Problem Statement











‘

‘

地理相关性

与传统推荐系统中的非空间项目（如常规推荐系统中的音乐）不同，LBSN用户需要与POI进行物理交互，以消费其提供的产品或服务，例如在餐厅吃饭或在电影院看电影。因此，POI的地理邻域在用户的签到行为中起着重要的作用。换句话说，近POI具有比远POI更强的地理相关性。因此，我们可以利用用户访问的POI和她未访问的POI之间的地理相关性来估计用户对未访问的POI的相关性得分。为了模拟POI之间的地理相关性，我们根据用户自己访问的POI估计每个用户的地理坐标的个性化签到分布情况。

到目前为止的POI推荐算法中多数采用非参数估计的方法（即固定带宽的核密度估计）来从用户签到的POI中学习得到分布，正是因为这样，固定带宽并不能反映用户签到数据中的事实：密集的城市地区的签到密度很高，人烟稀少的农村地区的签到密度较低。

预备知识：密度估计相关知识

密度估计经常在统计学中作为一种**基于有限的样本来估计其概率密度函数**的方法。

在研究随机变量的过程中，**随机变量的概率密度函数的作用是描述随机变量的特性。**但是在实际应用中，总体概率密度函数通常是未知的，那么如何来估计总体概率密度呢？一般，我们通过抽样或者采集一定的样本，可以根据统计学知识从样本集合中推断总体概率密度。这种方法统称为概率密度估计，即根据训练样本来确定随机变量的概率分布。一般概率密度估计方法方法大致分为两类：参数估计和非参数估计。

参数估计(Parametric Estimation)是根据对问题的经验知识，假设问题具有某种数学模型 ，随机变量服从某种分布，即**假定概率密度函数的形式**，**然后通过训练数据估计出分布函数的参数。**常见的参数估计方法有**最大似然方法**和**贝叶斯方法**。对于参数估计，根据样本中是否已知样本所属类别将参数估计又划分为**监督参数估计**和**非监督参数估计**。监督参数估计是由 已知类别的样本集对总体分布的某些参数进行统计推断 。而无监督参数 估计已知总体概率密度函数形式但未知样本所属的类别 ，要求推断出概率密度函数的某些参数 ，这种推断方法称之为非监督情况下的参数估计。

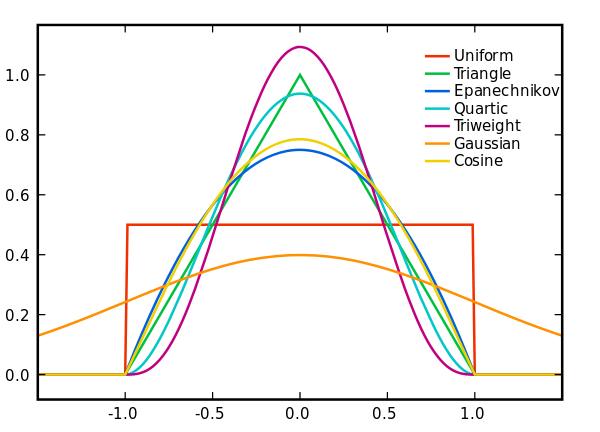
非参数估计(Nonparametric Estimation)则是在 **已知样本所属的类别不假定总体分布形式下 ， 基于大样本的性质，直接利用样本估计出整个函数。**在很多情况下，我们对样本的分布并没有充分的了解，无法事先给出密度函数的形式，而且有些样本分布的情况也很难用简单的函数来描述。在这种情况下，就需要用到非参数估计。但是，并不是非参数估计一定优于参数估计，因为非参数估计受训练样本影响，其完备性或者说是泛化能力不会很好；且这种估计只能用数值方法取得，无法得到完美的封闭函数图形。常用的非参数估计方法有**直方图法**，**核概率密度估计**等。

