模式识别-第二次作业

王健

201628015029018

1. 第一题
   1. 对于一维特征空间中的Parzen窗估计，其估计得到的概率密度函数为：

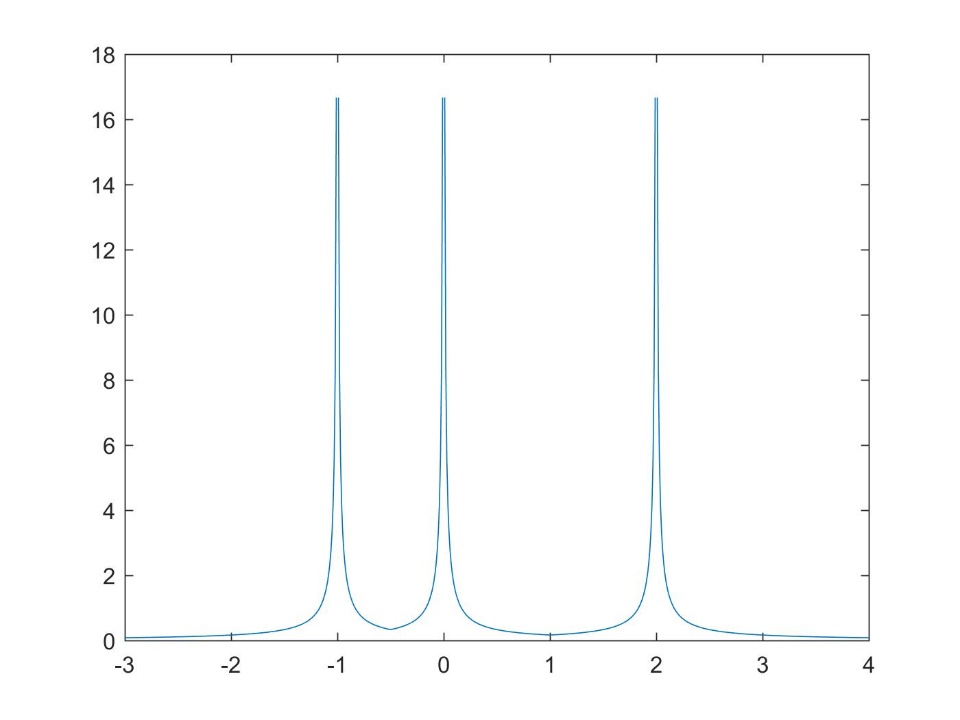


* 1. 对于一维特征空间中的最近邻估计，得到的概率密度函数为：



其中 表示x点与最近训练样本数据点之间的距离。

根据所给出的数据点，概率密度函数做图如下：



2、

（a）E step：





(b)最大化Q：





这个函数对于θ来说是单调的，则当Q最大的时候，θ2=3，此时Q=-8.52

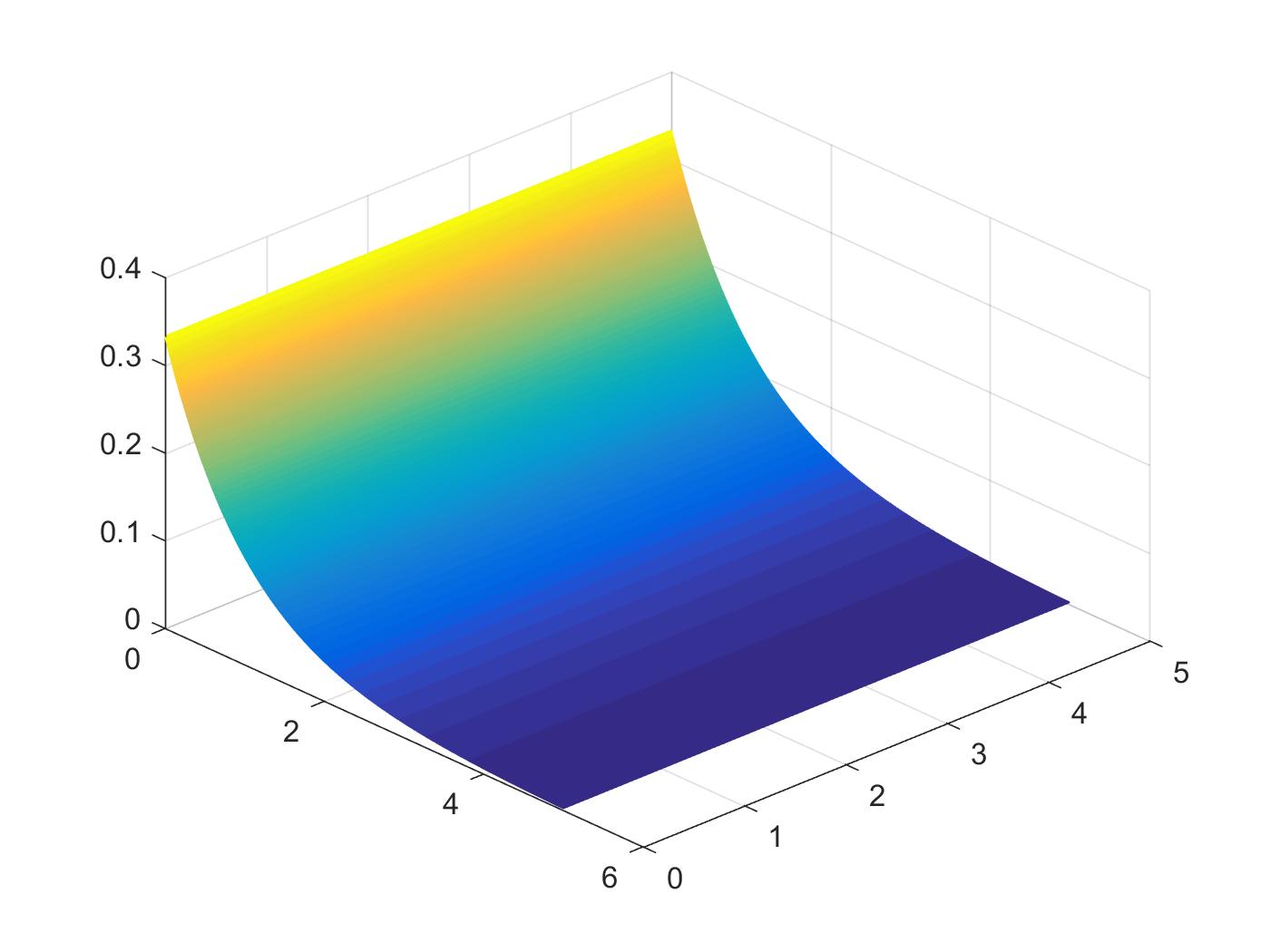




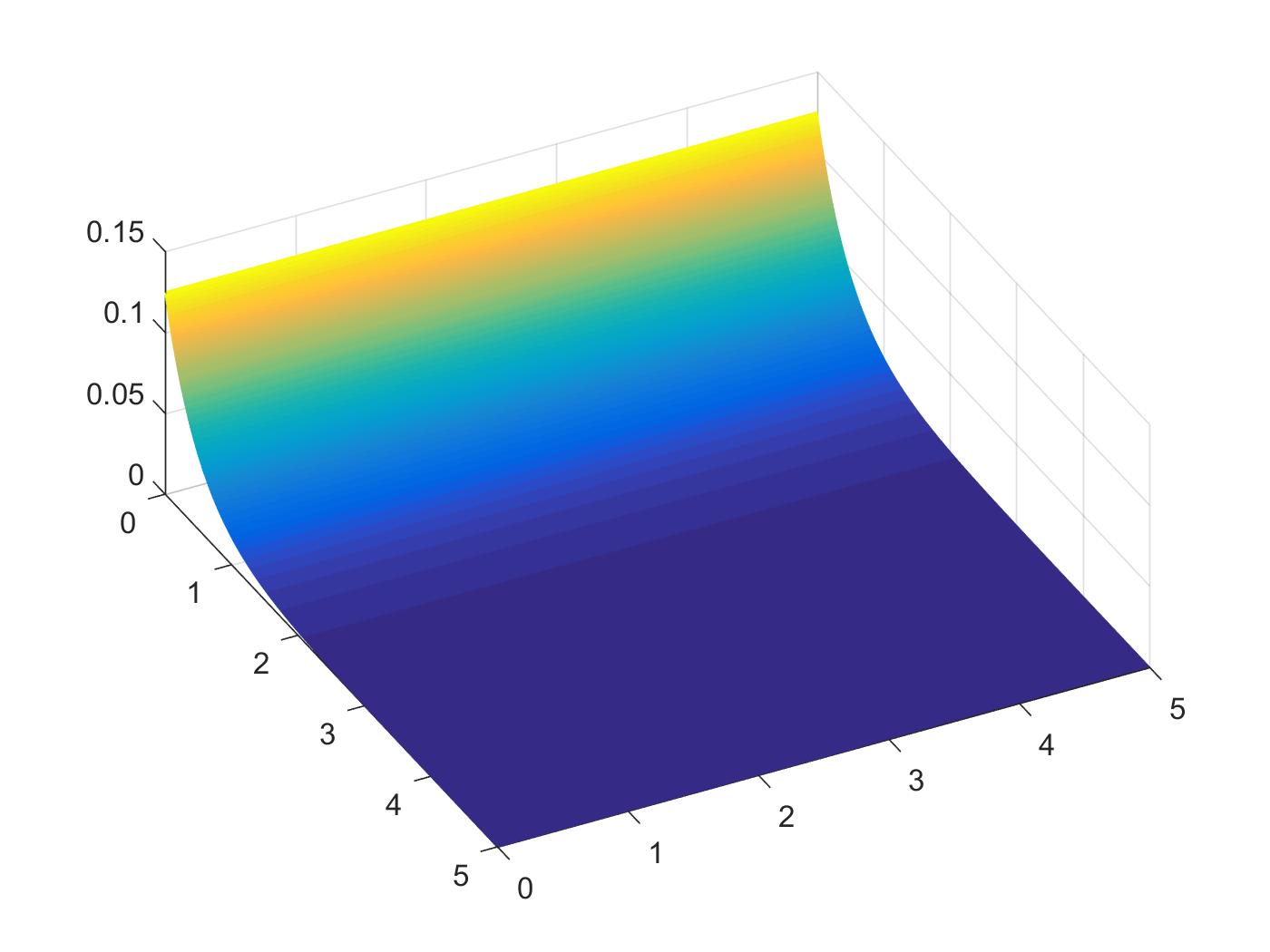
当Q最大的时候，θ2=4，此时Q=-10.16

综合来说，

（c）两个函数的图像为：



P(x,y)—θ=(1,3)



P(x,y)—θ=(2,4)

3、（1）表达式为：



（2）最大似然估计：



4、（1）1-NN规则：

对于测试样本点x，在已知的训练集合中寻找距离它最近的点，记为x‘，那么将点x分为x‘所属的类别

