Part III-B: Artificial Intelligence Outline

Lecture by 熊庆宇
Note by THF

2024 年 11 月 22 日

目录

Lecture 1

人工智能大事件

1. GPT(ChatGPT), 2022.11
2. ERINE(文心一言), 2023.3
3. GPT4(多模态，Sora)，2024.2

Notation. 历史上人工智能与人类对弈：

1. 1997.5，IBM DeepBlue vs 卡斯帕罗夫（国际象棋）
2. 2016.3，Google Alpha Go vs Lee & Ke
3. 2019.7，Facebook Pluribus vs 德州扑克世界冠军

算法案例化：

$$\begin{cases} \text{人心可测} \\ \text{路径导航} \\ \text{数码寻优} \\ \text{适者生存} \\ \text{蚁群觅食} \\ \text{性别预测} \\ \text{电影分类} \\ \cdots \end{cases}.$$

课程要求

32 学时，16 节课
教材：人工智能导论
课后作业：选修《人工智能导论》的动因、定位、设想，800-1000 字

Lecture 2

Notation. 课程有闭卷考试（60%），9-10 次作业和 2 次报告（40%）
考试基于课上内容

1 人工智能发展历程

人工智能发展开始：1956 年

孕育期：1956 年前

Notation. 1943 年麦克洛奇和皮兹建成第一个神经网络模型（MP 模型）
1949 年提出了 Hebb 规则（激发函数规则）

神经网络的一些标准：神经元层数、个数，激发函数，连接方式（全连接/非全连接），**权重**，……

第一次低谷期：1957-1973

形成期：1974-1980

黄金期：1980-1987

专家系统出现：MYCIN,PROSPECTOR,XCON 等
AI 被引入市场：Rumelhart 提出 BP（反向传播）算法，实现多层神经网络学习

第二次低谷期：1987-1993

专家系统难以使用、升级、维护，AI 未能完成既定目标

平稳期：1993-2011

蓬勃期：2012 至今

小结

Notation. 图灵测试：在封闭的房间中，一个人分别对两个对象询问并获得答案，两个对象分别是 AI 和人类，判断 AI 是否具备人类的特征

Notation. 人工智能三大学派：

1. 符号学派
2. 连接主义
3. 行为主义

Lecture 3

Notation. 行为主义的代表性成果：蚁群算法、粒子群算法

比较三种主流方法：
第一章作业：1-19 题

表 1: 学习模式

符号主义	连接主义	行为主义
与人类逻辑类似	直接从数据中学习	从经验中持续学习

2 人工智能的知识表示

2.1 概述

研究人工智能的目的：使其得以模拟、延伸、扩展，

Notation. 人是一个物理符号系统

为使人工智能达到相应的功能：将知识破译、重新编码、建立相应的符号系统

Notation. 知识的层次：

现象 $\Longrightarrow$ 数据 $\Longrightarrow$ 信息 $\Longrightarrow$ 知识 $\Longrightarrow$ 智慧

数据：一些无关联的现象

数据 $\rightarrow$ 信息：组织、分析

信息 $\rightarrow$ 知识：解释、评价

知识 $\rightarrow$ 智慧：理解、归纳

Example. 数据：下雨了，温度下降至 15 度

信息：地面水蒸发，遇冷暖峰过境

知识：理解下雨、蒸发、空气状况、地形、风向等及其中的作用机理

智慧：模拟天气变化，人工天气可控化

Notation. 知识的特性：

1. 相对正确性
2. 不确定性
3. 可表示性和可利用性

Question. 如何将人类知识形式化/模型化

对知识的一种描述或约定：转化为机器可接受描述的形式

$$\text{知识表示}\begin{cases}\text{符号主义}\begin{cases}\text{谓词逻辑}\\ \text{产生式系统}\\ \text{框架系统}\end{cases}\\ \text{经验主义}\begin{cases}\text{状态表示}\\ \text{特征表示}\end{cases}\\ \text{连接主义}\begin{cases}\text{语义变量}\\ \text{网络权重}\end{cases}\end{cases}.$$

