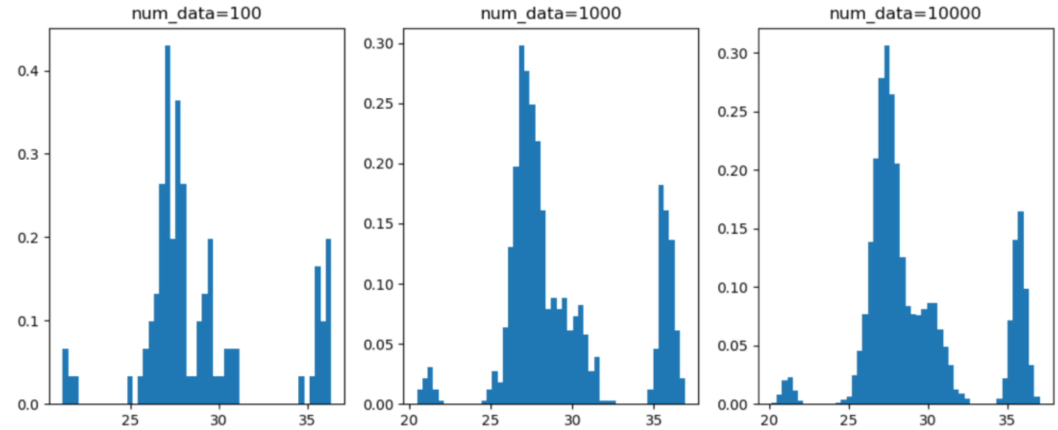
作业一报告

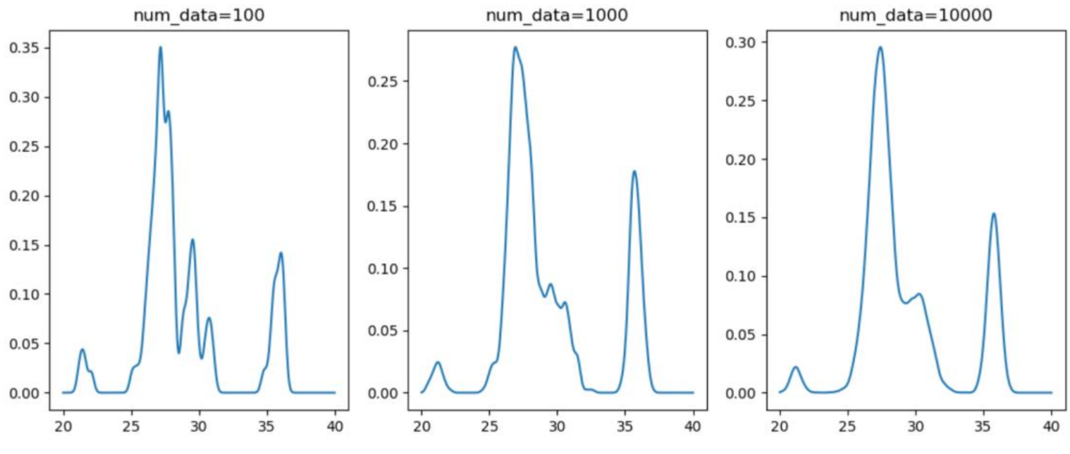
一、　不同的样本数对估计的影响

总的来说，随着样本数据的增大，三种算法估计效果变好，估计效果对参数的选择的敏感度降低。具体来说：

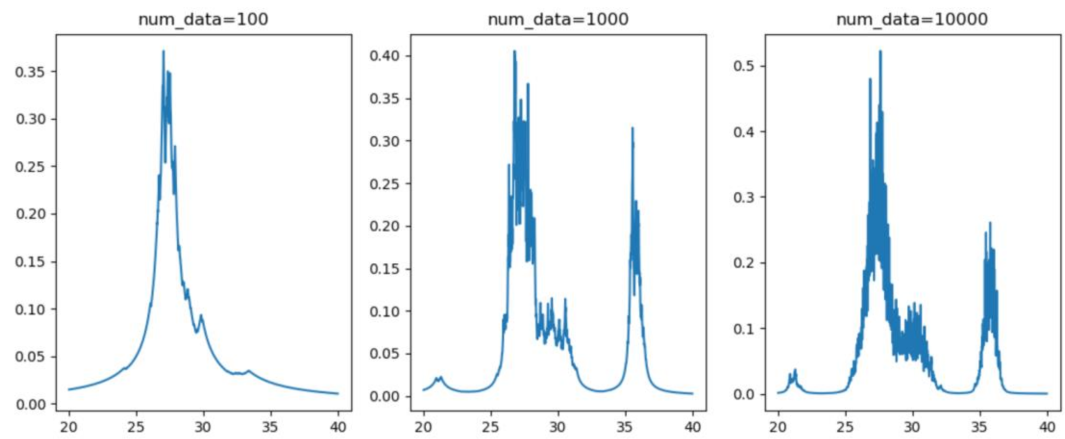
对于 bins=50 的直方图估计，分别使用 num_data=100,1000,10000 进行估计，可以看到随着样本数据的增大，直方图的概率密度估计更为连续，也更贴近真实分布。



