Assignment 1 Report

常朝坤

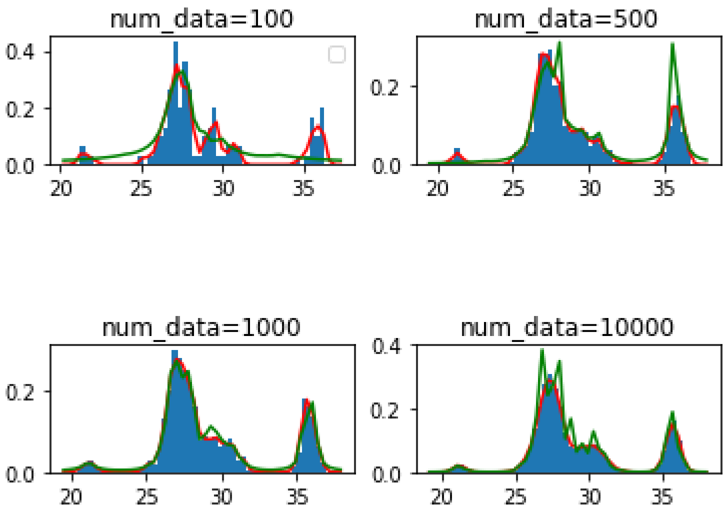
16307130138

Task 1

参数设置

- Histogram method: 分桶数量bins=50
- KDE method: kernel function区间长度h=0.2
- KNN method: k=20

估计结果



注：图中蓝色为Histogram estimation，红色为Kernel density estimation，绿色为KNN estimation

分析

当数据量为100个点时，通过图我们可以发现三个估计趋势大致相同，峰值数量差不多，但是KDE和KNN明显不够平滑，而Histogram也显得十分抖动，有些点的估计有较大差距（如图中27左右）

随着数据量的增加，KDE的曲线变得越来越平滑，且与Histogram相似度也变大。KNN曲线在`num_data`为1000时表现较好，但是当`num_data`为10000时却发生明显的抖动，峰值数量增加。原因在于随着数据量的增加，满足k近邻的区间长度越来越小，导致部分点的估计值陡增，造成图示状况。

Task 2 - Histogram Method

参数设置

- 样本点个数： 200
- 分桶数量：不定

bins 调参

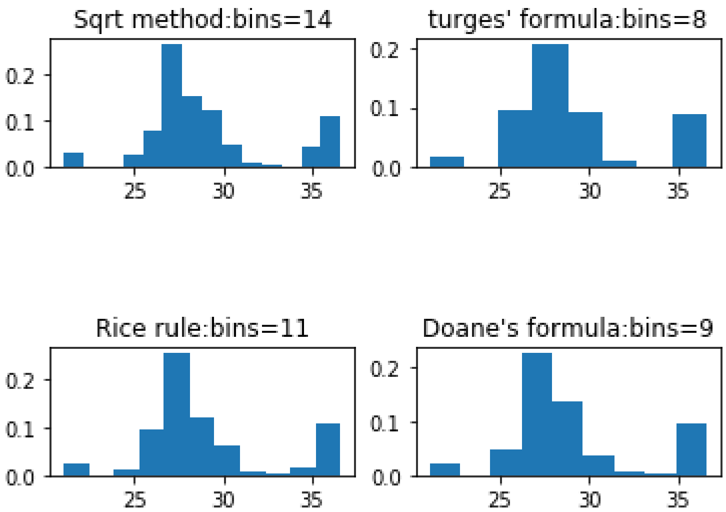
在1-20之间，以2为补偿为bins赋值；在20-200之间，以20为步长为bins赋值，通过画出的图形探究bins对估计结果的影响。

```
    sample_data = get_data(200)
    data_size = len(sample_data)
    for bins in range(2, 20 ,2):
        plt.title(bins)
        histogram_method(sample_data, bins)
    for bins in range(20,data_size, 20):
        plt.title(bins)
        histogram_method(sample_data, bins)
```

可以发现，在0-20之间随着bins数目的增长，图形渐渐得出现峰，且峰的数量也有增加，当bins的数量为20左右时，图形已经呈现出了某种分布的样子。随着bins数目的继续增长，分布估计变得更加精细，但是图象也变得十分的spiky，峰的数目增多，而且出现了越来越多的空值区间，很难确定分布，这种变化在bins=60附近发生。

