**实验报告**

实验二

Stage1

假设先验概率和满足高斯分布，由极大似然估计容易得到:

, 

, 

可求得:

, ，

, ，

推导和，如下：





所以，最后得到：









在最大概率贝叶斯决策情况下，the classifying decision boundary function为:



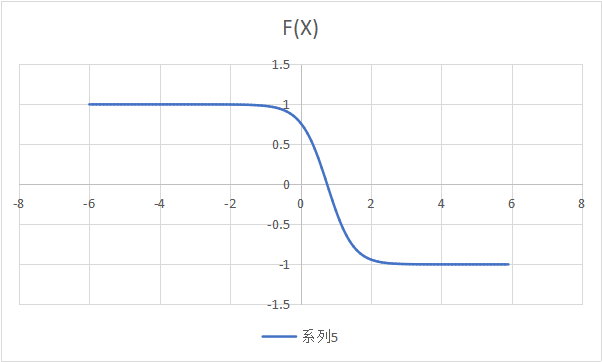
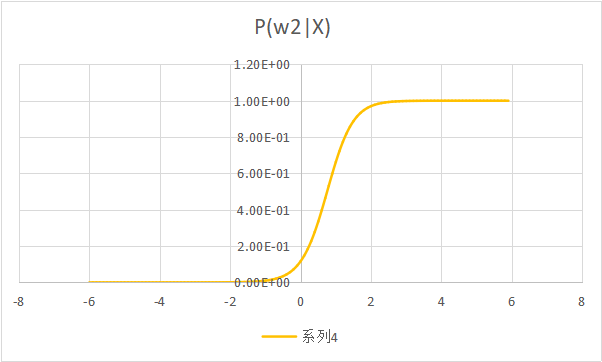
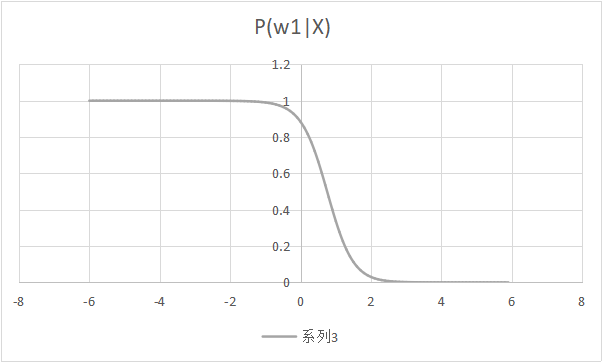
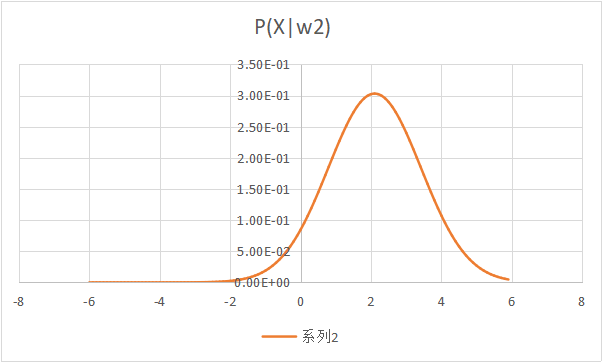
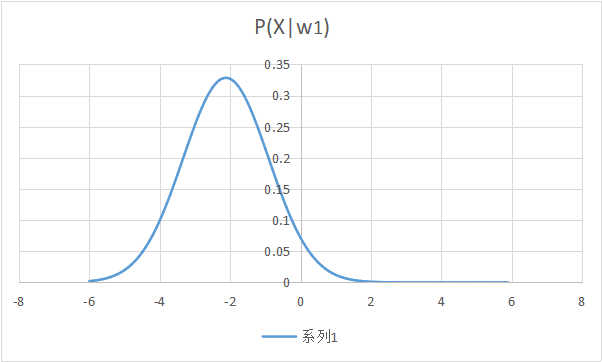
当F(X)>0时，预测为类1；当F(X)<0时，预测为类2

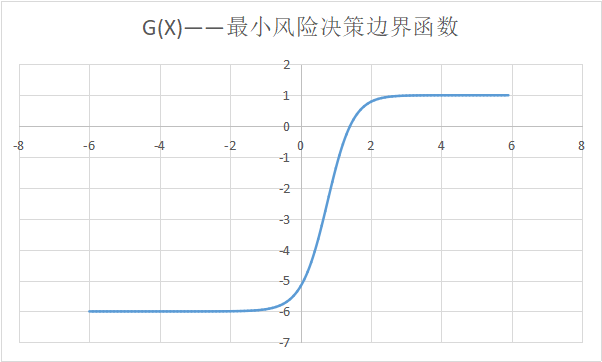
在最大贝叶斯决策情况下，the classifying decision boundary function为:



