Part III-B: Medicine AI

Lecture by None

Note by THF

2024 年 10 月 17 日

目录

Learn 1

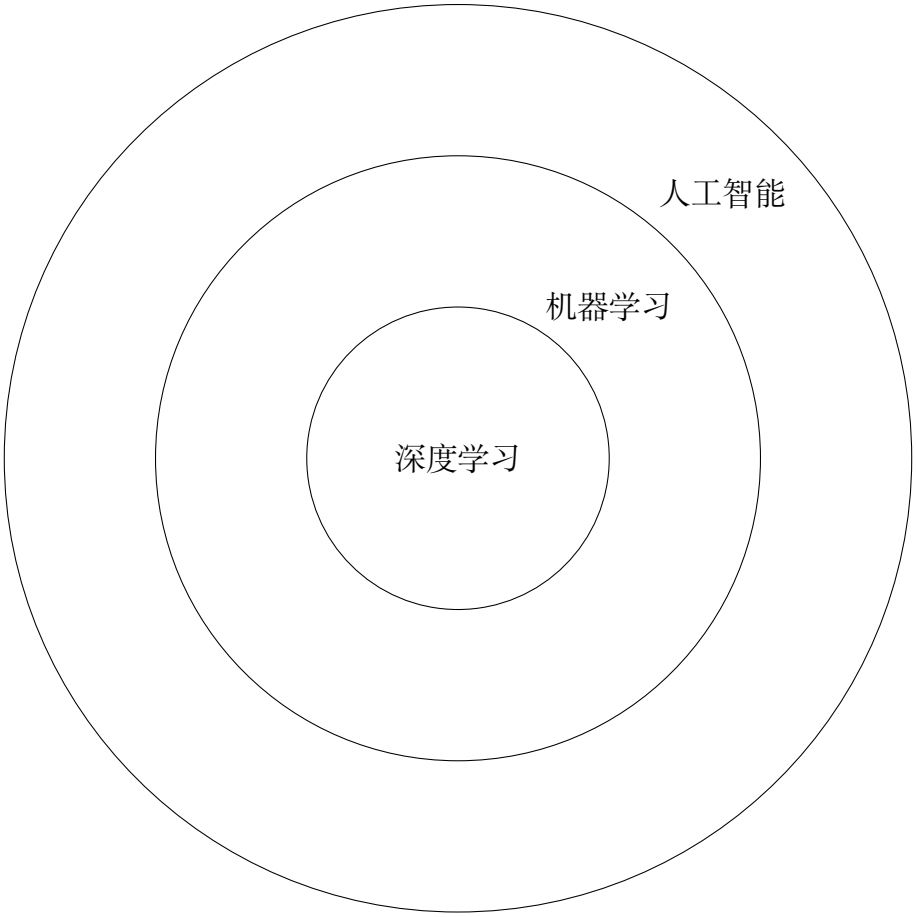
10.07

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. Fe^{3+}: [Fe+++]

C^{-} : [C-]

Pt^{6+} : [Pt++++++]

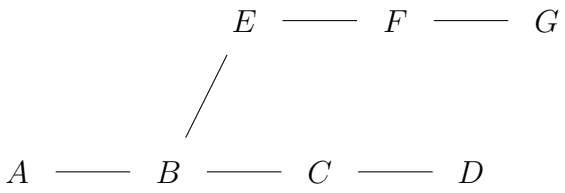
H 原子：省略

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}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

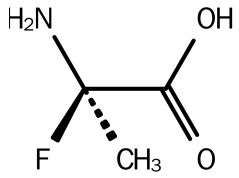


图 1: S&R

手性异构：@ 表示 S，@@ 代表 R

Example. $-CH_3$ 最小，放在最后，对基团大小比较：

$$F > NH_2 > COOH.$$

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

$$\text{主层}\begin{cases}\text{化学式}\\ \text{原子连接}\\ \text{氢原子}\end{cases}.$$

Learn 2

10.17

Notation. 氨基酸组成和二肽组成

基础知识：组成人体的二十种氨基酸

表 1: 20 amino acids

Alanine(A)	Arginine(R)	Asparagine(N)
Asparticacid(D)	Cysteine(C)	Glutamine(Q)
Glutamicaci(E)	Glycine(G)	Histidine(H)
Isoleucine(I)	Leucine(L)	Lysine(K)
Methionine(M)	Phenylalani(F)	Proline(P)
Serine(S)	Threonine(T)	Tryptophan(W)
Tyrosine(Y)	Valine(V)	

表 2: 20 种基本氨基酸

丙氨酸,A	精氨酸,R	天冬酰胺,N
天冬氨酸,D	半胱氨酸,C	谷氨酰胺,Q
谷氨酸,E	甘氨酸,G	组氨酸,H
异亮氨酸,I	亮氨酸,L	赖氨酸,K
甲硫氨酸,M	苯丙氨酸,F	脯氨酸,P
丝氨酸,S	苏氨酸,T	色氨酸,W
酪氨酸,Y	缬氨酸,V	

除此外还有用于终止密码子的硒半胱氨酸、吡咯赖氨酸（U）

Notation. 氨基酸组成的公式：

$$f(k) = \frac{N_k}{N}, k = 1, 2, \ldots, 20.$$

其中 N_k 表示第 k 种氨基酸的数量，N 表示氨基酸序列长度

Notation. 二肽组成的公式：

$$f(k,s)=\frac{N_{ks}}{N-1}, k,s=1,2,\ldots,20.$$

同理：N_{ks} 为第 k 种和第 s 种氨基酸形成的二肽数量

Notation. 蛋白质独热编码

使用 $20\times L$ 的矩阵表示蛋白质的序列信息，L 为蛋白质的序列长度

Example. 含 556 个氨基酸的蛋白质序列可以用 20×556 的矩阵表示，纵向量为二十种氨基酸，横向量为蛋白质在某位置的氨基酸种类

Notation. CTD 描述符

组成、转换与分布（Composition, Transition and Distribution, CTD）根据蛋白质序列中残基的特性编码蛋白质

CTD 编码分类方式 {
- 疏水性
- 范德华体积
- 极性
- 可极化性
- 带电性
- 表面张力
- 二级结构
- 溶剂可及性
- ...

}.

将氨基酸残基分为三类：

表 3: CTD 分类

性质	A	B	C
疏水性	亲水	中性	疏水
范德华体积	(0,2.78)	(2.95,4)	(4.43,8.08)
极性	(0,0.456)	(0.6,0.696)	(0.792,1)
可极化性	(0,0.108)	(0.128,0.186)	(0.219,0.409)
带电性	正电	中性	负电
表面张力	(-0.2,0.16)	(-0.52,-0.3)	(-2.46,-0.98)
二级结构	螺旋	折叠	卷曲
溶剂可及性	包埋	中等	暴露

Notation. 蛋白质二级结构及蛋白质溶剂可及性

1. 蛋白质二级结构 (PSS)