Problem 4: Programming for Error Rate Estimation

4.1 由题知

$p(w1)=p(w2)=0.5, p(x|w1)=N(\mu_1,\sigma_1), p(x|w2)=N(\mu_2,\sigma_2), \mu_1=(-1,0), \mu_2=(1,0), \sigma_1=(1,0;0,1), \sigma_2=(2,0;0,1)$

数值计算贝叶斯错误率。概率密度函数分布如下图 4-1 所示，要计算错误率则需要计算下图 4-2 两部分的体积。采用积分思想，在二维平面上采样 500*500 个小方块，近似计算体积得到理论错误率为 0.1963，若提高采样精度，能进一步提高理论错误率的计算精度达到 0.2。

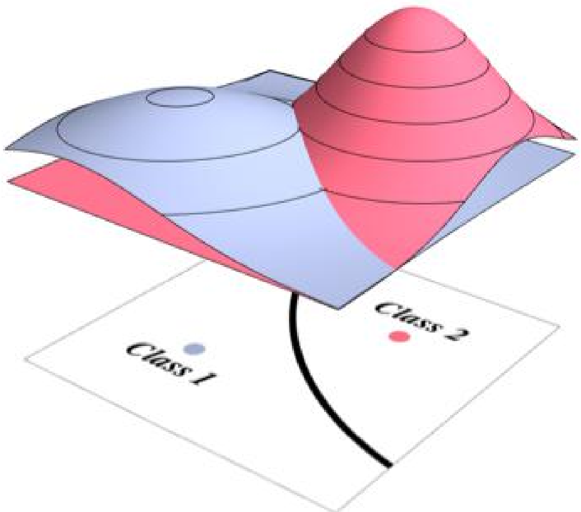

图 4-1：概率密度函数

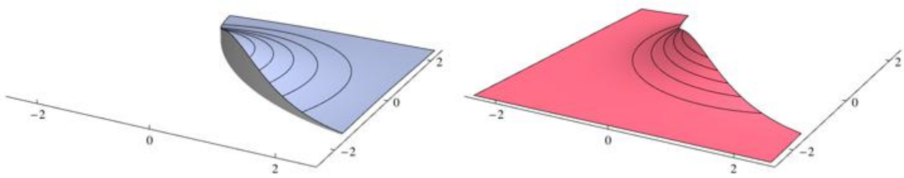

图 4-2：错误率部分

4.2 从 $p(x|w1)=N(\mu_1,\sigma_1), p(x|w2)=N(\mu_2,\sigma_2)$中分别采样 n=10000 个样本，包括标签。将其 9：1 划分为训练集和测试集，用 Parzen 窗法分别估计$p_n(x|w1), p_n(x|w2)$，用其构造一个贝叶斯分类器如图 4-3 表示，在测试集上检测其错误率。其中单位超立方体窗函数重复计算 5 次得到错误率为[0.2005, 0.21,0.2125,0.2085,0.2045]，均值为 0.2072，方差为 2.245E-5. 高斯函数（mu=0，Sigma=I，h=1）重复计算 5 次得到错误率为[0.205, 0.1985, 0.201, 0.2, 0.182]，均值为 0.1973，方差为 7.895E-5。

表 4-1：parzen 窗-超立方体函数错误率

0.2005	0.21	0.2125	0.2085	0.2045	均值 0.2072	方差 2.245E-5

表 4-2：parzen 窗-高斯函数 h=1 错误率

0.205	0.1985	0.201	0.2	0.182	均值 0.1973	方差 7.895E-5

$$\text{若 } l(\vec{x}) = \frac{P(\vec{x}|\omega_1)}{P(\vec{x}|\omega_2)} \begin{array}{c} > \\ < \end{array} \frac{P(\omega_2)}{P(\omega_1)}, \text{则} \vec{x} \epsilon \begin{cases} \omega_1 \\ \omega_2 \end{cases}$$

图 4-3：贝叶斯分类准则

与 4.1 中最优分类器理论错误率值比较，两者错误率差不多，parzen 估计法略大于理论值。这是因为前者只包括贝叶斯误差，后者除了贝叶斯误差，还有模型误差和估计误差。

4.3 从 4.2 可知，高斯窗函数比单位超立方体窗函数估计效果更好。现在改变高斯窗函数的参数（mu=0，Sigma=I，h=4），得到错误率为[0.2115, 0.2055, 0.198, 0.2015, 0.1965]，均值为 0.2026，方差为 3.68E-05；高斯窗函数的参数（mu=0，Sigma=I，h=0.5），得到错误率为[0.1835, 0.192, 0.221, 0.2, 0.1915]，均值为 0.198，方差为 2E-04。由此分析可知，在待估计的模型为高斯分布的情况下，选用准确的高斯窗函数能够降低模型误差。在此基础上，要选择适当的参数，本题中二维问题采样 100*100 给定的情况下，h=1 比 h=0.5, 4 更好，因为 h=4 时均值大，h=0.5 时方差大不稳定。

