PRML Assignment 1 报告

16307130076 赵伟丞

一、使用说明

使用-m 或—method 来指定使用的概率密度估计方法，可使用 hist（直方图），kde（核密度估计），nearest（最近邻方法）。

使用-n 或—numdata 来指定使用的样本数量。

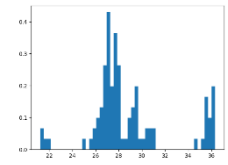
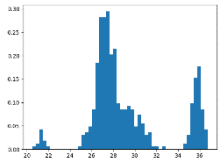
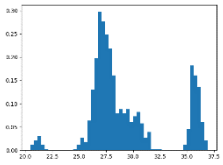
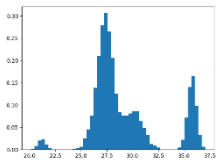
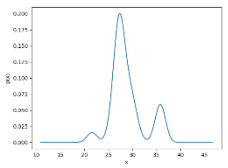
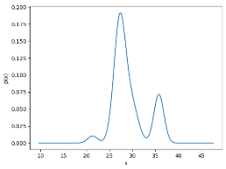
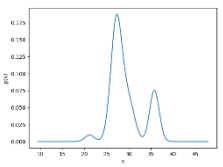
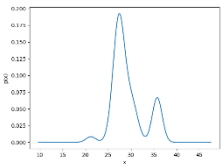
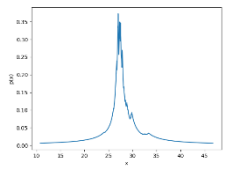
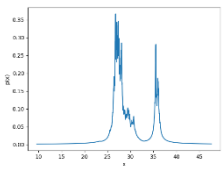
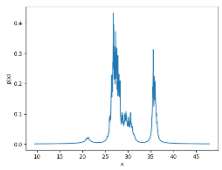
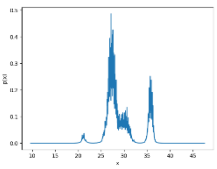
使用-pb 或—parabins 来指定直方图方法中的 bins 的数量。

使用-ph 或—parah 来指定核密度估计中的参数 h。

使用-pk 或—parak 来指定最近邻方法中的参数 K。

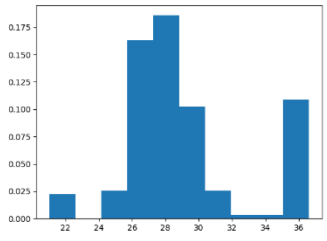
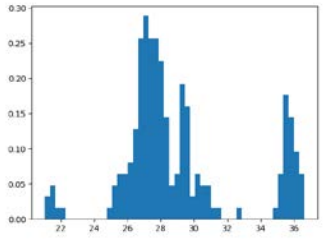
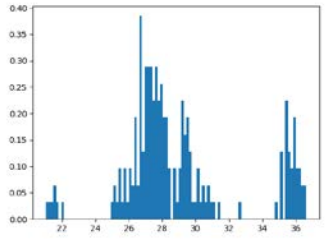
二、样本数量的影响

对于三种方法，样本数量越多，估计的结果越平滑，越有可能接近真实的概率密度分布函数。

N=100	N=500	N=1000	N=10000
			
			
			

三、直方图方法中 bins 数量的影响

bins 不能过多也不能过少，过多则会出现大量的值为 0 的区间，导致结果不平滑；若过少，则会导致过于平滑，导致部分特征消失。在 N=200 的情况中，我认为选择 50 左右的 bins 数量是恰当的（也就是总样本数的四分之一左右）。