Lecture 3

Notation. 亚里士多德提出了“三段论”演绎推理方法

莱布尼茨在 17 世纪提出二进制，乔治贝尔提出用简单符号表示逻辑命题，产生了“贝尔代数”：适于机器使用的数学规律

概念理论由概念名、概念内涵和概念外延组成

Notation. 命题：一个非真即假的陈述句

$$\text{命题分类}\begin{cases}\text{真命题}\\ \text{假命题}\\ \text{在一定条件下为真，一定条件下为假}\end{cases}.$$

对于 $R: x<8$ 由于 R 的真假依赖于 x 的取值，因此无法判断

2.2 命题逻辑

Notation. 蕴含连结词：“若 p 则 q” 称为 p 对 q 的蕴含式：$p \to q$

Lecture 4

表 2: 命题的五种连结词

$\wedge$(and)	$\vee$(or)	$\neg$(not)	$\Rightarrow$ (implies)	$\Leftrightarrow$(forth and come)

表 3: 命题真值表

p	q	$\neg p$	$p \wedge q$	$p \vee q$	$p \Rightarrow q$	$p \Leftrightarrow q$
F	F	T	F	F	T	T
F	T	T	F	T	T	F
T	F	F	F	T	F	F
T	T	F	T	T	T	T

Notation. 充分条件：$p \subset q$即$p \Rightarrow q$

Example. 符号化：

1. 铁和氧化合但铁和氮不化合：

$$p = \text{铁和氧化合}，\ q = \text{铁和氮化合}.$$

$$\text{Org: } p \vee (\neg q)\,.$$

2. 小张或小明通过了 CET6

$$p = \text{小张通过 CET6},\ q = \text{小明通过 CET6}.$$

$$\text{Org: } p \vee q.$$

3. 如果我下班早，就去商城看看，除非我很累

$$p = \text{我下班早},\ q = \text{我很累},\ r = \text{我会去商城}.$$

$$\text{Org: } (p \wedge (\neg q)) \Rightarrow r.$$

Notation. 命题逻辑的优劣：

优点：能把客观世界的各种事实转化为逻辑命题

缺点：不适合表达复杂问题、细节缺省

2.3 谓词逻辑

Definition. 谓词逻辑：一种形式语言，接近自然语言，方便计算机处理

谓词：用于刻画个体的性质、状态和个体之间关系的成分

Example. x 是 A 类型的命题使用 $A(x)$ 表达

x 大于 y 可表达为 $B(x,y)$

$A(x)$ 称为一元谓词，$B(x,y)$ 称为二元谓词

用 $P(x_1,x_2,\ldots,x_n)$ 表示一个 n 元谓词公式，P 为 n 元谓词，$x_1,x_2,\ldots,x_n$ 为客体变量或变元

定义谓词 $U(x)$ 表示 x 为大学生，该谓词可以记录相关的属性

语法元素：

1. 常量（个体符号）：通常是对象的名称
2. 变量符号：小写字母
3. 函数符号：小写英文字母 f,g 等

Example. 我喜欢音乐和绘画：

$$\text{Like}(\text{I},\text{Music}) \wedge \text{Like}(\text{I},\text{Painting})\,.$$

连词：

1. 与/合取：$\text{Like}(\text{I},\text{Music}) \wedge \text{Like}(\text{I},\text{Painting})$
2. 或/析取

Notation. 全称量词 $\forall$

Lecture 4

Example. 所有机器人都是灰色的:

$$\forall (x) [\text{Robot} (x) \to \text{Color} (x, \text{gray})].$$

Notation. 存在量词 $\exists$

Example. 1 号房间有一个物品:

$$\exists (x) \, \text{InRoom} (x, R_1) .$$

函数和命题的区别:
函数是定义域到值域的映射
命题是定义域到 $\{\text{True}, \text{False}\}$ 的映射

Example. 符号化 "所有数的平方是非负的":

1. 个体 x
2. 函数符号 f : 某数的平方
3. 谓词 Q : 某个数是非负的
4. 符号化: $(\forall x) \, Q (f (x))$

第二种:

1. 个体 z : 表达一个数
2. 谓词 R : x 是一个实数
3. 函数符号 f
4. 谓词 Q
5. 符号化: $(\forall z) [R (z) \to Q (f (z))]$

