统计学习课程设计报告

# 摘要

报告主要分为数据分析、LDA降维分析、测试数据预测3个部分。在数据分析部分，我分别用SVM和KNN的方法，用交叉校验的方式进行分析，给出了模型选择曲线；在LDA降维分析部分，我使用降维LDA的方式，将数据降到2维，找出了最优降维投影矩阵；在测试数据预测部分，我对测试数据进行了一系列预处理，并使用前面得到的模型预测测试数据的标签。

# 数据分析

数据分析我主要针对训练数据，首先使用PCA对数据进行降维，然后使用SVM和KNN方法进行分析，利用交叉验证画出模型选择曲线（曲面）。

我认为，心电信息最丰富的位置应该在R波波峰附近区域。因此，对于所有心跳数据，我以R波波峰为中心，取200个采样点，因此数据维数为200 \* 12 = 2400维。

为了便于处理，我将标签用数字来表示，以下是标签与数字的对应关系列表。

|  |  |  |  |
| --- | --- | --- | --- |
| 正常心电图 | 1 | 窦性心律\_提示左室肥厚 | 6 |
| 窦性心动过缓 | 2 | **窦性心律\_完全性右束支传导阻滞** | 7 |
| 窦性心动过速 | 3 | **窦性心律\_完全性左束支传导阻滞** | 8 |
| 窦性心律\_不完全性右束支传导阻滞 | 4 | **窦性心律\_左前分支阻滞** | 9 |
| 窦性心律\_电轴左偏 | 5 | **窦性心律\_左室肥厚伴劳损** | 10 |

## 个性化问题

根据训练数据的格式，可以推测一组数据里面包含的多个心跳是同一个人连续测量的结果。如果将每个心跳的数据全部集合在一起并打乱，进行交叉验证，则有很大的可能训练集与测试集分别包含了这组数据的不同心跳。在这种情况下，会低估模型的泛化误差。所以在进行交叉验证时，我特意确保同一个人的心跳数据不同时出现在训练集和测试集中。根据实际测试，经过这样的处理之后，模型泛化误差估算从5%左右上升到40%，而后者才是对泛化误差合理的估计。可见个性化问题确实对模型有很大的影响。

## 归一化

观察心电数据，可以发现心跳时心电振幅差异比较大，并且在一段较长时间里，心电存在基线漂移。因此，我首先对心电数据进行归一化，将心电振幅范围归一化到[-1, 1]的区间之内。

## PCA降维

按照前面的定义，数据的维数为2400维。处理高维数据不仅会花费更多的时间，而且模型的泛化能力也受到影响。因此，我首先使用PCA对数据进行降维。为了确定需要下降到多少维，我将数据协方差矩阵的特征值列出如下表。

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Order | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | … |
| Latent | 171.6 | 32.2 | 12.6 | 10.5 | 7.0 | 6.5 | 6.4 | 3.9 | … |

