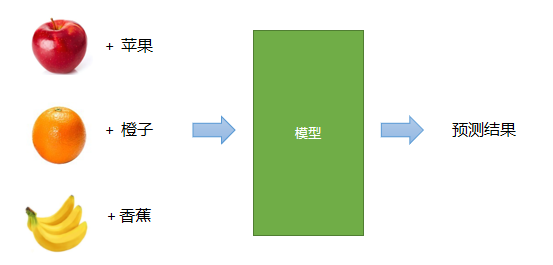
数据挖掘与机器学习算法考试

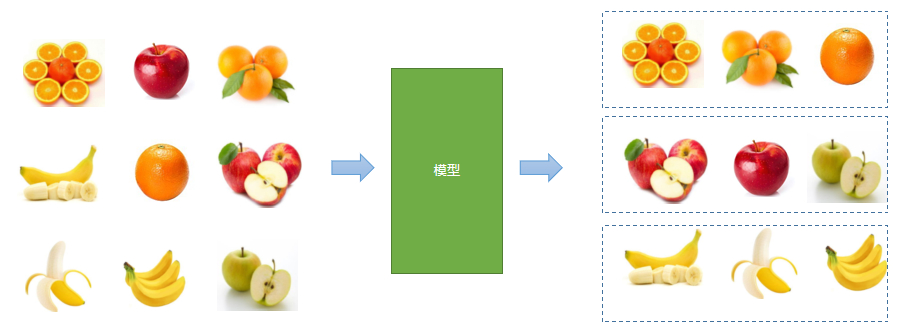
1. 有监督学习和无监督学习的概念和区别
2. 有监督学习：

在已知数据输出(经过标注的)的情况下对模型进行训练，根据输出进行调整、优化的学习方式。



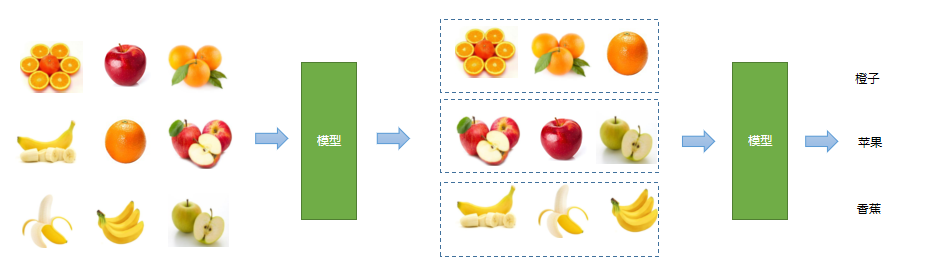
1. 无监督学习：

没有已知输出的情况下，仅仅根据输出的相关性，进行类别的划分。



1. 半监督学习：

先通过无监督学习划分类别，再人工标记通过有监督学习方式来预测输出。例如先对相似的水果进行聚类，再识别是哪个类别。



1. 决策树的概念和意义

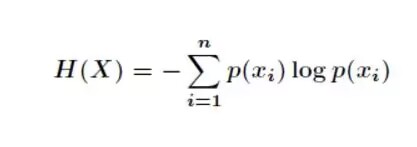
利用样本特征进行决策归类，将具有相同属性的样本划入一个子节点；用作分类器、回归器。

1. 常用的分类算法有哪些

逻辑回归、决策树分类、支持向量机、朴素贝叶斯。

1. 数据预处理有哪些常用的方式，具体实现的原理
2. 数据清洗：对数据进行去重、缺失值处理、异常值处理、数据类型转换等操作，以保证数据的准确性和完整性
3. 特征选择：根据特征重要性和相关性等指标，选择最具代表性和区分度的特征，以提高模型的精确性和泛化能力
4. 特征缩放：对特征进行缩放操作，使得不同特征之间具有相同的尺度，以便于模型的训练和优化
5. 特征编码：将文本型特征进行编码，以便于 模型的处理和计算，常见的编码方法包括one-hot编码、标签编码、二进制编码等
6. 数据归一化：将数据进行归一化处理，以使得数据分布在[0,1]或[-1,1]之间，以便于模型的训练和优化
7. 数据降维：对高维数据进行降维处理，以减少特征数量和计算复杂度，常用的降维方法包括主成分分析法(PCA)、线性判别分析法(LDA)等
8. 数据集划分：将数据集划分为训练集、验证集和测试集，以便于模型的训练、评估和优化
9. 信息熵、信息增益、增益率、基尼系数