谓词逻辑推理形式化

Example. 所有人都要死，孔子是人，所以孔子会死:

$$(\forall x) (A (x) \to B (x)) \wedge A (\text{Confucious}) \to B (\text{Confucious}) .$$

Notation. 谓词逻辑的优点: 自然性、精确性、易实现
缺点: 不能表示不确定性知识，过于自由而兼容性差
应用:

1. 自动问答系统
2. 机器人行动规划系统
3. 机器博弈系统
4. 问题求解系统

作业: 第二章 1-18 题

Lecture 4

2.4 产生式知识表示法

Notation. 确定性规则知识产生式:

$$P \to Q.$$

不确定性规则知识产生式:

$$P \to Q\,(\mathrm{Conf})\,.$$

确定性规则知识产生式表示:

$$(\mathrm{Relate}, a, b)\,.$$

不确定性规则知识产生式表示:

$$(\mathrm{Relate}, a, b, \mathrm{Conf})\,.$$

Lecture 5

Notation. 产生式规则和谓词逻辑的区别:

1. 产生式规则额外包含各种操作、规则、转换、算子、函数等
2. 产生式可以不精确表示知识

产生式规则缺点: 效率较低
应用: 专家系统

Notation. 专家系统组成:

第一层: 人机交互界面
第二层: 知识获取、推理机、解释器
第三层: 知识库、综合数据库

专家系统的局限性:

1. 知识获取的瓶颈
2. 规则“跷跷板”问题
3. 知识动态化困难

作业: 第二章 19-23 题

Lecture 6

10.10

2.5 框架式表达方法

Definition. 框架: 对某种知识的整体认识, 描述所论对象属性的数据结构

通过框架可以生成表格

Example. 框架: 课程

框架理论使用层次化结构表达知识

表 4: 上课

XXX 课程	
课程需求	需求 1 需求 2 需求 3
课程内容	内容 1 内容 2 内容 3
...	...

表 5: 框架

框架名			
槽名 1	侧面名 11 侧面名 12 侧面名 13	值 111，值 112，... 值 121，值 122，... 值 131，值 132，...	约束条件 ...
槽名 2	侧面名 21 侧面名 22 侧面名 23	值 211 值 212 值 213	
...	...	...	
关联框架	< 框架名 1, 关系 >,< 框架名 2, 关系 >...		

Example. 例：教师

框架名：教师

1. 姓名（VARCHAR(12)）
2. 年龄（INT）
3. 性别（男、女）
4. 职称（教授、副教授、讲师、助教）
5. 部门:（系、教研室）
6. 住址:（VARCHAR(64)）
7. 工资（INT）
8. 开始工作时间（DATETIME）
9. 截止时间（DATETIME，DEFAULT DATE(CURRENT_TIMESTAMP)）
10. 框架关联：教职工，教师

Notation. 框架表达的特点：

1. 结构性
2. 继承性

3. 自然性

作业：习题 24,25

Lecture 7

10.14

2.6 状态空间表示法

Notation. 回忆：

命题 → 谓词 → 产生式 → 框架 → 状态

Definition. 状态空间表示法：表示问题及其搜索过程

Example. 与空格相连的棋子可以移动到空格中：

表 6: 初始状态

2	8	3
1		4
7	6	5

如何将某一初始状态变成目标状态：

表 7: 目标状态

1	2	3
8		4
7	6	5

Example. 渡河问题：三个传教士 M 和三个野人 C 过河，只有一条能装下两个人的船，在河的一方或船上，如果野人的人数大于传教士的人数，那么传教士会有危险，如何使所有人安全地过河

$$\text{状态空间适用的场景}\begin{cases}\text{调度}\\ \text{分配}\\ \text{导航}\\ \text{路径规划}\\ \text{游戏}\\ \cdots\end{cases}.$$

Notation. 状态空间法主要包括：

1. 状态集：其中的每个元素表示一种状态
2. 操作算符集：连结状态间的条件
3. 状态空间：包括状态集、操作算符集、目标状态集

Example. 某棋局：

表 8: 棋盘 S

X_1	X_2	X_3
X_8	X_0	X_4
X_7	X_6	X_5

用 $S = (X_0, X_1, \ldots, X_8)$ 表示状态，0 代表空格
如：表 6 表示为：$S_0 = (0, 2, 8, 3, 4, 5, 6, 7, 1)$
表 7 表示为 $S_8 = (0, 1, 2, 3, 4, 5, 6, 7, 8)$
将表 6 中棋子 4 左移，状态变为：$S_1 = (4, 2, 8, 3, 0, 5, 6, 7, 1)$