表 4-3：parzen 窗-高斯函数 h=4 错误率

0.2115	0.2055	0.198	0.2015	0.1965	均值 0.2026	方差 3.68E-5

表 4-4：parzen 窗-高斯函数 h=0.5 错误率

0.1835	0.192	0.221	0.2	0.1915	均值 0.198	方差 2E-4

4.4 从混合高斯分布$p(x) = p(w1) * p(x|w1) + p(w2) * p(x|w2)$中采 2n=20000 的样本，无标签。采用 EM 算法估计$\mu_1, \mu_2, \sigma_1, \sigma_2$，得到$p_{2n}(x|w1), p_{2n}(x|w2)$。由 4.5 错误率可知，采用 EM 算法估计得到的$p_{2n}(x|w1), p_{2n}(x|w2)$更准确。理论有参数估计（EM 算法）比无参数估计（parzen 窗）用到了更多的先验知识，所以估计也会更准确。

4.5 4.4 估计得到的$p_{2n}(x|w1), p_{2n}(x|w2)$, 再通过 4.2 中同样的贝叶斯分类器, 得到错误率为[0.199, 0.2035, 0.2005, 0.2015, 0.195]，均值 0.1999，方差 1.0175E-05. 非常接近最优贝叶斯分类器的理论值，同时可以观察到 EM 算法的错误率在几种方法中方差最小。

表 4-4：EM 算法错误率

0.199	0.2035	0.2005	0.2015	0.195	均值 0.1999	方差 1.0175E-5

4.6 综上，EM 算法的估计方法和相关的有先验知识的贝叶斯分类器效果最好。且采样的时候不需要标签，无监督学习对样本的要求更低，实际操作中更方便。

Problem 5: Programming for Perceptron Algorithm

5.1 用 sklearn 的 make_blobs 包生成两簇聚类数据共 200 个，一簇标为 1，另一簇标为-1，并保证这两簇数据是线性可分的。分布如图 5-1 所示。

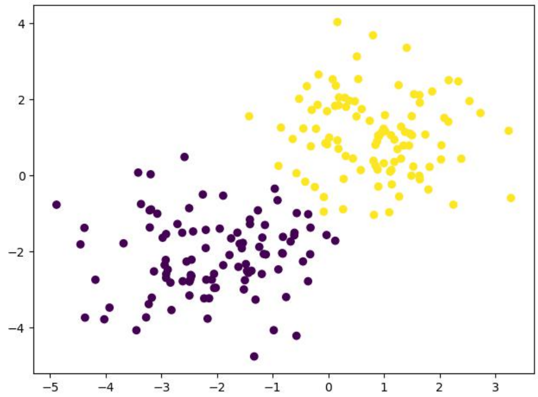

图 5-1：2D 数据点分布图

5.2 采用经典感知器算法，在数据上画出分界线，如图 5-2 所示。

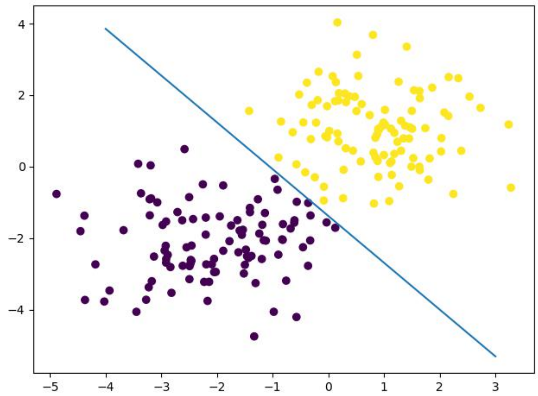

图 5-2：经典线性感知器算法

5.3 采用 margin 线性感知器算法，在数据上画出分界线，如图 5-3 所示。其中蓝线为图 5-2 中经典感知器的分界线，红色为 gamma=1 时分界线，绿色为 gamma=10 时分界线，继续增大 gamma 产生的分解线与绿线基本重合，肉眼难以区分。由此，在一定范围里 gamma 增大使得分界线往更优的方向偏转，即分类的两边到分界线的距离更接近。gamma 太大的时候会导致算法不收敛。程序运行时间也随 gamma 增大而增大。

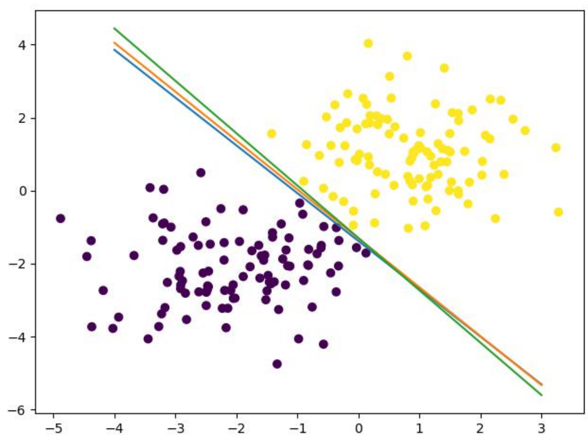

图 5-3：margin 线性感知器算法