一、

我判断，100 个数据，直方图；500 个，1000 个；10000 个，直方图和高斯核密度和 k 近邻都基本拟合。

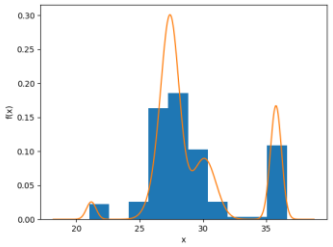
二、

直方图区间个数不应该太少，也不能过多，太少则一个区间内集中了太多的数据，曲线的特征完全被掩盖了，太多则曲线太尖锐，中间太多取 0 的区间，失去了曲线的形态，变成了针形。

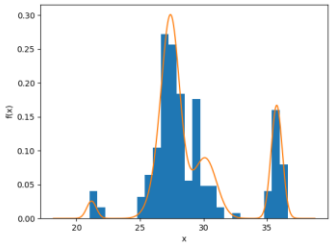
经过进一步尝试，发现对于同一个区间数，增大数据量，大区间数时的尖锐形态消失了，直方图更加接近曲线。因此区间数和数据量也有关。

我认为应该大约把区间数取到数据量的 1/10 到 1/5 左右。

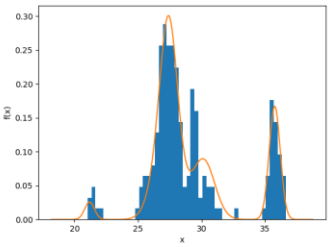
10:



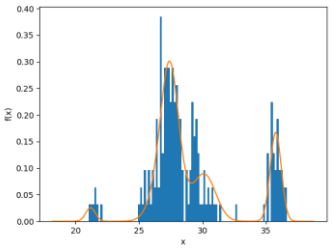
25:



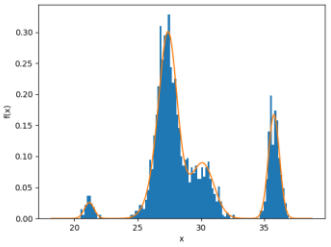
50:



100:



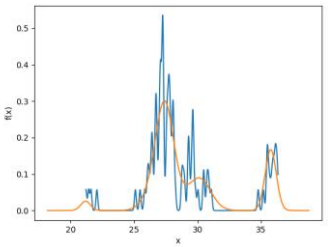
2000 个数据，100:



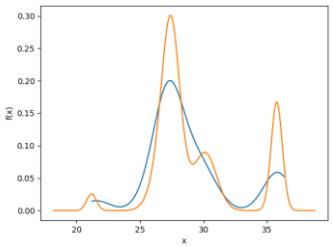
三、

调整 h，h 较小时，曲线过于陡峭，h 过大时，曲线接近于平的

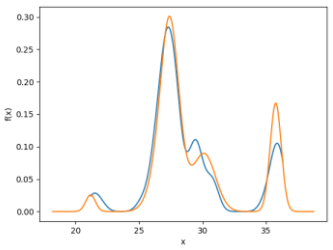
H 过小:



h 过大

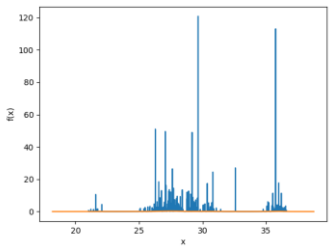


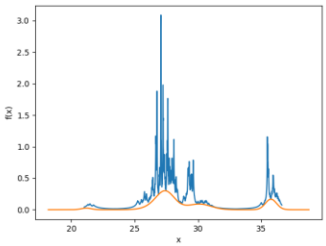
通过不断的调整 h，测试出相对较好的 h，大约为 0.45

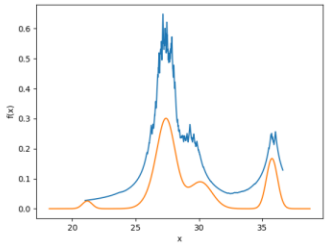


四、

K=1 K=5 K=30







证明：根据经验，k 近邻的每一点概率大于原曲线，原曲线相加为 1，因此 k 近邻的概率的相加应该大于 1.