Part IV: Introduction to Big Data

Lecture by 柳玲
Note by THF

2025 年 3 月 25 日

目录

Lecture 4

03.21

Notation. 数据脱敏的方法：**数据替换、无效化、随机化、偏移和取整、对称加密、平均值**

0.1 常用数据库

NewSQL、NoSQL：速度快，扩展性好，数据模型多，云计算集成
HDFS、GFS：分布式，规模大

1 数据挖掘和机器学习

包括：**聚类、分类、神经网络等方法**

1.1 分类算法

使用已有的样本预测新样本的所属类别

Example. 邮件分类为垃圾邮件；患者诊断为患病

常用算法：朴素贝叶斯、逻辑回归、**决策树**、随机森林、支持向量机；使用有监督学习（标记样本）

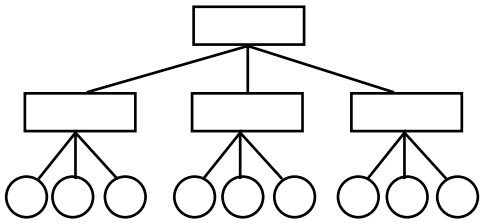

图 1: 决策树示例

决策树

1.2 聚类算法

无监督学习（不标记样本），有**划分聚类**、**层次聚类**、**基于密度聚类**、**基于网格聚类**；要求根据 k 个数据划分，每个簇至少有一个数据，每个数据有且仅属于一个簇

***K*-均值聚类**

首先随机从数据点中取 k 个数据作为初始中心，确定分类为 k 组数据

回归分析

$$\boldsymbol{w} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}.$$

关联规则

人工神经网络

简称神经网络（Neural Network），有多种分支，如卷积神经网络 CNN

一个神经元至少包含三个部分：输入、计算、输出。输入到神经元的数据首先会通过权重调整，然后传入神经元；传出时需要经过非线性函数对输出归一化

Lecture 5

03.25