模式识别导论第二次作业——编程题

杨登天-202028015926089-微电子研究所

1. 现有一维空间的 50 个样本点（实际上，这些样本点是在 Matlab 中按如下语句生成的：mu=5; std_var = 1; X=mvnrnd(mu, std_var, 50);）。现需要采用 Parzen 窗方法对概率密度函数进行估计。请分别编程实现方窗和高斯窗情形下的概率密度函数估计；请讨论窗宽的影响，并画出几种不同窗宽取值下所估计获得的概率密度函数曲线。

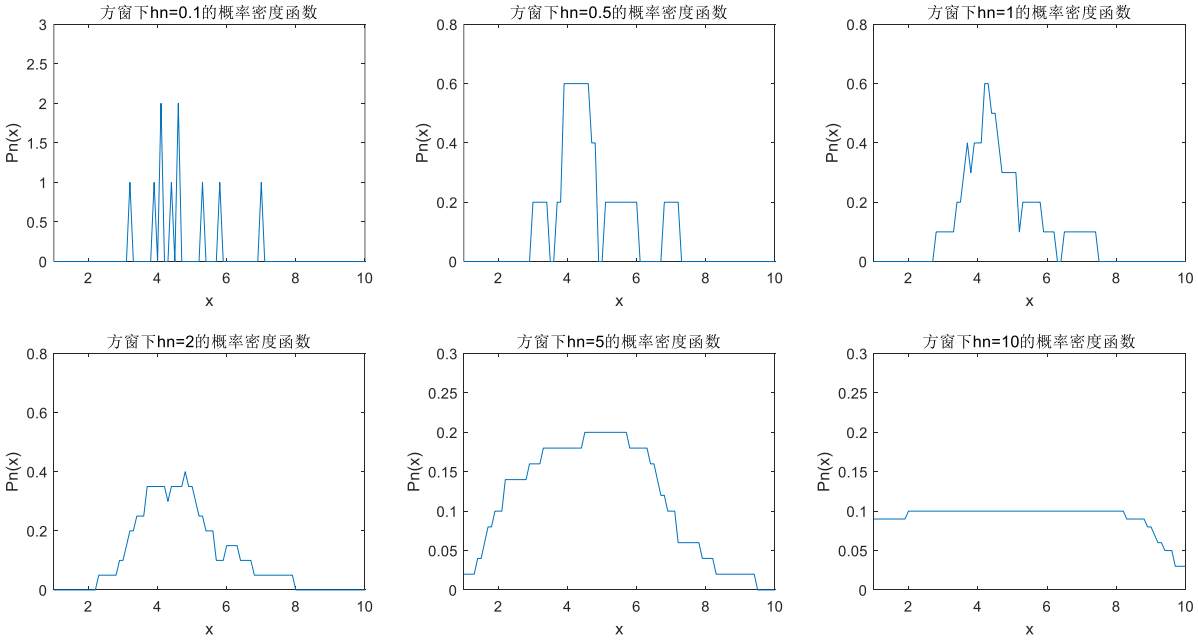
   解：首先根据方窗表达出概率密度函数

$$p_n(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_n}, |x_i - x| \le \frac{h_n}{2}$$