当G(X)<0时，预测为类1；当G(X)>0时，预测为类2

图像分别为：

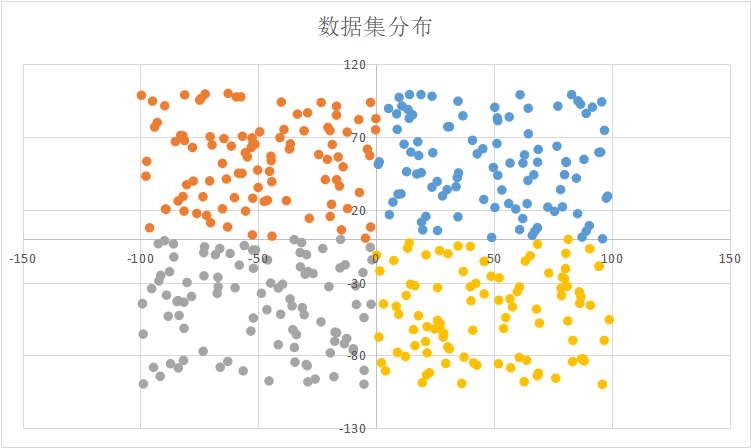




**Stage2**

**训练集**

数据集共400个二维点，横、纵坐标范围均为(-100,100)，但不包括原点。数据点分为4类，各自分别随机分布在坐标轴的第一、二、三、四象限，分别记为类1、类2、类3、类4，如图所示：



假设数据点的X、Y坐标没有相关性，各个类的X、Y坐标满足正态分布，各个类出现的概率分别为15%,20%,30%,35%，损失函数设置如下：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| real class |  | w1 | w2 | w3 | w4 |
| decision  action | a1 | 0 | 1 | 4 | 1 |
| a2 | 1 | 0 | 1 | 4 |
| a3 | 4 | 1 | 0 | 1 |
| a4 | 1 | 4 | 1 | 0 |

(同一象限为0,相邻象限为1，对角象限为4)

同理Stage1中的推导，可以得到:

先验概率：









后验概率：







**按照最大概率贝叶斯决策**：

**边界函数**共有6个，分别为：













当F1(X)>0, F2(X)>0, F3(X)>0时，即P(w1|X)最小时，判定为1类。以此类推。

**按照最小风险贝叶斯决策**：

**风险函数**为:



i = 1, 2, 3, 4

此时**边界函数**也为6个，分别为









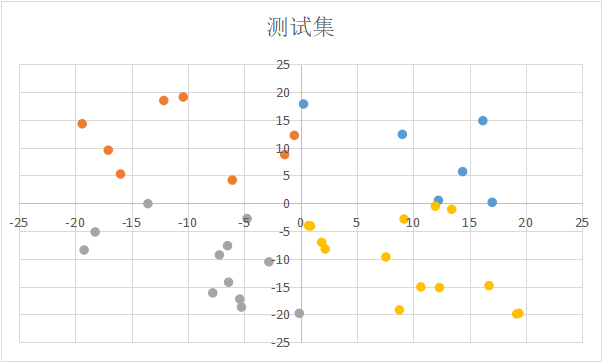




当F1(X)<0, F2(X)<0, F3(X)<0时，即A1(X)最小时，判定为1类。以此类推

**验证集验证**

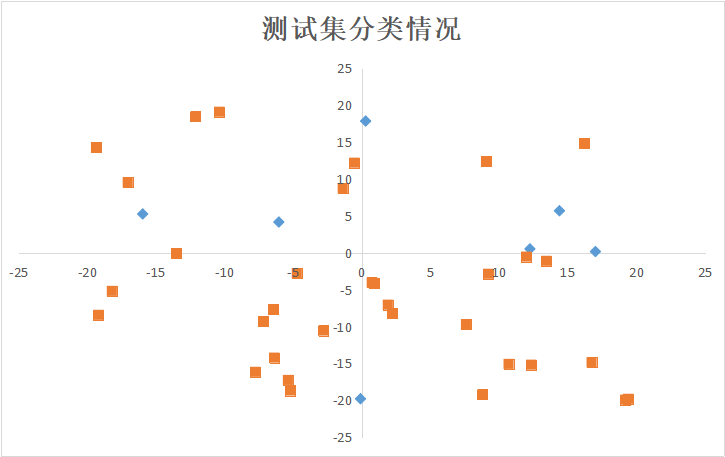
验证集中共包含了40个数据点，属于类1、类2、类3、类4的个数分别为6个、8个、12个、14个，但此时给的点的横、纵坐标范围为(-20,20)。数据点如图所示：



