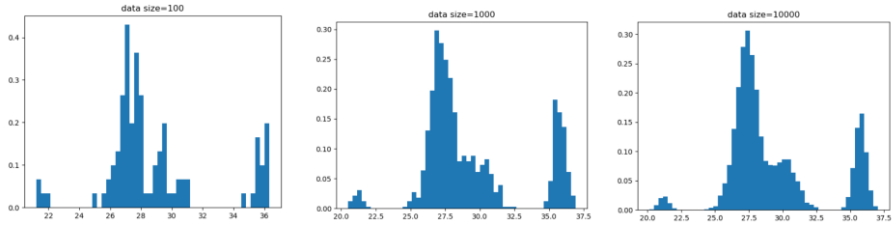
Assignment1 报告

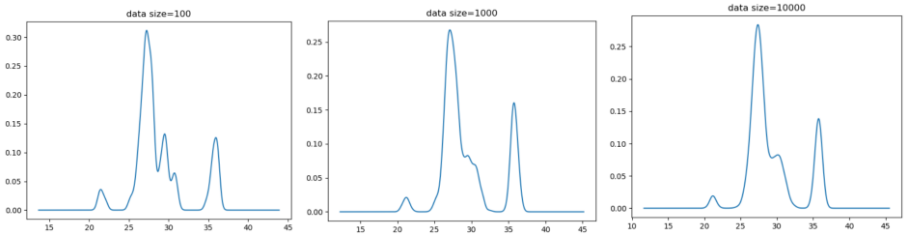
一、 不同样本数对估计的影响

一般来说，样本数据越多三种算法估计结果越准确。

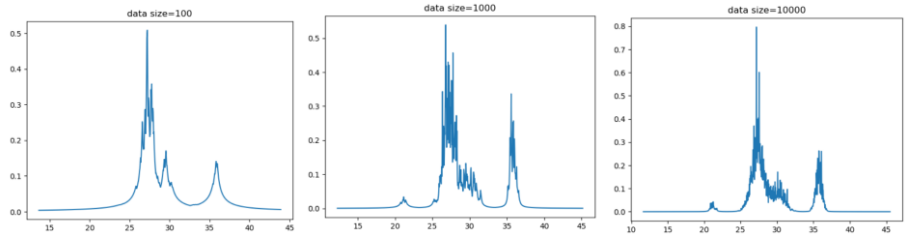
当直方图$\text{bins} = 50$时，样本数越多直方图的概率密度估计越连续。



核密度估计中 h=0.3 时，数据越大曲线总体越平滑，越符合真实分布。

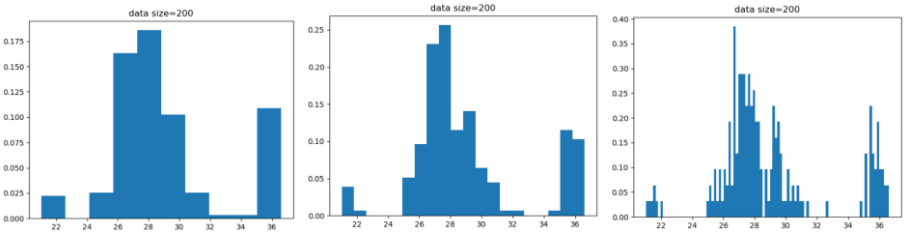


K 近邻方法中，取$\text{K} = 10$，数据较小时分布与真实分布的性质差距较大，随着数据量增大，曲线显示出真实分布应有的性质，其中性质可以由曲线的锋看出。但是当数据量特别大而K值不变时，真实概率分布高的部分由于样本点非常密集会导致个别点数值非常大产生异常。

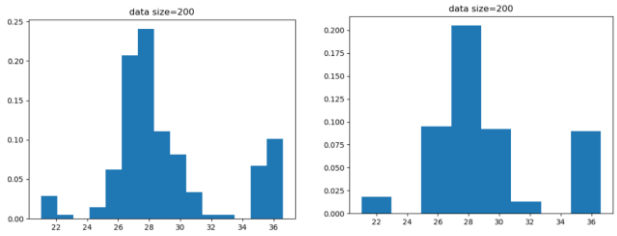


二、 直方图估计 bins 选取

取$\text{N} = 200$，分别选取$\text{bins} = 10,20,100$，画出直方图:



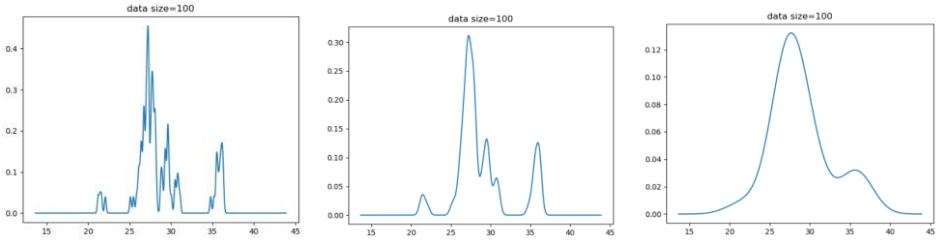
我发现：当 bins 值太小的时候，直方图并不能很好地显示真实的分布。当 bins 值太大的时候，许多值的概率分布为 0 或$\frac{1}{N \times d}$(其中d为组距)，图像呈现出锯齿状，并且很容易受到干扰。因此需要选取一个适中的 bins。可以通过选取多个 bins 进行经验性判断，也可以采用一些经验公式，例如$\text{bins} = \sqrt{\text{N}}$或$\text{bins} = \log_2 N + 1$。当$\text{N} = 200$时，分别采用$\text{bins} = \sqrt{\text{N}} \approx 15$和$\text{bins} = \log_2 N + 1 \approx 8$画出直方图。



