专题 1：练习题

1. 在基于记忆的学习中，确定邻域是算法的第一步。邻域的确定需要借助距离函数。常用的距离有多种形式，比如曼哈顿街区距离、欧氏距离、闵可夫斯基距离(Minkovski)、切比雪夫距离、以及马氏(Mahalanobis)距离等。请尝试完成：

a) 请选择 3 种几何上的距离，要求写出距离函数 $d\left(\mathbf{x},\mathbf{y}\right)$ 的表达式，其中 $\mathbf{x},\mathbf{y}\in R^m$；

b) 假设 $\mathbf{x}\in R^2$，请画出各个距离函数的单位圆 $d\left(\mathbf{x},0\right)=\left\|\mathbf{x}\right\|=1$；

c) 哪个“距离”最另类？为什么？

提示: 可借助 MATLAB

2. 在基于记忆的学习中，确定邻域是第一步。第二步要基于邻域内样本点作出决策。在请尝试不同的决策方式：

a) k-NN 分类法： $w_i=1,i=1,....,k$

b) 基于距离加权的 k-NN 分类法：对邻域内的 k 个样本点，根据距离远近进行加权计算，比如 $w_i=\begin{cases}\dfrac{d^{(k)}-d^{(i)}}{d^{(k)}-d^{(1)}}, & d^{(k)}\neq d^{(1)}\\ 1, & d^{(k)}=d^{(1)}\end{cases}$，其中 $d^{(k)}$ 和 $d^{(1)}$ 为距离样本 x 第 k 远的距离和最近的距离；

c) 基于线性表示诱导的 k-NN 分类法：与距离加权类似，权值采用如下的二次规划问题计算

$$\min_{\{w_j\}}\left\|\mathbf{x}-\sum_{j=1,...,k}w_j\mathbf{x}_j\right\|_2^2\quad\textbf{s.t.}\sum_{j=1,...,k}w_j=1,\ w_j\geq 0\ 。$$

请在数据集 IRIS_4x150_3class.mat, MNIST784x50x10.mat 和 YaleB_32x32.mat 上完成分类实验，比较每种方法在不同数据集上的性能，并给出简要分析。要求采用 50%样本训练 50%样本测试(即分别从每个类别中随机选择 50%的样本作为训练集剩余样本作为测试集)，重复 10 次，给出参数 k=1，20，50 时的平均识别率(和标准差)。

3. 密度估计算法与实现:

a) 尝试推导密度估计的经验公式，给出核密度估计法和 k-近邻密度估计法的基本步骤;

b) 在 cvpr16_id_dataset.txt 数据集上，用两种方法完成密度估计，把得到的密度估计结果分别画出来；

c) 请调整局部参数的大小，观察局部参数对密度估计结果的影响，分别画出 3 种局部参数下的的密度估计结果。

4. 请完成非线性回归模型的推导过程,其中采用平方误差损失函数作为目标泛函，即：

$$\min_{f} \varepsilon(f)$$

其中 $\varepsilon(f)=\mathbf{E}\left[Y-f(X)\right]^2=\iint\left\{y-f(\mathbf{x})\right\}^2 p(\mathbf{x},y)d\mathbf{x}dy$ 。要求给出计算和推导步骤。

5. [变分法] 请通过适当的计算回答：对于离散随机变量，何种分布的熵最大？ 提示：以熵为目标函数，加上规范化和非负性作为约束条件，构造泛函最优化问题。

6. [变分法] 请通过适当的计算回答：对于一个一阶矩和二阶矩均有限的连续型随机变量，何种分布的熵最大？ 提示：对于连续型随机变量，需要计算微分熵；以微分熵为目标函数，以规范化、非负性、一阶矩和二阶矩为约束条件，构造泛函最优化问题。

7. 关于最近邻分类器，请查阅文献尝试：

a) 推导 1-NN 分类法的错误率与贝叶斯错误率的关系；

b) 推导 k-NN 分类法的错误率与贝叶斯错误率的关系。

8. 为了加快在一个大规模的高维数据集上进行最近邻搜索的速度，人们提出了多种搜索近似最近邻的策略。请查阅文献，介绍目前的主要方法。