Assignment-1 report

Part 1:

将数据量 N = 100，500，1000，10000 所对应的 12 张图绘制出来后，进行组间和组内的对比分析，**结论如下：**

1、直方图方法的特点是随着数据量的上升准确性越来越好，当数据量较少时，如 N = 100 时，会出现本应是高密度的 x 值处出现相当低的密度。当数据量达到最大值 10000 时，产生了最佳的密度估计。

2、核密度估计方法的特点是对数据量的变化非常迟钝。当 N 从 100 变化到 10000 时，产生的分布曲线几乎没有任何变化。

3、最近邻方法的特点是对数据量的变化非常敏感。当 N = 100 时，无法体现分布曲线的一些特征，而当 N = 500 时又过量反应了分布曲线的特征。通过尝试 N = 200 时，能产生相对较好的密度估计。

4、通过三种方法的组间对比发现，核密度估计方法对数据量的变化十分迟钝，从后面的观察来看，主要是参数 h 决定核密度估计方法的准确性。直方图方法和最近邻方法对数据量有一定的要求，在有合适的数据量的情况下，能产生较佳的密度估计。

Part 2:

绘制出 bins_num = 10, 20, 50, 100, 200 所对应的 5 张图，进行观察，发现：

当 bins_num 较小时，如为 10，20 时，基本不能反映出分布曲线的特征。当 bins_num 适中时，如 50 时，能较好反应出分布曲线的特征。当 bins_num 较大时，如为 100，200 时，会出现原本密度估计相当高甚至最高的 x 值处，密度估计值突然变 0。

对寻找较好 bins_num 的思考：

主要要避免的问题应该是 bins_num 较大时，某些密度估计本来应该较大的范围中，出现某个 bin 中没有一个点落入，密度估计值突然变 0 的情况。具体应对思路大致是，对 sample_data 进行排序后得到相邻两值之间的距离数列，在取值较小的距离中，如[0, g](g 较小)，得到一个最大的距离值 max_gap，通过 floor((max_range-min_range)/max_gap)得出较好的 bins_num。

Part 3:

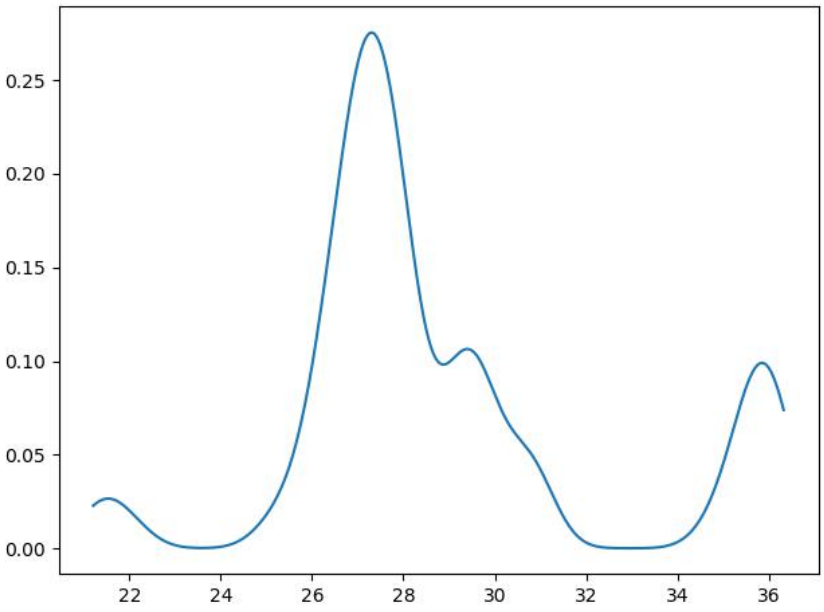
绘制出 h = 0.1，0.2，0.5，1，2 所对应的 5 张图，进行观察，发现：