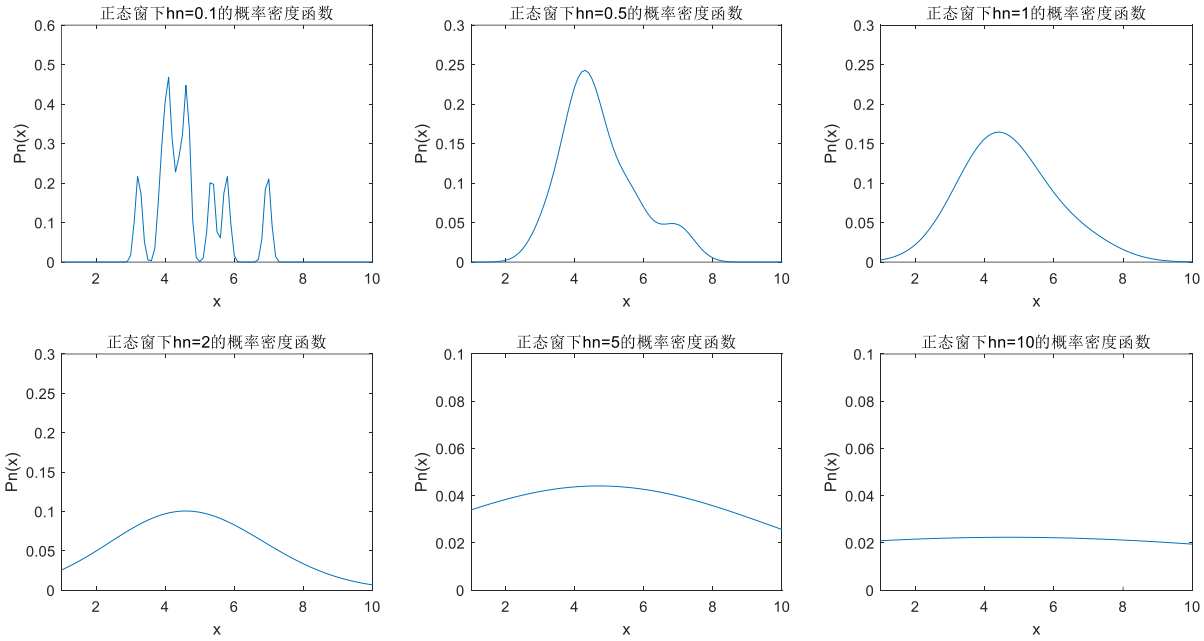
   其次根据高斯窗表达出概率密度函数

$$p_n(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_n}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x_i-x}{h_n}\right)^2}$$

   根据 MATLAB 求出的方窗情况下概率密度函数随着h_n的变化情况如下



   根据 MATLAB 求出的正态窗情况下概率密度函数随着h_n的变化情况如下



随着窗宽的变化发现，无论是方窗还是正态窗，当窗宽过小时，概率密度函数多尖峰，过拟合；当窗宽过大时，概率密度函数平滑，欠拟合。

***特别说明一点：因为考虑到在概率密度函数的求解是需要归一化，但是上述两个函数积分的求解需要遍历坐标轴。现以正态窗为例予以推导

$$p_n(x)=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_n}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x_i-x}{h_n}\right)^2}$$

