模式识别导论上机题2-非参数估计

薛犇 1500012752

1. 程序实现说明

本次实验采用Matlab作为编程语言，使用的版本为2016b。

在实验的一开始，利用importdata函数读取hw2_data.txt中的数据，保存在向量x中。

```
x = importdata('hw2_data.txt');
```

（1） 用parzen窗方法求概率密度分布

假设parzen窗的宽度为h_n，核函数为$K(x)$，数据样本数为n，那么概率密度分布为

$$p_n(x) = \frac{1}{nh_d}\sum_{i=1}^{n} K(\frac{x-x_i}{h_d})$$

本次实验中，选择高斯函数作为核函数，也即：

$$K(x) = \frac{1}{\sqrt{2\pi}}exp(-\frac{||x||^2}{2})$$

在代码中的实现为（第42-42行）：

```
function y = K(xi)
    y = exp(-0.5 * xi.^2) / (2*pi)^0.5;
end
```

之后，在已有数据x的range范围之内用linspace函数取一个散列x_0，利用x_0计算对应的概率分布y_0。

在本次实验中，考虑到样本数值的数量级在10^5左右，所以考虑选取窗的大小为100,1000，10000。

```
h = [100, 1000, 10000];
colors=['r-', 'g-', 'm-'];
h_len = size(h, 2);
x_0 = linspace(min(x), max(x));
x_1 = repmat(x_0, n, 1);
for i=1 : h_len
    y_0 = sum(K((x_1 - x)/h(i)))/(n*h(i));
    fig = plot(x_0, y_0, char(colors(i)));
end
```

（2） k_n——近邻估计方法

此方法固定窗覆盖的点的个数，改变窗的大小。

假设窗的大小为V_n，每个窗覆盖k_n个点，样本总数为n，那么概率密度分布为：

$$p_n(x) = \frac{k_n}{nV_n}$$

为了求出恰好覆盖k_n个点的窗的大小，我对数据做如下处理：首先求出每个数据点到待求点x的欧拉距离，存放在dist数组中，然后对这个dist数组排序，取出前k_n个点在原样本x中的index，再用这个index取回这前k_n个点在原样本中的数值。最后求这k_n个数值的极差，也就是窗的大小。

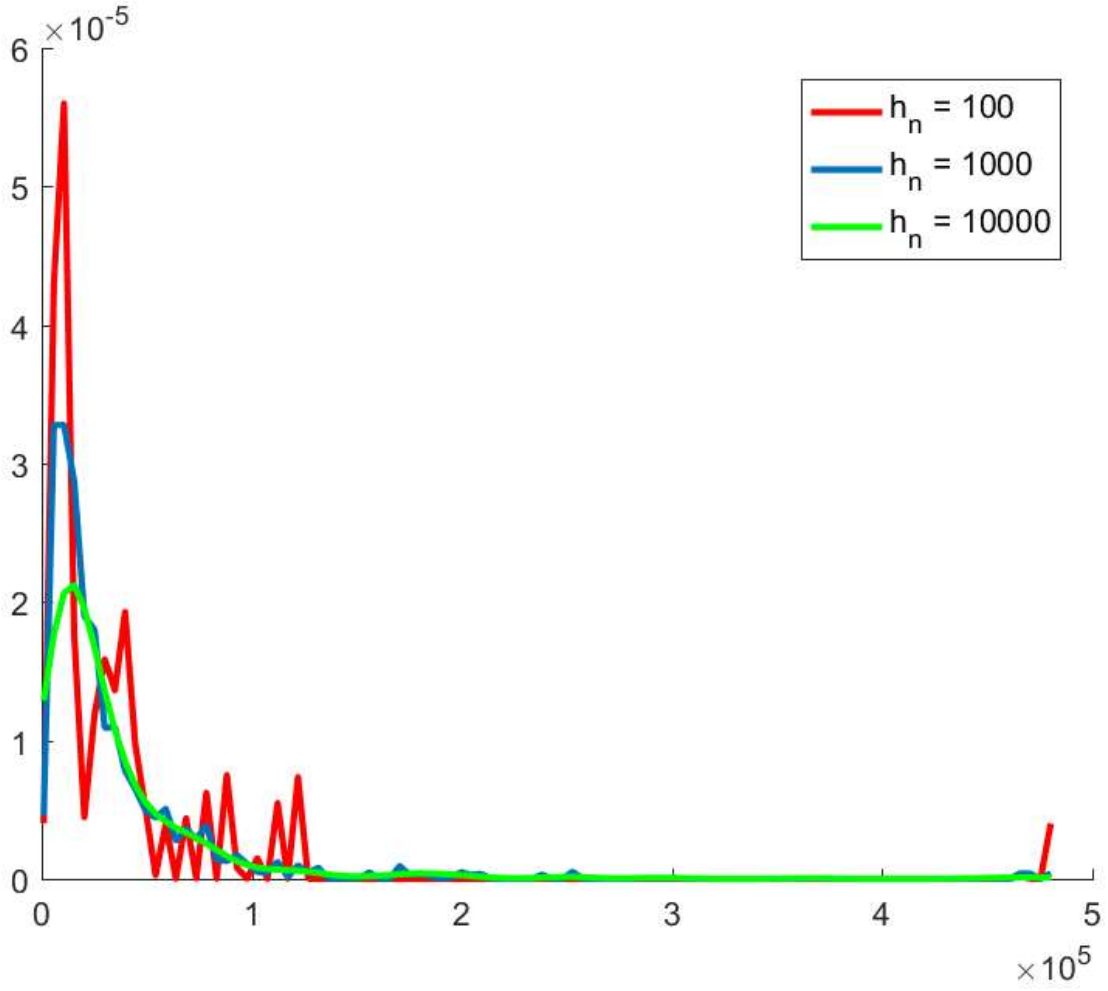
实现代码如下：

```
k = [3, 5, 10];
k_len = size(k, 2);
figure;
hold on;
for i = 1 : k_len
    y_1 = zeros(size(x_0));
    for j = 1: size(x_0, 2)
        dist = abs(x_0(j)-x);
        [sort_dist, index] = sort(dist);
        k_n_neighbour = x(index(1:k(i)));
        v = max(k_n_neighbour) - min(k_n_neighbour);
        y_1(j) = k(i)/(n*v);
    end
    fig2 = plot(x_0, y_1, char(colors(i)));
end
```

