

Part III-B: Medicine AI

Lecture by None

Note by THF

2024 年 10 月 23 日

目录

0.1	数据预处理	1
0.1.1	标准化	1
0.1.2	插补缺失值	2
0.2	模型评估和性能度量	4

Learn 4

10.20

0.1 数据预处理

0.1.1 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:

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

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:

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

0.1.2 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

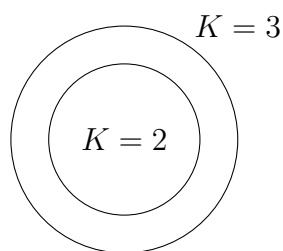
Notation. KNN(*K-nearest neighbor*) 缺失值插补: 找到与含缺失值样本相似的 K 个样本, 使用这 K 个样本在该缺失变量上的均值填充

***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：

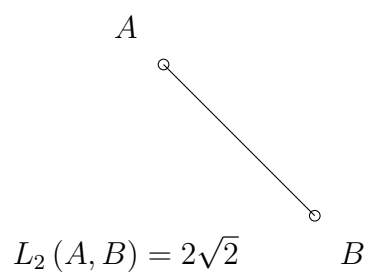
$$\begin{aligned}\mathbf{x}_i, \mathbf{x}_j &\in \mathbf{X}. \\ \mathbf{x}_i &= (x_i^1, x_i^2, \dots, x_i^n)^T. \\ \mathbf{x}_j &= (x_j^1, x_j^2, \dots, x_j^n)^T.\end{aligned}$$

特征点之间的距离定义为：

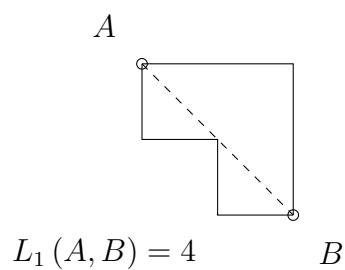
$$L_p(\mathbf{x}_i, \mathbf{x}_j) = \left(\sum_{l=1}^n |x_i^l - x_j^l|^p \right)^{\frac{1}{p}}.$$

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

0.2 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: $D = S \cup T, S \cap T = \emptyset$