表 9: S1

2	8	3
1	4	
7	6	5

继续移动，直至找到一套操作得到状态 S_8：$S_0 \to S_1 \to \ldots \to S_8$
该问题称为八数码难题

Notation. 八数码难题的算符：
仅为空格制定操作：空格上下左右移动，空格的约束条件为不能移出棋盘

表 10: S0->S1

2		3
1	8	4
7	6	5

$\xrightarrow{\text{空格向左移}}$

	2	3
1	8	4
7	6	5

Example. 此时空格有 3 种移动方式

Notation. 状态空间图：
把初始状态可达到的各状态所组成的空间设想为由各种状态对应的节点组成的图（有向图）
图的节点表示状态
图的边表示操作算符

Lecture 7

Example. 二阶汉诺塔问题:

$S_i(a,b)$ 表示状态盘 A 在 a 柱上，盘 B 在 b 柱上

算符: $A(i,j)$ 表示将 A 盘从 i 柱移动到 j 柱

$B(i,j)$ 同理

汉诺塔问题的状态图可以是双向图，限制为: A 不能在 B 下方

Example. 渡河问题:

初始状态: $(0,0,0)$

目标状态: $(3,3,1)$

状态格式: (右岸传教士数量, 右岸野人数量, 船的位置)

算符:

Move-1m1c-lr: 将一个传教士和一个野人从左边传到右边

Move-2c-lr: 将两个野人从左边移到右边

Move-1m-rl: 将一个传教士从右边移动到左边

操作:

$(0,0,0) \xrightarrow{\text{Move-1m1c-lr}} (1,1,1) \xrightarrow{\text{Move-1c-rl}} (1,0,0) \xrightarrow{\text{Move-2c-lr}} (1,2,1) \xrightarrow{\text{Move-1c-rl}} (1,1,0) \to \cdots$

作业: 第 26-29 题

3 搜索求解策略

Notation. 早期搜索策略: 图搜索、盲目搜索、启发式搜索

高级搜索技术: 规则演绎系统、产生式系统

Definition. 搜索技术: 根据问题的实际情况，不断寻找可利用的知识，构造出一条代价较少的推理路线

搜索技术是 AI 的基本技术之一

搜索好的标准:

1. 搜索空间小
2. 解最佳

Example. 爬山路径:

1. 问题全状态空间: 整座山
2. 搜索空间: 山的路
3. 解: 爬山路径

$$\text{一般搜索算法}\begin{cases}\text{无信息搜索（盲目搜索）}\begin{cases}\text{宽度/广度优先搜索（BFS）}\\ \text{深度优先搜索（DFS）}\\ \text{等代价搜索}\end{cases}\\ \text{有信息搜索}\begin{cases}A\text{算法}\\ A^{\star}\text{算法}\end{cases}\end{cases}.$$