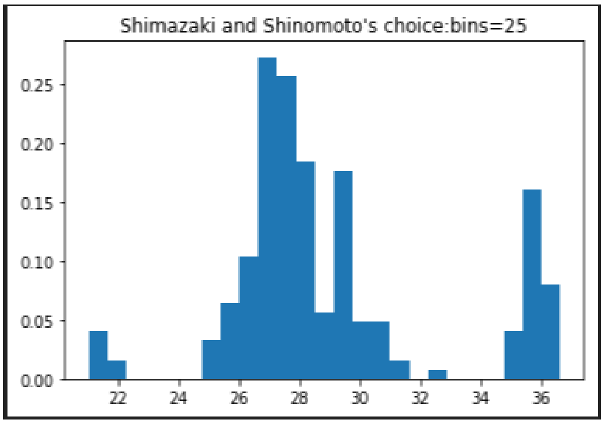
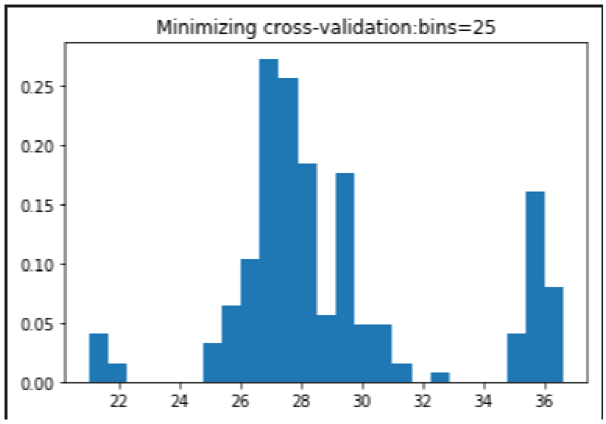
寻找 optimal bins

参考维基百科对Histogram的叙述，本实验尝试了六种不同的方法去寻找合适的`bins number`。其中前四种对正态分布效果较好，但对本实验的数据效果并不太理想，如下图所示：



由于这四个效果不好，所以不作详述，具体可参考"参考文献1"

后两种分别是"基于Loss函数的Shimazaki和Shinomoto选择方法"以及最小化交叉验证的方法，二者得到的结果相同。如下图所示：

Shimazaki和Shinomoto选择方法	最小化交叉验证
	

前者的计算公式为：

$$\arg\min_h \frac{2\bar{m}-v}{h^2}$$

$$\bar{m}=\frac{1}{k}\sum_{i=1}^{k} m_i, v=\frac{1}{k}\sum_{i=1}^{k}(m_i-\bar{m})^2$$

后者的计算公式为：

$$\arg\min_h \hat{J}(h)=\arg\min_h\left(\frac{2}{(n-1)h}-\frac{n+1}{n^2(n-1)h}\sum_k N_k^2\right), N_k\text{表示第}k\text{个}bin\text{中元素个数}$$

Task3 - KDE Method

参数设置

- 样本点个数：100
- 画图点个数：(max-min)/0.1
- 带宽：不定

h 调参

range(1,20)

h	峰数	其他特点
= 1	3	非对称
= 2	2	非对称
>=3	1	逐渐趋于对称

本实验采用的kde方法的kernel函数为高斯分布函数，其中h即为高斯分布中的标准差。而高斯分布的特点就是大部分的值集中在均值附近3*h的范围内，且均值附近h范围内占68.26%，所以当h比较大时，高斯分布图象区域扁平，每一个样本点对估计的贡献范围都很大，平滑过度，掩盖了很多底层的分布结构，导致最后的估计结果也区域高斯分布。

从Task 1中num_data = 10000时的结果可以看出，最后的结果应有4个峰，所以将范围调至0.1-1.0之间观察现象。

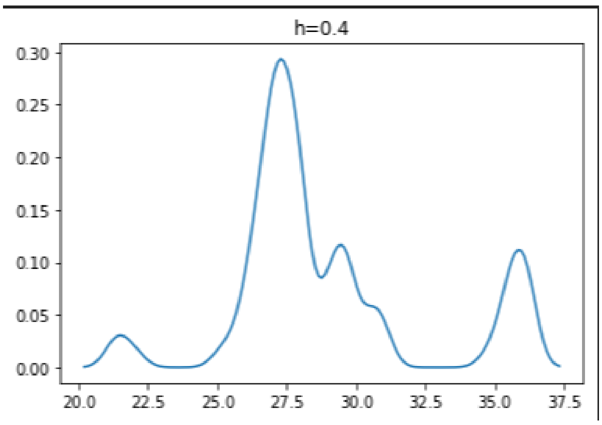
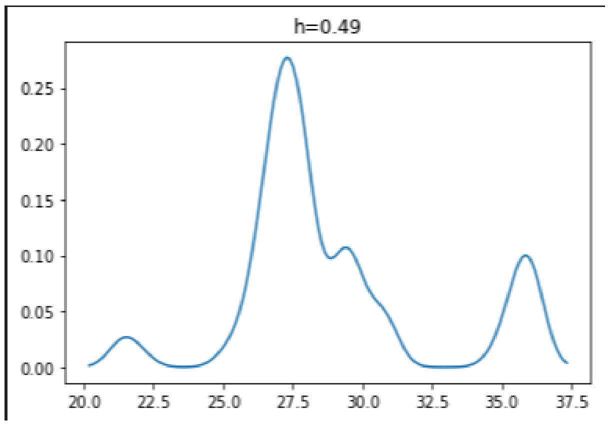
range(0.1, 1.0)