对于核密度估计（假定 h=0.2），同样可以看出随着样本数据增大，曲线明显更加平滑连续。在实际画图时发现，一个较好的参数 h 受数据量影响小一些，比如后文 h=0.39 时虽然 num_data 只有 100，但依然很接近真实分布的图像。



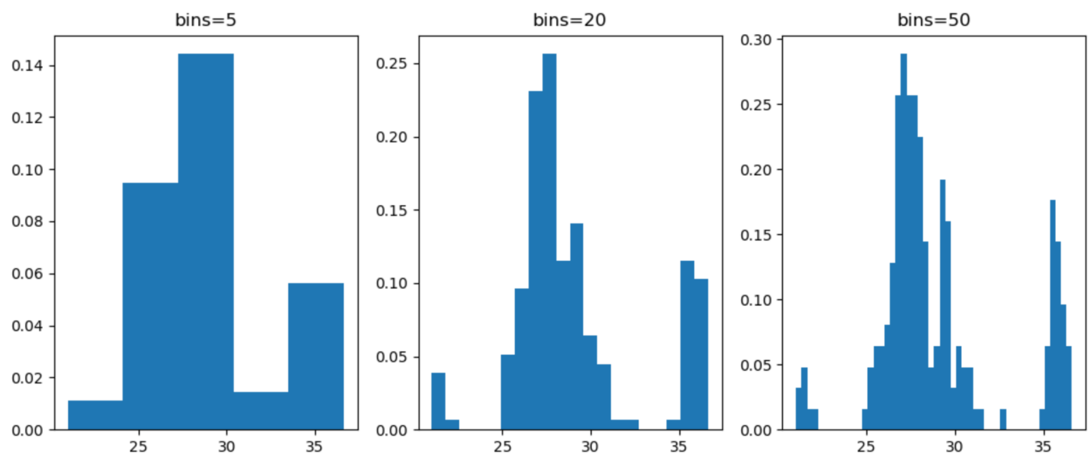
对于 k 近邻方法（假定 k=20），可以看出随着数据量的增多，在相同参数 K 下估计更能反应原有分布的性质（可以由峰的形状看出）。



以上图中的参数并不是对应估计的最优参数，但是在 num_data=10000 的情况下依然能较好的反映出原有分布的性质。因此采样数据越大，参数的选择对估计效果的影响会大致变小。

二、 直方图估计 bins 选取

对比 bins=5,20,50 的估计图可以发现，bins 大小选择与估计效果密切相关：



如果 bins 较小，会导致最终估计出来的概率密度函数非常粗糙；如果 bins 较大，可能会导致有些区域内根本没有样本或者样本非常少，这样会导致估计出来的概率密度函数很不连续。所以，随着样本数的增加，bins 应足够大，同时又必须保证区域内有一定的样本。

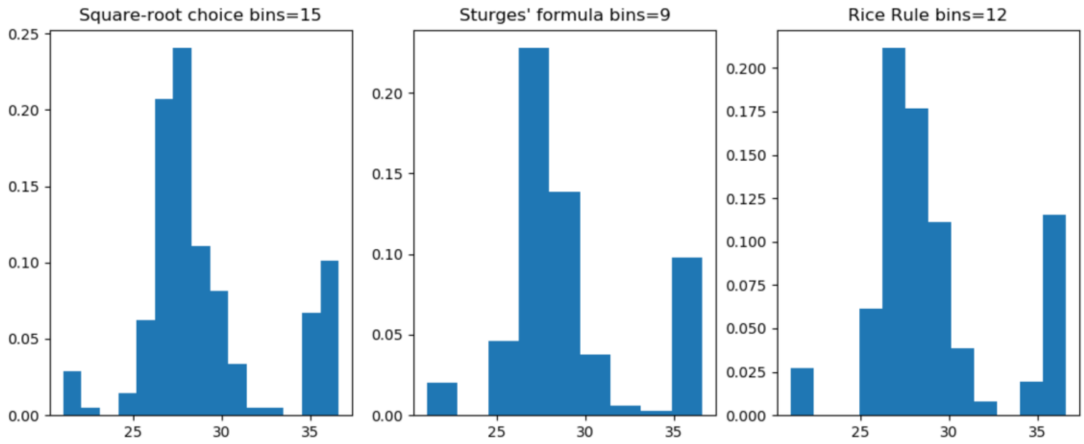
因此在实际选取 bins 时可以通过每个区间里样本数量判断，也可以根据概率密度的连续程度进行判断。另外也可以根据以下一些经验性的估计公式：

