一、简答题

1. M=9 时为何能精确匹配10个点？

M=9包含了低阶多项式。

sin(2πx) 的幂级数展开包含所有阶数

2. M=9 时为什么泛化能力差？

更⼤的M值的更灵活的多项式被过分地调参，使得多项式被调节成了与⽬标值的随机噪声相符

1. 机器学习方法有哪几类？

有监督学习（ supervised learning ）

分类（classification） ： 输出离散的特定类别

回归（regression） ： 输出一个或多个连续实变量

…

无监督学习（unsupervised learning）

聚类（clustering） : 相似样本分组

密度估计（density estimation）： 确定输入空间中数据的分布

半监督学习

弱监督学习

增强学习（reinforcement learning）

AlphaGo…

1. 数据拟合过程是如何的？

提示：数据分为train和test……

1. 如何判断是否过拟合？如何解决？

过拟合：非常强的拟合效果会使得样本点之外的函数值远远偏离期望的目标，反而降低分类器的性能的现象。

解决：1.增大数据集的规模；2.正则化（regularization）：给误差函数增加⼀个惩罚项，使得系数不会达到很⼤的值。

1. 似然函数（likelihood function）如何表示？



1. 为什么要剪枝？

决策树分支过多，过拟合；

预剪枝（prepruning）

决策树生成过程中，每个结点划分前确定是否提升泛化性能；

后剪枝（postpruning）