可见数据前10维即可保持大部分能量，因此我选择把数据降到了10维。

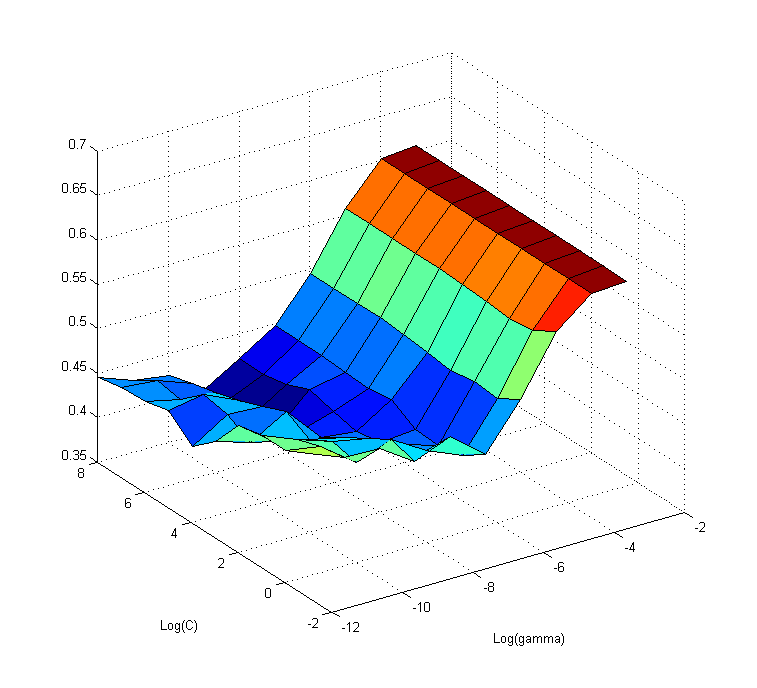
由于PCA降维并不依赖于数据标签，所以在交叉验证之前做PCA是合理的。

## SVM

对于SVM，我使用RBF核，RBF核的和参数作为模型选择的参数。根据LibSvm的文档，我采用指数的方式来搜索合适的参数。使用5重交叉验证，选择不同的和，泛化误差估计结果如下。

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Log Gamma  Log C | -12 | -11 | -10 | -9 | -8 | -7 | -6 | -5 | -4 | -3 |
| -1 | 0.5120 | 0.4960 | 0.4953 | 0.4690 | 0.4857 | 0.4546 | 0.5050 | 0.5710 | 0.6017 | 0.6047 |
| 0 | 0.4983 | 0.4917 | 0.4626 | 0.4770 | 0.4530 | 0.4366 | 0.4900 | 0.5547 | 0.6027 | 0.6057 |
| 1 | 0.4983 | 0.4827 | 0.4686 | 0.4606 | 0.4436 | 0.4313 | 0.4917 | 0.5520 | 0.6021 | 0.6061 |
| 2 | 0.4927 | 0.4630 | 0.4543 | 0.4456 | 0.4286 | 0.4326 | 0.4870 | 0.5534 | 0.6024 | 0.6061 |
| 3 | 0.4576 | 0.4456 | 0.4690 | 0.4283 | 0.4246 | 0.4486 | 0.4907 | 0.5560 | 0.6024 | 0.6061 |
| 4 | 0.4350 | 0.4596 | 0.4580 | 0.4179 | 0.4296 | 0.4450 | 0.4943 | 0.5574 | 0.6024 | 0.6061 |
| 5 | 0.4583 | 0.4660 | 0.4480 | 0.4009 | 0.4383 | 0.4453 | 0.4963 | 0.5574 | 0.6024 | 0.6061 |
| 6 | 0.4523 | 0.4436 | 0.4396 | 0.4013 | 0.4266 | 0.4483 | 0.4950 | 0.5574 | 0.6024 | 0.6061 |
| 7 | 0.4513 | 0.4473 | 0.4330 | 0.4026 | 0.4206 | 0.4480 | 0.4950 | 0.5574 | 0.6024 | 0.6061 |
| 8 | 0.4453 | 0.4313 | 0.4253 | 0.3986 | 0.4223 | 0.4480 | 0.4950 | 0.5574 | 0.6024 | 0.6061 |

