4.4.2的统计直方图可以发现，各类对应特征的数值分布都比较接近，均值基本为0。所以各个类在不同特征下的类条件概率也非常接近，即，在本次实验中贝叶斯判决实际上绝大部分取决于类先验概率。综合第一点原因，可知最后结果会以极大概率预测成第二类。