根据上图通过经验觉得取$\text{bins} = \sqrt{\text{N}} = 15$效果是比较好的。

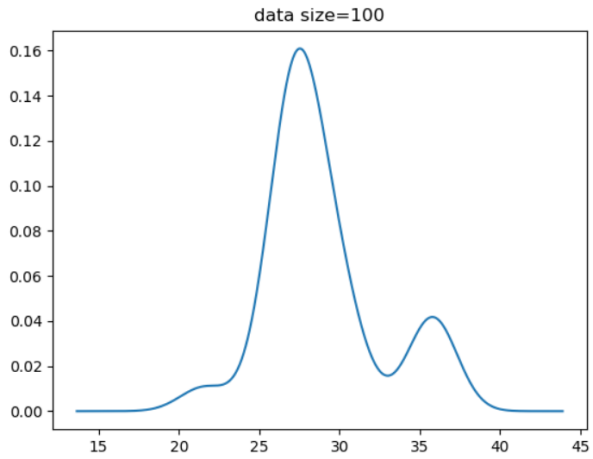
三、 核密度估计中h的选择

由于不知道真实分布，当$\text{N} = 100$时，分别选取$\text{h} = 0.1,0.3,2.0$画出核密度估计的分布图

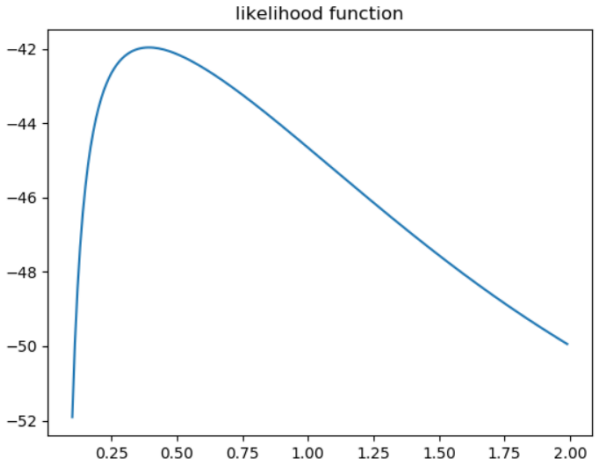


我们发现，当h过小时，曲线非常不平滑，很容易受到个别点的干扰，当h过大时，曲线过于平滑会掩盖掉真实分布的某些特征。因此要选取一个适中的h值。

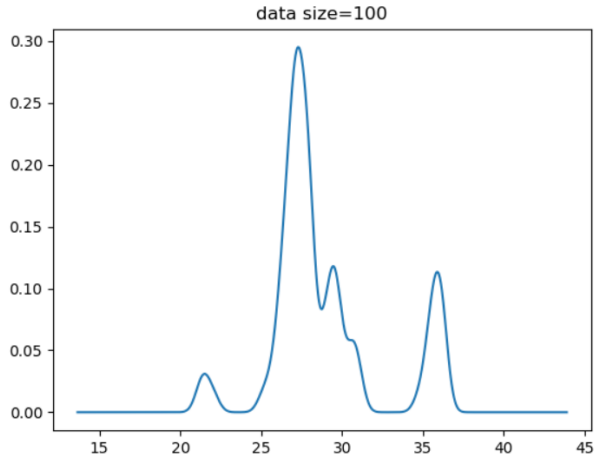
我采用一个经验公式 silverman's rule:$\text{h} = \frac{4}{3\times N}^{0.2} \times \sigma$，其中$\sigma^2$为样本方差。在这里$\text{h} = 1.47$，概率分布如下：



我认为这还是不够好的，因为某些特征如第一个峰没有很明显的显示出来。因此我们可以考虑交叉验证，在这里我采用最大似然估计方法，使得似然函数$\text{L(h)} = \prod_{\text{i=1}}^{\text{n}} p(x_i, h)$最大，划分训练集和测试集为 8:2。由于数据集较小，我采用交叉验证使其更加准确。经过测试，似然函数取对数计算后的图像如下：



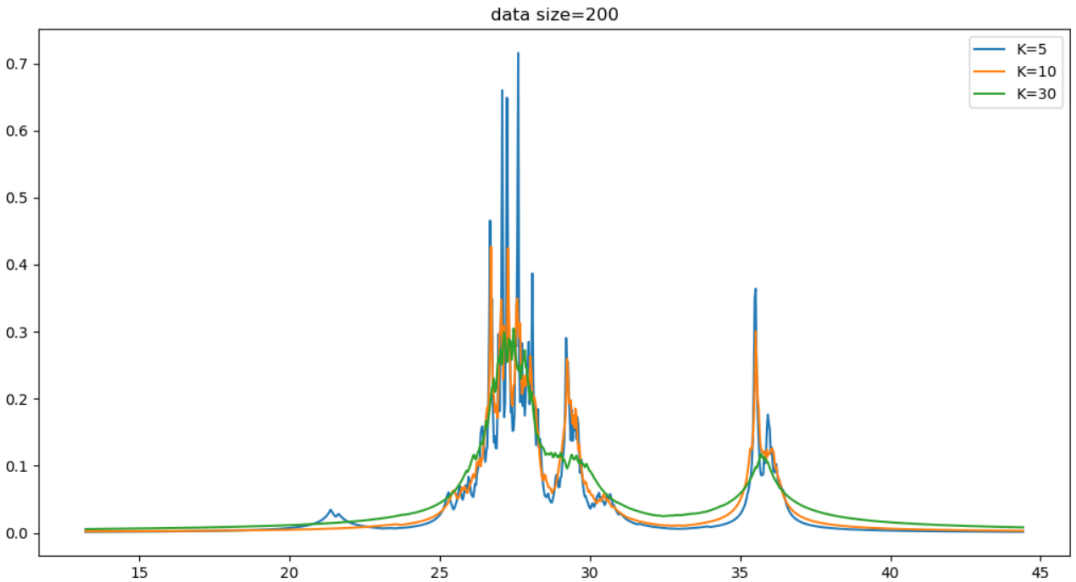
当h = 0.39时似然函数取得最大值，图像如下:



可以看出当N较小时采用最大似然估计并通过交叉验证来确定h是一个比较好的办法。

四、 K 近邻方法

取N = 200，这里我选择了不同的K，其中K = 5,10,30。下图为采用 K 近邻方法的图像:



当K比较小时，曲线容易产生较多凸起，当K比较大时，左边的峰就被掩盖掉了，无法正确估计真实的概率分布。

显然，$\mathrm{K} \le \mathrm{N}$，设样本最大值为R，最小值为L，有

$$\int_{-\infty}^{+\infty} p(x)dx = \int_{-\infty}^{+\infty} \frac{K}{\mathrm{N} \times \mathrm{V}} dx \ge \int_{-\infty}^{+\infty} \frac{K}{\mathrm{N} \times 2 \times |R - L|} dx = \int_{-\infty}^{+\infty} C dx = +\infty$$

其中C为常数。

所以 K 近邻方法得到函数并不是概率密度函数。

五、 程序使用方法

执行程序后按照输入提示输入相应参数即可。