1. 复杂的模型在训练过程中，通常会产生过拟合的现象。试从以下三种模型：（1）支持向量机，（2）决策树，（3）神经网络中以一种模型为例，简单说明如何去避免过拟合现象。

2. K-mediods 算法描述：

a) 首先随机选取一组聚类样本作为中心点集

b) 每个中心点对应一个簇

c) 计算各样本点到各个中心点的距离（如欧几里德距离），将样本点放入距离中心点最短的那个簇中

d) 计算各簇中，距簇内各样本点距离的绝度误差最小的点，作为新的中心点

e) 如果新的中心点集与原中心点集相同，算法终止；如果新的中心点集与原中心点集不完全相同，返回 b)

试着：

a) 阐述 K-mediods 算法和 K-means 算法相同的缺陷

b) 阐述 K-mediods 算法相比于 K-means 算法的优势

c) 阐述 K-mediods 算法相比于 K-means 算法的不足

3. Apriori 算法使用产生—计数的策略找出频繁项集。通过合并一对大小为 k 的频繁项集得到一个大小为 k+1 的候选项集（称作候选产生步骤）。在候选项集剪枝步骤中，如果一个候选项集的任何一个子集是不频繁的，则该候选项集将被丢弃。假定将 Apriori 算法用于表中所示数据集，最小支持度为 30%，即任何一个项集在少于 3 个事务中出现就被认为是非频繁的。

事务 ID	购买项
1	{a,b,d,e}
2	{b,c,d}
3	{a,b,d,e}
4	{a,c,d,e}
5	{b,c,d,e}
6	{b,d,e}
7	{c,d}
8	{a,b,c}
9	{a,d,e}
10	{b,d}

(a) 画出表示表中所示数据的项集格，用下面的字母标记格中的每个结点。

- **N：** 如果该项集被 Apriori 算法认为不是候选项集。一个项集不是候选项集有两种可能的原因：它没有在候选项集产生步骤产生，或它在候选项集产生步骤产生，但是由于它的一个子集是非频繁的而在候选项集简直步骤被丢掉
- **F:** 如果该候选项集被 Apriori 算法认为是非频繁的
- **I：** 如果经过支持度计数后，该候选项集被发现是非频繁的

(b) 频繁项集的百分比是多少？（考虑格中所有的项集）

(c) 对于该数据集，Apriori 算法的剪枝率是多少？（剪枝率定义为由于如下原因不认为是候选的项集所占的百分比：在候选项集产生时未被产生，或在候选剪枝步骤被丢掉）

(d) 假警告率是多少？（假警告率是指经过支持度计算后被发现是非频繁的候选项集所占的百分比）

4. 已知正例点 $x_1 = (2.5, 2.5)^T$，$x_2 = (5, 2)^T$，负例点 $x_3 = (1.5, 1.5)^T$，试用 SVM 对其进行分类，求最大间隔分离超平面，并指出所有的支持向量。

5. 请评价两个分类器 M1 和 M2 的性能。所选择的测试集包含 26 个二值属性，记作 A 到 Z。

表中是模型应用到测试集时得到的后验概率（图中只显示正类的后验概率）。因为这是二类问题，所以 P(-)=1-P(+),P(-|A,…,Z)=1-P(+|A,…,Z)。假设需要从正类中检测实例

(a) 画出 M1 和 M2 的 ROC 曲线（画在一幅图中）。哪个模型更多？给出理由

(b) 对模型 M1，假设截止阈值 t=0.5.换句话说，任何后验概率大于 t 的测试实例都被看作正例。计算模型在此阈值下的 precision，recall 和 F-score

(c) 对模型 M2 使用相同的截止阈值重复（b）的分析。比较两个模型的 F-score，哪个模型更好？所得结果与从 ROC 曲线中得到的结论一致吗》

(d) 使用阈值 t=0.1 对模型 M2 重复（b）的分析。t=0.5 和 t=0.1 哪一个阈值更好?该结果和你从 ROC 曲线中得到的一致吗？

实例	真实类	P(+|A,…,Z,M1)	P(-|A,…,Z,M2)
1	+	0.73	0.61
2	+	0.69	0.03
3	-	0.44	0.68
4	-	0.55	0.31
5	+	0.67	0.45
6	+	0.47	0.09
7	-	0.08	0.38
8	-	0.15	0.05
9	+	0.45	0.01
10	-	0.35	0.04

6. 下表是一个由 15 个贷款申请训练数据，数据包括贷款申请人的四个特征属性：分别是年龄，是否有工作，是否有自己的房子以及信贷情况，表的最后一列为类别，是否同意贷款。

ID	年龄	有工作	有自己的房子	信贷情况	类别
1	青年	否	否	一般	否
2	青年	否	否	好	否
3	青年	是	否	好	是

4	青年	是	是	一般	是
5	青年	否	否	一般	否
6	中年	否	否	一般	否
7	中年	否	否	好	否
8	中年	是	是	好	是
9	中年	否	是	非常好	是
10	中年	否	是	非常好	是
11	老年	否	是	非常好	是
12	老年	否	是	好	是
13	老年	是	否	好	是
14	老年	是	否	非常好	是
15	老年	否	否	一般	否

1）请根据上表的训练数据，以错误率作为划分标准来构建预测是否进行放贷的决策树。
2）按照所构建的决策树，对属性值为（中年，无工作，无自己的房子，信贷情况好）的申请者是否进行放贷
3） 在构建决策树的时候，可能会出现过拟合的问题，有哪些方法可以避免或者解决？
4） 对于含有连续型属性的样本数据，决策树有哪些处理方法？