h	峰数
0.1	many
0.2 / 0.3	5
0.4 / 0.5	4
>=0.6	3

当h=0.1时，图象变得十分尖锐，峰很多，平滑效果很差，显然不是我们想要的结果，根据预估，最终的值应该在0.3-0.6之间。可以将h再次调整，观察现象。

range(0.3, 0.6)

通过此次调节，可以发现在h=0.4附近（如下左图），第三个峰右侧的峰逐渐消失（同时第三个峰也逐渐变小），当h=0.5附近时，有图象有三个峰，且图象较为平滑，当h继续增大，第三个峰边逐渐消失了。所以最佳的参数应该在0.4-0.5之间，通过观察，当h=0.49时效果较好（如下右图）。

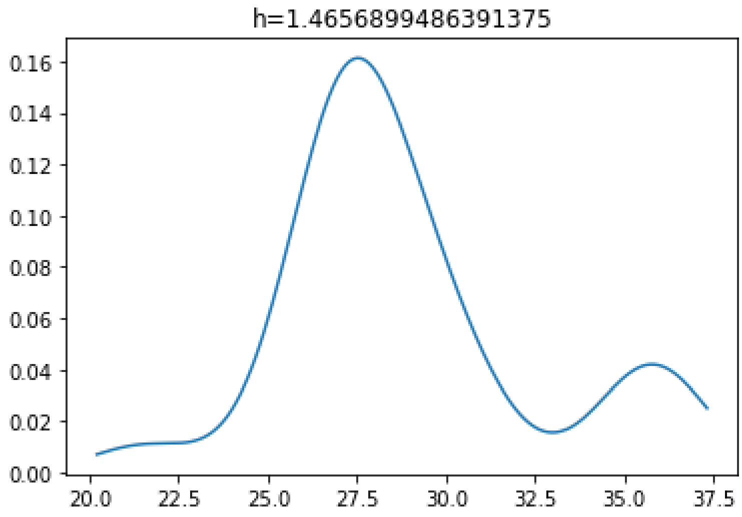
h=0.4	h=0.49
	

理论分析

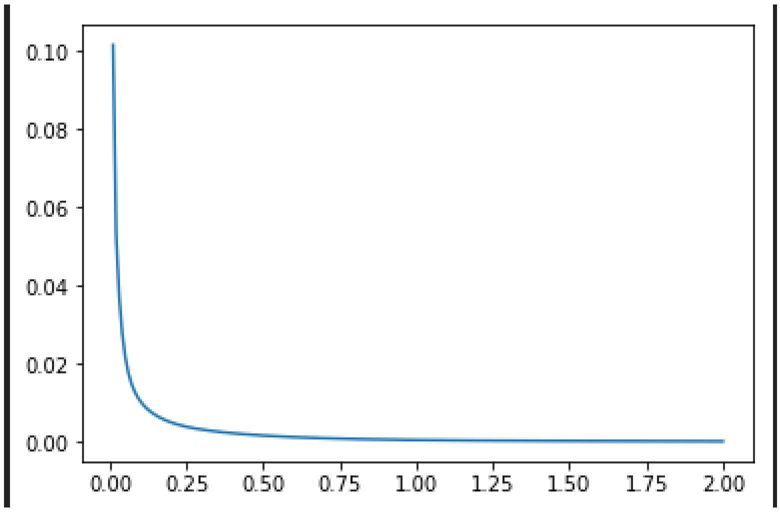
若使用rule-of-thumb bandwidth estimator进行预估，可计算h=1.465，计算公式如下：

$$h = \left(\frac{4\hat{\sigma}^5}{3n}\right)^{\frac{1}{5}} \approx 1.06\hat{\sigma}n^{-1/5},$$

