

Part III-B: Medicine AI

Lecture by None

Note by THF

2024 年 10 月 24 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
1.2	核酸物质表征	9
1.2.1	碱基	9
1.3	数据预处理	10
1.3.1	标准化	10
1.3.2	插补缺失值	11
1.4	模型评估和性能度量	13
1.5	模型性能度量	14

Learn 1 10.07

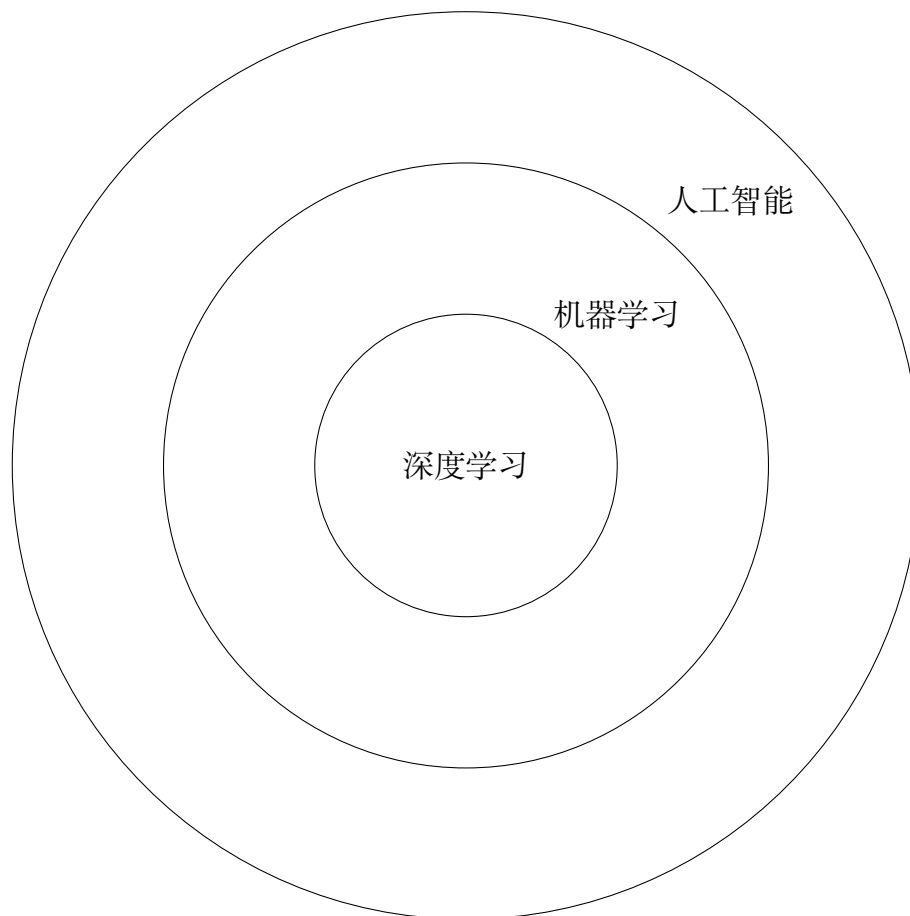
1 导论

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：

Learn 1

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+++++]

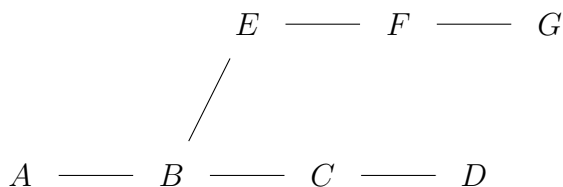
H 原子: 省略

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

手性异构: @ 表示 S, @@ 代表 R

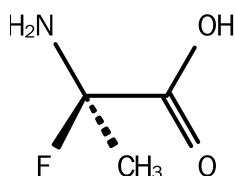
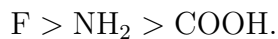


图 1: S&R

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

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

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)	

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：

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

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：

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

同理： N_{ks} 为第 k 种和第 s 种氨基酸形成的二肽数量

Learn 2

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	

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
- ...

.

