Assignment 1

非参数化方法

一、简介

本次作业使用 3 种非参数化算法——直方图法，核密度估计，K 近邻算法，对 1 个未知的数据分布进行概率密度估计，并在这过程中呈现了这 3 种算法的特性和参数的选择方法。

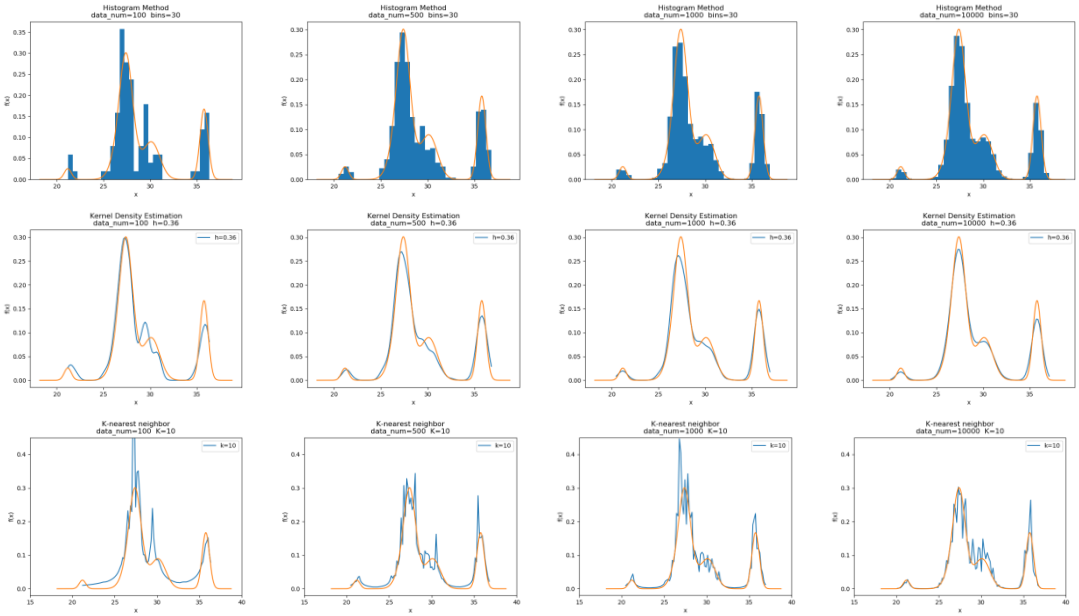
展示算法结果的 Python 程序需通过以下格式的命令启动：

```
python3 source.py --hist data_num bins
python3 source.py --kde data_num h
python3 source.py --knn data_num k
python3 source.py --cv h_min h_max （交叉验证法选取 kde 算法中的 h 值）
```

