第二次课后作业

1、朴素贝叶斯分类器解决了什么障碍？ 朴素贝叶斯分类器解决了什么障碍？它的关键假设是什么？

- 解决了类条件概率 P(x|c)难以从有限的训练样本直接估计而得。
- 朴素贝叶斯分类器采用了属性条件独立性假设，对已知类别，假设所有属性相互独立。

2、请简述局部极小与全局极小。

- 局部极小：参数空间中某个点，其邻域点的误差函数值均不小于该点的误差函数值。
- 全局极小：参数空间中所有店的误差函数值均不小于该点的误差函数值。

3、什么是监督学习和非监督学习，请说明它们的区别并各举一个例子 ； 请说明分类和回归问题的区别。

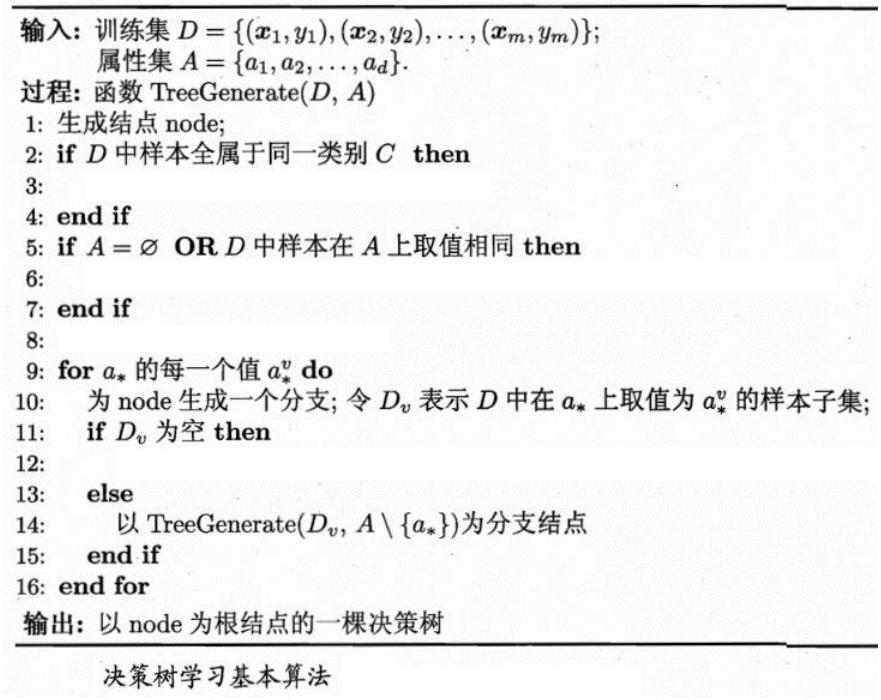
- 监督学习：对于数据集中的每个数据，都有相应的正确的答案，算法就是基于这些来进行预测，我们知道了输出应该是什么样子的。比如给定房价数据集，对于里面的每一个例子，算法都知道正确的房价，即这个房子实际卖出的价格，算法的结果就是计算出更多的正确的价格。
- 非监督学习：没有给定事先标记过的训练范例，自动对输入的资料进行分类或分群。例如谷歌新闻利用聚类算法把不同的主题放在一起。
- 分类问题：分类是预测离散类标签的任务；可以使用准确度评估分类预测，而回归预测则不能；分类问题的误差是预测错误的数量
- 回归问题：回归是预测连续数量的任务；可以使用 R2 来评估回归预测，而分类预测则不能；回归问题的误差可以是均方误差

4.请简述随机森林的生成方法以及其随机性体现在哪里？

- 生成方法：
  - 对于一个样本容量为 N 的样本集，我们做有放回的抽取 N 次，每次抽取 1 个样本，那么最终就形成了 N 个样本。
  - 假设样本特征数目为 a，对 n 个样本随机选择 a 中的 k 个特征，用建立决策树的方式获得最佳分割点
  - 重复 m 次，产生 m 棵决策树
  - 多数投票机制来进行预测

- 随机性：随机性体现在在构建基学习器的过程中，随机选择样本、随机选择特征、从所有特征中随机选取

5、请为以下决策树算法的步骤 3 ，6 ，8 ，12 填写为代码



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 节点标记为 C 类叶节点，return;
6：将 node 节点标记为叶节点，其类别标记为 D 中类别最多的类，return;
8：从 A 中选择最优划分属性 a*
12：将分支节点标记为叶节点，其类别标记为 D 中类别最多的类，return;

6、请阐述机器学习中欠拟合 和 过拟合现象，并结合偏差 （ bias ） 和 方差 （ variance ） 解释其出现的原因。以人工神经网路学习为例，请给出 至少两 种解决其过拟合的方法。

- 欠拟合：
  - 欠拟合是指模型不能在训练集上获得足够低的误差。换句换说，就是模型复杂度低，模型在训练集上就表现很差，没法学习到数据背后的规律。
  - 训练集的预测结果就不准，偏差较大。但对于不同训练集，训练得到的模型都差不多（都不太准，对训练集不敏感），因此预测结果差别不大，方差小。
- 过拟合;
  - 过拟合是指训练误差和测试误差之间的差距太大。换句换说，就是模型复杂度高于实际问题，模型在训练集上表现很好，但在测试集上却表现很差。
  - 模型完全学习训练集的信息，训练集偏差较小，测试集偏差较大。此外，模型对与训练样本分布不同的测试集上表现不一，预测结果相差大，方差大。

- 过拟合解决方法：
  - 加入正则化系数
  - 降低模型复杂度，减少深度或者神经元数
  - 获取和使用更多的数据