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)
2. 氨基酸溶剂可及性 (PSA)

Learn 3

10.18

编码规则:

$$\left\{ \begin{array}{l} \text{二级结构} \left\{ \begin{array}{l} \text{H: } \alpha\text{螺旋} \rightarrow (0, 1, 0) \\ \text{E: } \beta\text{折叠} \rightarrow (1, 0, 0) \\ \text{C: 其他结构} \rightarrow (0, 0, 1) \end{array} \right. \\ \text{溶剂可及性} \left\{ \begin{array}{l} \text{b: buried (包埋)} \rightarrow (1, 0) \\ \text{e: exposed (暴露)} \rightarrow (0, 1) \end{array} \right. \end{array} \right. .$$

Example. 有一条 10 氨基酸长度的蛋白质序列:

PSSSA 使用 5×1000 的矩阵编码蛋白质, 每一个氨基酸由一个 5 维向量表示

用 PSSSA 编码时, 一般取序列羧基的一侧开始的 1000 个氨基酸编码, 如不满 1000 个使用 0 向量补齐

Prot.		M	V	L	S	P	A	D	K	T	N
Sec.		C	C	C	C	E	H	E	E	H	H
PSSSA	PSS	0	0	0	0	1	0	1	1	0	0
		0	0	0	0	0	1	0	0	1	1
	PSA	1	1	1	1	0	0	0	0	0	0
		0	0	1	0	0	0	1	1	0	0
		1	1	0	1	1	1	0	0	1	1
S.A.		e	e	b	e	e	b	b	e	e	e

1.2 核酸物质表征

Notation. 基本知识：碱基与核酸

1.2.1 碱基

表 4: 常见碱基

种类	DNA	RNA
嘌呤族 (R)	腺嘌呤 (A)	
	鸟嘌呤 (G)	
嘧啶族 (Y)		胞嘧啶 (C)
	胸腺嘧啶 (T)	尿嘧啶 (U)

Notation. 碱基配对方式：

$$\left\{ \begin{array}{l} \text{DNA} \left\{ \begin{array}{l} A = T \\ C \equiv G \end{array} \right. \\ \text{RNA} \left\{ \begin{array}{l} A = U \\ C \equiv G \end{array} \right. \end{array} \right. .$$

Notation. K-mer

K: DNA 或 RNA 中一个长度为 K 的序列

以该序列为子序列，遍历核酸序列，计算该长度的所有子序列组合出现的频率

Example. 长度为 K 的 K-mer 种类共有 4^k 种可能

如长度为 3 的子序列，子序列每个位置有 A,G,C,U 四种选择，共 4^3 种组合
一段 15 个核酸的 RNA 序列如下：

表 5: 3-mer RNA

C	A	T	C	G	G	T	A	A	C	C	C	A	T
---	---	---	---	---	---	---	---	---	---	---	---	---	---

所有可能的长度为 3 的子序列及其频率：

表 6: 3-mers

	RNA seq.	freq.
1	CAT	0.111
2	ATC	0.056
3	TCG	0.056
4	CGG	0.056
...
12	CCA	0.056
13	ATG	0
...
64	...	0

Learn 4

10.20

1.3 数据预处理

1.3.1 标准化

Notation. 变量离差标准化：标准化后所有变量范围都在 $[0,1]$ 内

$$y_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}}.$$

Learn 4

Example. 一组变量如下：

$$X = (1.5, 1.7, 2.2, 1.2, 1.6, 1.4, 1.1).$$

易得 $x_{\min} = 1.1, x_{\max} = 2.2$

$$\begin{aligned} y_i &= \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \\ &= \frac{x_i - 1.1}{2.2 - 1.1} \\ &= \frac{x_i - 1.1}{1.1} \\ &= \frac{x_i}{1.1} - 1 \end{aligned}$$

得 $Y = (0.364, 0.545, 1, 0.091, 0.455, 0.273, 0)$

Notation. *Z-score* (变量标准差) 标准化

经过标准化后平均值为 0，标准差为 1

$$z_i = \frac{x_i - \bar{x}}{s} \quad s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}.$$