二、数据集规模的影响

由大数定理可知，当训练集数据量越大时，生成的概率密度模型与真实分布越接近。

使用以下 3 个问题中得到的表现较好的参数，用 100，500，1000，10000 的数据规模训练模型，与真实分布进行对比，结果如下：



可以很明显的看出，对这 3 种算法，当训练集规模增大时，得到的概率模型越来越接近真实分布。

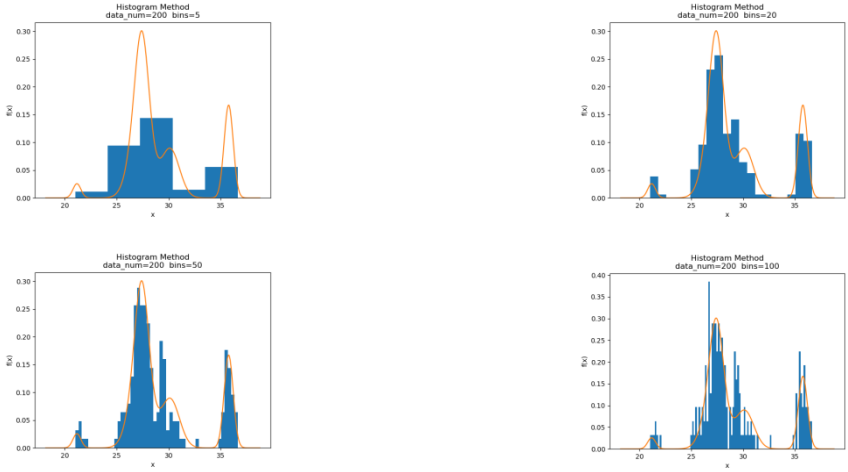
然而对于 KNN 算法，数据规模的增大并不能明显使得模型曲线中的**尖峰数量**减少，而尖峰的存在会影响模型的泛化性能，这也许是 KNN 算法的特点。我认为这与 KNN 算法得到的概率模型在整个空间积分发散有关。因为空间中往往存在数据密度特别大的区域，存在**K 个点几乎聚集在同一个位置**的情况，这时通过 KNN 算法得到的该位置的概率密度会相当大，甚至达到无穷大。这便是出现许多尖峰的原因。（相关证明在第 4 个问题的回答中）

三、直方图法

直方图法将数据变量 x 的取值范围划分为宽度为 Δ 的若干区间（Bins），对落在每个区间中的 x 的观测数量进行计数。对于第 i 个区间 Δ_i，我们可以描述它的概率：

$$\mathrm{P_i} = \frac{n_i}{N \bullet \Delta_i}$$

在每个区间内概率密度是常数。我们选取不同的区间宽度（区间数），对未知分布进行估计（与真实分布曲线进行对比）：



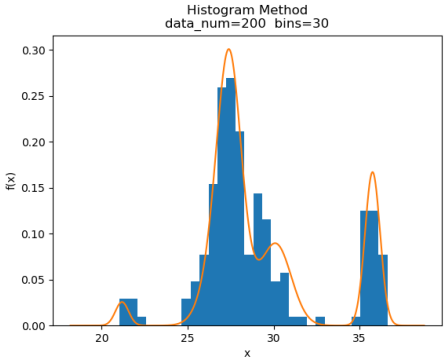
可以发现，当选取的区间**宽度较大**时，直方图呈现的概率模型**过于平滑**，缺失了很多信息；当选取的区间**宽度较小**时，直方图中会有很多**尖刺**，且很多区间中**没有点落入**，即发生了**“过拟合”**。当选取的区间宽度适中时，我们会得到较好的概率模型。但如何选取最佳的区间宽度呢？

我们可以找到一些**经验规则**来计算区间宽度，例如”Freedman－Diaconis rule”。但是这些经验规则往往需要数据分布的信息，或者在大数据集下不适用。因此我们最好进行**对比实验**，选取若干不同的区间宽度（区间数），绘制直方图，根据这些直方图的一些特征，来选择最佳的区间宽度。

1. 为避免过拟合，其中一点就是避免区间之间巨大的不连续性。我们需要**适当减少空区间(empty bins)的出现**。因此我们可通过**空区间数占总区间数的比例**来评估直方图的好坏。

2. 在**空区间占比较低**的直方图之间，为避免模型过于平滑，我们选择**区间宽度较小，**的直方图，但同时也要避免**孤立尖峰数量过多**。由于随着区间数的上升，空区间占比往往是呈现上升趋势，不能直接得出一个极值点，所以 1、2 中的特征要综合考虑，同时避免欠拟合和过拟合。

