模式识别导论第二次作业——计算题

杨登天-202028015926089-微电子研究所

1. 设一维特征空间中的窗函数$\varphi(u)=\frac{1}{\sqrt{2\pi}}e^{-\frac{u^2}{2}}$，有$n$个样本$x_i, i=1,2\ldots,n$，采用宽度为$h_n$的窗函数，请写出概率密度函数$p(x)$的$Parzen$窗估计$p_n(x)$。

   解：将考察样本x_i放在以x为中心、以h_n为棱长的超立方体内，则落在该超立方体内的样本点个数k_n为

$$k_n=\sum_{i=1}^{n}\varphi(\frac{x-x_i}{h_n})$$

   则根据离散样本点的概率密度函数$p(x)$的$Parzen$窗估计$p_n(x)$表达式并且由于一维空间

$$p_n(x)=\frac{k_n}{nV_n}=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{V_n}\varphi(\frac{x-x_i}{h_n})=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_n}\varphi(\frac{x-x_i}{h_n})=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_n}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{x-x_i}{h_n})^2}$$

   其中一维空间下，$V_n=h_n$。

***但是严格意义上需要证明窗函数在全空间内的积分为 1(窗函数>0 是明显的，故而不再证明)下面给出全空间积分值为 1 的证明。

$$I=\int_{-\infty}^{+\infty}\frac{1}{h_n}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-x_i}{h_n}\right)^2}d(x-x_i)=2\int_{0}^{+\infty}\frac{1}{h_n}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x}{h_n}\right)^2}dx=2\int_{0}^{+\infty}\frac{1}{\sqrt{\pi}}e^{-t^2}dt$$
$$=\frac{2}{\sqrt{\pi}}\Gamma\left(\frac{1}{2}\right)=1$$

   其中$t=\frac{x}{h_n\sqrt{2}}$，$\Gamma\left(\frac{1}{2}\right)=\int_0^{+\infty}e^{-t^2}\times t^0dt=\frac{\sqrt{\pi}}{2}$。

   解答完毕

2. 给定一维空间三个样本点{-4,0,6}，请写出概率密度函数$p(x)$的最近邻(1-NN)估计，并画出概率密度函数曲线图。

   解：根据一维情况下的概率密度函数$p(x)$的(k-NN)最近邻估计$p_n(x)$

$$p_n(x)=\frac{k_n}{nV_n}=\frac{k_n}{2^d n\prod_{i=1}^{d}|x^i-x_{KNN}^i|}$$

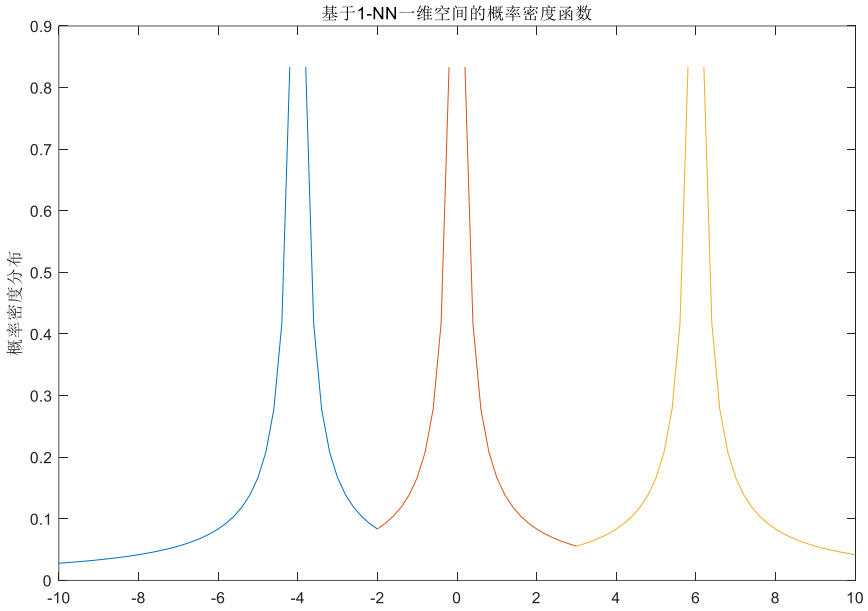
   那么很自然根据要求，当$K=1$时，$p_n(x)$有如下表达式

$$p_n(x)=\begin{cases}\dfrac{1}{2\times3\times|x-(-4)|}=\dfrac{1}{6|x+4|} & x<-2\\ \dfrac{1}{2\times3\times|x-0|}=\dfrac{1}{6|x|} & -2<x<3\\ \dfrac{1}{2\times3\times|x-6|}=\dfrac{1}{6|x-6|} & x>3\end{cases}$$

