

# Part III-B: Medicine AI

Lecture by None

Note by THF

2024 年 10 月 16 日

## 目录

1	导论	1
1.1	监督学习 . . . . .	2
1.1.1	数据挖掘 . . . . .	3
1.1.2	数据选择 . . . . .	3
1.1.3	数据表征 . . . . .	3

Learn 1

10.07

## 1 导论

Notation. 机器学习的流程:

1. 确立目标

2. 收集数据

3. 数据预处理

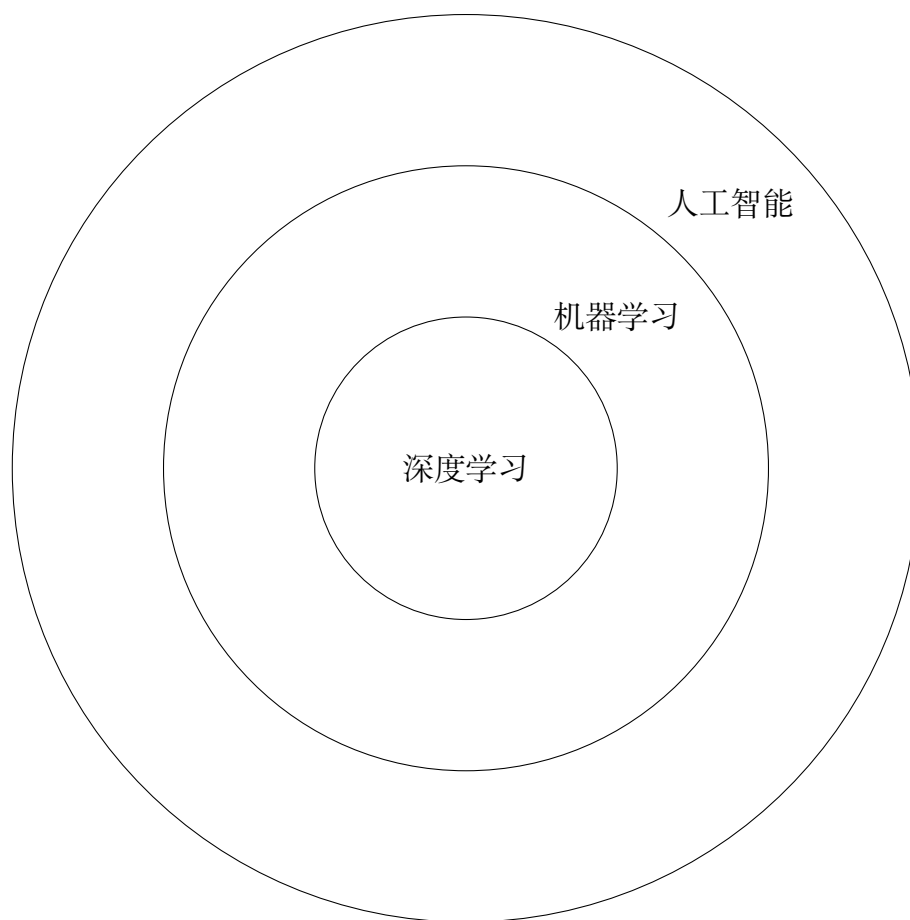
4. 数据分析

5. 模型训练

6. 模型评估优化

7. 预测

机器学习和人工智能的关系：



机器学习算法包含：无监督学习、监督学习、强化学习

## 1.1 监督学习

**Notation.** 机器学习选择数据要求：

1. 了解数据类型、属性、量纲
2. 分析分布特性
3. 选择高可信度数据
4. 进行数据表征（将原始数据转换为计算机可识别数据）

**Example.** 医药领域对小分子、蛋白质、核酸进行特征数字化方法

### 1.1.1 数据挖掘

1. 通过数据分析与统计学规律
2. 通过爬虫与自动化程序

### 1.1.2 数据选择

通过一部分数据来体现总体数据

### 1.1.3 数据表征

**Example.** 分子指纹:

首先提取分子结构特征（官能团等），使用分子结构特征生成比特向量，每个比特元素对应一种分子片段，通过对比比特向量的相似度来记录分子特征

分子指纹分类：基于子结构、拓扑或路径、药效集团的分子指纹和圆形分子指纹

**Notation.** SMILES/简化分子线性输入规范:

SMILES 是一种 ASCII 字符串，具体规则如下

## SMILES RULE

### 1. 简单规则

原子：原子缩写符号

**Example.** Au, Pt, C, N

离子：原子加上电荷数，外接中括号

**Example.**  $\text{Fe}^{3+}$ : [Fe+++]

$\text{C}^-$ : [C-]

$\text{Pt}^{6+}$ : [Pt+++++]

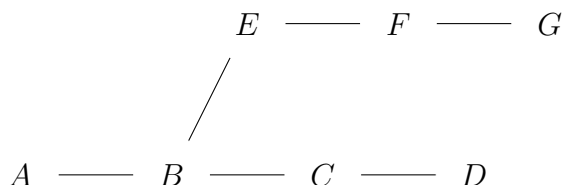
H 原子：省略

相邻原子：直接连接

**Example.** Dodecane: CCCCCCCCCCCC (12 Carbons)

分支：以小括号表示

**Example.** Write in git style:



SMILES: AB(EFG)CD

单键：直接省略

双键：“=”

三键：“#”

芳香键 = 单键（直接省略）

**Notation.** 部分软件芳香键使用单双键交替表示

芳香原子使用小写字母

**Example.** hex-2-en-4-yne/戊-2-烯-4-炔（不分顺反）: CC=CC#CC

toluene: Cc1ccccc1

## 2. 立体结构

环状结构：将环断开形成线性结构，以数字标记断开的原子

**Example.** Cyclohexane: C1CCCCC1

同位素：[核电荷数 + 元素符号]

**Example.**  $^{13}\text{C}$ : [13C]

Z/E 构象：使用 “/” 和 “\” 代表单键方向

**Example.** (2E)-hex-2-en-4-yne: C/C=C/C#CC

(2Z)-hex-2-en-4-yne: C/C=C\C#CC

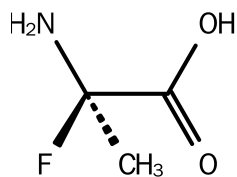
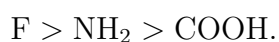


图 1: S&amp;R

手性异构: @ 表示 S, @@ 代表 R

**Example.**  $-\text{CH}_3$  最小, 放在最后, 对基团大小比较:



为 R 构型, 即: N[C@@](F)(C)C(=O)O

### 3. 算法与生成

**Notation.** 大部分 SMILES 生成算法为商业算法, 如 Morgan 算法、Canonical SMILES 算法等

生成 SMILES 主要使用深度优先搜索 (DFS) 算法遍历分子图

**Notation.** InChI: 国际化合物标识, 是规范的线性表示法、基于规范命名法则的唯一标识符

通过分层符号 “/” 将表示小分子的字符串分层, 前三层简化连接表的信息, 其他层处理额外问题

## InChI RULE

### 1. 主层

主层可包括三个子层: 化学式、原子连接、氢原子

主层  $\left\{ \begin{array}{l} \text{化学式} \\ \text{原子连接} \\ \text{氢原子} \end{array} \right.$