《模式识别》第2 次作业

姓名：谷绍伟　　学号：202418020428007

1　计算和简答题

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 $p_n(x)$。

答：

$$p_n(x)=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{V_n}\varphi\left(\frac{x-x_i}{h_n}\right)$$

$$h_n=\frac{h}{\sqrt{n}}$$

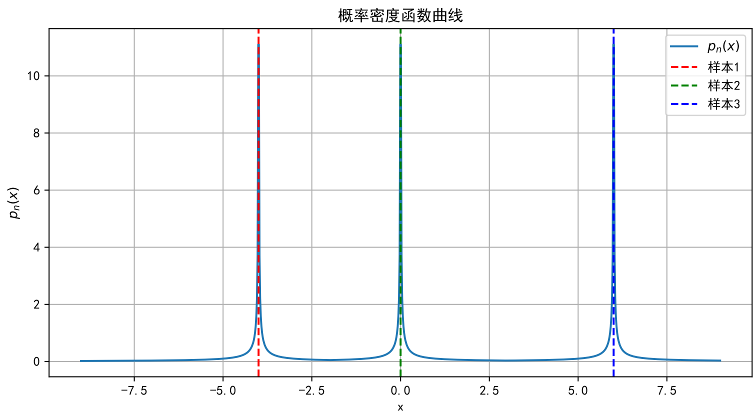
其中，h 代表初始窗宽，是一个可调节的参数，V_n 为超立方体体积。

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。

答： 概率密度函数 $p(x)$ 的最近邻 (1-NN) 估计为：

$$p_n(x)=\frac{k_n}{nV_n}\begin{cases}\frac{1}{10|x+4|}, & if \quad x<-2\\ \frac{1}{10|x|}, & if \quad -4<x<3\\ \frac{1}{10|x-6|}, & if \quad x>3\end{cases}$$

画出的概率密度函数曲线图如下：



3. 针对概率密度估计问题，请简述 EM 算法的基本步骤。

答： EM 算法的任务是对给定的数据集 $X=\{x_1,x_2,\dots,x_n\}$，估计观测数据概率密度的参数。基本要素如下：

1. 观测数据：$X=\{x_1,x_2,\dots,x_n\}$（不完全数据）；
2. 隐含数据：$Z=\{z_1,z_2,\dots,z_n\}$；
3. 观测数据的概率密度函数：$p(x|\theta)$；
4. 完全数据的联合概率密度函数：$p(x,z|\theta)$；
5. 观测数据的对数似然函数：$\ln\prod_{i=1}^{n}p(\mathbf{x}_i\mid\theta)=\sum_{i=1}^{n}\ln p(\mathbf{x}_i\mid\theta)$；
6. 完全数据的对数似然函数：$\ln\prod_{i=1}^{n}p(\mathbf{x}_i,\mathbf{z}_i\mid\theta)=\sum_{i=1}^{n}\ln p(\mathbf{x}_i,\mathbf{z}_i\mid\theta)$；

EM 算法的基本步骤如下：

- 初始化参数 θ^{old}；
- E step：基于当前的参数 θ^{old} 和样本，估计隐变量的后验分布 $p(\mathbf{z_i}\mid\mathbf{x_i},\theta^{old})$；
- M step：基于当前所估计的 $p(\mathbf{z_i}\mid\mathbf{x_i},\theta^{old})$ 更新参数 θ：

$$\begin{aligned}\theta^{new}=\arg\max_{\theta}Q(\theta,\theta^{old})&=\sum_i E_{p(\mathbf{z}_i|\mathbf{x}_i,\theta^{old})}\Big[\ln\big(p\big(x_i,z_i\mid\theta\big)\big)\Big]\\&=\sum_i\sum_{z_i}p(z_i|\mathbf{x}_i,\theta^{old})\ln\big(p\big(x_i,z_i\mid\theta\big)\big)\end{aligned}$$

- 重复迭代 E step 和 M step，直到参数收敛或达到目标要求。

4. 对混合高斯模型参数估计问题，在 EM 优化的框架下，请给出其中的 $Q(\theta,\theta^{old})$ 的基本形式。

答： 基本形式为：

$$\begin{aligned}Q(\theta,\theta^{old})&=\sum_i\sum_{z_i=1:K}p(z_i|\mathbf{x}_i,\mathbf{0}^{old})\ln\big(p\big(\mathbf{x}_i,z_i\mid\mathbf{0}\big)\big)\\&=\sum_i\sum_{z_i=1:K}p(z_i|\mathbf{x}_i,\mathbf{0}^{old})\ln\left(\pi_{z_i}\mathcal{N}(\mathbf{x}_i\mid\mu_{z_i},\ _{z_i})\right)\\&=\sum_i\sum_{z_i=1:K}p(z_i|\mathbf{x}_i,\mathbf{0}^{old})\left(\ln\pi_{z_i}+\ln\mathcal{N}(\mathbf{x}_i|\mu_{z_i},\ _{z_i})\right)\\&=\sum_i\sum_{z_i=1:K}\big(p(z_i|\mathbf{x}_i,\mathbf{0}^{old})\ln\pi_{z_i}+p(z_i|\mathbf{x}_i,\mathbf{0}^{old})\ln\mathcal{N}(\mathbf{x}_i\mid\mu_{z_i},\ _{z_i})\big)\\&=\sum_i\sum_{z_i=1:K}p(z_i|\mathbf{x}_i,\mathbf{0}^{old})\ln\pi_{z_i}+\sum_{k=1:K}\sum_i p(z_i=k|\mathbf{x}_i,\mathbf{0}^{old})\ln\mathcal{N}(\mathbf{x}_i\mid\mu_k,\mathbf{\Sigma}_k)\end{aligned}$$