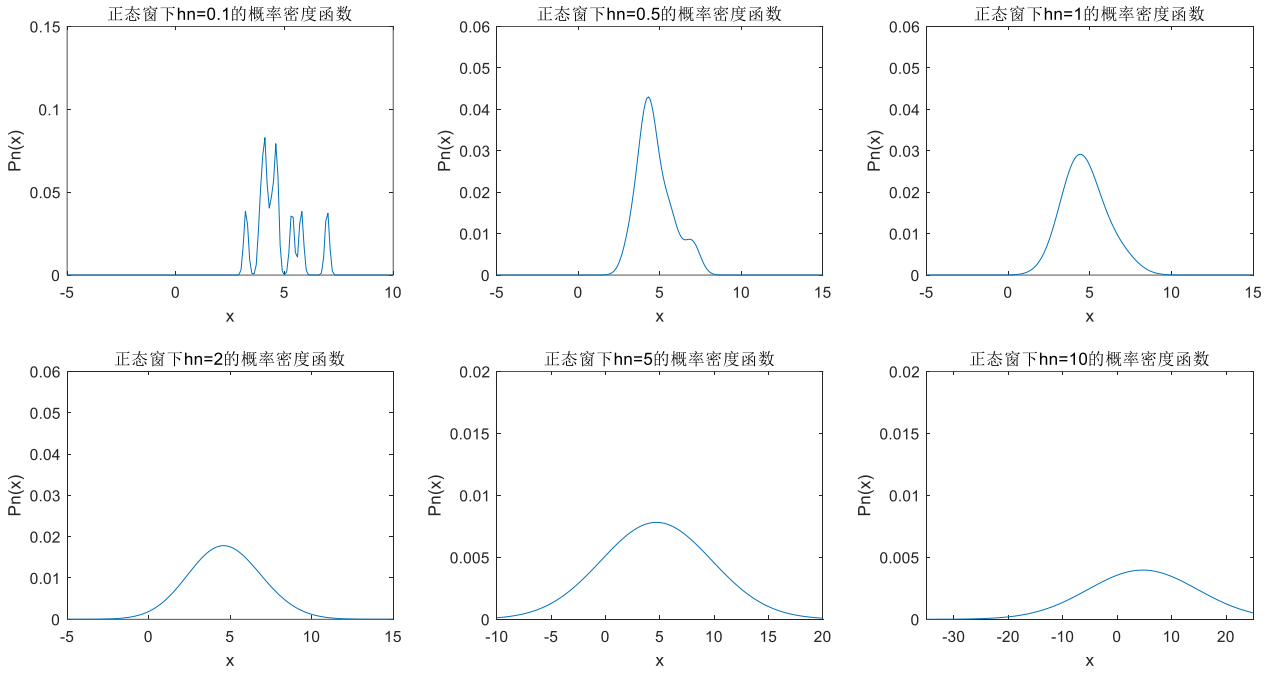
$$p_{n\_normalize}(x)=\frac{p_n(x)}{\int p_n(x)\,dx}$$

但在计算机计算中，以离散形式表达，则有

$$p_{n_{normalize}}(x_j)=\frac{p_n(x_j)}{\sum_{j=1}^{N}p_n(x_j)}=\frac{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_n}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x_i-x_j}{h_n}\right)^2}}{\sum_{j=1}^{N}\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_n}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x_i-x_j}{h_n}\right)^2}}$$

$$=\frac{\sum_{i=1}^{n}\frac{1}{h_n}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x_i-x_j}{h_n}\right)^2}}{\sum_{j=1}^{N}\sum_{i=1}^{n}\frac{1}{h_n}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x_i-x_j}{h_n}\right)^2}}，$$

得到如下图的正态窗概率密度函数



从上图中可以看出在趋势上类似概率密度函数和概率密度函数无差异，均遵循：随着窗宽的变化发现，无论是方窗还是正态窗，当窗宽过小时，概率密度函数多尖峰，过拟合；当窗宽过大时，概率密度函数平滑，欠拟合。

解答完毕

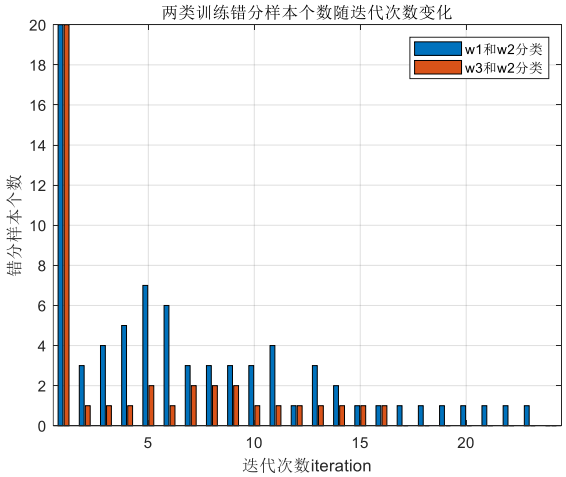
2. 本题关于线性分类器的构造与训练，所使用的四类二维样本（共 40 个）如下：

