

山东大学机器学习课程

实验报告

——实验三：无参数估计

姓名：陈竞帆

学院：软件学院

班级：软件八班

学号：201500150180

**一、实验目的：**

（1）熟悉python实验软件及相关函数

（2）学习以parzen窗方法和k-邻近方法进行无参数估计的思想与算法

（3）根据已给测试数据，求得的概率密度进行分类

**二、实验环境：**

（1）硬件环境：

英特尔® 酷睿™ i5-7500U 处理器

256 GB PCIe® NVMe™ M.2 SSD

8 GB LPDDR3-1866 SDRAM

（2）软件环境：

Windows10家庭版64位操作系统

Python3.6

**三、实验内容**

**3.1无参数估计方法概述**

无参数技术中涉及到的方法，可以在不知道概率密度分布的情况对其进行合理的估计。我们可以知道一个事实，一个向量x落在区域R中的概率为：



如果我们假设p（x）是连续的，并且区域R足够小，以至于在这个区间中p几乎没有变化，那么有：



其中V是区域R所包含的体积。

结合上述的两个公式，我们可以得到p（x）的估计为：



我们将结合这个公式利用两种方法对p（x）进行估计。

Parzen窗：



φ表示Parzen窗函数，在实验中我们采用了球形的高斯函数：

C:\Users\dell\Documents\Tencent Files\951718744\Image\C2C\{61133841-0319-E0DA-40C2-E2810860D348}.png

Vn在实验中表示区域R的体积，是球状的。x为测试数据，xi为训练数据集，h为球的半径。

我们事先拥有某一个类别w下的训练数据集，对每一个测试数据集，进行该测试数据集在当前分类下的概率密度，而公式则可以看做是概率密度的累加求和最后再求平均值。

k-邻近法：

k-邻近方法让体积成为训练样本的函数，为了估计p（x），我们以x为中心不断扩大直到包含kn个最近邻的样本点。



我们对一个测试样本x与训练数据集中的每一个xi求欧式距离，并对这些欧式距离进行从小到大排序，然后看我们的kn的取值，按递增顺序从排好序的距离值中获得前kn个，就是所谓的在kn个最近邻的样本点，带入公式即可求得p（x）

**3.2两种无参估计的算法**

设计程序的伪代码，实现图1所示：

Parzen窗方法

1.读取json文件中的训练数据集

2.训练数据和测试数据x代入窗函数求得每一个分类下的p(X)

3.比较p（x）的大小，最大的所在的分类即对应的概率密度

k-近邻方法：

1.读取json文件中的训练数据集

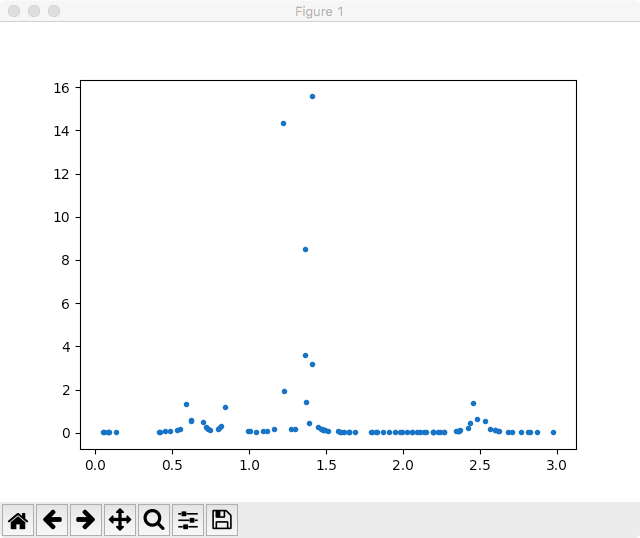
2.随机数生成100个测试数据集，并利用k-近邻方法求得这100个测试数据集每一个对应的p（x）