3-NN规则：对于测试样本点，在已知的训练集合中距离最近的三个点，记为x1,x2,x3，三个点中若存在两个或两个以上点属于某一类别，那么判别x为那个类别

（2）重新设计：

1-NN规则：

对于测试样本点x，在已知的训练集合中寻找距离它最近的点，记为x‘，那么将点x分为x‘所属的类别，如果有两个或者多个不同类别的点与x距离相同且最近，那么拒绝判别

3-NN规则：对于测试样本点，在已知的训练集合中距离最近的三个点，记为x1,x2,x3，三个点中若存在两个或两个以上点属于某一类别，那么判别x为那个类别。如果在距离点x前三近的点不止一个，那么可以将它们都考虑进来，设为x1,x2,…,xn，如果有恰好半数个点属于某一类，另外半数个点属于另一类，那么拒绝判别，否则则判别为多数的类别。

（3）优缺点：

1、优点  
简单，易于理解，易于实现，无需估计参数，无需训练  
适合对稀有事件进行分类  
特别适合于多分类问题  
  
2、缺点

对测试样本分类时的计算量大，内存开销大  
可解释性较差

5、本次使用了1-NN，3-NN，5-NN，三种方法对图像进行了分类，代码为python，程序见附件

三种分类方法的分类结果如下表所示：

表1 分类方法及其准确率

|  |  |
| --- | --- |
| 分类方法 | 准确率 |
| 1-NN | 0.9691 |
| 2-NN | 0.9713 |
