Part III-B: Medicine AI

Lecture by None

Note by THF

2024 年 10 月 20 日

目录

Learn 3

10.18

编码规则：

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

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

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
		1	1	1	1	0	0	0	0	0	0
	PSA	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

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

0.1 核酸物质表征

Notation. 基本知识：碱基与核酸

0.1.1 碱基

表 1: 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：

$$\begin{cases} \text{DNA} \begin{cases} A = T \\ C \equiv G \end{cases} \\ \text{RNA} \begin{cases} A = U \\ C \equiv G \end{cases} \end{cases}.$$

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 2: 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：

Learn 4

10.20

0.2 数据预处理

0.2.1 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

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

Example. 一组变量如下：

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

表 3: 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

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:

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

Learn 4

0.2.2 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

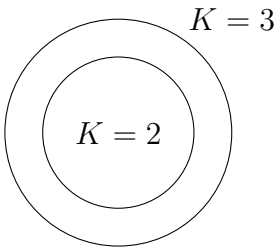
Notation. KNN(*K-nearest neighbor*) 缺失值插补：找到与含缺失值样本相似的 K 个样本，使用这 K 个样本在该缺失变量上的均值填充

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

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

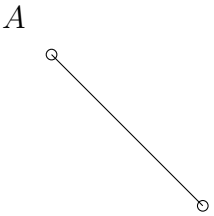
$$\boldsymbol{x}_i = \left(x_i^1, x_i^2, \ldots, x_i^n\right)^T.$$
$$\boldsymbol{x}_j = \left(x_j^1.x_j^2, \ldots, x_j^n\right)^T.$$

特征点之间的距离定义为:

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

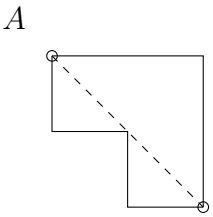
Example. 代入 $p = 2$ ，易得 $L_2\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right)$ 为平面上两点间的距离公式，该距离又称为欧氏距离:

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



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

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