   通过 MATLAB 画出的概率密度函数曲线图如下

***需要特别注意的是：概率密度函数本该满足全空间积分为 1，但是由于可近似表达为概率密度函数$p(x)$的最近邻估计$p_n(x)$存在第二类积分间断点，导致全空间积分不收敛。因此从严格意义上来讲，该最近邻估计$p_n(x)$并非满足概率密度函数的定义(未归一化)。

   但是此处利用最近邻估计函数$p_n(x)$的目的在于大致估计坐标轴上所有点的类属及其被分配到该类的概率，因此也可也勉强作为概率密度函数。



解答完毕

3. 现有 7 个二维向量:$x_1=(1,0)^T, x_2=(0,1)^T, x_3=(0,-1)^T, x_4=(0,0)^T, x_5=(0,2)^T, x_6=(0,-2)^T, x_7=(-2,0)^T$。这里上标$T$表示向量转置。假定前三个为$\omega_1$类，后四个为$\omega_2$类。画出最近邻法决策面。

   解：从点分布可以知道，**判别直线是任意两点的垂直平分线**，因为根据最近邻的原理，将平面上所有点分配到距离上述 7 个二维向量最近的二维向量所在的类别。

   现依次表达出构成判别边界的线段斜率及必过点。

   分析x_1、x_4——直线方程为$x=\frac{1}{2}$，必定经过$(\frac{1}{2},0)$;

   分析x_1、x_5——直线方程为$y=\frac{1}{2}\left(x-\frac{1}{2}\right)+1$，必定经过$(\frac{1}{2},1)$;

   分析x_1、x_6——直线方程为$y=-\frac{1}{2}\left(x-\frac{1}{2}\right)-1$，必定经过$(\frac{1}{2},-1)$;

   分析x_2、x_4——直线方程为$y=\frac{1}{2}$，必定经过$(0,\frac{1}{2})$;

   分析x_2、x_5——直线方程为$y=\frac{3}{2}$，必定经过$(0,\frac{3}{2})$;

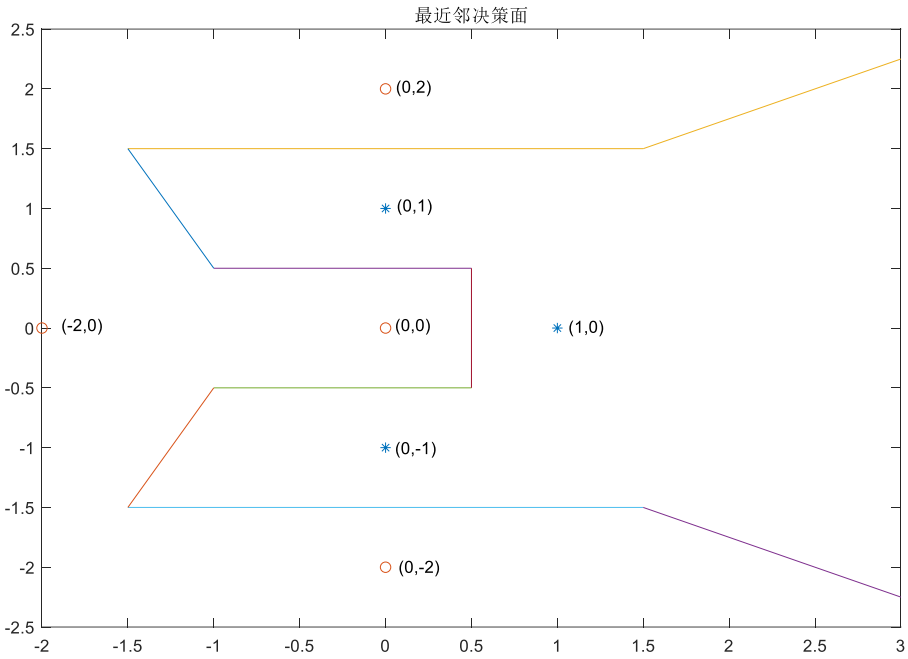
   分析x_2、x_7——直线方程为$y=-2(x+1)+\frac{1}{2}$，必定经过$(-1,\frac{1}{2})$;

   分析x_3、x_4——直线方程为$y=-\frac{1}{2}$，必定经过$(0,-\frac{1}{2})$;

   分析x_3、x_6——直线方程为$y=-\frac{3}{2}$，必定经过$(0,-\frac{3}{2})$;