| 3-NN | 0.9694 |

可以看出即使是最近邻方法，其分类准确率也相当高，而3-NN和5-NN对于准确率的提升效果不明显，而5-NN判断的准确率反而有所下降，可能是由于测试数据集中某些经过旋转或者平移的数字，与其欧氏距离近的数字类别多，例如5个最近的样本点的类别为[1,2,3,4,5]，在类似此种情况下，在欧式距离最近的5个样本点中选取最多的那个类别，则不是一个很好的策略，在这种情况下，选择距离最近的样本点所属的类别，可能相对比较准确。

针对1-NN方法进行分析，对于不同的数字分类准确率为：

表2 不同数字的分类准确率

|  |  |
| --- | --- |
| 数字 | 准确率 |
| 0 | 0.9929 |
| 1 | 0.9947 |
| 2 | 0.9612 |
| 3 | 0.9604 |
| 4 | 0.9613 |
| 5 | 0.9641 |
| 6 | 0.9854 |
| 7 | 0.9650 |
| 8 | 0.9446 |
| 9 | 0.9584 |

从表格中可以看出，数据对0,1此类容易分辨的数字，分类效果相当良好，但是对于8，9这种本身难于分辨的数字，分辨效果就稍差，考虑到在测试数据集中存在数字图像位置不在中心，数字图像有少许旋转等情况，分辨效果稍差也是可以理解的。