3.画出这些概率密度的分布图

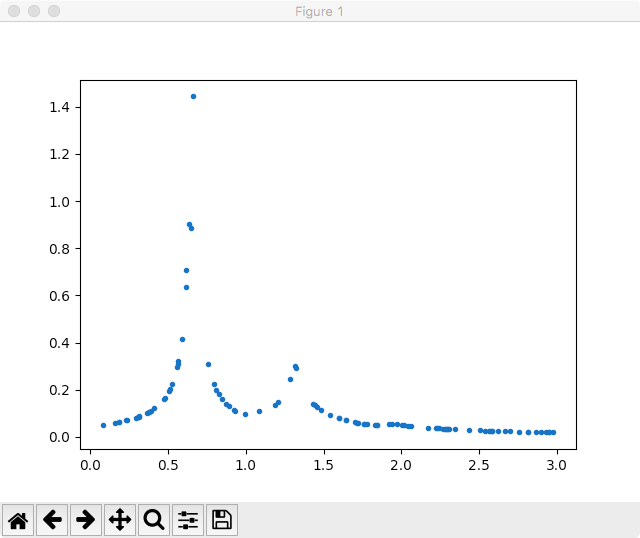
4.针对第三问的三个测试点进行每一个w下的p（x）的估计

**图1.无参数估计算法伪代码**

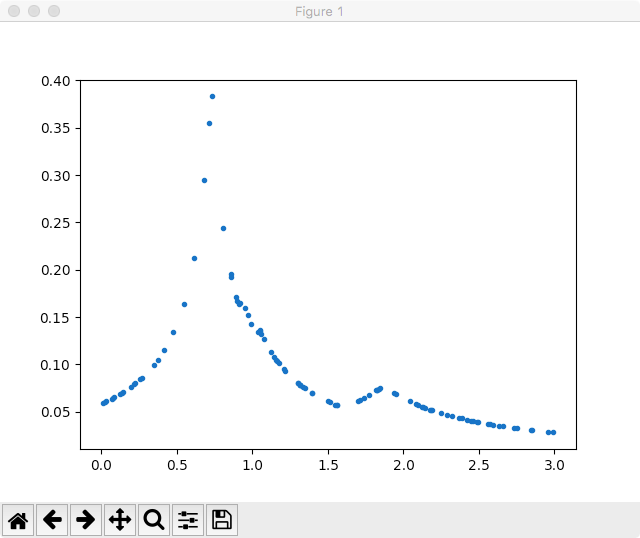
**四、实验结果**



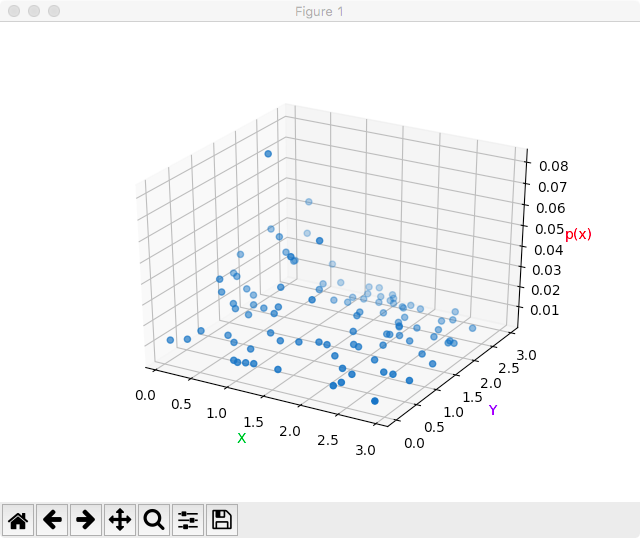
**图2. 在k=1时，针对w3一维数据k-近邻估计的p（x）分布图**



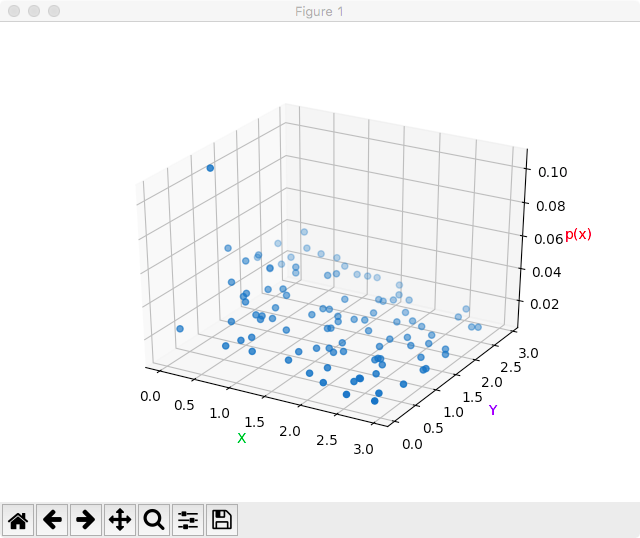
**图3. 在k=3时，针对w3一维数据k-近邻估计的p（x）分布图**



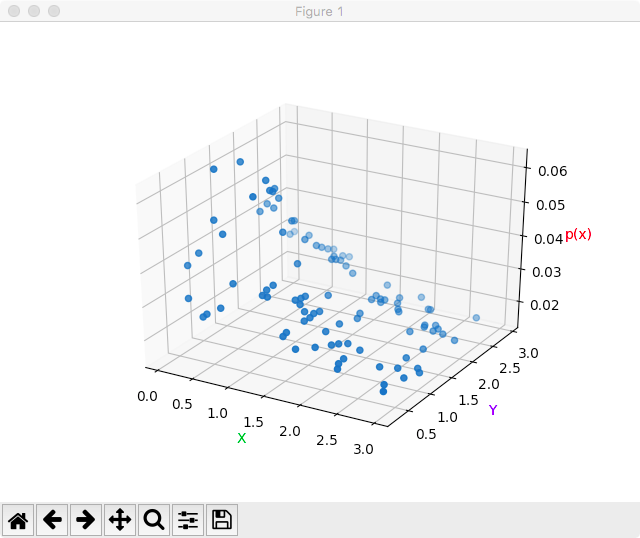
**图4. 在k=5时，针对w3一维数据k-近邻估计的p（x）分布图**



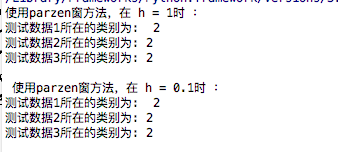