**分类结果（最小风险）**

用已经训练好的贝叶斯分类器（最小风险）进行分类，分类结果如下

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| X坐标 | Y坐标 | 实际分类 | 预测分类 | 预测是否正确 |
| 14.37919463 | 5.678009393 | 1 | 4 | FALSE |
| 9.026792935 | 12.37986487 | 1 | 1 | TRUE |
| 12.22931559 | 0.517321147 | 1 | 4 | FALSE |
| 0.258316126 | 17.83159797 | 1 | 2 | FALSE |
| 16.17613577 | 14.83817783 | 1 | 1 | TRUE |
| 17.00220417 | 0.156698224 | 1 | 4 | FALSE |
| -1.412549273 | 8.737893852 | 2 | 2 | TRUE |
| -19.38856459 | 14.27354326 | 2 | 2 | TRUE |
| -0.562155173 | 12.19721518 | 2 | 2 | TRUE |
| -6.066683417 | 4.154272371 | 2 | 3 | FALSE |
| -17.0775595 | 9.54014891 | 2 | 2 | TRUE |
| -12.14479032 | 18.46227561 | 2 | 2 | TRUE |
| -15.98616365 | 5.24861276 | 2 | 3 | FALSE |
| -10.41634128 | 19.09686445 | 2 | 2 | TRUE |
| -18.23443929 | -5.166519942 | 3 | 3 | TRUE |
| -13.55536655 | -0.075302815 | 3 | 3 | TRUE |
| -7.796296879 | -16.12824491 | 3 | 3 | TRUE |
| -5.252772716 | -18.66175379 | 3 | 3 | TRUE |
| -0.11664591 | -19.78530828 | 3 | 4 | FALSE |
| -19.21645039 | -8.4107992 | 3 | 3 | TRUE |
| -5.406950087 | -17.22690931 | 3 | 3 | TRUE |
| -6.397632602 | -14.20420697 | 3 | 3 | TRUE |
| -4.756984923 | -2.750048794 | 3 | 3 | TRUE |
| -2.817903589 | -10.53715606 | 3 | 3 | TRUE |
| -7.214173551 | -9.296957228 | 3 | 3 | TRUE |
| -6.490116892 | -7.635233877 | 3 | 3 | TRUE |
| 12.32272253 | -15.1590417 | 4 | 4 | TRUE |
| 2.179164662 | -8.231636713 | 4 | 4 | TRUE |
| 0.701142296 | -4.03124806 | 4 | 4 | TRUE |
| 16.7192594 | -14.82255017 | 4 | 4 | TRUE |
| 8.767727601 | -19.18812067 | 4 | 4 | TRUE |
| 7.566620462 | -9.66485116 | 4 | 4 | TRUE |
| 0.883992512 | -4.090375824 | 4 | 4 | TRUE |
| 19.3732331 | -19.78322866 | 4 | 4 | TRUE |
| 13.40906117 | -1.093179239 | 4 | 4 | TRUE |
| 10.67090777 | -15.05368436 | 4 | 4 | TRUE |
| 1.871556038 | -7.031918964 | 4 | 4 | TRUE |
| 11.95387694 | -0.522522371 | 4 | 4 | TRUE |
| 19.17022926 | -19.8981816 | 4 | 4 | TRUE |
| 9.187640357 | -2.846717788 | 4 | 4 | TRUE |



共有7个点被错分，分类的正确率为82.5%。错误分类的多数点位于分类边界。

**结论**

1. 贝叶斯分类器用到了贝叶斯定理，当样本的数据量越大时，预测的准确率越高。
2. 无论是分为几类，或者是几维数据，都可以用同一的贝叶斯公式解决。
3. 最小风险贝叶斯决策更关注风险的大小，分类的正确率可能比最大概率贝叶斯决策低。