本例中我们认为**较好的区间数量为 30**。



四、核密度估计方法

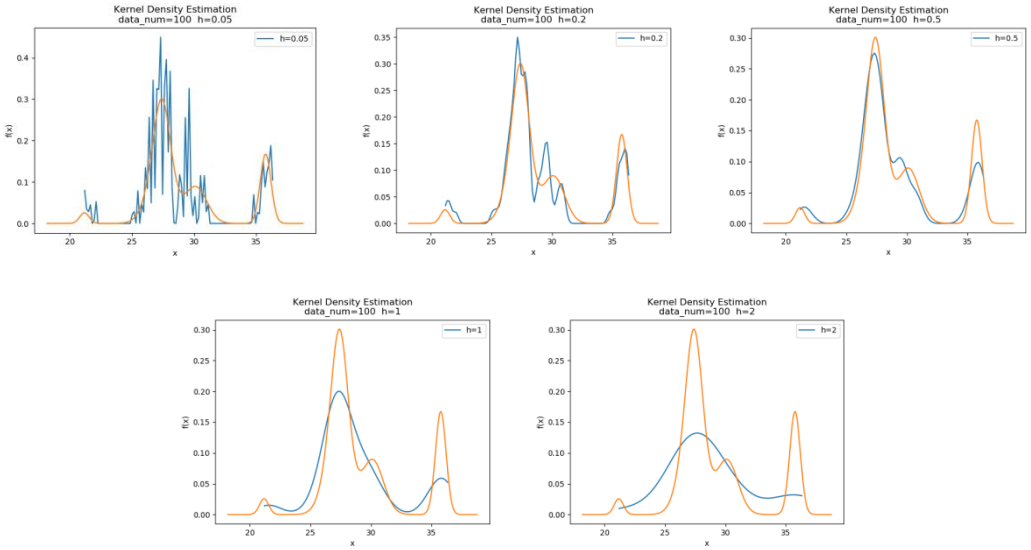
对于空间中某一个位置 x，核密度估计方法会根据训练集中所有数据的信息来估计该位置的概率密度，训练集中每个数据点的贡献和它与位置 x 的距离有关。用于评估的函数被称为核函数。该方法的基本公式为：

$$p(\boldsymbol{x}) = \frac{K}{NV}$$

为了得到更加平滑的概率密度模型，我们使用高斯核函数，概率密度公式如下：

$$p(\boldsymbol{x}) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{(2\pi h^2)^{\frac{D}{2}}}\exp\left\{-\frac{\|\boldsymbol{x}-\boldsymbol{x}_n\|^2}{2h^2}\right\}$$

其中 h 是高斯分布的标准差，在这里又称作带宽。我们很容易发现模型复杂度的重要因素是这个带宽 h。对 h 取不同值，得到如下若干概率密度模型：



可以发现，<u>当 **h 很小**时，模型曲线**有许多尖刺**，连续性差，泛化性能差；当 **h 较大**时，曲线又**过于平滑，丢失很多信息**。</u>因此我们要选择一个适中的带宽 h 来得到具有良好泛化性能的模型。

一种优化方法是**使误差函数（error function）最小**。在密度估计中常使用平均积分平方误差（mean integrated squared error）：

$$\mathrm{MISE}(h) = \mathrm{E}\left[\int (\hat{f}_h(x) - f(x))^2\, dx\right]$$

其中 f 为真实分布。当$\frac{\partial}{\partial h}MISE(h) = 0$时 h 的取值使模型最优化。但由于我们不知道真实分布 f，我们无法在数学上直接得到 h 的最优取值。至此我们有 2 种方法去选择 h：

1. **对真实分布进行近似假设**。前人的研究中，有将真实分布近似为高斯分布的方法，这样可以得到 h 的经验公式（Silverman's rule of thumb）：

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

其中$\hat{\sigma}$为样本的标准差。但由于我们这里的分布曲线至少有 2 个峰，这种近似得到的结果应该不会很好。

3. **基于数据集的带宽选择方法**。我们可以使用测试集对模型进行泛化性能的评估，从而选择出使模型最优化的带宽 h。在带宽选择问题中，常采用**交叉验证法**，将数据集分为 k 份，将(k-1)份作为训练集，余下 1 份作为测试(验证)集对模型进行评估。这样可进行 k 次训练。