Notation. 搜索算法：

必须记住哪些节点已经遍历（OPEN 表）

需给出下一步可以选择哪些节点（CLOSED 表）

必须记住从目标节点返回的路径

Notation. BFS: Breath-First Search

首先扩展根节点

然后扩展根节点的**所有后继节点**

以此类推，在第 n 层节点未完全遍历之前不进入第 $n+1$ 层的遍历

Lecture 8

10.17

Notation. DFS: Deep-First Search

优先朝长节点扩展：先进入开放表的节点先扫描

优点：搜索空间可以远小于宽度优先

缺点：忽略深度

修正：加入深度界限，在已知目标节点的深度范围时限制搜素深度

最坏情况：$o(n)$

应用：状态表 = 树状图

Notation. 等代价搜索/Dijkstra 算法：

BFS 的一种推广

$g(n)$ 代表从初始节点到节点 n 的代价

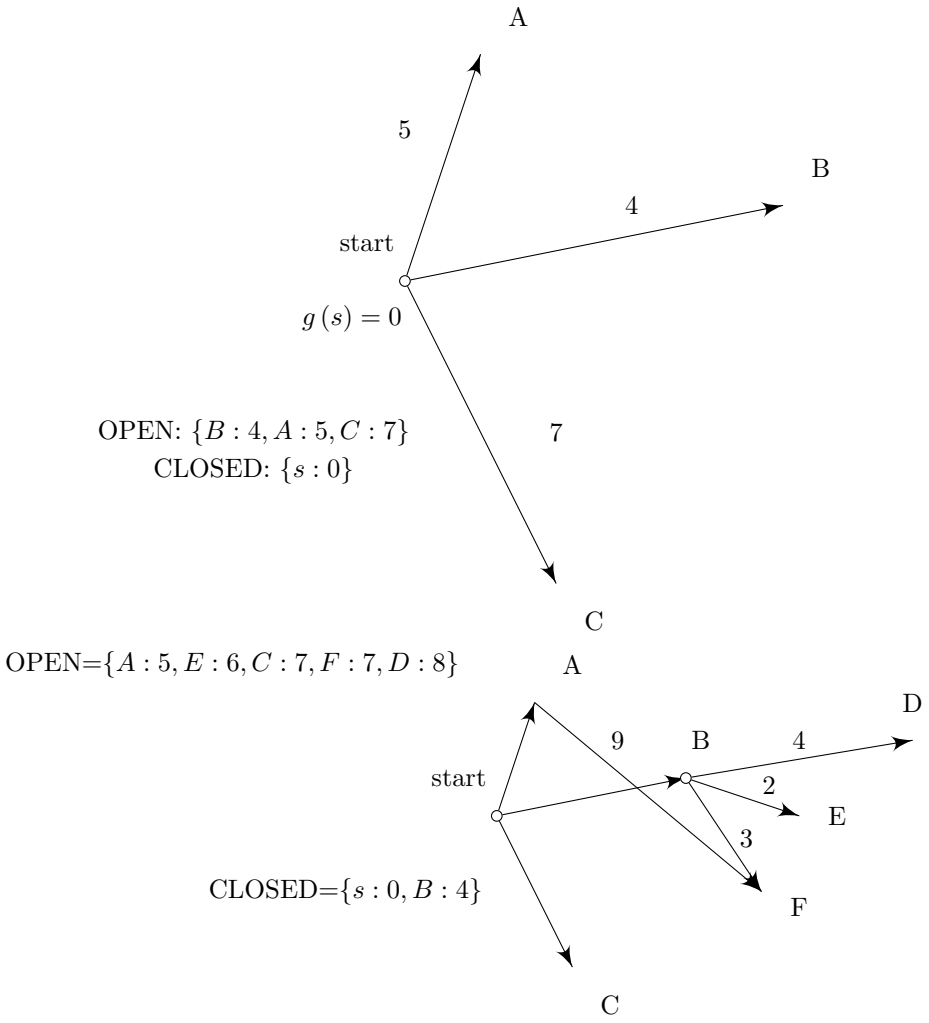
$c(n_1, n_2)$ 表示从 n_1 到 n_2 的代价

$g(n_2) = g(n_1) + c(n_1, n_2)$

优点：加入了状态图中的路径长短元素（走一步看一步），以等代价选择下一节点的选择

Lecture 9

10.21



$$g(D) = g(B) + g(B, D) = 4 + 4 = 8.$$

3.1 启发式搜索

Notation. 盲目搜索的不足：效率低、组合爆炸、产生大量无用节点

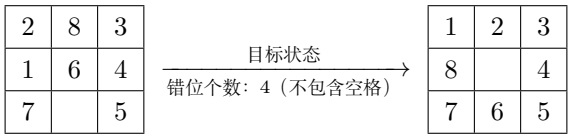
Notation. 启发式信息：与具体问题求解过程有关的，指导搜索过程朝最可能前进方向的数据

