核密度估计（kernel density estimation）

Table of Contents

有一些数据，想“看看”它长什么样，我们一般会画直方图（Histogram）。现在你也可以用核密度估计。

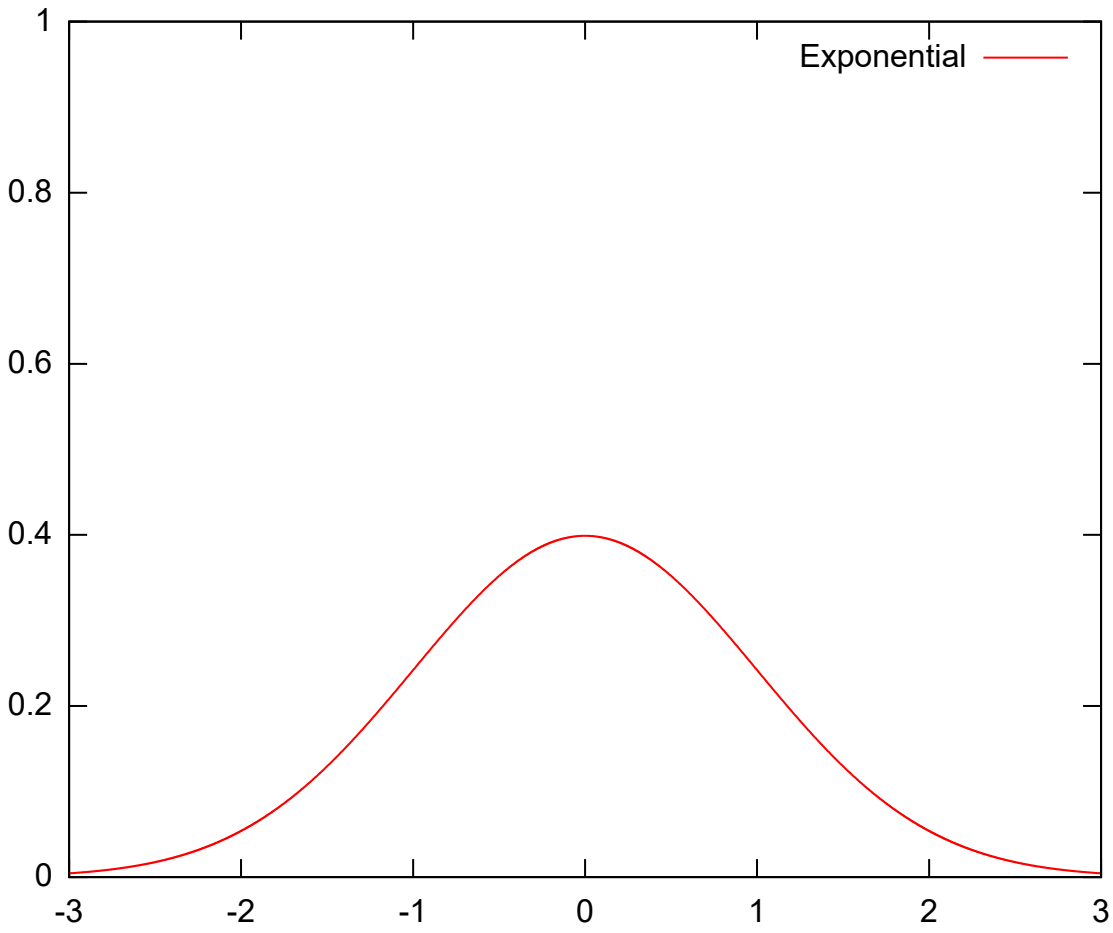
什么是“核”

如果不了解背景，看到“核密度估计”这个概念基本上就是一脸懵逼。我们先说说这个核 (kernel) 是什么。

首先，“核”在不同的语境下的含义是不同的，例如在模式识别里，它的含义就和这里不同。在“非参数估计”的语境下，“核”是一个函数，用来提供权重。例如高斯函数 (Gaussian) 就是一个常用的核函数。

让我们举个例子，假设我们现在想买房，钱不够要找亲戚朋友借，我们用一个数组来表示 5 个亲戚的财产状况：`[8, 2, 5, 6, 4]`。我们是中间这个数 `5`。“核”可以类比 成朋友圈，但不同的亲戚朋友亲疏有别，在借钱的时候，关系好的朋友出力多，关系不好的朋友出力少，于是我们可以用权重来表示。总共能借到的钱是：`8*0.1 + 2*0.4 + 5 + 6*0.3 + 4*0.2 = 9.2`。

