一：试析使用“最小训练误差”作为决策树划分选择的缺陷。

答：

1.最小训练误差不可靠。使用最小训练误差利用了贪心的思想，会出现过早陷入局部最优的问题。

2.决策树的生成是递归生成，并且是贪心地生成的，无法保证是全局最优树，所以本身就是一层层生成的，无法直接使用最小化训练误差(最小化训练误差只有树结构确定时使用)。

3最理想的情况是产生所有可能的子树，在子树上使用最小化训练误差来进行求解，但是是代价过高，对于时间代价与结果准确度上需要做取舍，所以选择局部最优贪心逼近。

二、试述将线性函数f(x)=wTx 用作神经元激活函数的缺陷。

答：

1.神经网络中在隐层，还是输出层都需要有非线性的激活函数。激活函数的作用是将神经元的输出值做进一步处理以区分出激活和抑制两种状态。激活函数的选取灵活，不同层可选择不同的激活函数。

2.如果单用线性函数作为激活函数，多层这样的神经元叠加起来的效果可以等效于单层神经网络，运算叠加就是一个线性运算，退化成一个线性回归，因此无法实现多层非线性神经网络的“深度”学习功能。

3.无法解决非线性问题，如线性不可分的分类问题。

三、 试述高斯核SVM与RBF神经网络的联系

答：

1. SVM的确与神经网络有密切联系：若将隐层神经元数设置为训练样本数，且每个训练样本对应一个神经元中心，则以高斯径向基函数为激活函数的RBF网络恰与高斯核SVM的预测函数相同,基本等价。

2.RBF网络的径向基函数与SVM都可以采用高斯核，也就分别得到了高斯核RBF网络与高斯核SVM。高斯核SVM的主要思想是先用RBF函数将原数据做转换, 然后再施加SVM.

RBF神经网络也是类似的思想, 先用RBF函数将原来的数据做转换, 然后线性加权组合输出。

3.神经网络是最小化累计误差，将参数作为惩罚项，而SVM相反，主要是最小化参数，将误差作为惩罚项。

4.非支持向量将得到很小的w使用LIBSVM对异或问题训练一个高斯核SVM得到α，固定β参数为高斯核SVM的参数，修改每个隐层神经元的中心为各个输入参数，得到结果w,w与α各项成正比例。

5.两者的目标函数、优化方式不同。

五、试述为什么Bagging难以提升朴素贝叶斯分类器的性能。

答：Bagging主要是降低分类器的方差，而朴素贝叶斯分类器中假设各特征相互独立，已经是很简化的模型了，所以其误差主要是在于偏差，没有方差可降。对全训练样本生成的朴素贝叶斯分类器是最优的分类器，不能用随机抽样来提高泛化性能。

1.从偏差—方差分解的角度看， Boosting 主要关住降低偏差，因此 Boosting 能基于泛化性能相当弱的学习器构建出很强的集成。从偏差—方差分解的角度看， Bagging 主要关注降低方差，因此它在不剪枝决策树、神经网络等易受样本扰动的学习器上效用更为明显。

2.统计学习框架下，Error = Bias + Variance。Error可以理解为模型的预测错误率，是有两部分组成的，一部分是由于模型太简单而带来的估计不准确的部分（Bias），另一部分是由于模型太复杂而带来的更大的变化空间和不确定性（Variance）。方差大（偏差低）的模型往往是因为对训练数据拟合得过好，模型比较复杂，输入数据的一点点变动都会导致输出结果有较大的差异，它描述的是模型输出的预测值相比于真实值的离散程度，方差越大，越离散，偏差大（方差低）的模型则相反，往往因为对训练数据拟合得不够，模型比较简单，输入数据发生变化并不会导致输出结果有多大改变，它描述的是预测值和和真实值直接的差距，偏差越大，越偏离真实值。

六、试析AGNES算法使用最小距离和最大距离的区别。

答：最大距离可以认为是所有类别先生成一个能包围所有类内样本的最小圆，然后所有圆同时慢慢扩大相同的半径，哪个类圆能完全包围另一个类则停止，并合并这两个类。由于此时的圆已经包含另一个类的全部样本，所以称为全连接。

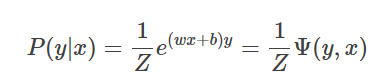
最小距离则是扩大时遇到第一个非自己类的点就停止，并合并这两个类。由于此时的圆只包含另一个类的一个点，所以称为单连接。

七、 比较条件随机场和对率回归，试析其异同。

答：同：两者都是判别式模型，是对条件分布进行建模，学习的是条件概率分布P（Y|X）

不同点：

1.对率回归是一维变量对所有自变量x的条件概率，只有一个节点，即为只有一个输出变量y,公式如下：



2.条件随机场是每一个自变量xi对应一个yi,即为多维变量y对自变量的条件概率。多元变量形式多变，可为链式结构、树形结构。

3.在求条件概率时，对率回归需要考虑所有的变量，而条件随机场仅需要考虑当前与之对应的变量xi。

4.条件随机场可以看做是广义图连接上对率回归模型，而对率回归模型可以看做是条件随机场的一个特例。