可以看出 s 为原数据的标准差， z_i 值其实等同于标准正态分布中的 u 值：

$$u = \frac{x - \mu}{\sigma} \quad y = \frac{1}{\sigma\sqrt{2\pi}} e^{-u^2/2}.$$

1.3.2 插补缺失值

Notation. 均值插补

1. 数值性变量：采用平均值插补
2. 离散型：采用众数插补

Notation. 同类均值插补：使用层次聚类方法归类缺失值的样本，用该类别的特征均值插补

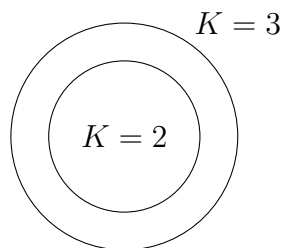
Notation. KNN(*K-nearest neighbor*) 缺失值插补：找到与含缺失值样本相似的 K 个样本，使用这 K 个样本在该缺失变量上的均值填充

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：

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

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

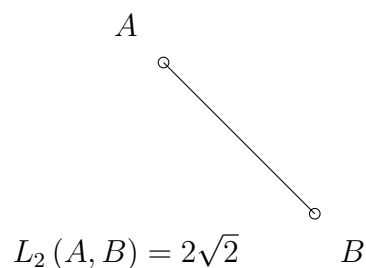
$$\mathbf{x}_j = (x_j^1, x_j^2, \dots, x_j^n)^T.$$

特征点之间的距离定义为：

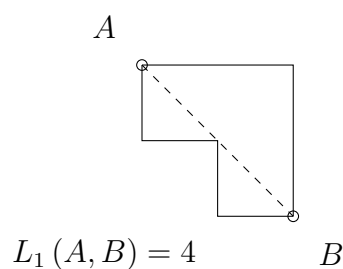
$$L_p(\mathbf{x}_i, \mathbf{x}_j) = \left(\sum_{l=1}^n |x_i^l - x_j^l|^p \right)^{\frac{1}{p}}.$$

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 $p = 1$: $L_1(x_i, x_j)$ 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

Learn 5

10.23

1.4 模型评估和性能度量

Notation. 留出法 (*hold-out*):

将原始数据集 D 分为两个互斥的子集 S, T , S 作为训练数据集, T 作为测试数据集: $D = S \cup T, S \cap T = \emptyset$

在划分任务时要尽量保证 S 和 T 中的样本类别比例相似

Example.

$$D(a, b) \rightarrow S(\lambda a, \lambda b) \cup T((1 - \lambda)a, (1 - \lambda)b).$$

该过程称为分层采样法, 其中 $\lambda \in [\frac{2}{3}, \frac{4}{5}]$

使用 S 训练模型, T 进行模型测试, 多次随机划分 a, b 在 S 和 T 内的内容, 多次实验取测试结果平均值

Notation. 交叉验证法/ k 折交叉验证 (*cross validation/ k -fold cross validation*):

$$D = D_1 \cup D_2 \cup \dots \cup D_k \text{ 且 } D_i \cap D_j = \emptyset (i \neq j).$$

此处 $\forall D_i$ 由 D 分层采样得到

每次实验使用 $k-1$ 个子集的并集训练, 剩下的一个子集作为测试集:

$$S = \sum_{i=1}^{m-1} D_i + \sum_{i=m+1}^k D_i \quad T = D_m.$$

取不同的 m 值共可以得到 k 组“训练集-测试集”, 得到 k 个结果, 取 k 个结果的平均值

Example. 5折交叉验证的数据划分:

D_1	D_2	D_3	D_4	D_5
D_1	D_2	D_3	D_4	D_5
D_1	D_2	D_3	D_4	D_5
D_1	D_2	D_3	D_4	D_5
D_1	D_2	D_3	D_4	D_5

 $\Rightarrow \begin{cases} Res_1 \\ Res_2 \\ Res_3 \\ Res_4 \\ Res_5 \end{cases} \xrightarrow{\text{Avg.}} \text{Result}$

Notation. 若样本量 m 等于子集数 k , 交叉验证法等同于留一法 (*leave one out, LOO*)

留一法的优点: 训练结果更准确

缺点: 样本量太大的时候消耗过多资源

1.5 模型性能度量

Notation. 错误率:

$$E = \frac{1}{m} N(f(x_i) \neq y_i).$$

准确率:

$$\text{Acc} = \frac{1}{m} N(f(x_i) = y_i).$$

m 为样本总数, $N(f(x_i) = y)$ 表示符合特征 $f: x \rightarrow y$ 的样本数量

Notation. 二分类问题: