1. 决策树

假设是一个离散型随机变量，其概率分布函数为，则的信息量为：概率对数的负数，概率越小，信息量越大



信息熵：对于某个事件，有n种可能性，每一种可能性的概率为，则信息熵为所有信息量的期望：



联合熵：如果随机变量有多个，则信息熵就会变化成联合熵：



条件概率：



<https://www.cnblogs.com/fantasy01/p/4581803.html>

条件熵：为已知随机变量的条件下，随机变量的不确定性：



信息增益：待分类的集合的熵-按某个条件分类后的条件熵



在决策树算法中，分类效果越好，就会越小，相应的信息增益就会越大。

<https://blog.csdn.net/baimafujinji/article/details/50554664>

<https://www.cnblogs.com/maybe2030/p/4585705.html>

<https://www.cnblogs.com/fionacai/p/5894142.html>

Bootstrap抽样：“有放回地全抽”，（其实样本量也要视情况而定，不一定非要与原样本量相等），抽取的Bootstrap样本量与原样本相同，只是在抽样方式上采取有放回地随机抽，这样对于一个原始数据集就可以抽样多次，每次抽样的数据集就很有可能有重复数据，相应的也会剔除部分数据。

随机森林：为了解决决策树过拟合的问题，如果数据集的特征非常多，决策树算法就容易陷入过拟合（特征很多，会根据不同特征将训练集分隔成多个方块，不具备泛化能力）。

1. 采用Bootstrap抽样对训练集进行抽样；
2. 从样本的M个特征维度中，随机选取m（m<<M）个特征，只根据这个m个特征建立决策树；
3. 对根据之前两个步骤建立的多个决策树进行投票分析，得票最多的就是最终预测类别。

决策树的一些超参数：

1. 最大深度（max\_depth）：当决策树深度为时，它最多可以拥有片叶子。较大的深度往往会导致过拟合，这是因为过深的决策树可以记忆数据。而较小的深度会使得模型过于简单，导致欠拟合。
2. 每片叶子的最小样本数（min\_samples\_leaf）：当分裂结点时，每片叶子允许的最小样本数。当每片叶子的样本数量较小时，叶子上的样本数量也有可能过于稀少，此时模型将记忆数据，也就是过拟合。当每片叶子的样本数量较大时，决策树能够获得足够的弹性进行构建，这也许会导致欠拟合。
3. 每次分裂的最小样本数（min\_samples\_split）：结点允许分裂最小样本数，即如果结点样本数>min\_samples\_split，则允许分裂，否则不允许。
4. 最大特征数（max\_features）：寻找最佳分裂时，考虑的最大特征数目。
5. 朴素贝叶斯分类：假设各特征之间相互独立。

贝叶斯公式：



为先验概率：即没有条件下，事件A的概率；

为后验概率：条件B发生后，事件A的概率；

<https://www.cnblogs.com/marc01in/p/4775440.html>

<https://www.cnblogs.com/by-dream/p/7884606.html>

例：

一所学校里面有60%的男生，40%的女生。男生总是穿长裤，女生则一半穿长裤一半穿裙子。

P(男)=0.6 P(女)=0.4

P(男|长裤)=1 P(女|长裤)=0.5

根据贝叶斯公式求P(长裤|男)，即穿长裤的人是男生的概率。

P(长裤|男)=P(男)P(男|长裤)/P(长裤)

P(长裤)=P(男,长裤)+P(女,长裤)=P(男)P(男|长裤)+P(女)P(女|长裤)=0.8

则：

P(长裤|男)=0.6\*1/0.8=0.75

P(长裤|女)=0.4\*0.5/0.8=0.25

朴素贝叶斯分类方法：

1. 设为一个待分类项，其中为的特征属性；
2. 设类别集合为，表示总共有n类。
3. 求出，，…
4. 如果，则。

朴素贝叶斯分类的核心就是计算第3步的条件概率。根据贝叶斯公式：



对于所有类别而言，是一个常数，可以忽略掉。

根据贝叶斯假设，的各个特征之间相互独立，所以：



拉普拉斯平滑：上述的公式会存在一个问题，如果某个特征没有再某个分类中出现过，如特征没有在中类别出现过，则为0，从而导致整个公式结果为0，显示不能如此计算。拉普拉斯平滑就是在计算不为0，而是赋一个很小的概率。具体操作如下：

令为在垃圾邮件类别中，单词出现的概率；定义：

令为垃圾邮件中，单词出现的个数。

令为垃圾邮件，所有单词总数目。则：



拉普拉斯平滑：分子+1，分母加特征总数，这样可以保证概率和为1