A)信息熵本质上是对可能性的一种度量：一个事件的不确定性越大，我们从该事件的发生中获得的信息就越多。信息熵的数学定义为：



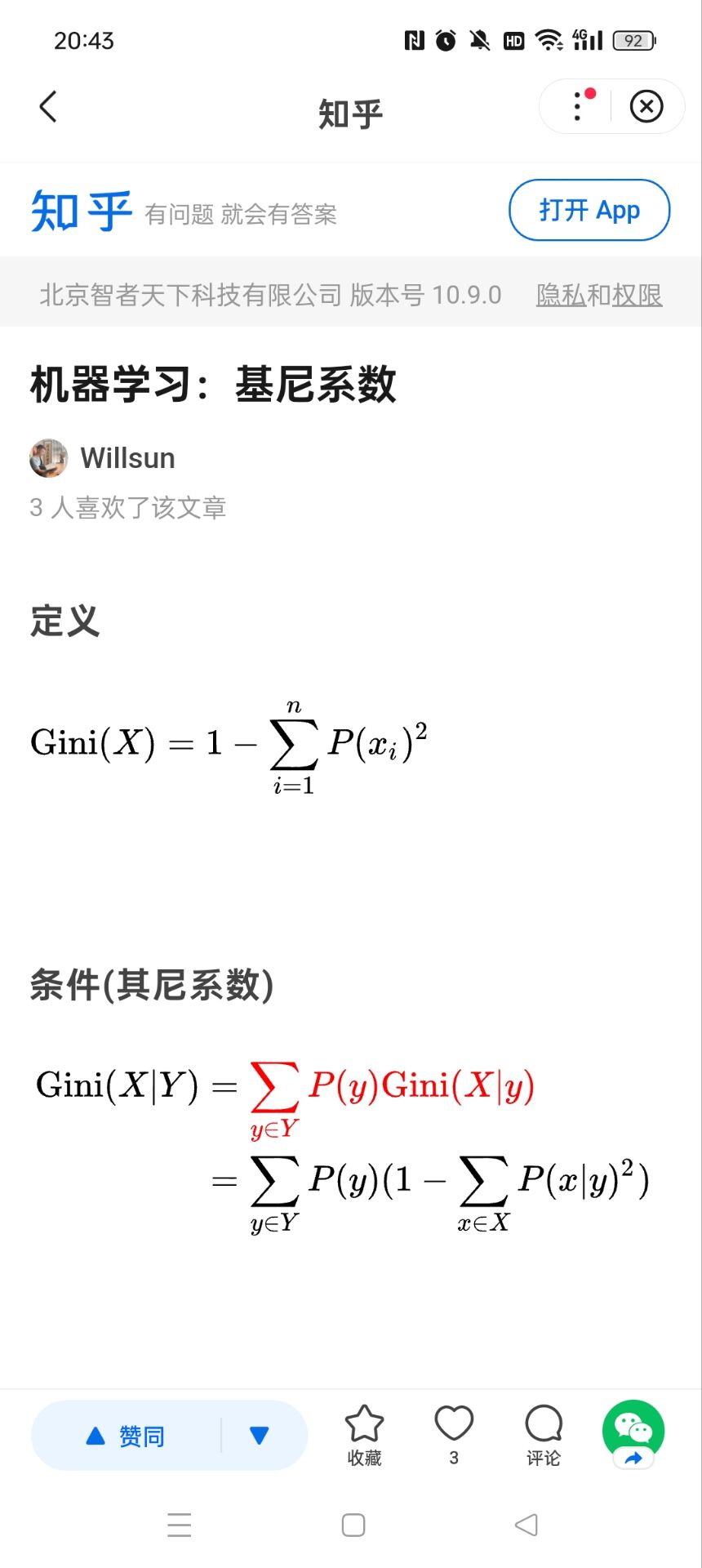
B)条件熵是在给定某个特征的情况下，对于分类结果的不确定性的度量。条件熵的数学定义为：