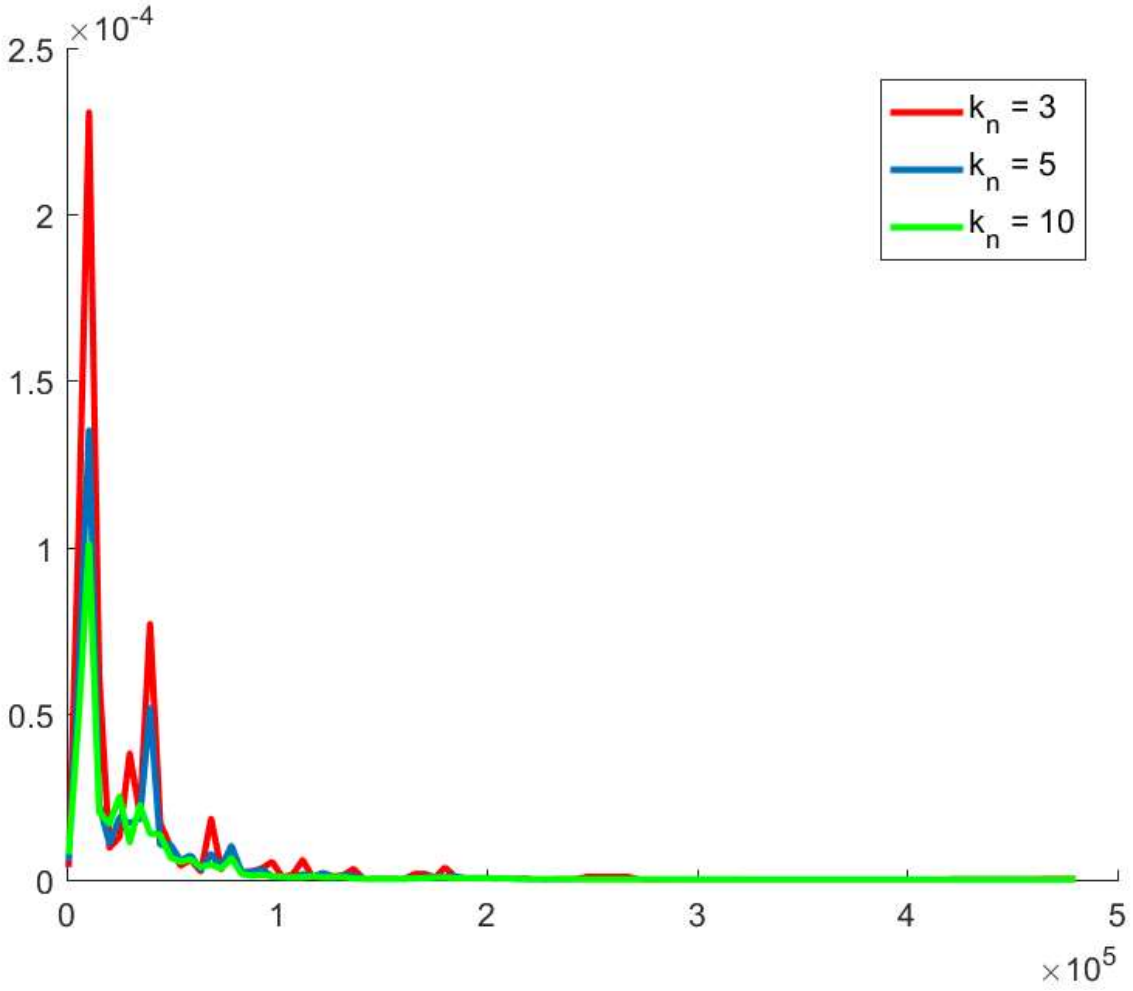
注明：k的取值为3，5, 10。

2. 实验结果

parzen窗方法结果



k_n——近邻方法结果



3. 实验结果分析

（1）parzen窗

可以看到，随着窗的大小的增加，拟合结果越来越平滑。

（2）k_n——近邻方法

随着k_n的增加，可以看到拟合结果趋于平缓，没有较大幅度的升降，但是仍然存在许多非常不光滑的点。