Notation. A 算法：

引入估价函数：$f(n) = g(n) + h(n)$

$g(n)$ ： 从起始状态到当前状态已实际付出的代价

$g(n)$ 从当前状态到目标状态的估计代价（启发函数）



Example. 错位个数：与目标状态的比较差别

$$g(n)=0 \quad h(n)=4.$$

可得 $f(n)=g(n)+h(n)=4$

类似于等代价算法，通过比较估价函数值即可减少遍历节点数

Notation. $A^\star$ 算法：对函数进行限定，使其一定可以找到最优解

$$A^\star=g(n)+h(n).$$

$g(n)$ 为起点到 n 点已走过距离

$g^\star(n)$ 是起点到 n 点的最短路径

$g(n)$ 是对 $g^\star(n)$ 的估计

$h(n)$ 为引导从 n 点到目的地的参照距离，一般为欧氏距离 $L_2(\boldsymbol{x}_i,\boldsymbol{x}_j)$

$h^\star(n)$ 为从 n 点到目的地的实际最短距离，$h(n)\le h^\star(n)$

Example. 百度地图：一直有一条红线引导方向，该红线即是 $h(n)$

确定的路线为绿色，为 $h^\star(n)$

Example. 八数码难题：$h_1(n)$ 表示不在位置上的数字数量

$h_2(n)$ 表示节点 n 到目标位置的曼哈顿距离之和

易得 $0\le h(n)\le h_1(n)\le h_2(n)$

Lecture 10

10.24

Notation. $A^\star$ 算法的搜索效率：$h(n)\le h^\star(n)$ 的前提下 $h(n)$ 越大越好

4 智能计算

4.1 对物质适者生存能力的解读

时间维度：进化智能

空间维度：群体智能

Lecture 11

10.28

4.2 遗传算法

Notation. 如何模拟物种繁殖：

1. 种群中选择两个个体
2. 随机确定编码序列断裂点
3. 交换编码片段

基因发生突变的概率称为遗传算法的**突变算子**

Notation. 如何模拟竞争与选择：**适应度**

Example. 求函数 $f(x)=x^2$ 的最大值，$x\in[0,31], x\in\mathbb{Z}$

使用 5 个二进制码表示取值：$0\sim 31 \Rightarrow$ 0b00000 $\sim$ 0b11111

定义 32 条染色体：

表 11: 染色体

X_o	X_b
0	0b00000
1	0b00001
...	...
31	0b11111

假设初始种群数量 $N=4$ ，随机产生 20 位的二进制串，每 5 个一组，得到 4 个初始个体

如：0b**00110**10010**1001**101010

适应度函数为给定问题 $f(x)=x^2$

表 12: 选择算子

编号	个体编码	个体	适应度	$\frac{f}{\sum f}$	$\frac{4f}{\sum f}$	生存数
S_1	00110	6	36	0.043	0.175	0
S_2	10010	18	324	0.394	1.578	2
S_3	10011	19	361	0.439	1.759	2
S_4	01010	10	100	0.121	0.487	0
适应度总和			821	平均适应度		205.25

得到第一代种群：(0b10010, 0b10011), (0b10010, 0b10011)：避免近亲相交

令交叉概率 $P_s=1$ ，变异概率 $P_m=0.01$ （每五代变异一个基因），通过生存数生成新的种群，配对后随机选择断裂点位交叉配对，完成第一代遗传：第二代种群的适应度更高

生成第五代种群后，通过变异算子随机挑选一个基因进行改变

经过数代遗传后，种群趋于稳定，适应度不再提升：$X=31$

表 13: 第一次交叉

交叉前	交叉后	个体	适应度	生存数
0b100**10**	0b10011	19	361	1
0b100**11**	0b10010	18	324	1
0b1**0010**	0b10011	19	361	1
0b1**0011**	0b10010	18	324	1
适应度总和			821 $\Rightarrow$ 1370	