   2.1 子问题 11 编写程序实现批感知器算法，以$a = 0$作为起始的解向量，并训练ω_1和ω_2，记录下收敛的步数。

   2.2 子问题 12 运用程序训练ω_3和ω_2，同样记录下收敛的步数。

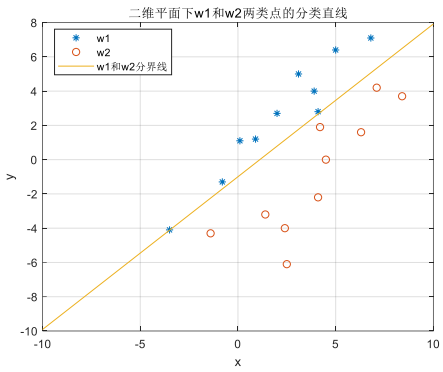
   2.3 子问题 2 请写一个程序，实现 MSE 多类扩展方法，每一类用前 8 个样本构造分类器，用后两个样本做测试，给出正确率。

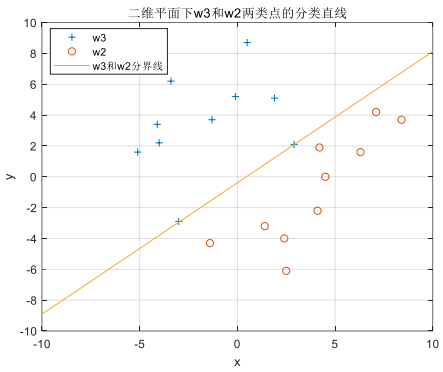
   解答：关于子问题 11 和子问题 12，实现原理相同，根据批感知器伪代码实现，得到如下训练结果。该结果基于$\eta_k = 1$



可以得到关于ω_1和ω_2与ω_3和ω_2的训练次数分别为 24 和 17，并得到对应的解向量分别

为$a=(-30.4,34.1,34)^T$和$a=(-41.4,48.6,19)^T$
现从平面划分的角度直观观察点的分类情况。





其中平分ω_1和ω_2的直线方程为

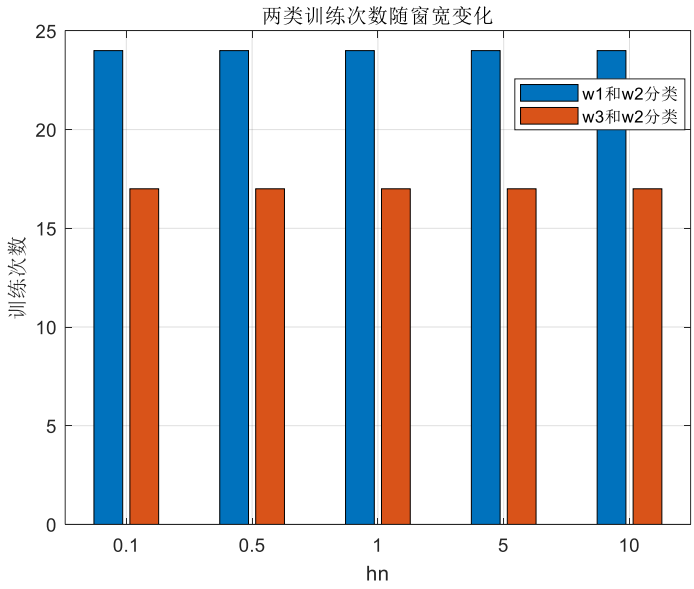
$$Y=\frac{30.4}{34.1}X-\frac{34}{34.1}$$

关于靠近直线的ω_1的两点坐标分别为$(-3.5,-4.1)$和$(4.1,2.8)$，而实际落在直线上的点坐标分别为$(-3.5,-4.1173)$和$(2.9,2.6581)$，因此靠近直线的ω_1两点并未落在直线上，而是在直线上方。

其中平分ω_3和ω_2的直线方程为

$$Y=\frac{41.4}{48.6}X-\frac{19}{48.6}$$

关于靠近直线的ω_3两点坐标分别为$(-3.0,-2.9)$和$(2.9,2.1)$，而实际落在直线上的点坐标分别为$(-3.0,-2.9465)$和$(2.9,2.0794)$，因此靠近直线的ω_3两点并未落在直线上，而是在直线上方。
接下来尝试考虑不同η_k的情况下，两类样本的收敛次数变化。



