Part III-B: Artificial Intelligence Outline

Lecture by 熊庆宇
Note by THF

2024 年 11 月 30 日

目录

0.0.1 归一化 1
0.0.2 决策树 2
0.1 无监督学习 2

Lecture 13

11.04

KNN 算法的局限

- 对参数选择很敏感
- 计算量大

当 K 值较小：易发生过拟合，受噪声影响较大
当 K 值太大：无法区分不同样本

0.0.1 归一化

Notation. 为何归一化：某些数据在未归一化之前占比过大，如年龄-存款

归一化处理：

$$M_j = \max_{i=1,2,\ldots,m} x_{ij} - \min_{i=1,2,\ldots,m} x_{ij}.$$

归一化后的距离计算：

$$L_2(A,B) = \sqrt{\sum_{j=1}^{n} \left(\frac{a_j - b_j}{M_j}\right)^2}.$$

特征值标准一致时无需归一化

Notation. 欧几里得距离：

$$S = \sqrt{\sum_{i=1}^{n} \left(x_i^{(P)} - x_i^{(Q)}\right)^2}.$$

表 1: 分类

样本名	x_1	x_2	x_3	类型	S_n 距离
S_1	39	0	21	K_1	$\sqrt[3]{4133} \approx 16.05$
S_2	3	5	65	K_2	$6\sqrt[3]{5^2}\sqrt[3]{19} \approx 46.81$
S_3	21	17	5	K_1	$2\sqrt[3]{3^2}\sqrt[3]{14} \approx 10.03$
...					
S_n	23	3	17	?	0

曼哈顿距离:

$$S = \sum_{i=1}^{n} \left| x_i^{(P)} - x_i^{(Q)} \right|.$$

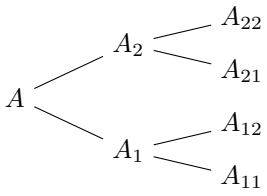
切比雪夫距离:

$$S = \max_{l} \left(\left| x_i^{(P)} - x_i^{(Q)} \right| \right).$$

0.0.2 决策树

Definition. 树形结构，由节点和边组成

基本思想：一个 if-then 的规则集合
可以分为树形或细胞型



Example. ID3 算法

0.1 无监督学习

Notation. 区别：有监督学习中提供样本的标签，无监督学习中机器自行提取样本的相似性

通过样本可以提取颜色、纹理、频率等特征
无监督函数通过定义相似度计算函数来提取特征的相似性，根据选择的相似度函数来分类

Notation. K-均值聚类算法

监督学习补充：线性回归 *Linear regression*

Definition. 回归与分类：挖掘和学习输出变量和输入变量之间的潜在关系模型
回归为连续、分类为离散

Example. 高尔顿提出衰退 (regression, 回归) 效应，指出：

$$y = 33.73 + 0.516\frac{x_1 + x_2}{2}.$$

其中 x_1, x_2 为父母身高 (单位：inch)，y 为经过回归后的下一代身高

Notation. 最小二乘法：求出使残差平方和最小的 a, b

Lecture 14

11.07

$$a = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \quad b = \bar{y} - a\bar{x}.$$

Notation. 无监督学习技术：Clustering 聚类

无监督学习因素：相似度函数

Notation. K 均值聚类：

设定随机中心，通过欧氏距离判断中心和数据间的相似性

结课

期末考试：11.24，第 2 节课