Square-root choice：$\text{bins} = \lceil\sqrt{n}\rceil$

Sturges' formula：$\text{bins} = \lceil log_2 n\rceil + 1$

Rice Rule：$\text{bins} = \lceil 2n^{1/3}\rceil$

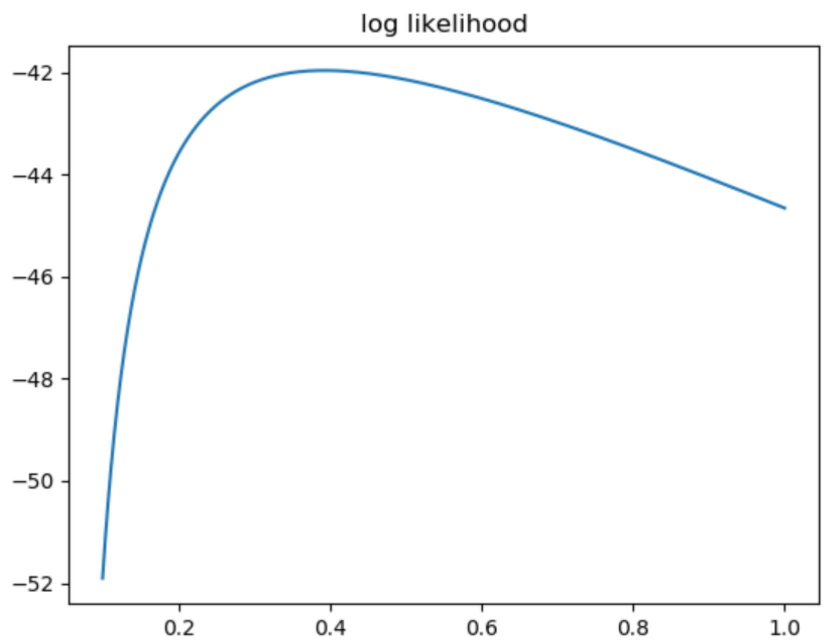
对于 num_data=200，可以得到下图，从结果来看感觉 bins=15 稍好一点。