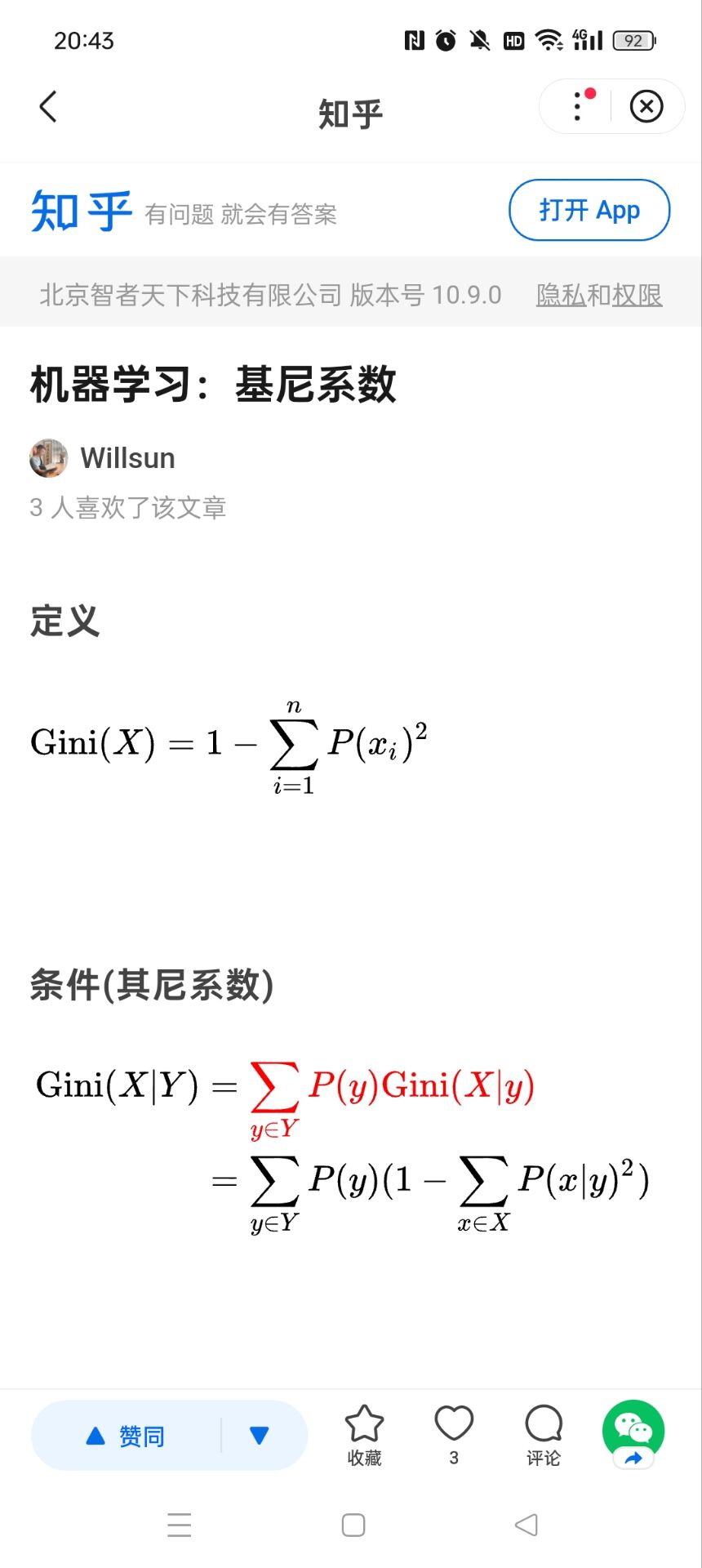
C)信息增益=信息熵-条件熵

Gain(D,A) = H(D) - H(D|A)

1. 基尼系数



条件基尼系数



1. 逻辑回归是什么，如何实现

逻辑回归是线性分类器(线性模型)——主要用于二分类问题

1. 先验概率和后验概率的联系

事件还未发生，为先验概率；事件发生了，为后验概率；先验概率为后验概率的基础。

1. 什么是贝叶斯定理

贝叶斯定理由英国数学家托马斯-贝叶斯提出，用来描述两个条件概率之间的关系，定理描述为：P(A|B)=P(A)P(B|A)/P(B)

1. 回归模型的评估指标
2. RMSE——平方根误差
3. MAE——平均绝对误差
4. MSE——平均平方误差
5. R^2 score——决定系数
6. 支持向量机SVM的超平面有哪些要求
7. 正确性：对大部分样本都可以划分类别，避免特殊点
8. 安全性：支持向量，即离分类边界最近的样本之间的距离最远
9. 公平性：支持向量与分类边界的距离相等
10. 简单性：采用线性方程(直线、平面)表示分类边界，也称分割超平面。如果在原始维度中无法做线性划分，那么就通过升维变换，在更高维度空间寻求线性空间超平面，从低纬度到高纬度空间的变换通过核函数进行
11. 什么是核函数，常用的有哪些

核函数是特征转换函数，它可以将数据映射到高维特征空间中，从而更好地处理非线性关系。核函数的作用是通过计算两个样本之间的相似度(内积)来替代显式地进行特征映射，从而避免了高维空间的计算开销。

常见的核函数包括：线性核函数、多项式核函数、高斯核函数、sigmoid核函数。

1. K-means模型算法的步骤
2. 选择K个初始中心点：随机选择数据集中的K个数据点作为初始中心点(质心)
3. 分配数据点到最近的中心点：对于数据集中的每一个点，计算其与所有中心点的距离，并将其分配给最近的中心点
4. 更新中心点：重新计算每个集群的中心点，通常是该集群内所有点的平均值
5. 迭代直至收敛：重复步骤2和步骤3，直至中心点不再显著变化或达到预设的迭代次数
6. 机器学习中，“距离”作为度量样本的相似度的指标，具体具备什么条件

距离越小，相似度越高；距离越大，相似度越低。

1. 机器学习进行段落语义预测的步骤
2. 文本预处理
3. 特征提取
4. 机器学习模型训练
5. 模型评估与优化
6. SVM支持向量机算法的应用步骤
7. 数据预处理：将数据集划分为训练集和测试集，并进行特征缩放(对数据进行标准化)
8. 构建模型：选择合适的核函数和惩罚系数，构建SVM模型
9. 训练模型：使用训练集对模型进行训练，通过最大化间隔来找到最优的超平面
10. 预测：使用训练好的模型对测试集进行预测
11. 随机森林算法的应用步骤
12. 假设训练集T的大小为N，特征数目为M，随机森林的大小为K
13. 遍历随机森林的大小K次
14. 从训练集T中有放回抽样的方式，取样N次形成一个新子训练集D
15. 随机选择m个特征，其中m<M
16. 使用新的训练集D和m个特征，学习出一个完整的决策树