简答题（5 道*10 分=50 分)

1. 简述假设空间和版本空间。

①假设空间：样本中所有属性可能产生的所有假设所组成的集合.
②版本空间：与训练数据集一致的“假设集合”.

2. 简述训练误差、测试误差、泛化误差。训练误差很小，测试误差一定很小吗?

①训练误差：模型在训练集上产生的误差，反映的是模型自身的拟合能力;
②测试误差：模型在测试集上产生的误差；反映的是模型的泛化能力;
③泛化误差：模型在新样本上产生的误差，反映的是模型的表现性能。
训练误差小，测试误差不一定很小。因为若模型发生过拟合现象，则训练误差会很小，但模型的泛化能力会降低，测试测试误差就有可能会很大。

3. 对数几率回归解决的是回归问题还是分类问题?主要通过什么方法训练样本?

①对数几率回归解决的是分类问题
②主要基于均方误差、用极大似然估计法估计参数来训练样本；常见的优化算法有梯度下降法和牛顿法。

4. 朴素贝叶斯主要解决了什么障碍?其关键假设是什么?

①朴素贝叶斯主要解决了在有限的训练集上估计 $p(x,c)$ 的联合概率分布计算 $p(x|c)$ 的障碍，例如属性组合爆炸、样本稀疏性的问题；

②关键假设：所有属性条件独立性假设，即各属性间互不干扰。

5. 简述局部极小和全局最小。

①局部极小：给定参数空间的一点 A，若在 A 点邻域附近点对模型产生的期望误差均不小于 A，则称 A 为局部极小点；
②全局最小：给定参数空间的一点 A，若对于空间中其他任意一点的期望误差均不小于 A，则称 A 为全局最小点；
数学描述：
记E为均方误差
对于一点$(w^*,\theta^*),\exists\varepsilon>0$，$\forall(w,\theta)\in\{(w,\theta)\|\ (w^*,\theta^*)-(w,\theta)\|\le\varepsilon\}$,
都有$E(w^*,\theta^*)\le E(w,\theta)$,则称$(w^*,\theta^*)$为局部极小点；
若对空间中任意一点(w,θ)，都有$E(w^*,\theta^*)\le E(w,\theta)$，则称$(w,\theta)$为全局最小点

算法题（2 道*10 分=20 分)

1. 为 k-means 算法的伪代码添加注释。

简要描述：
输入样本集 D={x1,x2,....xn}，划分簇数 k
①从 D 中选取 k 个初始均值向量{u1,u2.....uk}
②repeat:
For i =1:k
Ci = 空集 //初始化所有簇为空
For j=1:n
计算每一个样本 xj 到各个均值向量 ui 的距离 dji=欧氏距离
根据最小均值向量距离划分 xj 的簇类
将 xj 后加入对应的 C
For i=1:k
重新计算每个簇的均值向量
若均值向量发生变化，更新
否则保持不变
Until 所有均值向量保持不变

2. 补全决策树的伪代码。

输入： 训练集 $D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)\}$;
属性集 $A = \{a_1, a_2, \ldots, a_d\}$.
过程： 函数 TreeGenerate(D, A)
1: 生成结点 node;
2: **if** D 中样本全属于同一类别 C **then**
3:
4: **end if**
5: **if** $A = \varnothing$ **OR** D 中样本在 A 上取值相同 **then**
6:
7: **end if**
8:
9: **for** a_* 的每一个值 a_*^v **do**
10: 为 node 生成一个分支; 令 D_v 表示 D 中在 a_* 上取值为 a_*^v 的样本子集;
11: **if** D_v 为空 **then**
12:
13: **else**
14: 以 TreeGenerate(D_v, $A \setminus \{a_*\}$)为分支结点
15: **end if**
16: **end for**
输出： 以 node 为根结点的一棵决策树

决策树学习基本算法

3：将 node 结点划分为类别为 C 的叶节点，return
6：将 node 结点划分为叶节点，其类别为 C 中样本数最多的类别

8：从 A 中选取最有属性 a*
12: 将分支结点标记为叶结点，其类别为 C 中样本数最多的类别

综合题（2 道*15 分=30 分)

1. 支持向量机基本型解决的是回归问题还是二分类问题?如果超平面无法在训练集样本中进行划分，请问还可以用支持向量机基本型吗?如果不可以用，请问可以使用什么改进方法?改进的基本原理是什么?

①支持向量机基本型解决的是二分类问题
②不可以用基本型。改进方法：
1. 将原始空间通过核函数映射到一个更高维的特征空间，即“再生核希尔伯特空间”，使其线性可分。
原理：如果一个训练集是有限维，即属性的个数是有限的，那么一定存在一个更高维的空间使其线性可分。
2. 还可以使用软间隔：允许一些样本划分错误
原理: 最小化误差的同时最大化软间隔

2 2..集成学习的性能一定会提升吗?集成学习的关键影响因素是什么?是如何影响的呢?请枚举一种可以提高集成学习的性能的方法。

①集成学习的性能不一定会提升，取决于个体学习器的选取
②关键因素：
2.1 个体学习器的准确率：个体学习器的准确率越高，那么集成的学习器性能往往会越好，一般来说，个体学习器的准确率至少不低于弱学习器。
2.2 个体学习器的多样性：多样性指个体学习器之间的差异，例如可以采用不同类型的个体学习器，改变学习器的模型参数等增加多样性；多样性越好，则集成学习性能越好。
③提升性能的方法：增强多样性：
数据样本扰动：如 bagging 通过自助采样的方法
输入属性扰动：如随机森林中决策树的每个结点仅选取部分属性进行划分
输出表示扰动：翻转法(改变划分结果),输出调制法(分类转回归),拆解任务为若干个子任务(如 ECOC 法)
算法参数扰动:如负相关法等