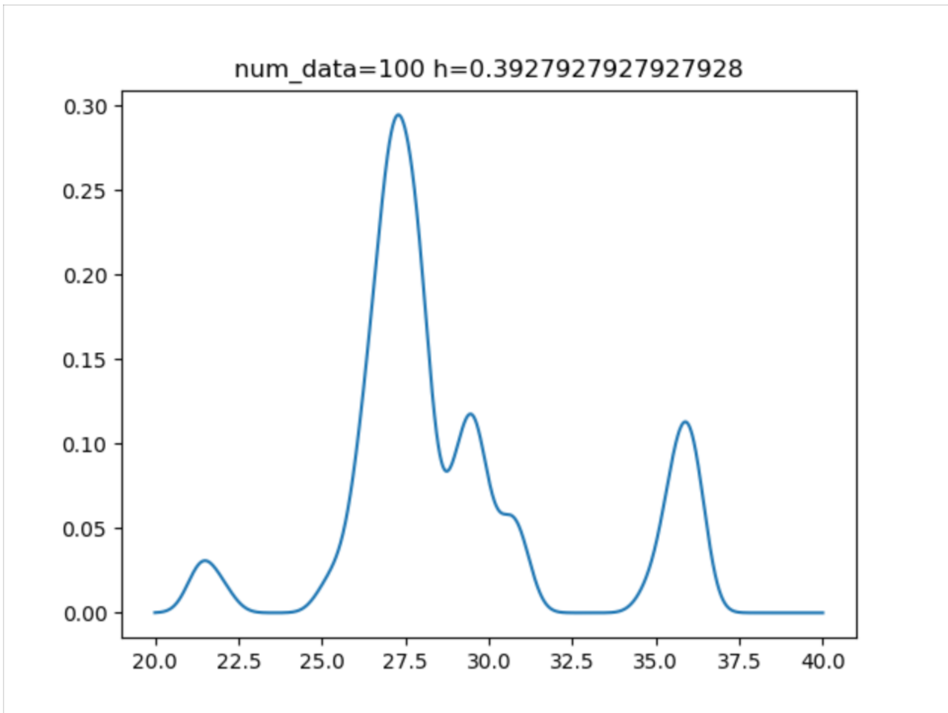
三、 核密度估计

对于 h 的估计，考虑到由于并不知道真实分布，因此运用最大似然估计：对于 num_data=100 的采样数据，划分训练集和测试集，寻找在测试集下使得似然函数$L(h)=\prod_{i=1}^{n}p(xi;h)$ 最大的 h。考虑到不好直接通过似然方程直接求出，因此考虑画出 L(h)的图像找出一个近似最优的 h。

考虑到在本问题下采样数据 num_data=100，数据量较少，因此采用交叉验证会更为准确，实验中采用 5 折交叉验证。首先经过一些大致的取值可以发现一个较好的 h 应大致在(0.1,1)之间，之后在(0.1,1)之间选取 1000 个点绘制对数似然函数的图像，如下图：



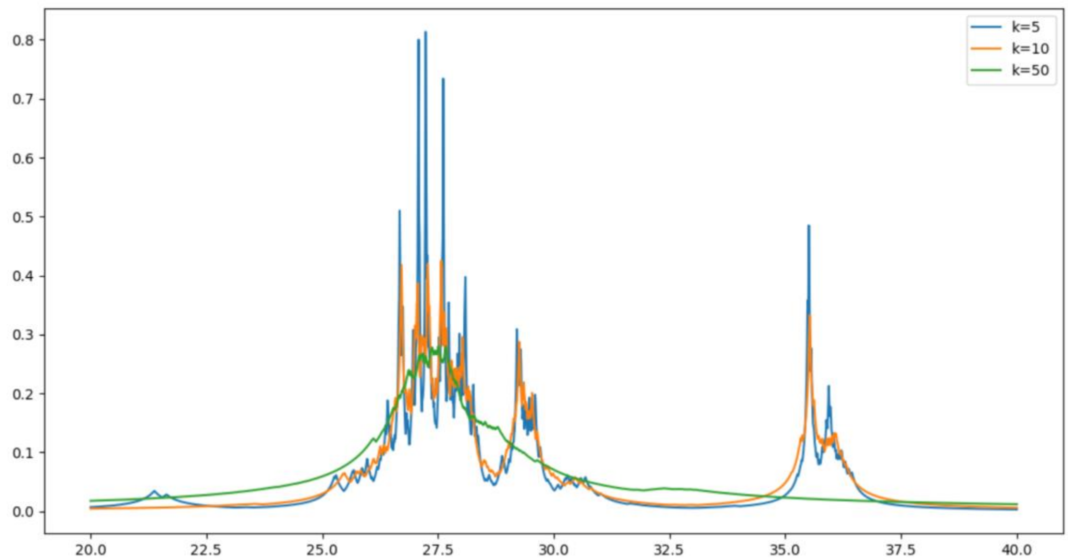
可以看到最优的 h 大约在 0.39 左右，因此可以得到此时核密度曲线估计图：



四、 k 近邻方法

首先确定公式中 V 的计算。按照 V 是以 x 为中心正好包含 k 个数据点的球体体积的定义，在一维情形下，V 应该为 x 到第 k 近数据点距离的两倍。

分别尝试 k=5, 10, 50，画出 k 近邻估计的图像：



可以看出，较小的 k=5 会产生较大的噪声密度模型，而较大 k=50 会过度平滑以至于原有分布的一些特性被忽视，比如右侧的峰就被平滑掉了。

关于 k 近邻估计的收敛性，假设样本数据中最大值为 r，最小值为 l，则：

$$\int_{-\infty}^{\infty} p(x)\,dx \geq \int_{r}^{\infty} p(x)dx \geq \int_{r}^{\infty} \frac{K}{N*2(x-l)}\,dx = \frac{K}{2N}\int_{r-l}^{\infty} \frac{1}{x}dx \to \infty$$

因此 k 近邻方法不能得出一个概率分布。

五、 程序说明

源代码见 source.py。运行后通过 input()输入 1~4 分别表示四个问题对应的输出结果。(因为任务较为简单没有采用命令行传参的方式)