可以从理论证明训练次数与窗宽长度无关，因为$a=0$为起始训练向量，假设有任意两个非零η_{k1}和η_{k2}，那么对于第一次所有样本测试均得到错分类，所以有

$$a_1 = 0 + \eta_{k1}(\sum_{i=1}^{10} y_i)_1 = \eta_{k1}(\sum_{i=1}^{10} y_i)_1$$

$$a_2 = 0 + \eta_{k2}(\sum_{i=1}^{10} y_i)_2 = \eta_{k2}(\sum_{i=1}^{10} y_i)_2$$

那么用这两个得到a_1和a_2进一步判断错分类样本，比如

$$g_1(y_i) = {a_1}^T y_i$$

$$g_2(y_i) = {a_2}^T y_i = \frac{\eta_{k2}}{\eta_{k1}} {a_1}^T y_i = \frac{\eta_{k2}}{\eta_{k1}} g_1(y_i)$$

在η_{k1}和η_{k2}均为正数的情况下，a_1和a_2对于样本的错分判断$g(y_i)$仅仅只是倍数上的差异而不存在正负号的差异，因此不存在错分样本的差异，所以在进一步通过错分样本更新a_1和a_2时，得到

$$a_1 = 0 + \eta_{k1}(\sum_{i=1}^{10} y_i)_1 + \eta_{k1}(\sum_{i=1}^{10} y_i, y_i \in error)_2$$

$$= \eta_{k1}[\left(\sum_{i=1}^{10} y_i\right)_1 + \left(\sum_{i=1}^{10} y_i, y_i \in error\right)_2]$$

$$a_2 = 0 + \eta_{k2}(\sum_{i=1}^{10} y_i)_1 + \eta_{k2}(\sum_{i=1}^{10} y_i, y_i \in error)_2$$

$$= \eta_{k2}[\left(\sum_{i=1}^{10} y_i\right)_1 + \left(\sum_{i=1}^{10} y_i, y_i \in error\right)_2]$$

进一步分析可以得到，直至无错分样本，必然得到如下关系

$$a_2 = \frac{\eta_{k2}}{\eta_{k1}} a_1$$

因此可以得到如下结论：

1) 当以$a = 0$为起始训练向量，那么无论取非零非负的η_k均得到相同的训练次数；
2) 当以$a = 0$为起始训练向量，那么取非零非负的η_k之间最后得到的解向量不同，但解向量之间的比率与对应η_k之间的比率完全相同；
3) 当以$a = 0$为起始训练向量，那么取非零非负的η_k之间最后得到的判决平面完全相同（三维增广坐标），投影得到的判决直线方程完全相同(二维坐标)。

关于子问题 2

采用 MSE 多分类方法，先确定判决矩阵$\widehat{W}$

$$\widehat{W} = (\hat{X}\hat{X}^T)^{-1}\hat{X}Y^T$$

其中Y是 4×8 的矩阵，$\hat{X}$是 3×32 的矩阵，得到判决矩阵以后检测其正确性

$$\hat{X}_{result} = \widehat{W}^T\hat{X}_{test}$$

其中$\hat{X}_{test}$是 3×8 的矩阵，$\hat{X}_{result}$是 4×8 的矩阵，之后只需要按照

$$if \quad j = \arg\max(\hat{X}_{result}(:, i)), \quad then\ x_{test\_i} \in j$$

因此通过 MATLAB 直接数值求解矩阵，并根据上述判决准则得到如下表的结果

测试样本	样本 1	样本 2	样本 3	样本 4	样本 5	样本 6	样本 7	样本 8
实际分类	1	1	2	2	3	3	4	4
测试分类	1	1	2	2	3	3	4	4

可以发现测试分类的结果与实际分类结果完全相同，因此正确率为 100%。

解答完毕