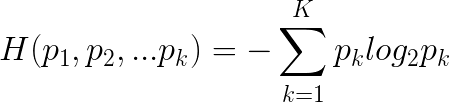
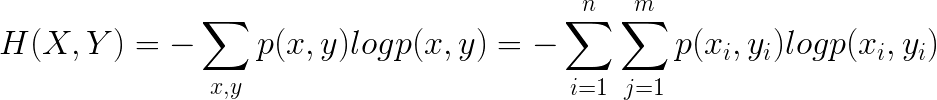
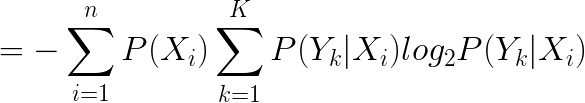
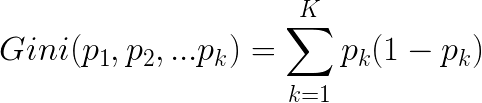
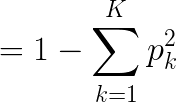
【任务3 - 决策树算法梳理】时长：2天

1. 信息论基础（熵 联合熵 条件熵 信息增益 基尼不纯度）

<https://blog.csdn.net/xbmatrix/article/details/56691137>

信息增益比https://www.cnblogs.com/muzixi/p/6566803.html

****熵****：原本物理学中的定义，后来香农将其引申到啦信息论领域，用来表示信息量的大小。信息量大（分类越不“纯净”），对应的熵值就越大，反之亦然。  
信息熵的计算公式如下：  
  
****联合熵****：一维随机变量分布推广到多维随机变量分布。  
  
****条件熵****： H(Y|X) 表示在已知随机变量 X 的条件下随机变量 Y 的不确定性。条件熵 H(Y|X) 定义为 X 给定条件下 Y 的条件概率分布的熵对 X 的数学期望。  
  
  
****信息增益****：以某特征划分数据集前后的熵的差值。即待分类集合的熵和选定某个特征的条件熵之差。  
IMG_260  
****基尼不纯度****：即基尼指数  
  


1. 决策树的不同分类算法（ID3算法、C4.5、CART分类树）的原理及应用场景

决策树通常包括三个步骤: ****特征选择，决策树生成和决策树剪枝****  
决策树的生成过程就是：使用满足划分准则的特征不断的将数据集划分为纯度更高，不确定性更小的子集的过程。  
****ID3****：采用信息增益划分数据。计算使用所有特征划分数据集，得到多个特征划分数据集的信息增益，从这些信息增益中选择最大的，因而当前结点的划分特征便是使信息增益最大的划分所使用的特征。  
****不足****：信息增益偏向取值较多的特征。  
****C4.5****：采用信息增益比划分数据，弥补ID3的不足  
IMG_256  
****不足****：只能进行分类  
****CART****：采用基尼系数划分数据，可针对离散和连续型，可以做分类和回归

1. 回归树原理

划分的准则是平方误差最小化

1. 决策树防止过拟合手段

当某个模型过度的学习训练数据中的细节和噪音，以至于模型在新的数据上表现很差，我们称过拟合发生了。这意味着训练数据中的噪音或者随机波动也被当做概念被模型学习了。而问题就在于这些概念不适用于新的数据，从而导致模型泛化性能的变差。可以从以下两个方面着手：  
****预剪枝****：是在决策树的生成过程中，对每个结点在划分前先进行估计，若当前结点的划分不能带来决策树泛化性能提升，则停止划分即结束树的构建并将当前节点标记为叶结点。  
****后剪枝****：是先从训练集生成一棵完整的决策树，然后自底向上地对叶结点进行考察，若将该结点对应的子树替换为叶结点能带来决策树泛化为性能提升，则将该子树替换为叶结点。泛化性能的提升可以使用交叉验证数据来检查修剪的效果，通过使用交叉验证数据，测试扩展节点是否会带来改进。如果显示会带来改进，那么我们可以继续扩展该节点。但是，如果精度降低，则不应该扩展，节点应该转换为叶节点。

1. 模型评估

评估分类器性能

自助法（bootstrap）：

训练集是对于原数据集的有放回抽样，如果原始数据集N，可以证明，大小为N的自助样本大约包含原数据63.2%的记录。当N充分大的时候，1-（1-1/N）^(N) 概率逼近 1-e^(-1)=0.632。抽样 b 次，产生 b 个bootstrap样本，则，总准确率为（accs为包含所有样本计算的准确率）：

accboot=1b∑i=1b(0.632×εi+0.368×accs)accboot=1b∑i=1b(0.632×εi+0.368×accs)

准确度的区间估计：

将分类问题看做二项分布，则有：

令 X 为模型正确分类，p 为准确率，X 服从均值 Np、方差 Np（1-p）的二项分布。acc=X/N为均值 p，方差 p（1-p）/N 的二项分布。acc 的置信区间：

P(−Zα2≤acc−pp(1−p)/N−−−−−−−−−√≤Z1−α2)=1−αP(−Zα2≤acc−pp(1−p)/N≤Z1−α2)=1−α

P∈2×N×acc+Z2α2±Zα2Z2α2+4×N×acc−4×N×acc2−−−−−−−−−−−−−−−−−−−−−−−−−−−√2(N+Z2α2)P∈2×N×acc+Zα22±Zα2Zα22+4×N×acc−4×N×acc22(N+Zα22)

<https://blog.csdn.net/longgb123/article/details/52972604>

原文链接：https://blog.csdn.net/longgb123/article/details/52972604

1. sklearn参数详解，Python绘制决策树

<https://blog.csdn.net/li980828298/article/details/51172744>

官方函数解释https://scikit-learn.org/dev/modules/generated/sklearn.tree.DecisionTreeClassifier.html#sklearn.tree.DecisionTreeClassifier