先生成完整决策树，然后自底向上考察是否能将非叶结点的子树删除，以带来泛化性能提升；

1. 如何确定泛化性能？

数据集->训练集+验证机 ， 交叉验证

1. 递归返回条件

1. 当前结点包含的样本属于同一类别，无需划分；

2. 当前属性集为空，或所有样本在所有属性上取值相同；

3. 当前结点包含的样本集为空，不能划分；

1. 递归对应操作：

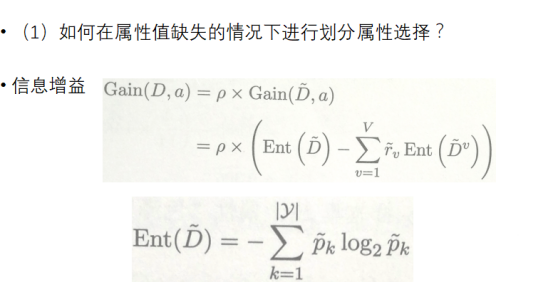
1. 标为该类别的叶结点

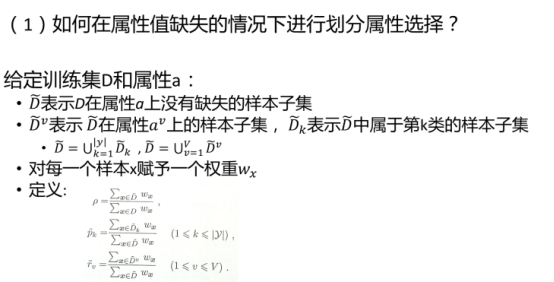
2. 标为叶结点，设定为该结点所含样本最多的类别； （后验概率）

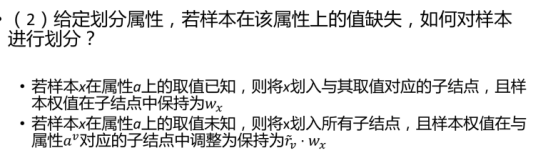
3. 标为叶结点，设定为该结点父结点所含样本最多的类别； （先验概率）

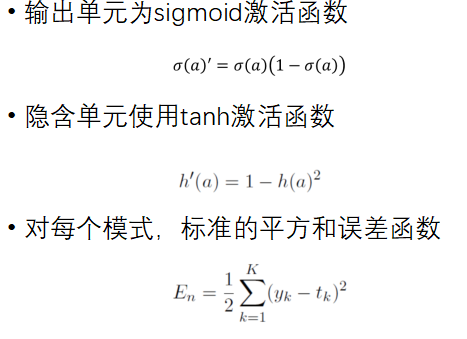
1. 不考虑最优划分属性， 怎么构造决策树？

二、填空题



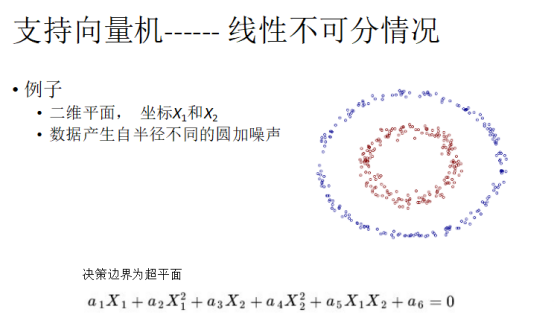


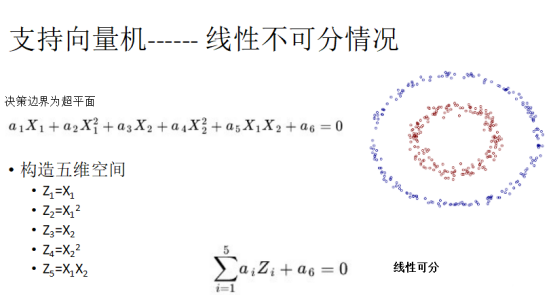


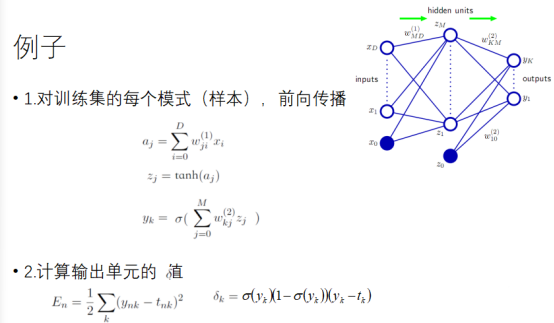


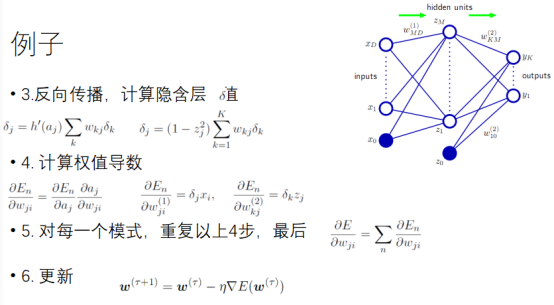
属性选择依据

* 信息增益(ID3)
* 信息增益率(C4.5)
* 启发式选择
* 基尼系数（Gini index）(CRAFT)