**图5. 在k=1时，针对w2二维数据k-近邻估计的p（x）分布图**



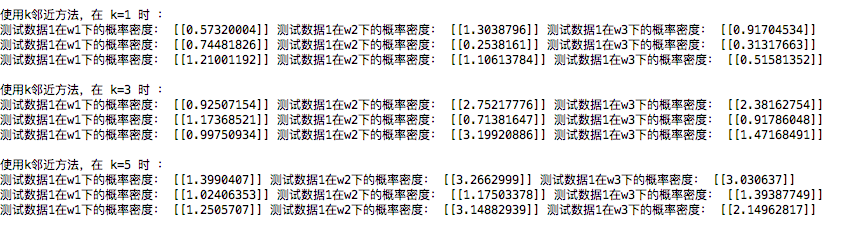
**图6. 在k=3时，针对w2二维数据k-近邻估计的p（x）分布图**



**图7. 在k=5时，针对w2二维数据k-近邻估计的p（x）分布图**



**图8. Parzen窗方法下的实验结果**



**图9.k-近邻三维数据下的实验结果**

**五、总结与归纳**

**本次实验是对无参数估计方法的一个掌握，Parzen窗方法是从所有训练数据集的方法下固定体积进行无参数估计，而k-近邻则是对体积进行变化，作为训练数据集的函数，我们可以从k-近邻的图片中看到，估计的分布表现良好，但是我们也可以看到，当训练数据集数量较少时，结果误差会比较大，在理论的分析中我们也可以得出这个结论。最后，这次实现巩固了我对无参数估计方法的掌握。**