这里说说核密度估计方法

核密度估计的形式：



这里是核函数（非负，积分为1，均值为0，符合概率密度的性质），h>0是带宽。有很多种核函数，uniform,triangular, biweight, triweight, Epanechnikov,normal,等。各种核函数的图形如下：



由于高斯内核方便的数学性质，也经常使用 K(x)= ϕ(x)，ϕ(x)为标准正态概率密度函数。核密度估计与直方图很类似，但相比于直方图还有光滑连续的性质。

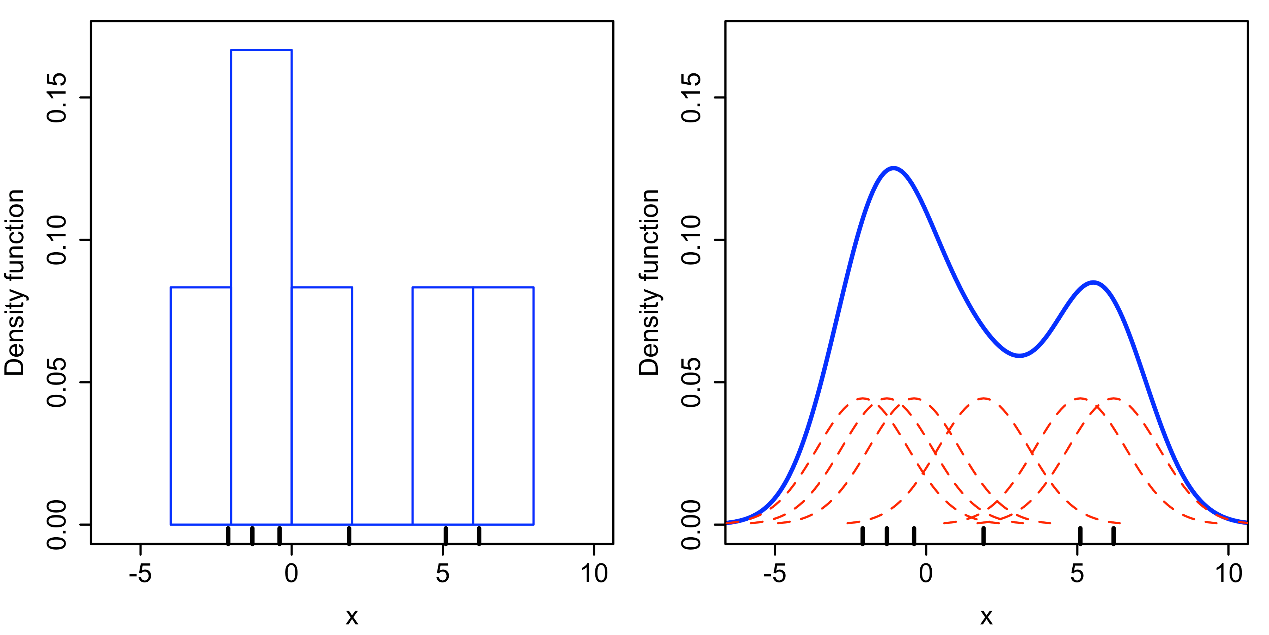
举例理解

已知：6个数据点*x*1 = -2.1，*x*2 = -1.3，*x*3 = -0.4，*x*4 = 1.9，*x*5 = 5.1，*x*6 = 6.2

对于直方图，首先将水平轴划分为覆盖数据范围的子间隔或区段。在这种情况下，我们有6个宽度为2的矩形。每当数据点落在此间隔内时，我们放置一个高度为1/12的矩形。对于核密度估计，我们在每个数据点xi上放置方差2.25（由红色虚线表示）的正态核函数。叠加一起得到核密度估计的结果，蓝色线表示。

很明显，直方图得到的密度估计平滑程度比使用核密度估计得到的密度函数要差很多.

那么问题就来了，如何选定核函数的“方差”呢？这其实是由h来决定，不同的带宽下的核函数估计结果差异很大。

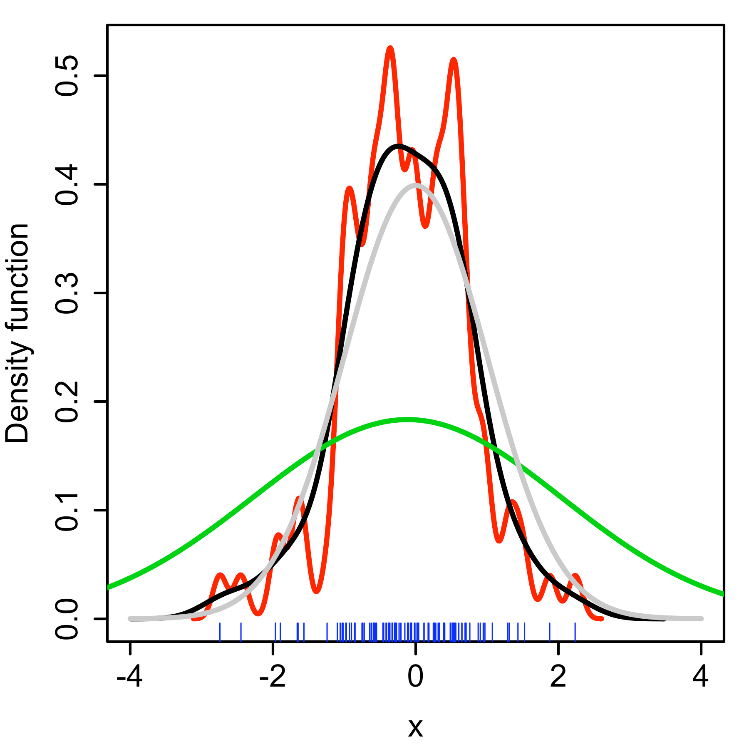


核带宽的选择

带宽是一个自由参数，对所得到的估计值有很大的影响。为了说明效果，举个例子：

下图是从标准正态分布中抽取的随机样本（横轴上的蓝色的点点代表样本点）灰色曲线是真是的概率密度（正态密度，均值0，方差1）。相比之下，红色曲线是使用了过小的带宽h（=0.05）得出的概率密度曲线，可见其波折陡峭；绿色曲线过于平滑，因为它使用了过大的带宽h（=2），掩盖了数据大部分基础结构。

那么对于h的选择可以使用最小化L2风险函数（即均值积分平方误差）。



如果使用高斯核函数进行核密度估计，则*h*的最优选择（即使平均积分平方误差最小化的带宽）为



这里{\ displaystyle {\ hat {\ sigma}}}是样品的标准差。这种近似称为*正态分布近似*，高斯近似，或*Silverman（1986）经验法则*。虽然这个经验法则很容易计算，但应谨慎使用，因为当密度不接近正态时，可能会产生泛化极差的估计。

推导过程详见：

Silverman, B.W.(1986).Density Estimation for Statistics and Data Analysis. London: Chapman & Hall/CRC. p. 48.ISBN 0-412-24620-1.

这里带宽的作用简述：

1.在数据可视化的相关领域中，带宽的大小决定了核密度估计函数（KDE）的平滑（smooth）程度，带宽越小越undersmooth，带宽越大越oversmooth。（详细解释）

2.在本篇文章中带宽h的设置主要与分析尺度以及地理现象特点有关。较小的带宽可以使密度分布结果中出现较多的高值或低值区域，适合于揭示密度分布的局部特征，而较大的带宽可以在全局尺度下使热点区域体现得更加明显。另外，带宽应与兴趣点的离散程度呈正相关，对于稀疏型的兴趣点分布应采用较大的带宽，而对于密集型的兴趣点则应考虑较小一些的带宽。。

