

## Part III-B: Artificial Intelligence Outline

Lecture by 熊庆宇

Note by THF

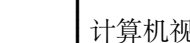
2024 年 11 月 7 日

# 目录

0.0.1	监督学习	2
0.0.2	归一化	4
0.0.3	决策树	4
0.1	无监督学习	5

## Lecture 12

10.31



```

graph LR
    ML[机器学习] --- CV[计算机视觉]
    ML --- NLP[自然语言处理]
    ML --- SMC[社交媒体计算]
    ML --- EF[经济金融]
    CV --- HR[手写识别]
    CV --- PR[行人再识别]
  
```

机器学习 { 计算机视觉 { 手写识别  
行人再识别  
自然语言处理  
社交媒体计算  
经济金融

## 机器学习的四个部分：

$$\left\{ \begin{array}{l} \text{T: Task} \\ \text{A: Algorithm} \\ \text{E: Experience} \\ \text{P: Performance} \end{array} \right.$$

**Example.** 人脸识别:

A: 线性回归

E: 以标定身份的人脸图片数据

P: 人脸识别准确率

## 机器学习的基本过程

从给定的数据中学习规律  $\rightarrow$  学习方法, 建立模型  $\rightarrow$  预测  $\rightarrow$  测试匹配度

## 机器学习分类

$$\left\{ \begin{array}{l} \text{半监督学习} \left\{ \begin{array}{l} \text{监督学习} \\ \text{无监督学习} \end{array} \right. \\ \text{强化学习} \end{array} \right. .$$

### 0.0.1 监督学习

**Definition.** 根据已知的输入和输出训练模型, 预测未来输出

监督学习的数据存在样本标签, 有训练集和测试集

**Example.** 学习书籍内容, 设定标签: 艺术/政治/科学等, 找出训练文字和标签的映射关系

**Notation.** 分类方法:  $K$ -nearest neighbour, 决策树, 支持向量机, 朴素贝叶斯

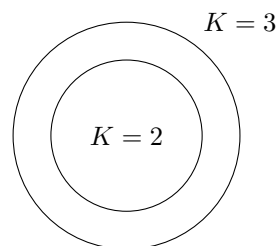
回归方法: 线性、树、支持向量回归, 集成方法

## ***$K$ -nearest neighbor***

### 基本思路

找到与新输入的待预测样本最临近的  $K$  个样本, 判断这  $K$  个样本中绝大多数的所属类别作为分类结果输出

条件: 已经具有较大的样本量



**Notation.** KNN 算法的基本要素: 距离度量、 $K$  值、分类决策规则

### 距离度量

**Notation.** KNN 算法能够分类: 特征空间内的样本点之间的距离能够反映样本特征的相似程度

设有两个样本点  $\mathbf{x}_i, \mathbf{x}_j$  , 以  $n$  维向量空间作为特征空间, 将这两个点表示为:

$$\mathbf{x}_i, \mathbf{x}_j \in \mathbf{X}.$$

$$\mathbf{x}_i = (x_i^1, x_i^2, \dots, x_i^n)^T.$$

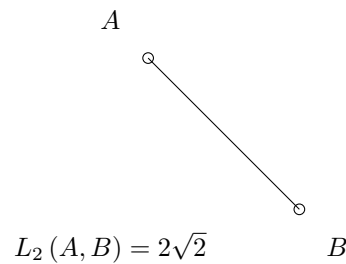
$$\mathbf{x}_j = (x_j^1, x_j^2, \dots, x_j^n)^T.$$

特征点之间的距离定义为:

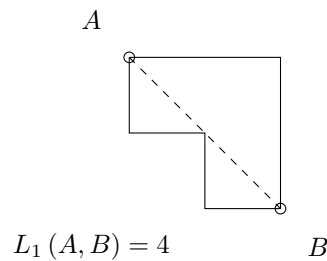
$$L_p(\mathbf{x}_i, \mathbf{x}_j) = \left( \sum_{l=1}^n |x_i^l - x_j^l|^p \right)^{\frac{1}{p}}.$$

**Example.** 代入  $p = 2$  , 易得  $L_2(\mathbf{x}_i, \mathbf{x}_j)$  为平面上两点间的距离公式, 该距离又称为欧氏距离:

$$L_2(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(x_{i_1} - x_{j_1})^2 + (x_{i_2} - x_{j_2})^2}.$$



代入  $p = 1$  :  $L_1(\mathbf{x}_i, \mathbf{x}_j)$  称为曼哈顿距离:



## K 值的选择

使用交叉验证方法确定最合适的  $K$  值

## 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

### 0.0.2 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:

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

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

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

**Notation.** 欧几里得距离:

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

曼哈顿距离:

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

切比雪夫距离:

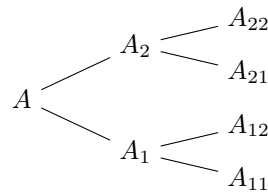
$$S = \max_l \left( \left| x_i^{(P)} - x_i^{(Q)} \right| \right).$$

### 0.0.3 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