图象只有两个峰，但十分平滑，如下图所示：



估计的偏差与结果相差甚远，原因在于本方法更加适用于正态分布的估计，对本实验的分布预估效果不佳。



使用k折交叉验证的方法，可以发现loss随着h的增大而减小，但是实际上h一旦超过1效果会变差，所以这个方法也不可行。

虽然两个方法的估计效果都不好，但是第一种方法得到结果后可以缩小尝试的范围，对估计也是有很大帮助的。而对于第二种方法，我们也可以看出，在h>1时偏差的变化几乎微乎其微。

Task 4 - KNN Method

参数设置

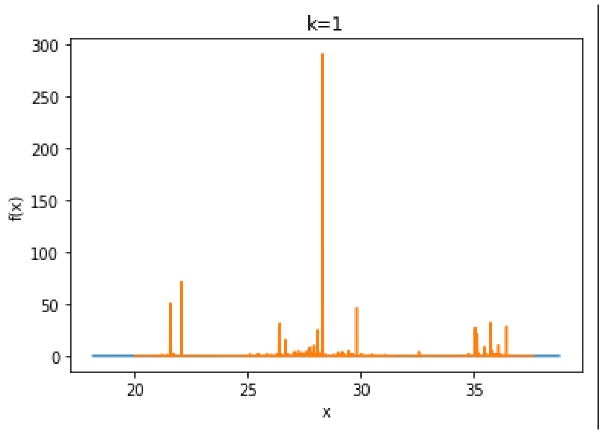
- 样本点个数：200
- 画图点个数：(max-min)/0.01
- K：不定

k 调参

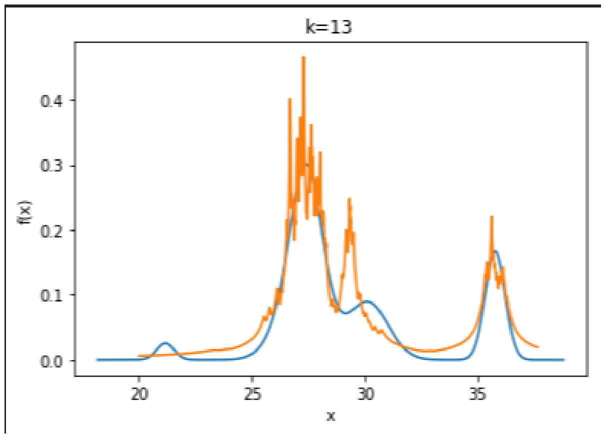
注：在画图时，在图像上取的点数为(max-min)/0.01,约为1500个

在区间[1,50]按步长2调节k，可以发现，当k比较小时，图象十分尖锐，如同噪声声波；随着k的增大，各个峰开始渐渐集聚起来，形成宏观上的更大的峰区间，当k=13左右时，图象的特征便已经比较明显，可以通过图象所形成的面积的趋势判断出整体图形的变化。当k=40左右时，有一些峰开始消失，渐渐集中在中间最高的那个峰周围，形成类似正态分布的分布。从图中也可以发现，KNN估计对于较高的峰估计较好，而较矮的峰就相对差些。