5. 针对离散状态和离散观测情形的一阶 HMM，请描述其学习问题的基本任务。

答: 对于离散状态和离散观测情形的一阶 HMM，其学习问题的基本任务指给定观测序列 $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, 如何调整模型参数 [A,B,$\pi$] 使该序列出现的概率 $P(\mathbf{x} \,|\mathbf{A}, \mathbf{B}, \pi)$ 最大即。如何模型使其能够最好地描述观测数据。

2 编程题

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。50 样本点如下：

4.6019,	5.2564,	5.2200,	3.2886,	3.7942,
3.2271,	4.9275,	3.2789,	5.7019,	3.9945,
3.8936,	6.7906,	7.1624,	4.1807,	4.9630,
6.9630,	4.4597,	6.7175,	5.8198,	5.0555,
4.6469,	6.6931,	5.7111,	4.3672,	5.3927,
4.1220,	5.1489,	6.5319,	5.5318,	4.2403,
5.3480,	4.3022,	7.0193,	3.2063,	4.3405,
5.7715,	4.1797,	5.0179,	5.6545,	6.2577,
4.0729,	4.8301,	4.5283,	4.8858,	5.3695,
4.3814,	5.8001,	5.4267,	4.5277,	5.2760

答： 分别选择窗宽为 0.2、0.5、1、2，利用矩形窗和高斯窗估计概率密度函数得到的曲线图如图 1和图 2所示。从图中可以看出，窗宽较小，每个样本点对附近区域的概率密度函数估计会有较大的影响，估计的概率密度函数曲线会相对尖锐；窗宽较大，每个样本点的影响范围会更广，因此估计的密度函数会更加平滑，有利于去除早上。当窗宽取值过大时，但同时也可能导致丢失一些重要的局部特征，导致估计结果不准确。

编程代码见附件中的 code.py。

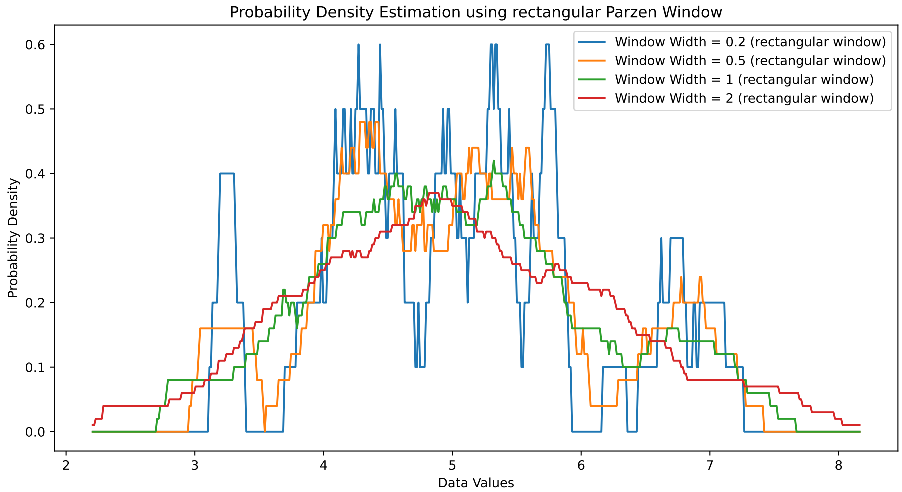


Figure 1: 矩形窗估计概率密度函数

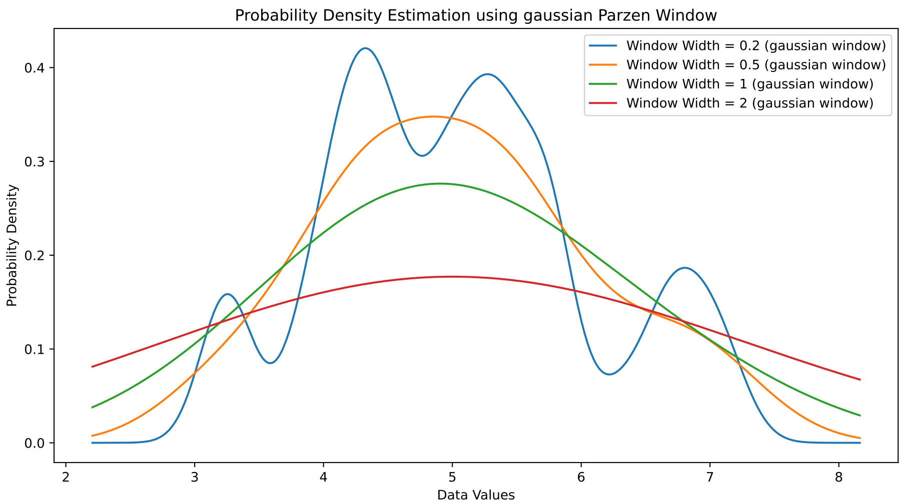


Figure 2: 高斯窗估计概率密度函数