那么“核”的作用就是用来决定权重，例如高斯函数（即正态分布）：



如果还套用上面的例子的话，可以认为在 3 代血亲之外的亲戚就基本不会借钱给你了。

最后呢，一般要求核函数有下面两个性质：

- 归一化：$\int_{-\infty}^{+\infty} K(u)du = 1$
- 对称性：对所有 u 要求 $K(-u) = K(u)$

最后的最后：一些常用的核

核密度估计

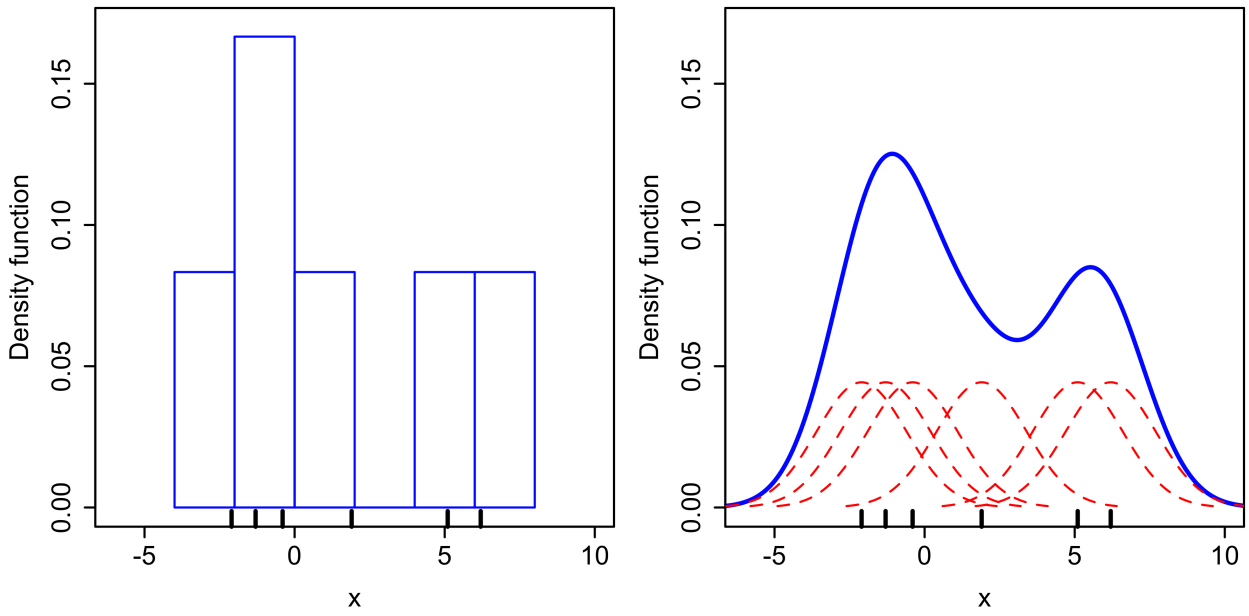
理解了“核”，核密度估计就容易理解了。

如果我们画直方图，其实目的是画出“概率密度函数”，而直方图本质上是认为频率等于概率。但这种假设不是必然的。核密度函数就是一种“平滑(smooth)”的手段。相当于是“我说我很牛逼你可能不信，但你可以听听我的朋友们是怎么评价我的，加权平均下就能更好地了解我了”。于是乎：

设 $(x_1, x_2, \ldots, x_n)$ 是独立同分布的 n 个样本点，它的概率密度函数是 f，于是我们的估计：

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x - x_i}{h})$$

上面式子中 `h` 是人为指定的，代表“朋友圈”的大小，正式的叫法是“带宽”(bandwidth)。而 $x - x_i$ 就是自己与朋友的亲疏程度，当然最后要正归化到 `[-1, 1]` 之间。下图是直方图和核密度估计的一个对比：



选择合适的带宽

选择不同的带宽，核密度估计的结果也大不相同，因此人们研究了一些算法来选择带宽。这方面对理解 KDE 本身没有什么太重要的意义，并且常见的算法在 scipy 里也已经都实现了，这里就不细说了，有兴趣的看看 wiki 吧。

参考

- http://blog.csdn.net/pipisorry/article/details/53635895
- 一维数据可视化：核密度估计(Kernel Density Estimates)

4 Comments

Login ▾

G

Join the discussion…

LOG IN WITH

OR SIGN UP WITH DISQUS ?

Name

♡ 2

Best Newest Oldest

T

Tiffany

5 years ago

看了这么多 你讲的最明白了。赞一个

1 0 Reply •

刘

刘洋

2 years ago

博客很简洁，赞一个

0 0 Reply •

Mealie

3 years ago

讲得很不错，多谢

0 0 Reply •

W

wei wu

4 years ago

很易懂的讲解，谢谢

0 0 Reply •

Subscribe Privacy Do Not Sell My Data