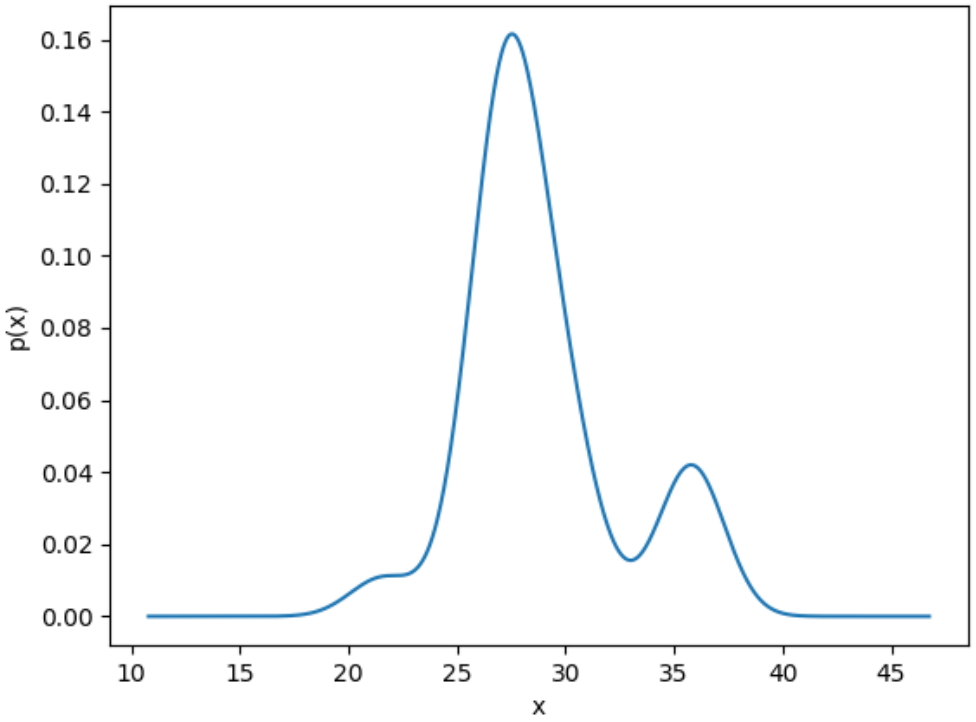
Bins=10	Bins=50	Bins=100
		

四、核密度估计方法中 h 的影响

h 过大，曲线过于平滑，会丢失信息；h 过小，曲线不够平滑，出现过拟合现象，为

了找到合适的 h，我们使均方误差最小，即 $\mathbf{MISE}(h) = \mathbf{E}\left[\int (\hat{f}_h(x) - f(x))^2\, dx\right]$. 最小。根据 Silverman, B.W. (1986). Density Estimation for Statistics and Data Analysis. London: Chapman & Hall/CRC. p. 48.中指出的对于高斯核的 h 公式：

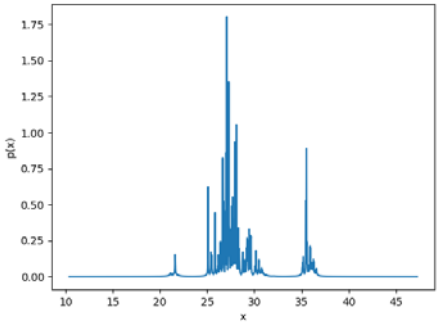
$h = \left(\frac{4\hat{\sigma}^5}{3n}\right)^{\frac{1}{5}} \approx 1.06\hat{\sigma}n^{-1/5}$，我们可以推算在这种情况下（N=100，std 约 3.473），h 最优解约 1.464.作出的图如下：

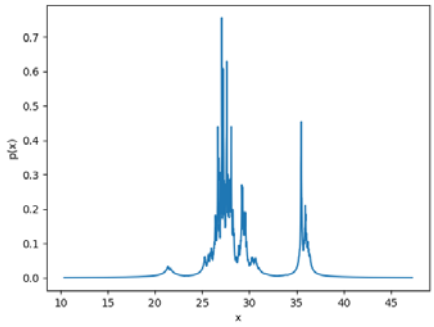
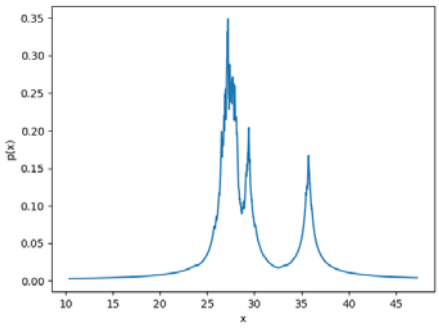
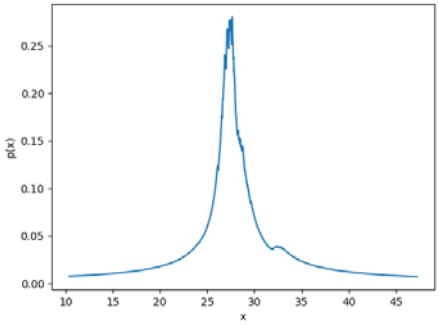


五、最近邻方法中 K 的影响

K 较小时，有过拟合的现象，K 较大时，曲线过于平滑，出现信息丢失的状况。

由于 V 的取值为 x 到 K-近邻点距离的两倍，是一个关于 x 的一次分段函数，因此在整个空间上求 p(x)对 x 的积分，会得到若干个形同 ln(||xi-xj||)的式子的和，这完全取决于样本点的具体分布，在据大多数情况下，都不会收敛到 1.

K=1	

k=5	
k=20	
k=50	
K=100