根据以上结果，可以画出采用RBF核的SVM模型的模型选择曲面。

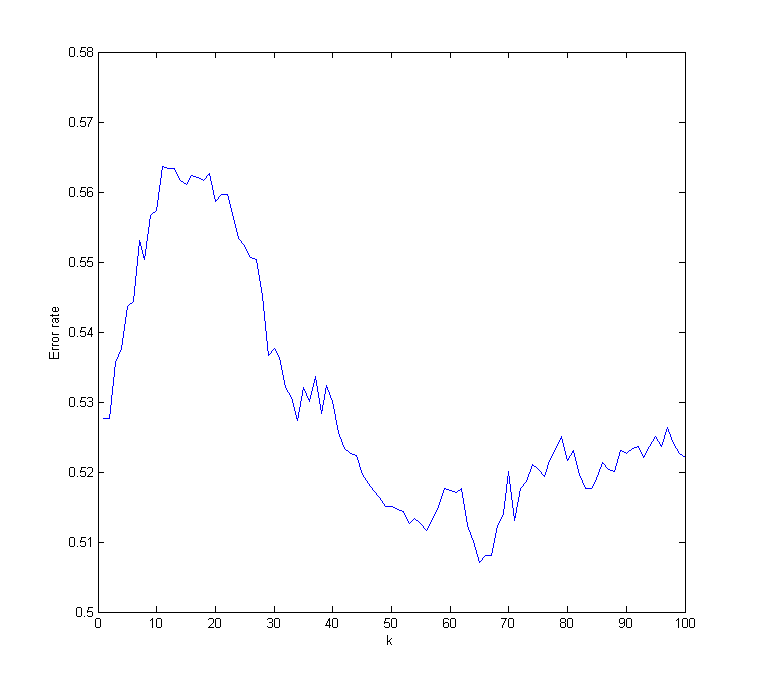


## KNN

对于KNN模型，其复杂度由参数k决定。k值越大，模型复杂度就越小。同样使用5重交叉验证，结果如下表（由于k值范围较大，只展示其中部分值）。

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| K | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 | 100 |
| Error | 0.5574 | 0.5587 | 0.5377 | 0.5300 | 0.5150 | 0.5173 | 0.5200 | 0.5217 | 0.5227 | 0.5220 |

根据以上结果，可以画出KNN模型的模型选择曲线。

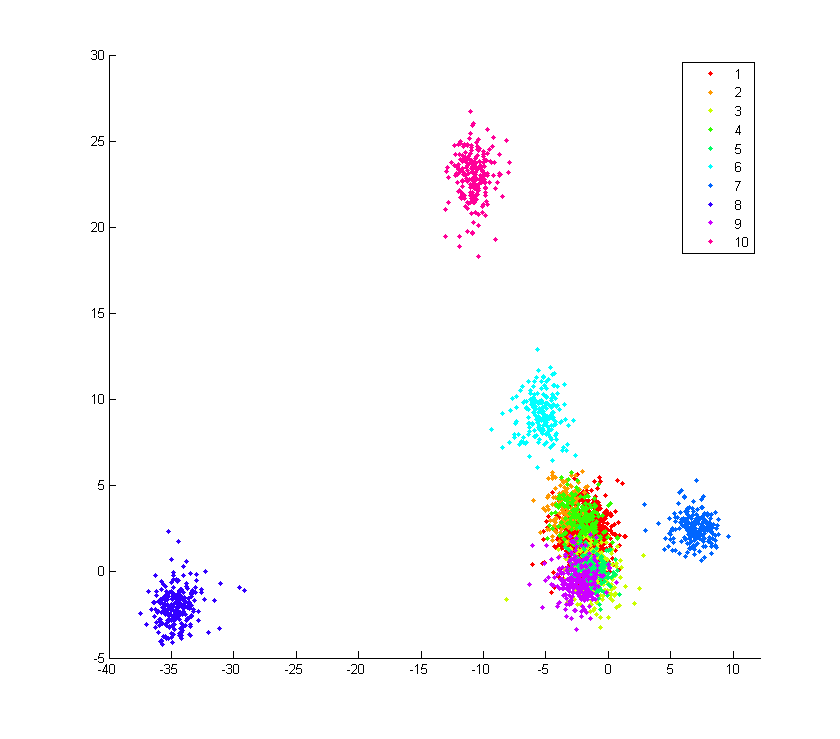


## 小节

从上面的数据可以看出，在选择了适当的参数之后，SVM和KNN都可以取得比较好的结果。PCA降维在其中起到了比较关键的作用。在PCA降维之前，我尝试过直接用SVM分类，效果非常糟糕。

# LDA降维分析

编写了LDA降维的代码之后，我对训练数据整体进行了降维。降到2维之后，根据最优投影矩阵，投影结果如下图。

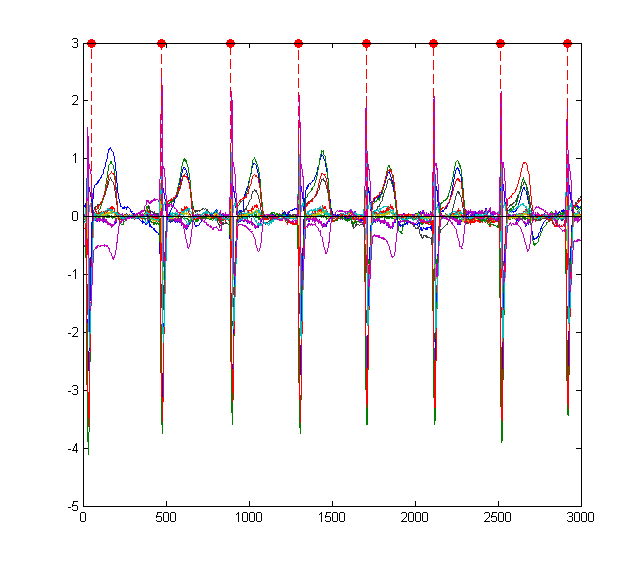


从LDA降维投影后的结果来看，标签6（窦性心律\_提示左室肥厚）、标签7（窦性心律\_完全性右束支传导阻滞）、标签8（窦性心律\_完全性左束支传导阻滞）和标签10（窦性心律\_左室肥厚伴劳损）与其余类别有较好的区分度。