如果带宽不是固定的，而是根据样本的位置而变化，则会产生一种特别有力的方法，称为**自适应或可变带宽内核密度估计**。由于密集的城市地区的签到密度很高，人烟稀少的农村地区的签到密度较低。就是说不同位置应该采取不同的分析尺度，因此本文采用不固定的带宽来进行核密度估计。

1. GeoSoCa:利用地理，社交，分类相关性的兴趣点推荐

2.主要贡献

三个相关性模型：地理相关性、社交关系相关性、分类相关性

GeoSoCa，第一个融合这三种相关性来进行兴趣点推荐的框架

使用两个真实数据集显示GeoSoCa在兴趣点推荐上面有显著的性能。

3.目标

给定五个已知量：

兴趣点的地理坐标，签到矩阵，社交关系矩阵，分类矩阵以及人气矩阵，

预测用户对未访问过的兴趣点的偏好评分，取top-k推荐给该用户。

分别解释上面的五个已知量：

**兴趣点的地理坐标**，很好理解，即所有的POI的经纬度坐标(x,y)。

**签到矩阵**，使用来自LBSN的POI用户的历史签到数据，我们就能构建出签到矩阵，在这个矩阵中，每一个实体代表用u在兴趣点l上的签到频数，这里大写的U表示用户集合，大写的L表示兴趣点集合，注意，矩阵R中多数的实体是0，因为在LBSN中用户仅仅到过很少的兴趣点。

**社交关系矩阵**，利用来自LBSN的用户之间的社会联系，很容易构建出一个社会关系矩阵，在这个矩阵中，如果两个不同用户u和u‘存在一个社会关系，也就是说，他们是好友，那么对应的值就是1，否则是0。

**分类矩阵**，利用用户在POI的历史签到数据以及来自LBSN的POI类别数据，就可以构建出分类矩阵，在这里，每一个实体表示用户u去过的兴趣点（多个）属于类别c的频数，这里大写的C表示在LBSN中预先定义好的POI类别的集合。注意，一个POI可以属于多个类别。

**人气矩阵**，构建一个人气矩阵，每一个实体表示类别c中兴趣点l的受欢迎程度，也就是说，在兴趣点l上，对于所有用户来说，其被签到的频数，若兴趣点l不是c类别的，那么该频数就为0。P中多数的实体是为0的，因为一个兴趣点只属于那么一两个类别。

知道了这些量，目标是预测用户对于没有访问过的兴趣点的偏好程度，取top-K推荐给该用户。

4.三种重要的相关性

在本文中，利用用户和未访问POI之间的三种重要类型的相关性来预测任何目标用户对她未访问的POI的偏好分数。

**地理相关性**。有一条地理定律[17]指出：“Everything is related to everything else, but near things are more related than distant things.”例如，在现实中，一个人经常访问一些POI比如博物馆，然后他也会前往其附近的POI，例如餐馆、商店或者电影院。也就是说，距离比较近的POI比距离比较远的POI具有更强的地理相关性。从而，我们可以根据用户访问过的POI和未访问的POI之间的地理相关性来估计未访问POI的用户相关性得分。文中利用了概率估计来度量用户对其未访问的POI的偏好得分。在地理相关性这块，作者使用核密度估计来估计用户对未访问的POI的偏好得分。

**社会相关性**。在现实世界中，一个人可能会喜欢她朋友强烈推荐的POI。例如，经常会和朋友一起去博物馆，餐馆或商店。换句话说，好友比陌生人更可能分享共同的爱好。同样，在LBSN中，用户会建立彼此的社交联系来分享他们访问POI的经历。因此，对于同一个POI，某一个用户没有访问过，而他的朋友访问过，计算这两个用户之间的社交相关性，然后根据这个相关性来估计这个用户与这个POI之间的偏好得分。

**分类相关性**。LBSN通常预先定义一组兴趣点的类别。比如：在火车站的麦当劳，就属于快餐类别，属于快餐类别的兴趣点还有很多，可能是中街的麦当劳，也可能是火车站的肯德基，等等。实际上，人们对POI类别有不同的偏爱：一个吃货经常访问餐馆品尝各种食物，而驴友通常在全球旅游景点旅行。因此，可以根据用户访问过的POI类别和未访问的POI的类别之间的相关性来计算出用户与未访问POI之间的偏好得分。

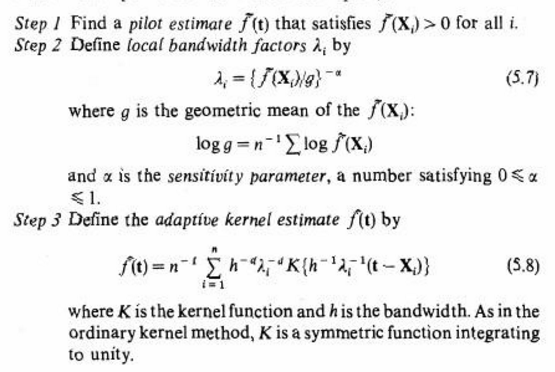
5.地理相关性

这部分的目标是：给出用户u喜欢兴趣点l的概率密度。

分为三个步骤，这三个步骤是自适应带宽的核密度估计的标准步骤，在silverman的一本书（Silverman, B.W. (1986). Density Estimation for Statistics and Data Analysis. London: Chapman & Hall/CRC. p. 48. ISBN 0-412-24620-1.）里完整的给出了这三个步骤。这篇文章是套用了这个密度估计的步骤。

首先，初步估计。使用固定带宽的核密度估计来进行初步估计。这里是概率密度估计函数，是核函数，该核函数为高斯核函数，其全局带宽H是由经度和纬度两个全局带宽组成。

第二步，局部带宽的确定。这里并不是直接使用第一步得出的概率密度估计函数来估计分数。因为对于不确定的数据分布来说，使用固定带宽的核密度估计可能会出现估计不准确的情况，因而，这里使用自适应带宽的核密度估计函数来进行估计，由第二步和第三步共同完成。第二步，确定局部带宽。这里使用初步估计得出的概率密度函数来确定局部带宽h\_i，这个h\_i是针对每一个被签到的POI的。这里\alpha是敏感因子，取值在0到1之间，\alpha越大，局部带宽h\_i就越敏感，g是为了保证所有的h\_i的几何平均值是1.(为什么是这个形式，论文里没有给出，我也没看懂，不过我知道这是一个成熟的理论，在Silverman的书里有讲到。)



第三步是自适应带宽的核密度估计，用于获取地理评分。这里使用上一步的h\_i作为自适应带宽，使用第一步的H作为全局带宽，那么自适应带宽的核密度估计就如公式所示，这里K\_{Hh\_i}代表自适应带宽的核函数。

这样使用自适应带宽的方法目的是，当POI l\_i附近的签到密度较大的时候，会得出一个比较小的局部带宽h\_i，这样得出一个在l\_i附近陡峭的走势来描述数据的局部特性；相反，当POI l\_i附近的签到密度较小的时候，会得出一个比较大的局部带宽h\_i，这样得出一个在l\_i附近平坦的走势来描述数据的局部特性。

6.举个例子

数据来自Foursquare和Yelp，都是用户签到的数据，alpha=0.5,图是某一个用户签到的地理分布轮廓线。

第一个图，是用户在全世界范围内的（基本上是两个区域），世界级别的；而第二幅图是用户只在亚利桑那州签到的兴趣点分布轮廓线（大概三个区域）。这就体现了局部带宽的作用。如果都是使用全局带宽的，那么可能用户在他所长期处于的范围内不能够被充分的描述出来。

7.社交相关性

前面说过，社交相关性会影响用户的签到行为。

也分为三个步骤

第一步是使用用户u的所有好友在POI l上签到的频数来描述用户u对其未访问POI l的关联得分。这就如同传统的协同过滤的思路一样。