当 h 较小时，如为 0.1，0.2 时，产生的分布曲线在一些局部会出现较为强烈的振荡现象。当 h 适中时，如为 0.5 时，能产生**可以说是迄今为止见到的最好的分布曲线**。当 h 较大时，如为 1，2 时，产生的分布曲线特征逐渐消失。

对寻找最好 h 的思考：

当 h 较小时，容易出现较强烈的振荡现象，一定程度上扰乱了分布曲线特征的呈现。为了避免这一现象，可以在 x 的取值范围内等间距取值，得出一个密度估计数列，如果这个数列出现较为强烈的震荡现象，可以适当调大 h 的值，就可以得到较好的 h。

产生的最佳的密度估计（h = 0.5）：

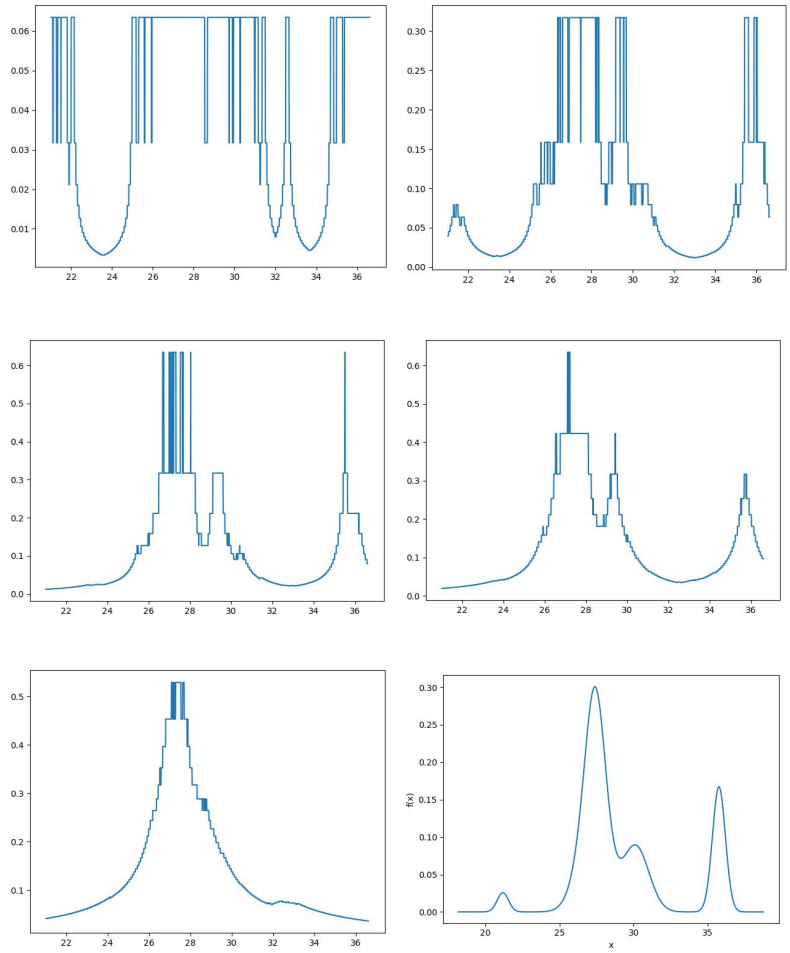


Part 4:

绘制出 K = 1，5，10，20，50 所对应的 5 张图，进行观察，发现：

当 K 较小的时候，如为 1 时，出现了较为强烈的震荡现象。当 K 适中时，如为 5 时，能产生较好的分布曲线。当 K 较大时，如为 10，20，50 时，产生的分布曲线的特征逐渐消失。

图对比（从前到后依次是 K = 1，5，10，20，50 的情况，最后一张图是真实分布曲线）：



最近邻方法不总是能够产生有效的分布：

使用 np.sum(px)计算和。

当 K 取 1 时，输出 inf，显然未能产生有效的分布。