# 测试数据预测

## 心跳切分

训练数据中已将R波波峰标出，然而测试数据中并没有给出R波波峰位置，并且一条测试数据包含了多个心跳。为了确保测试数据与训练数据格式一致，我也将心电数据按R波波峰进行切分。由于准确地检测R波波峰比较困难，所以我采用了一个简单近似，认为在一定时间窗内（50个采样点），心电数据各导差值大于某个阈值（阈值由整段心电数据确定），则波峰在该时间窗内。切分效果如下图。其中红色圆点代表检测出的波峰位置。

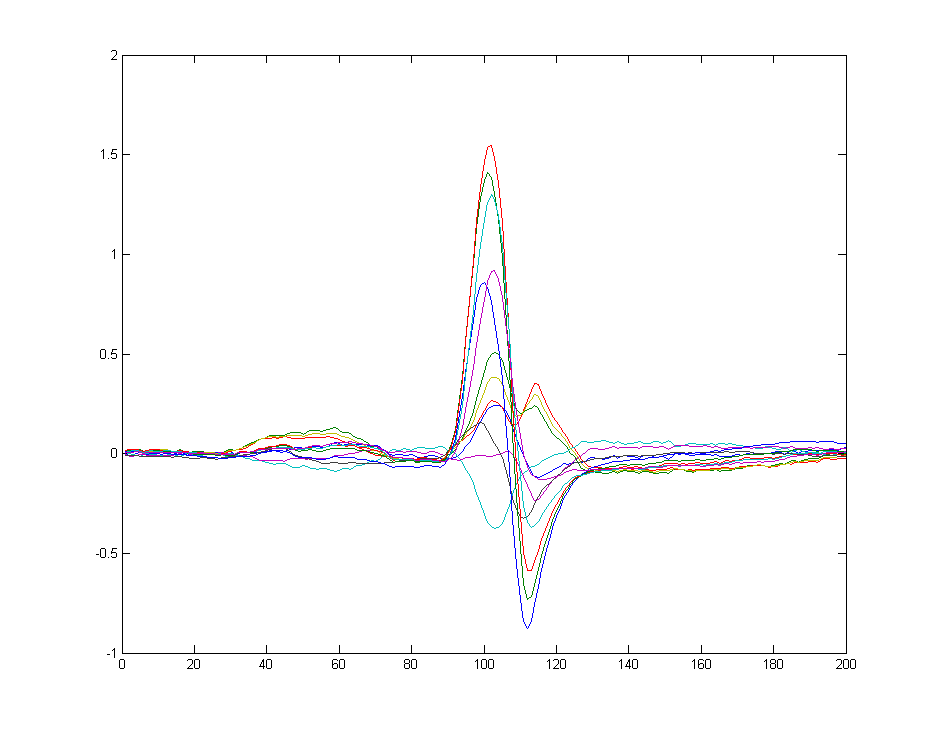


可见该方法能正确地近似检测出R波波峰位置。

## 心电数据融合与平滑

对于一段心电数据，虽然有多个心跳，但对应的标签只有一个。为了便于处理，我将各个心跳的数据进行融合，求均值。这同时也起到了平滑数据的作用。

对某段测试数据进行这样的处理之后，结果如下图所示。



可见这段心电数据已经融合为了一个心跳，并且得到了适当的平滑。

## 测试数据预测

在预测测试数据标签时，我使用了前面已经分析过的SVM模型，在所有训练数据上训练完毕之后，进行预测。预测结果存放在TestReult.mat文件中，其中filenames变量存放测试心跳文件名，labels变量存放预测结果，两者一一对应。预测结果按前面的定义，用数字来标记。