4.3 蚁群算法

Notation. 蚂蚁的智能程度非常低，单个觅食随机性很大；但组合成群体后可以完成复杂的任务，且可以适应环境变化

Notation. 蚂蚁依靠**信息素**寻找最短路径

信息素：蚂蚁自身释放的易挥发的物质

在该道路上经过的蚂蚁越多，信息素浓度越高；浓度越高，这条道路就越容易被选择

信息素会随着时间推移而消散

Notation. 正反馈机制：

在寻找到较短路径后，蚂蚁释放的信息素会增加蚂蚁选择该路径的概率，同时后续的蚂蚁释放的信息素会进一步加强信息素浓度

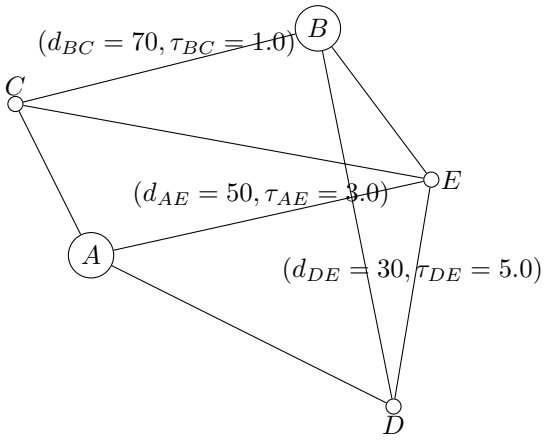
模拟蚂蚁觅食的 4 个抽象部分：

模拟蚂蚁

$$\begin{cases} \text{相同目标，相同速度运动} \\ \text{在到达目的地之前不回头、不转圈} \\ \text{根据相同的原则释放信息素、选择路径} \\ \text{记得自己走过的路径长度} \\ \text{种群中的个体数量不变} \end{cases}.$$

模拟地图

具有 N 个节点的全连通图，任意两点 X, Y 之间的距离 d_{XY} 设为已知，具有明确的起点与终点



Example. 旅行商问题

4.4 机器学习

Definition. 学习：

系统改进其性能的过程（西蒙）
获取知识的过程（专家系统）
技能的获取（心理学家）
事物规律的发现过程

Lecture 12

10.31

$$\text{机器学习}\begin{cases}\text{计算机视觉}\begin{cases}\text{手写识别}\\ \text{行人再识别}\end{cases}\\ \text{自然语言处理}\\ \text{社交媒体计算}\\ \text{经济金融}\end{cases}.$$

机器学习的四个部分：

$$\begin{cases}\text{T: Task}\\ \text{A: Algorithm}\\ \text{E: Experience}\\ \text{P: Performance}\end{cases}.$$

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 → 学习方法，建立模型 → 预测 → 测试匹配度

机器学习分类

$$\begin{cases} \text{半监督学习} \begin{cases} \text{监督学习} \\ \text{无监督学习} \end{cases} \\ \text{强化学习} \end{cases}.$$

4.4.1 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

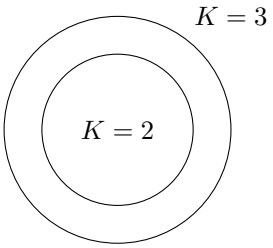
Notation. 分类方法：*K-nearest neighbour*，决策树，支持向量机，朴素贝叶斯
回归方法：线性、树、支持向量回归，集成方法

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 $\boldsymbol{x}_i, \boldsymbol{x}_j$ ，以 n 维向量空间作为特征空间，将这两个点表示为：

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

$$\boldsymbol{x}_i = \left(x_i^1, x_i^2, \ldots, x_i^n\right)^T.$$

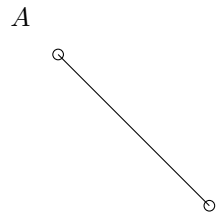
$$\boldsymbol{x}_j = \left(x_j^1.x_j^2, \ldots, x_j^n\right)^T.$$

特征点之间的距离定义为：

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

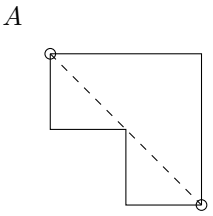
Example. 代入 $p = 2$ ，易得 $L_2\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right)$ 为平面上两点间的距离公式，该距离又称为欧氏距离：

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



$$L_2\left(A, B\right) = 2\sqrt{2}$$

代入 $p = 1$ ： $L_1\left(x_i, x_j\right)$ 称为曼哈顿距离：



$$L_1\left(A, B\right) = 4$$

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

4.4.2 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：

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

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

表 14: 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：

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

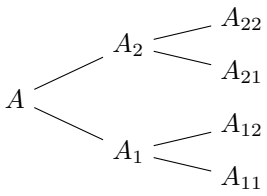
切比雪夫距离：

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

4.4.3 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

4.5 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