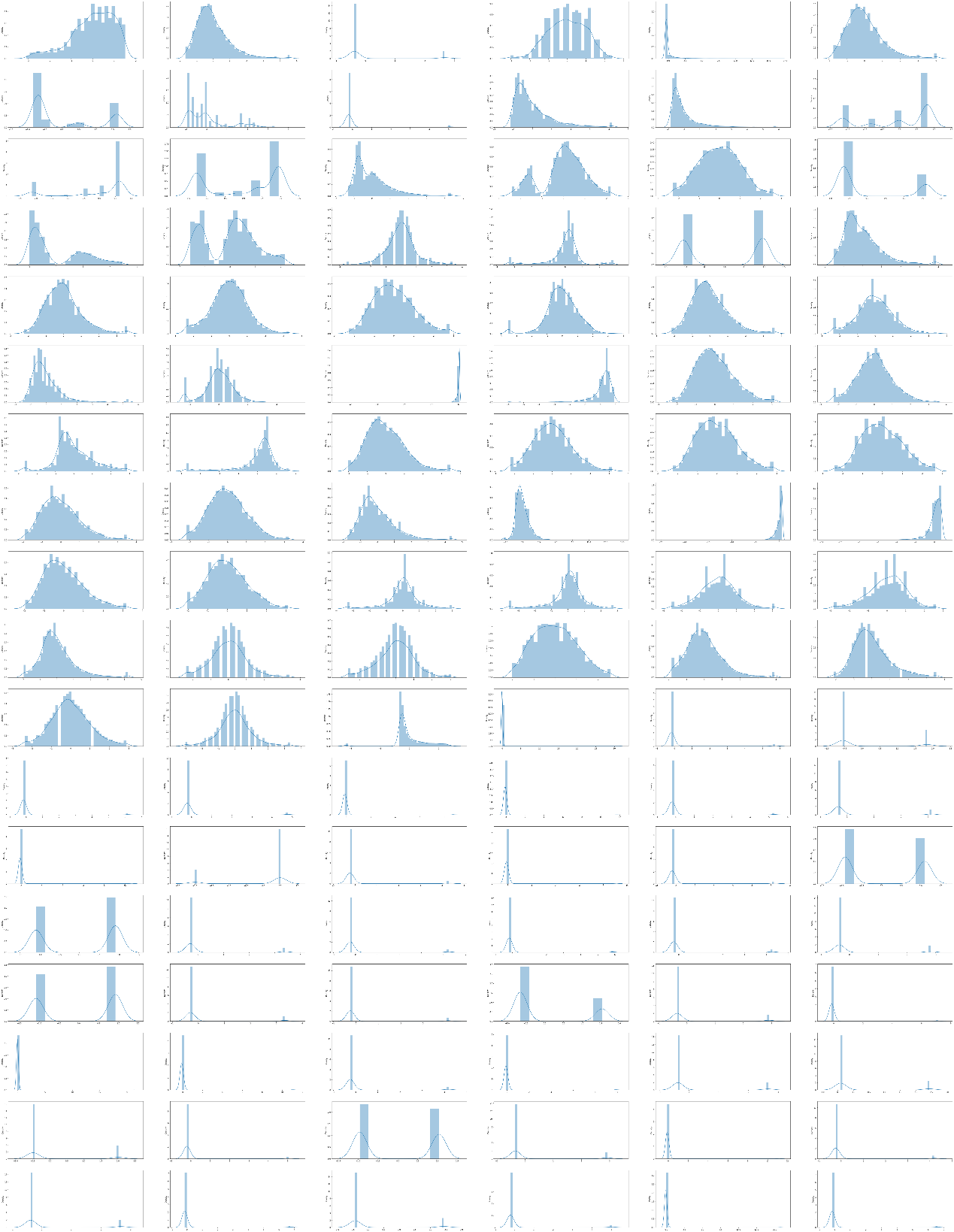
**Ex5 Report**

自硕21 崔晏菲 2021210976

1. 对于训练数据，我首先将他们进行标准化的放缩，即每个维度各自减去均值再除以标准差，这样就使得所有数据均值变为0，标准差变为1。

接着，为了判断每个维度的分布，我将108个维度的频率分布直方图画了出来，如下图所示。



接着使用parzen window方法进行估计。我假设这108个维度各自独立，用高斯核对分布进行了估计。

2. 我使用5折交叉验证，朴素贝叶斯分类器的结果如下：

朴素贝叶斯分类器训练和测试结果：

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Fold** | **train accuracy** | **train sensi** | **train speci** | **valid accuracy** | **valid sensi** | **valid speci** | **test accuracy** | **test sensi** | **test speci** |
| 1 | 0.512500 | 1.000000 | 0.005102 | 0.509000 | 0.994118 | 0.004082 | 0.501367 | 0.998172 | 0.007273 |
| 2 | 0.795000 | 0.759804 | 0.831633 | 0.768000 | 0.707843 | 0.830612 | 0.781222 | 0.751371 | 0.810909 |
| 3 | 0.793750 | 0.762255 | 0.826531 | 0.772000 | 0.745098 | 0.800000 | 0.787603 | 0.765996 | 0.809091 |
| 4 | 0.796500 | 0.754412 | 0.840306 | 0.762000 | 0.711765 | 0.814286 | 0.781222 | 0.751371 | 0.810909 |
| 5 | 0.796000 | 0.761275 | 0.832143 | 0.771000 | 0.788235 | 0.753061 | 0.777575 | 0.758684 | 0.796364 |
| Average | 0.738750 | 0.807549 | 0.667143 | 0.716400 | 0.789412 | 0.640408 | 0.725798 | 0.805119 | 0.646909 |

可见，除了fold-1因为某些原因导致训练结果不理想之外，其余四个fold都表现出了良好的性能。

3. 接下来做最小风险贝叶斯分类器。考虑到把死者误判为活人有错误，但是风险不大；但是把活人误判为死者的错误，造成的后果非常严重，因此得到的风险矩阵如下：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | **预测值** | **活** | **死** |
|  |  | **0** | **1** |
| **真实值** | |  |
| **活** | **0** | | 0 | 10 |
| **死** | **1** | | 1 | 0 |

使用5折交叉验证，朴素贝叶斯分类器的结果如下：

最小风险贝叶斯分类器训练和测试结果：

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Fold** | **train accuracy** | **train sensi** | **train speci** | **valid accuracy** | **valid sensi** | **valid speci** | **test accuracy** | **test sensi** | **test speci** |
| 1 | 0.512500 | 1.000000 | 0.005102 | 0.509000 | 0.994118 | 0.004082 | 0.501367 | 0.998172 | 0.007273 |
| 2 | 0.797000 | 0.823529 | 0.769388 | 0.754000 | 0.762745 | 0.744898 | 0.789426 | 0.813528 | 0.765455 |
| 3 | 0.797000 | 0.825490 | 0.767347 | 0.775000 | 0.819608 | 0.728571 | 0.785779 | 0.815356 | 0.756364 |
| 4 | 0.795250 | 0.814216 | 0.775510 | 0.774000 | 0.792157 | 0.755102 | 0.783045 | 0.802559 | 0.763636 |
| 5 | 0.797500 | 0.825980 | 0.767857 | 0.777000 | 0.831373 | 0.720408 | 0.787603 | 0.822669 | 0.752727 |
| Average | 0.739850 | 0.857843 | 0.617041 | 0.717800 | 0.840000 | 0.590612 | 0.729444 | 0.850457 | 0.609091 |

可以看到，虽然准确率变化不大，但是最小风险贝叶斯分类器的灵敏度提升了很多，这说明我加入的风险矩阵起到了作用。