   分析x_3、x_7——直线方程为$y=2(x+1)-\frac{1}{2}$，必定经过$(-1,-\frac{1}{2})$;

   因此由 MATLAB 求解并通过生成图可鼠标点击确定交点，进一步得到



其中图中显示的折线即为分离两类的判决线，其转折坐标点可以依次表达出，由上至下，由右至左为

$$(1.5,1.5)\ (-1.5,1.5)\ (-1,0.5)\ (0.5,0.5)\ (0.5,-0.5)$$
$$(-1,-0.5)\ (-1.5,-1.5)\ (1.5,-1.5)$$

解答完毕

4. 请给出 K 近邻分类器的优点和缺点。

解：优点和缺点如下所述

优点：

程序的易编辑性——算法理论简单易懂，程序容易实现；

算法的鲁棒性——准确性高，对异常值和噪声有较高的容忍程度；

算法的易操作性——支持多分类，而无需计算复杂的概率密度函数。

缺点：

不可重复利用性——每增加一个样本点，需要利用所有样本点重新构建决策面；

算法误差受分布制约——样本分布不平衡时，分类模型的误差大；

算法欠/过拟合——K 值过大时，容易欠拟合，K 值过小，容易过拟合。

解答完毕

5. 现有四个来自于两个类别的二维空间中的样本，其中第一类的两个样本为$(1,4)^T$和$(2,3)^T$，第二类的两个样本为$(4,1)^T$和$(3,2)^T$。这里上标T表示向量转置。若采用规范化增广样本表示形式，并假设初始的权向量$a=(0,1,0)^T$，其中向量a的第三维对应于样本的齐次坐标。同时，假定梯度更新步长η_k固定为 1。试利用批处理感知器算法求解线性判别函数$g(x)=a^Ty$的权向量a。（注：“规范化增广样本表示”是指对齐次坐标表示的样本进行规范化处理）。

解：首先对四个样本坐标规范化，并全部转变为第一类依次得到

$$y_1=(1,4,1)^T\quad y_2=(2,3,1)^T\quad y_3=-(4,1,1)^T\quad y_4=-(3,2,1)^T$$

其中转化方式为$y=(x^T,1)^T$

接下来用批处理感知器算法求解线性判别函数

第一次、 考虑错分的情况，需要依次求解判别函数数值

$$g(x_1) = a^T y_1 = (0,1,0) \cdot (1,4,1)^T = 4 > 0$$
$$g(x_2) = a^T y_2 = (0,1,0) \cdot (2,3,1)^T = 3 > 0$$
$$g(x_3) = a^T y_3 = -(0,1,0) \cdot (4,1,1)^T = -4 < 0$$
$$g(x_4) = a^T y_4 = -(0,1,0) \cdot (3,2,1)^T = -2 < 0$$

根据更新准则，有

$$a_1^T = a^T + \eta_k \sum_{y \epsilon Y_k} y = a^T + \eta_k[-(4,1,1)^T - (3,2,1)^T] = (-7,-2,-2)$$

第二次、 考虑错分的情况，需要依次求解判别函数数值

$$g(x_1) = a_1^T y_1 = -(7,2,2) \cdot (1,4,1)^T = -17 < 0$$
$$g(x_2) = a_1^T y_2 = -(7,2,2) \cdot (2,3,1)^T = -22 < 0$$
$$g(x_3) = a_1^T y_3 = -(7,2,2) \cdot [-(4,1,1)^T] = 32 > 0$$
$$g(x_4) = a_1^T y_4 = -(7,2,2) \cdot [-(3,2,1)^T] = 27 > 0$$

根据更新准则，有

$$a_2^T = a_1^T + \eta_k \sum_{y \epsilon Y_k} y = a_1^T + \eta_k[(1,4,1)^T + (2,3,1)^T] = (-4,5,0)$$

第三次、 考虑错分的情况，需要依次求解判别函数数值

$$g(x_1) = a_2^T y_1 = (-4,5,0) \cdot (1,4,1)^T = 16 > 0$$
$$g(x_2) = a_2^T y_2 = (-4,5,0) \cdot (2,3,1)^T = 7 > 0$$
$$g(x_3) = a_2^T y_3 = (-4,5,0) \cdot [-(4,1,1)^T] = 11 > 0$$
$$g(x_4) = a_1^T y_4 = (-4,5,0) \cdot [-(3,2,1)^T] = 2 > 0$$

可以得知此时没有被错分。

因此，根据批处理感知器算法求解得到的权向量$a = (-4,5,0)^T$

解答完毕