三、计算题

两个盒⼦，⼀个红⾊的，⼀个蓝⾊的

红盒⼦中有2个苹果和6个橘⼦，

蓝盒⼦中有3个苹果和1个橘⼦

假定： 1. 40%的时间中选择红盒⼦，在60%的时间中选择蓝盒⼦

2. 选择盒⼦中的⽔果时为等概率选择

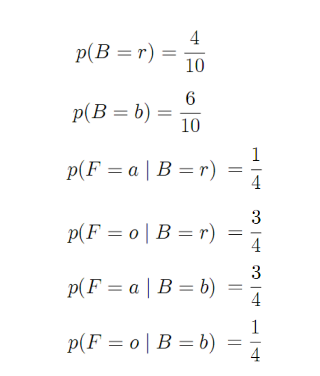
问： 1. 若随机选择一个盒子，会是红盒子还是蓝盒子？

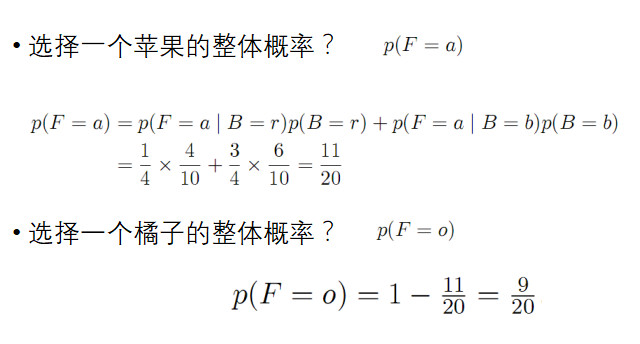
2. 若抽出来的是橘子，那你抽那个盒子的概率大？

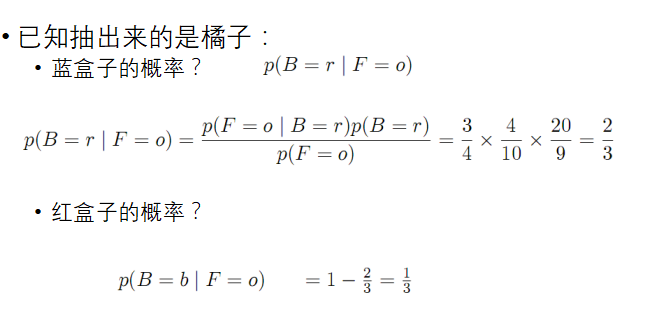
颜色变量 B ： r（对应红盒⼦）或b（对应蓝盒⼦）

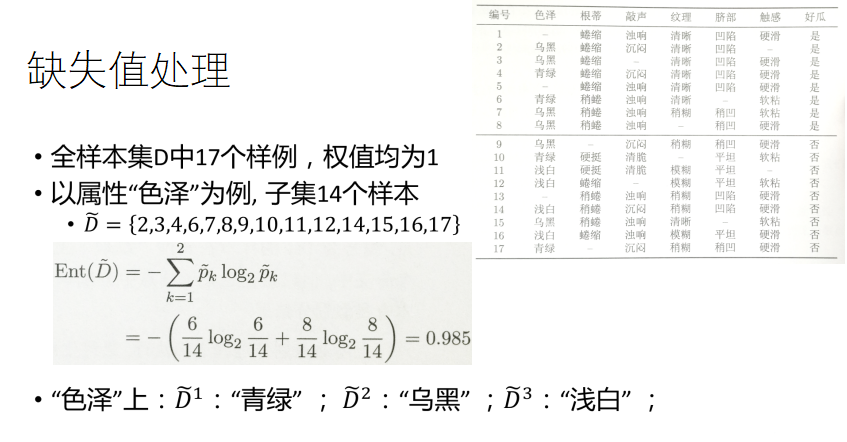
水果变量 F ： a（苹果）或者o（橘⼦）

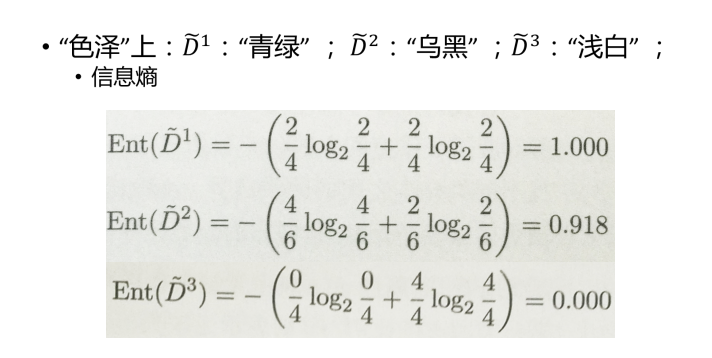
选择红盒⼦的概率为 4/10，选择蓝盒⼦的概率为 6/10











有一批同一型号的产品,已知其中由一厂生产的占 30% , 二厂生产的占 50% , 三厂生产的占 20%, 又知这三个厂的产品次品率分别为2% , 1%, 1%,问从这批产品中任取一件是次品的概率是多少？

解：设事件 A 为“任取一件为次品”，