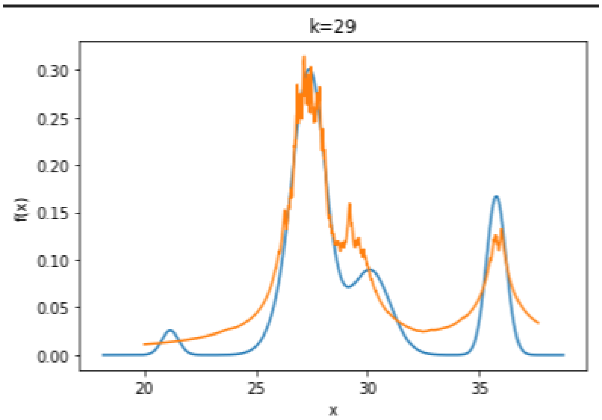
k=1



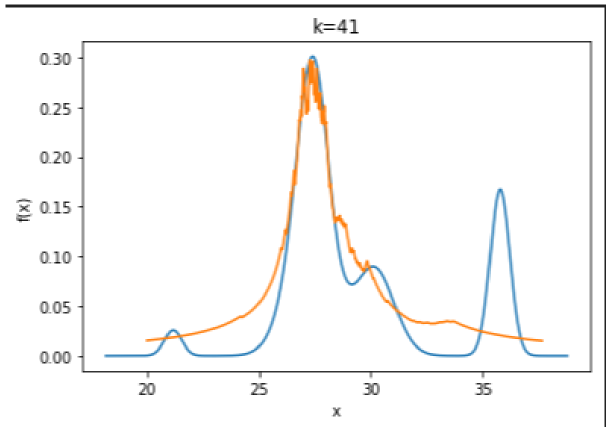
k=13



k=29



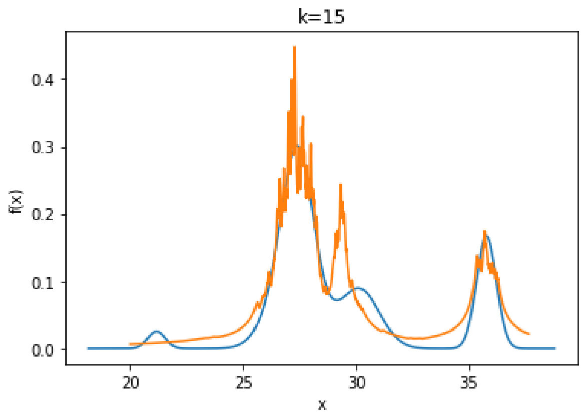
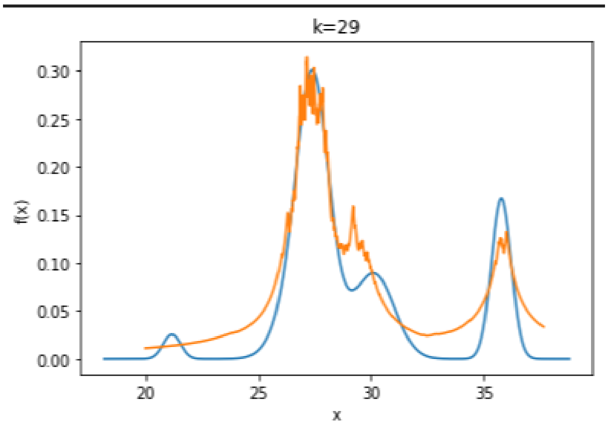
k=41



注：蓝色曲线为实际分布，橙色曲线为KNN估计结果，下同。

理论分析

从原理上看，当k比较小时，会导致图像上很多点的密度值离奇地高，从而产生大量噪声，但是当k比较大时，会导致在一定范围内很多点的密度值相似，导致部分峰丢失，过度平滑（不考虑噪声的情况下）。根据网上查阅的资料，k的经验取值为sqrt(n), n为样本大小。当num_data=200时，k=15，图象如下左图；若使用交叉验证的方法，可以得出最有的k=29，图象如下右图：

<table>
<tr><th>经验取值</th><th>交叉验证</th></tr>
<tr><td>
</td><td>
</td></tr>
</table>

两个取值各有特点，k=29时，对最高峰的估计小姑要好一些，而且面积与原分布更接近。

图象面积的证明

实验推论

使用不同的k画图，可以发现几乎所有使用KNN方法画出的图象的面积都不为1（与原分布面积不同），上述的五张图也显示出了这种现象，KNN方法的密度估计值比原来要更大一些（对大部分点而言），尤其在边界上值要明显偏大。

理论推导

从原分布可以看出，图象在边界上几乎为0，但是对于KNN估计算法，当X>Xn(n表示数据集的大小)时，有如下关系：

$$
\begin{gathered}
when\, x > x_n, \\
assume\, interval\, size\, around\, x\, is\, h\, ;\, number\, of\, data\, is\, N \\
then\, estimated\, probability\, is:\ p(x) = \frac{K}{N * h} = \frac{K}{N * (x - x_{n+1-k})} \\
so\, S_{x_n}^{+\infty} = \int_{x_n}^{+\infty} p(x)\mathrm{d}x = \frac{K}{N} \ln(x - x_{n+1-k})|_{x_n}^{+\infty} \\
S_{x_n}^{+\infty} \to +\infty \\
in\, the\, same\, way,\ S_{-\infty}^{x_1} \to +\infty
\end{gathered}
$$

可以看出边界以外的图象面积是发散的，最终必然导致总面积和不为1.

Reference

1. https://en.wikipedia.org/wiki/Histogram
2. Doane DP (1976) Aesthetic frequency classification. American Statistician, 30: 181–183
3. https://en.wikipedia.org/wiki/Kernel_density_estimation
4. https://zhuanlan.zhihu.com/p/24825503