避免直接采用复杂的积分平方误差函数推导出的风险函数（risk function）：

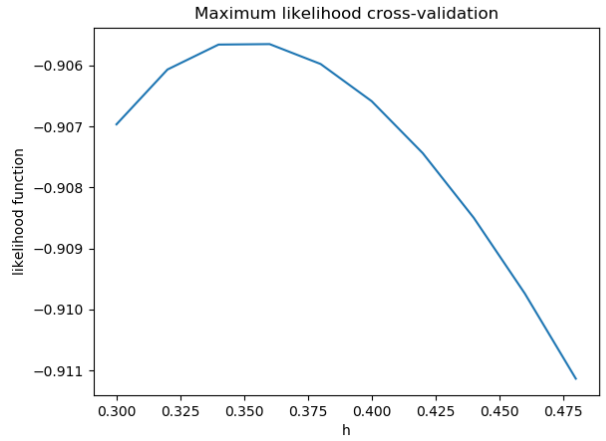
$$\widehat{R}(h) = \int \left(\widehat{p}_{(-i)}(x)\right)^2 dx - \frac{2}{n}\sum_{i=1}^{n}\widehat{p}_{(-i)}(X_i)$$

我们这里直接使用**最大似然方法进行留一法验证**，找出使对数似然函数log(p(x|h))最大的 h 值。公式如下：

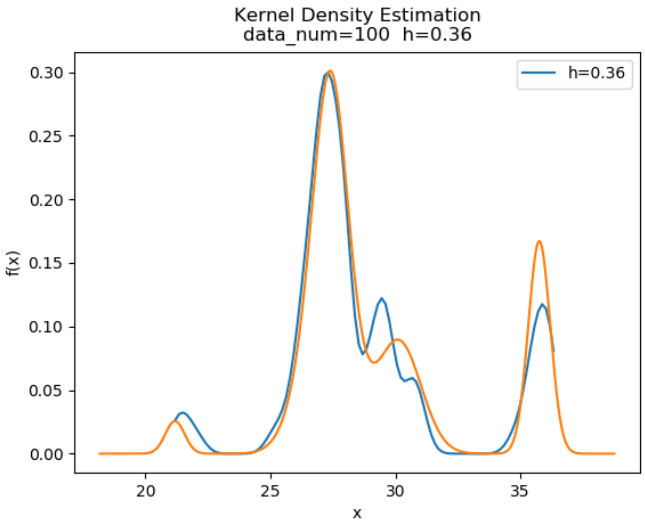
$$h^* = \arg\max\left\{\frac{1}{N}\sum_{n=1}^{N} log p_{-n}(x^{(n)})\right\}$$

$$where\ p_{-n}(x^{(n)}) = \frac{1}{(N-1)h}\sum_{\substack{m=1 \\ m\neq n}}^{N} K\left(\frac{x^{(n} - x^{(m}}{h}\right)$$

本例中我们将 h 的取值控制在[0.3，0.5]，以步长=0.02 进行“探测”。结果如下：



使得似然函数最大的 h=0.36。这是该精度下我们得到的 h 的最佳取值。将其代入高斯核函数，用所有的训练集数据重新训练得到概率密度模型如下：



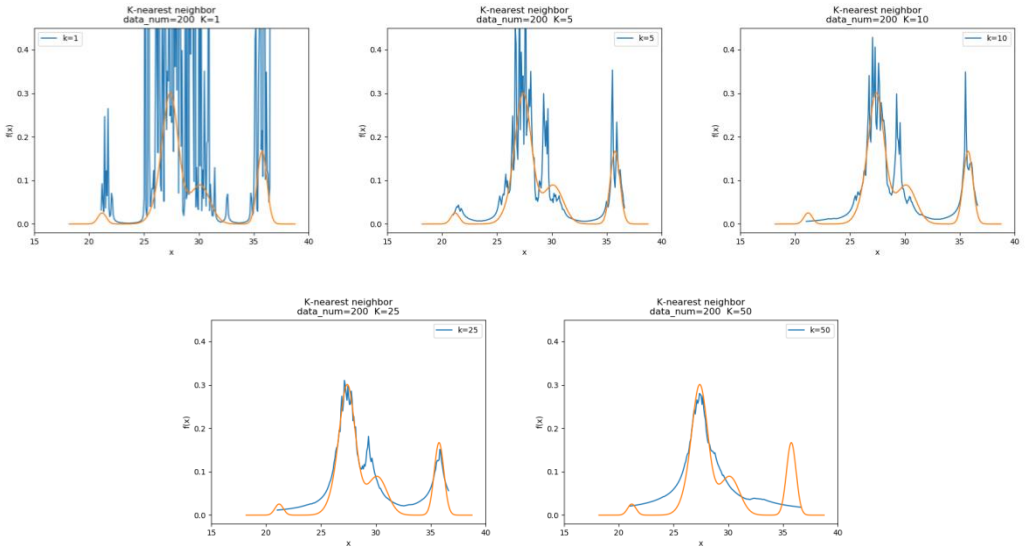
五、K 近邻方法

在核密度估计中，控制核大小的参数 h 对每个位置的核都是固定的，但由于空间中不同区域的数据密度不同，从局部角度看，h 的最优选择也不同。为了使核的大小能随着空间位置的改变而改变，根据概率密度公式：

$$p(\boldsymbol{x}) = \frac{K}{NV}$$

我们将核中数据点数目 K 固定，从而使得核体积 V 可变，得到一种可行的方法——K 近邻方法。

通过选取不同的 K 值，观察 K 对概率密度模型的影响：



可以发现，当 **K 值很小**时，曲线变得**十分陡峭**，模型**复杂度很大**，发生“**过拟合**”；当 **K 值很大**时，曲线变得**过于平滑**，模型变得**比较简单**，丢失了许多信息（比如丢失了一些峰）。因此我们仍需要选择一个适中的 K 值。本例中 K=10 时曲线尖刺较少，模型复杂度适中。

实际上这里的 K 值也可使用交叉验证法找出最佳选择。

证明：**K 近邻方法得到的概率模型在整个空间积分是发散的**。

根据概率密度公式：

$$p(\boldsymbol{x}) = \frac{K}{NV}$$

1. **特殊情况推断**。对于给定的 K 值，空间中某些位置可能本身就聚集了 K 个数据点，这时**概率密度公式中的体积 V=0**，因此**该点处概率密度为无穷大**。而对于所有点，$p(x) > 0$。所以总的积分是发散的。

2. **数学推导**。对于一维分布，假设训练集中共有 n 个数据点，从小到大依次为 x1，x2，x3，…，xn。对于一维空间中区间 $(-\infty, x1)$ 中的点 x，**它的第 k 个近邻点一定是 xk**。由于对空间中任意 x，$p(x) > 0$。所以可得：

$$\int_{-\infty}^{+\infty} P(x)dx > \int_{-\infty}^{x1} P(x)dx = \int_{-\infty}^{x1} \frac{K}{N * 2(x_k - x)} dx = \frac{K}{2N} \int_{x_k - x_1}^{+\infty} \frac{1}{x} dx$$

$$= \frac{K}{2N} * lnx|_{x_k - x_1}^{+\infty} = +\infty$$

即**积分发散**。易知该结论可推广到高维空间。

六、参考文献

1. Kernel density estimation, Wiki:
https://en.wikipedia.org/wiki/Kernel_density_estimation
2. Kernel Estimator and Bandwidth Selection for Density and its Derivatives, by Arsalane Chouaib Guidoum.
3. 一些课程讲义：
http://research.cs.tamu.edu/prism/lectures/pr/pr_l7.pdf
http://homepages.inf.ed.ac.uk/rbf/CVonline/LOCAL_COPIES/AV0405/MISHRA/kde.html
http://www2.stat.duke.edu/~wjang/teaching/S05-293/lecture/ch6.pdf
4. Silverman, B.W. (1986). Density Estimation for Statistics and Data Analysis. London: Chapman & Hall/CRC. p. 48.