练习题

一、 题目一

本题考虑逻辑回归模型

1. 假设 Y_i 是二分类的类别变量，$X_i \in \mathrm{R}^p$，$i = 1, \cdots, n$, 并且二者来自于某一逻辑回归模型。

   A. 写出逻辑回归模型的数学表达式

   B. 给出样本的似然函数和对数似然函数

   C. 该对数似然函数有显示解吗？给出两种能够求解对数似然函数的迭代算法并且写出其关键的迭代步骤

   D. 对于这组数据 $\{(Y_i, X_i)\}_{i=1}^n$, 为什么传统的线性回归方法可能不再适用？

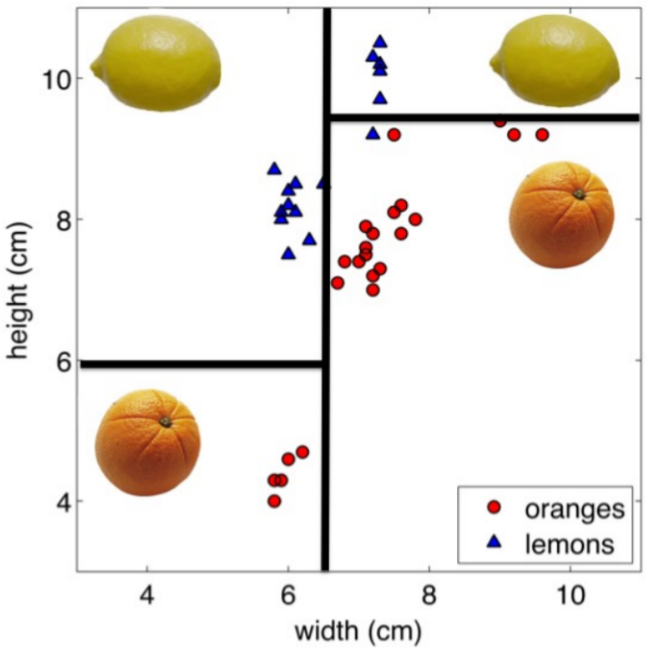
   E. 在逻辑回归中我们使用 sigmoid 变换 $\frac{\exp(x)}{1+\exp(x)}$，你还能给出一种其他变换方式吗？

   F. 对于这一模型, 如果自变量维数远大于样本量会出现哪些问题？如何处理这一情况，请给出至少两种解决方案

   G. 如果线性的逻辑回归模型拟合效果较差，请给出两种非线性的拓展方法 (需提供一定的数学细节)

   H. 对逻辑回归，简述向前法变量选择的过程。当我们选出 p 个模型后，能够选择使得似然函数值最大的那个模型吗？为什么？如果不能，有何可行选择方法。

二、 题目二



给出根节点不同的两种决策树。

三、 题目三

简述分层聚类（Hierarchical Clustering）方法。根据下述距离矩阵，画出分层聚类的树状图，如果想将数据聚成两类，两类所含样本序号分别是什么?

Distance Matrix:

	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

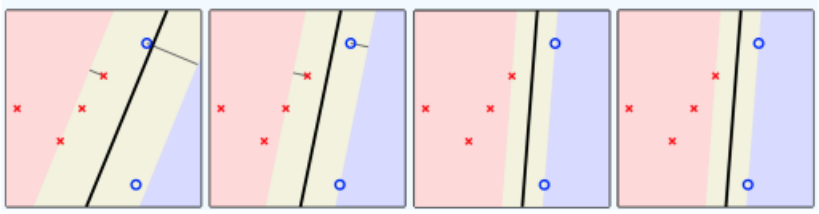
四、 题目四

简述感知机和支持向量机方法。写出 hard-margin SVM 的模型假设。考虑下述 4 个样本点，给出基于这组样本的 SVM 的参数估计。

$$X = \begin{bmatrix} 0 & 0 \\ 2 & 2 \\ 2 & 0 \\ 3 & 0 \end{bmatrix}, \mathbf{y} = \begin{bmatrix} -1 \\ -1 \\ +1 \\ +1 \end{bmatrix}$$

哪几个样本点是支持向量（support vector)?

写出 soft-margin SVM 的模型假设。这一模型中有一参数 C 控制容忍度，在下图中标出，哪一图对应较小的 C，哪一图对应较大的 C。



五、 题目五

结合本课程学过的某一模型，谈谈你对过拟合现象的理解。有哪些处理过拟合现象的方法?