Homework of Pattern classification II

*Name: Xue Yuan — Student number: 202228015926034

***Abstract*—This document is about the first homework for Pattern classification by LaTeX.**

I. CALCULATION

1) 一维特征空间中的窗函数为标准正态分布的概率密度函数$p(x)$的Parzen窗估计$p_n(x)$:

$$p_n(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(x-x_i)^2}{2}\right)$$

2) 给定一维空间三个样本点$\{-4, 0, 6\}$，请写出概率密度函数$p(x)$的最近邻$(1-NN)$估计,并画出概率密度函数曲线图:
(1)概率密度函数:
由：

$$p_1(\mathbf{x}) = \frac{k_n}{nV_n} = \frac{1}{2n\,|x-x_1|}$$

可得：

$$p_n(x) = \frac{k_n}{nV_n} = \begin{cases} \dfrac{1}{6|x+4|}, & \text{if } x < -2 \\ \dfrac{1}{6|x|}, & \text{if } -2 < x < 3 \\ \dfrac{1}{6|x-6|}, & \text{if } x > 3 \end{cases}$$

(2)概率密度图像：

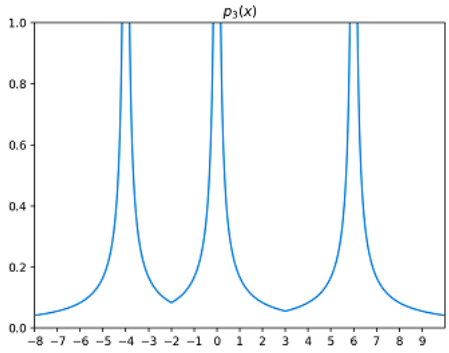

Fig. 1: Probability density function of 1-NN estimate

3) 现有7个二维向量:x_1=$(1,0)^T$,x_2=$(0,1)^T$,x_3=$(0,-1)^T$,x_4=$(0,0)^T$,x_5=$(0,2)^T$,x_6=$(0,-2)^T$,x_7=$(-2,0)^T$。这里上标T表示向量转置。假定前三个为ω_1类，后四个为ω_2类。画出最近邻法决策面。
由：

$$g_i(\mathbf{x}) = \arg\min_{\mathbf{x}_j\in\omega_j} d(\mathbf{x},\mathbf{x}_j), \quad i = 1, 2, \ldots, c$$

可作图：

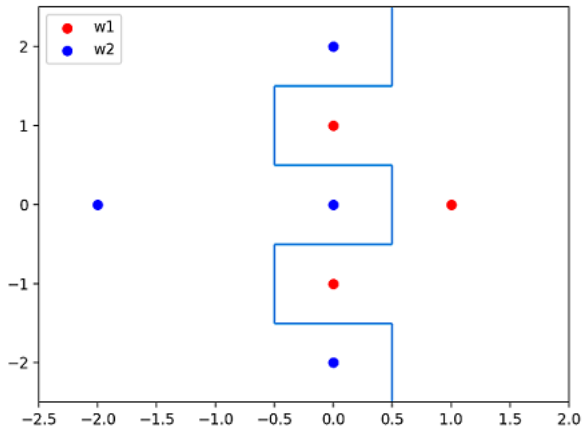

Fig. 2: Nearest neighbor decision surface

4) 请给出K近邻分类器的优点和缺点。
(1) 其优点在于:算法理论简单,容易实现:准确性更高,对异常值和噪声有较高的容忍度
(2) 其缺点在于:k近邻算法每预测一个"点"的分类都会重新进行一次全局运算,对于样本容量大的数据集计算量比较大.而且K近邻算法容易导致维度灾难,在高维空间中计算距离的时候,就会变得非常远;此外,它还存在着难以区分样本数据存在部分重叠时的情况等等。
5) 解答过程如下:

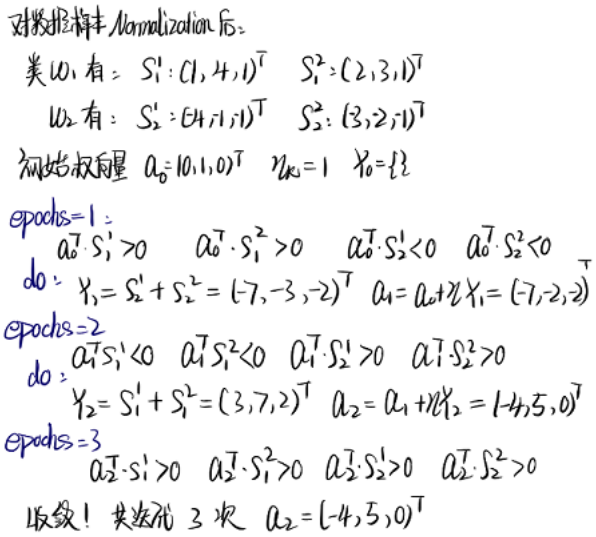

Fig. 3: Batch Perceptron flow chart

II. Programming

1) 不同窗宽取值下所估计获得的概率密度函数曲线:

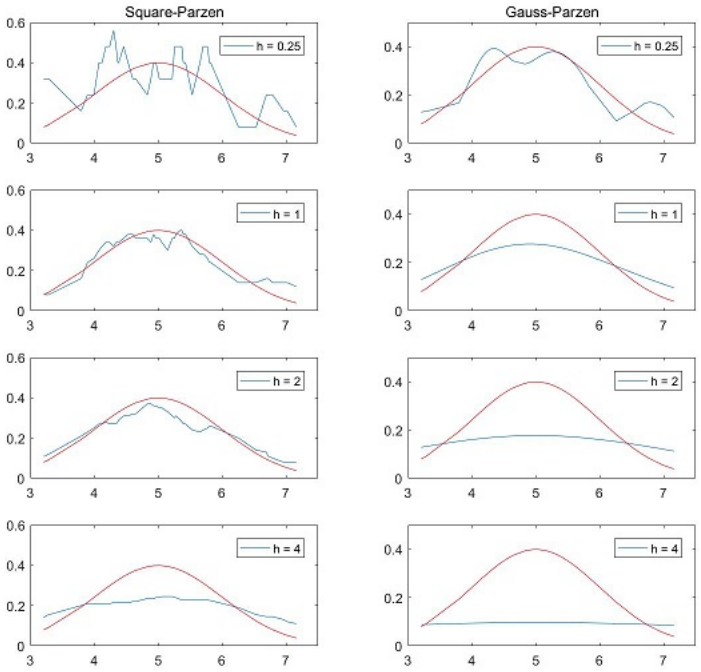

Fig. 4: Fitted images for two methods

图中蓝色线条是原始数据拟合得出,红色线条是$N(5,1)$的概率密度图像。可以看出窗大小(h)对拟合图像的好坏起到了十分明显的作用,并且不同的拟合方式对窗口大小的敏感成都也不同。本题所有代码均基于$Matlab$,源码放在附录中。

2) 线性分类器的构造与训练

(1)当采用$BatchPerception$时,程序运行结果如下:

```
(base) PS D:\VSCode_work> conda activate base
(base) PS D:\VSCode_work> & D:/Anaconda3/python.exe
Total iterations: 23
w1 and w2's Weight vectors is: [ 34.  -30.4  34.1]
Total iterations: 39
w3 and w4's Weight vectors is: [31.   5.1  5.2]
Total correct of MSE_expand methods is: 1
```

Fig. 5: Running result

可以看出,他们的迭代次数分别是23和39。此外,图中还给出了两种分类情况下的权矢量值。
他们整体上的运行分类结果表示在附录中
(2)如图5所示,当采用MSE_Expand方法时,其测试集正确率达到了100%! 本题目代码基于$python3.9$,相关程序也在附录中展示。

III. Appendix

本次作业中,所采用的拟合计算代码均是基于Matlab和Python3.9,相关的源码已经被开源于Github上:
https://github.com/Alexiopro/First-year-of-UCAS/tree/main/UCAS/Source%20Code%20of%20Pattern%20Classification
供读者查用。

图fig6和fig7是Programming部分题(2)的第一小问中采用$BatchPerception$方法的运行展示。可以清晰的看出,采用$BatchPerception$方法处理数据的方式能够较好的划分数据的真实类别。

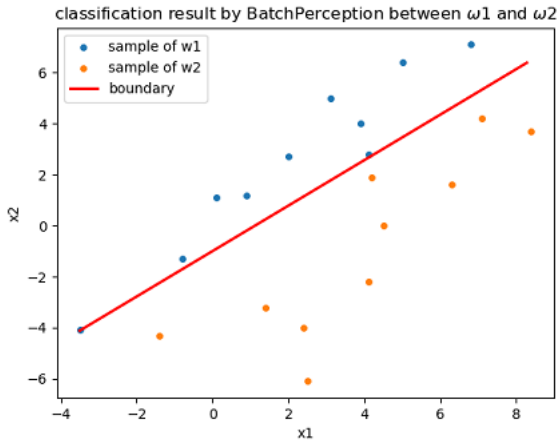

Fig. 6: W1 and W2

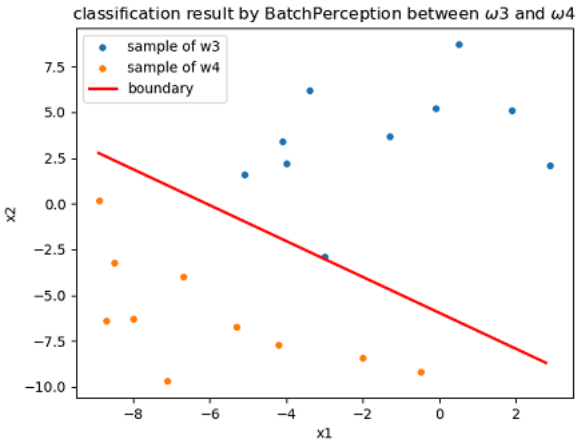

Fig. 7: W3 and W4